UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Bayesian Estimation Employing a Phase-Sensitive Observation Model for Noise and Reverberation Robust Automatic Speech Recognition

Von der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

Dipl.-Ing. Volker Sebastian Leutnant

Erster Gutachter:	Prof. Dr.-Ing. Reinhold Häb-Umbach
Zweiter Gutachter:	Prof. Bhiksha Raj
Tag der mündlichen Prüfung:	19.05.2015

Paderborn 2015

Diss. EIM-E/305

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fakultät für
Elektrotechnik,
Informatik und
Mathematik

Abstract of the PhD thesis

Bayesian Estimation Employing a Phase-Sensitive Observation Model for Noise and Reverberation Robust Automatic Speech Recognition

of Mr. Volker Sebastian Leutnant

Speech recognition technology has been emerging into everyday life. The acceptance of speech recognition systems is, however, still suffering from their lack of robustness w.r.t. acoustic environmental noise and reverberation. This problem is probably most severe when hands-free systems are employed to capture human speech. While allowing the user to move freely without the need of wearing a headset or holding a microphone, performance of hands-free systems is particularly highly sensitive to the acoustic conditions of the environment they are employed in.

The reason for this may be found in the increased distance of the speaker to the microphone compared to the use of a headset, which leads to a degradation of the acoustic signal. Since the training of a speech recognizer's acoustic model is often carried out with clean speech signals, the signal modification by reverberation and noise results in a mismatch between the statistics of the observed feature vectors at training and testing stage, and thus in an increased word error rate. But even in the case of matched noisy reverberant training the performance deteriorates, since the temporal feature correlations introduced by reverberation violate the conditional independence assumption inherent to hidden Markov model based speech recognition.

In this thesis a detailed (statistical) analysis of how reverberation and noise affect the speech signal and eventually the feature vectors passed to the recognizer is carried out to address those issues. The findings lead to the derivation of a novel statistical observation model which relates the features of the noisy reverberant speech signal to those of the underlying clean speech signal and the noise. It is eventually employed in the context of model-based Bayesian feature enhancement with subsequent speech recognition.

The derived observation model thereby generalizes both the observation model for noisy speech and the observation model for reverberant-only speech and extends previous models in two major directions: First, the contribution of additive background noise to the observation error is explicitly taken into account, and second, the vector of phase factors, which arises from the cross-term in the computation of the power spectrum carried out during the front-end feature extraction, is fully incorporated. The statistics of both the vector of phase factors and the observation error are thereby investigated in full detail along its derivation.

The Bayesian inference is further soundly embedded into the statistical framework of automatic speech recognition in terms of a novel uncertainty decoding scheme that renders the theoretically optimal solution to the robustness problem practically feasible.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fakultät für
Elektrotechnik,
Informatik und
Mathematik

Zusammenfassung der Dissertation

Bayesian Estimation Employing a Phase-Sensitive Observation Model for Noise and Reverberation Robust Automatic Speech Recognition

des Herrn Volker Sebastian Leutnant

Mit der zunehmenden Nutzung und Verbreitung von automatischen Spracherkennungssystemen steigen auch die Anforderungen an eben diese Systeme im Hinblick auf Robustheit gegenüber Nachhall und Hintergrundstörungen. Im besonderen Maße gilt dies für Freisprechsysteme. Zwar erhöhen diese den Bedienungskomfort für die Nutzer, sorgen aber auch dafür, dass das Sprachsignal auf verschiedene Arten gestört werden kann.

Da das Training des akustischen Modells eines automatischen Spracherkennungssystems oftmals mit ungestörten Sprachsignalen durchgeführt wird, sorgen Nachhall und Hintergrundstörungen dafür, dass es während der Erkennung zu einer statistischen Diskrepanz zwischen den gespeicherten Modellen und den beobachteten Merkmalsvektoren kommt. Als Konsequenz dieser Fehlanpassung lassen sich steigende Wortfehlerraten des Erkenners beobachten. Aber auch wenn bereits auf Merkmalsvektoren von verrauschten und verhallten Sprachsignalen trainiert wurde, kommt es zu Verschlechterungen der Erkennungsergebnisse. Diese lassen sich auf die Verletzung der so genannten ”conditional-independence” Annahme, auf welche die ”Hidden Markov Modell”-basierte Spracherkennung fußt, durch die, durch den Nachhall bedingten, verstärkten zeitlichen Korrelationen der Merkmalsvektoren zurückführen.

Um diese Probleme adressieren zu können, wird in dieser Arbeit eine detaillierte (statistische) Analyse der Auswirkung von Nachhall und Hintergrundstörungen auf das Sprachsignal und schlussendlich auf die Merkmalsvektoren, welche für die Erkennung verwendet werden, durchgeführt. Daraus wird dann ein neuartiges Beobachtungsmodell, welches die Merkmale des verrauschten und verhallten Sprachsignals mit denen des ungestörten Sprachsignals und denen der Hintergrundstörung in Beziehung setzt, entwickelt und im Rahmen der modellbasierten Bayes'schen Merkmalsverbesserung zur Erkennung eingesetzt.

Das vorgestellte Beobachtungsmodell kann dabei als Generalisierung des Beobachtungsmodells für verrauschte Sprache und des Beobachtungsmodells für verhallte Sprache gesehen werden und unterscheidet sich dabei von existierenden Beobachtungsmodellen in zwei wesentlichen Punkten: Zum einen berücksichtigt es explizit die Auswirkungen der Hintergrundstörung auf den Beobachtungsfehler und zum anderen trägt es dem Vektor von Phasenfaktoren, welche aus der Berechnung des Kurzzeit-Leistungsdichtespektrums während der Merkmalsextraktion resultieren, Rechnung. Sowohl die statistischen Eigenschaften der Beobachtungsfehler als auch die der Phasenfaktoren werden dabei während der Herleitung detailliert untersucht.

Die Bayes'sche Inferenz wird darüber hinaus durch einen neuartigen Ansatz des ”Uncertainty Decodings” elegant in den statistischen Rahmen der automatischen Spracherkennung eingebettet.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated. It has not been published yet or submitted in whole or in part for a degree at any other university.

Some of the work in this thesis has been published as journal articles [119, 120], as a book chapter [126] and in conference proceedings [123, 124, 125, 131].

Dedication

To Sonja
the sunshine of my life

Acknowledgements

I would like to express my gratitude to my supervisor Prof. Reinhold Häb-Umbach for his valuable guidance in this work and constant encouragement. Working with him at the Department of Communications Engineering of the University of Paderborn has been a pleasant, enriching and memorable experience. Thank-you also to my second supervisor Prof. Bhiksha Raj. His feedback greatly contributed to the success of my thesis.

I would also like to thank my research colleagues and friends at the Department of Communications Engineering. In particular Alexander Küger, Hai-Dang Tran-Vu, Jörg Schmalenströer, Maik Bevermeier, Sven Peschke, Oliver Walter, Florian Jacob and Aleksej Chinaev. Thank-you for both the numerous discussions on and around my work and for occasionally distracting me from it.

Thank-you also to the students working at the lab of the Department of Communications Engineering during the peak season of my work for patiently awaiting the end of my numerous simulations and to the *Deutsche Forschungsgemeinschaft* for partly funding this research under contract no. Ha3455/6-1.

The biggest thank-you goes to my family – I couldn't have made it without your love and support. Thanks for giving me roots and wings. "Ich liebe Euch!"

Contents

1 Introduction

Speech recognition technology has been emerging into everyday life, e.g., in terms of dictation systems on personal computers and command/text entry interfaces on mobile devices. The achieved recognition performance is, however, usually still much worse than that of human listeners. While this may partly be attributed to the paradigms underlying the development of state-of-the-art ***a**utomatic **s**peech **r**ecognition* (ASR) systems (e.g., regarding models and algorithms), the dominant cause may be found in their lack of robustness w.r.t. acoustic environmental noise, reverberation and any other kind of distortion.

This problem is probably most severe when hands-free systems are employed to capture human speech. While allowing the user to move freely without the need of wearing a headset or holding a microphone, they contribute to increased convenience as well as safety and may finally improve the acceptance of automatic speech recognition in many application areas. However, the increased distance of the speaker to the microphone compared to the use of a headset leads to a degradation of the acoustic signal. In particular, the signal-to-noise ratio is first degraded due to signal attenuation as a result of sound propagation through the air and second because it is likely that a distant microphone captures other noise sources, in addition to the desired speaker's voice. Further, the received signal is degraded by reverberation due to multi-path propagation.

Since the training of a speech recognizer's acoustic model is often carried out with clean speech signals, the signal modification by reverberation and noise results in a mismatch between the statistics of the observed feature vectors at training and testing stage, and thus in an increased word error rate. But even in the case of matched noisy reverberant training the performance deteriorates, since the temporal feature correlations introduced by reverberation violate the conditional independence assumption, which is inherent to the ***h**idden* **M**ARKOV ***m**odel* (HMM) based speech recognition also considered in this work.

Key to the success of any approach to the recognition of noisy reverberant speech is an understanding of how these two kinds of distortion affect the speech signal and eventually the feature vectors passed to the recognizer. This may be expressed by a so-called observation model, which relates the observed signal to the underlying clean, non-reverberant speech signal and that of the noise. The observation model is the core component of model-based feature compensation, a paradigm to environmentally robust ASR, which has lead to a variety of very effective noise compensation algorithms. In recent years, research has been undertaken to apply the same ideas to the recognition of reverberant speech. However, relatively few work has been devoted to the joint treatment of noise and reverberation.

Depending on where the effect of the distortions is addressed, it has to be formulated as a relationship among the signals, the features or the statistical models corresponding to the clean data, the noise and the observations. A system theoretic model for the effect of reverberation is the convolution of the source signal with the ***a**coustic **i**mpulse **r**esponse*

(AIR) from the source to the sensor. Although, in general, the AIR is infinite in length and time-variant in consequence of changes within the source-sensor enclosure, e.g., speaker movements or variations of temperature or humidity, it is usually assumed to be of finite length and time-invariant for the sake of simplicity.

This simplified model is at the outset of many speech enhancement algorithms which aim at dereverberating the signal prior to feature extraction by inverse filtering [1, 2], blind deconvolution [3, 4, 5] or the enhancement of the ***l**inear **p**rediction* (*LP*) residuum [6, 7]. Alternatively, dereverberation may be achieved by multi-step linear prediction [8] or by the exploitation of the harmonic structure of speech [9]. A comprehensive overview of signal dereverberation approaches may be found in [10]. The additional compensation of noise is addressed by, e.g., spectral enhancement techniques like [11, 12, 13], blind acoustic beamforming [14] or methods based on multi-step linear prediction like [15]. However, since the extension of dereverberation approaches to the compensation of noise is often not straight forward, the latter methods are distinctly more rare in existing literature. Moreover, most of them assume multi-channel data.

In the following only approaches that call for a single microphone will be considered, since this is also the scenario this work is based upon.

As the estimation of the AIR is a complicated task (even if a microphone array is available [16], and the more in the case of single-channel input) simplified observation models have been proposed which do not require the estimation of the full AIR but rather some of its characteristics. They represent the effect of reverberation in the spectral or log-spectral domain rather than in the time domain. The proposed simplified observation models can be broadly categorized into three groups:

i) Models that assume a linear, affine or additive relationship between clean and reverberant features in the logarithmic mel power spectral or cepstral domain, neglecting any temporal correlation introduced by reverberation [17, 18, 19, 20, 21].

ii) Models that describe the reverberation as an additive distortion in the power spectral domain [11, 10, 22, 23, 24].

iii) Models that describe reverberation by a convolution in the power spectral domain [25, 26, 27, 28, 29].

Despite being a very coarse approximation, most feature normalization approaches rely on models falling into the first category. One famous representative is the ***c**epstral **m**ean **n**ormalization* (CMN) [17] modeling the effect of reverberation to be multiplicative in the short-time spectral domain, and thus approximately additive in the cepstrum, using the ***m**ultiplicative **t**ransfer **f**unction **a**ssumption* (MTFA) [30]. However, the MTFA is only valid if the duration of the analysis window for the computation of the short-time spectrum is large compared to the duration of the AIR. Since the analysis windows used for feature extraction have a typical duration of about 25 ms, this condition is usually not met and CMN fails to compensate for reverberation. A possibility to circumvent this problem is to use longer analysis windows for normalization like in [18, 19], whereupon, in a subsequent step, the standard time-frequency resolution suitable for ASR has to be restored either directly or through signal re-synthesis and additional feature extraction. An extension of such normalization methods for the joint compensation of noise and reverberation is presented in [21]. CMN (if applied to the complete feature vector, containing also velocity and acceleration components) may also be considered a special case of affine transformations like ***c**onstraint **m**aximum **l**ikelihood **l**inear **r**egression* (CMLLR) [20]. Although the effect of reverberation is not well modeled with affine transformations, either, CMLLR may compensate for it to a certain degree through the dynamic features, which are computed for the duration of several frames and thus capture some of the temporal smearing introduced by reverberation [28].

An example of the use of observation models that fall into the second category are the single-channel speech dereverberation techniques which rely on the estimation of the ***l**ate **r**everberant **s**pectral **v**ariance* (LRSV). The LRSV is assumed to be an additive distortion which is uncorrelated with the direct path and early reverberant speech components. Suppression of the LRSV can be achieved by spectral enhancement techniques in the ***s**hort-**t**ime **d**iscrete* **F**OURIER ***t**ransformation* (STDFT) domain. LEBART, BOUCHER and DENBIGH were the first to derive an LRSV estimator [11]. They employed the model by POLACK, which describes the AIR as a realization of a zero-mean Gaussian random process multiplied by an exponentially decaying function [31] and developed a recursive estimator of the LRSV which does not require knowledge or the estimation of the AIR but only of its characteristic energy decay constant. Their model was later improved to better model the energy contribution of the direct sound [10, 22] and extended towards time-variant AIRs in [23]. If the observed signal was further degraded by background noise, a standard noise suppression algorithm was applied prior to dereverberation.

The BAYESIAN feature enhancement method by WOELFEL also assumed reverberation to be an additive distortion, here in the mel power spectral domain, whose contribution was estimated in the time domain using multi-step linear prediction [24]. In order to be used in a BAYESIAN framework, not only a point estimate of the LRSV was required, but a complete observation probability density, which was realized by employing particle filtering techniques.

The models falling into the third category are clearly closest to the physical model outlined earlier. The observation models to be presented in this work also fall in this category. They will be employed for BAYESIAN feature enhancement whereby the a posteriori ***p**robability **d**ensity **f**unction* (PDF) of the non-reverberant and noise-free logarithmic mel power spectral feature vector is estimated from the noisy reverberant input feature vectors and then forwarded to an ASR back-end. By enhancing the features rather than the signal one can take advantage of both a decimation in time, since the frame rate at which feature vectors are computed is much lower than the sampling rate, and in frequency, due to the mel filter bank applied to the power spectrum. Though rendering the derivation of a corresponding observation model quite challenging, it simplifies the estimation of a frequency domain representation of the AIR. Further, it is generally considered advantageous if the representation to be enhanced, here the features rather than the acoustic signal, is close to what is actually processed in the recognizer. The enhancement can then be tailored to the specifics of the recognizer rather than to those of a human listener.

For the absence of additive background noise an approximation for the relation between the ***me**l **p**ower **s**pectral **c**oefficients* (MPSCs) of the reverberant speech signal and that of the clean speech signal was presented in [27]. Due to the absence of a power compensation factor, the model, however, suffers from a systematic underestimation of the power spectrum of reverberant speech. Convolutive observation models were also employed in the static acoustic model adaptation technique presented in [26] and the model-based vector TAYLOR series compensation scheme [32]. In [25] a recursive observation model was employed for dynamically adapting an acoustic model to reverberation.

The aforementioned models express the feature vectors of the observable reverberant and noisy speech signal by means of the feature vectors of the noise and the clean speech in a purely deterministic way. However, such deterministic formulations in the feature domain are, due to loss of information inherent to the mixing process and the process of the feature computation, never exact. Instead of neglecting the remaining error between the model and the true observation, an improved modeling may be achieved by describing this error (and thus the observation) in a probabilistic way.

Such a probabilistic model has been presented in [24], where reverberation was described as additive in the mel power spectral domain (category ii) above). In [28] a model for the logarithmic mel power spectral domain has been developed, however assuming absence of an additive noise term. Later, the model was extended to the presence of additive noise [33]. However, that extension only considered the contribution of the noise term to the deterministic part of the observation model, while its effect on the observation error was neglected.

It is well known that modeling the effect of additive noise as additive in the power spectral domain is only an approximation, which breaks down at a ***i**nstantaneous **s**ignal-to-**n**oise **r**atio* (ISNR) close to 0 dB. Then, the cross term in the computation of the power spectrum of a signal consisting of a superposition of speech and noise can no longer be neglected [34] and has immediate consequences for the observation error. The characterization of the observation error is significantly more complicated in the presence of reverberation, which results from filtering the speech signal by the AIR. Consequently, it becomes highly non-stationary and therefore difficult to model.

2 Contribution and Organization

In this work, statistical observation models for noisy, reverberant and noisy reverberant speech are investigated in the context of Bayesian feature enhancement with subsequent speech recognition. The major focus is thereby laid on a novel statistical observation model for noisy reverberant speech including a more refined treatment of the observation error taking into account the contribution of additive noise and the vector of phase factors.

Since the observation models will be used in Bayesian feature enhancement with subsequent speech recognition, the statistical framework of ASR is described in the first part of this work, i.e., Ch. 3. After describing the building blocks of an ASR system, namely the feature extraction (Sec. 3.1), the acoustic and language modeling (Sec. 3.2 and Sec. 3.3) as well as the decoding (Sec. 3.4), Sec. 3.6 lays the ground for soundly embedding the inference of the a posteriori PDF of the clean speech feature vector into the statistical framework of ASR. The derived (partly novel) *__u__ncertainty __d__ecoding* (UD) schemes thereby aim at rendering the theoretically optimal solution to environmental robustness practically tractable.

Chapter 4 then goes into the details of inferring the a posteriori PDF of the clean speech feature vector given either noisy, reverberant or noisy reverberant observations. The conceptually optimal solution to the estimation problem is presented in Sec. 4.1. It identifies two key components. The first key component is the a priori model statistically describing the trajectory of the *__l__ogarithmic __m__el __p__ower __s__pectral __c__oefficient* (LMPSC) feature vectors of the clean speech signal and the noise, which is introduced in Sec. 4.2. The second key component is the observation model statistically relating these entities to the reverberant, noisy reverberant or noisy observation. Starting with a review of the existing observation model for reverberant-only speech and closing with a review of the existing observation model for noisy speech, Sec. 4.3 introduces the novel observation model for noisy reverberant speech. Special focus is thereby laid on a thorough statistical formulation of the observation model and the model for the observation error, which is shown to not only depend on approximation errors originating from the observation model for reverberant-only speech but also on the *__i__nstantaneous __r__everberant-to-__n__oise __r__atio* (IRNR) and the vector of phase factors. Thereby it is shown, that the derived observation model generalizes the existing observation models targeting either reverberation or additive background noise as the only distortion affecting the clean speech signal.

The observation model for noisy reverberant speech and reverberant-only speech require a representation of the AIR in the LMPSC domain, which may be computed from the AIR if the latter is known. To avoid a sensitive blind estimation of the AIR, a simplified model of it is introduced in Sec. 4.4 and applied to the observation models in Sec. 4.5. This model of the AIR not only requires only two parameters to be estimated but also allows to formulate recursive variants of the observation models for reverberant-only speech and noisy reverberant speech. These recursive observation models are also described in Sec. 4.5.

All observation models targeting a noisy environment, i.e., either noisy reverberant speech or noisy speech, involve the vector of phase factors resulting from the cross-term in the computation of the power spectrum carried out during the front-end feature extraction. A detailed analysis of its statistical properties, a parametric approximation to its PDF and an analytic solution to its central moments is given in Sec. 4.6.

A detailed analysis of the observation errors for the various observation models is carried out in Sec. 4.7. Special stress is thereby laid on highlighting the aforementioned sensitivity of the observation error on the IRNR and the vector of phase factors.

Practically realizable approaches to the inference of the a posteriori PDF of the clean speech feature vector employing the presented observation models are then considered in Sec. 4.8 where the pursued sub-optimal multi-model and model-specific inference schemes are discussed.

The performance of all considered inference schemes is finally investigated in speech recognition experiments on appropriate databases in Ch. 5. The inference schemes are thereby applied on both small and large vocabulary recognition tasks featuring both artificially distorted data and recordings in a real noisy reverberant environment.

The thesis eventually concludes by Sec. 6.

3 Statistical Framework of Automatic Speech Recognition

Automatic speech recognition in a statistical framework basically reduces to the application of Bayes' decision rule to a pattern classification problem. In its generic formulation, Bayes' decision rule aims at minimizing the a posteriori expected loss associated with the decision given a particular observation. Under the most prominent loss function – the $0/1$-loss function – Bayes decision rule turns into the ***m**aximum **a** **p**osteriori* (MAP) decision rule. It states *"Decide for that class $\hat{\Omega} \in \{\Omega_1, \ldots, \Omega_K\}$ that is most probable given the observed feature vector $\mathbf{o}$!"*, or, mathematically speaking,

$$\hat{\Omega} = \underset{\{\Omega_k\}}{\operatorname{argmax}} \Big\{ P_{\breve{\Omega}|\breve{\mathbf{o}}}(\Omega_k \,|\mathbf{o}) \Big\} \tag{3.1}$$

$$= \underset{\{\Omega_k\}}{\operatorname{argmax}} \left\{ \frac{p_{\breve{\mathbf{o}}|\breve{\Omega}}(\mathbf{o}\,|\Omega_k)\, P_{\breve{\Omega}}(\Omega_k)}{p_{\breve{\mathbf{o}}}(\mathbf{o})} \right\} \tag{3.2}$$

$$= \underset{\{\Omega_k\}}{\operatorname{argmax}} \Big\{ p_{\breve{\mathbf{o}}|\breve{\Omega}}(\mathbf{o}\,|\Omega_k)\, P_{\breve{\Omega}}(\Omega_k) \Big\}. \tag{3.3}$$

The maximization is carried out over all possible realizations Ω_k, $k \in \{1, \ldots, K\}$ of the ***r**andom **v**ariable* (RV) $\breve{\Omega}$. The parameter K thereby denotes the cardinality of the set of all possible classes considered for the classification problem. Here and in the following, the overscript $\breve{(\cdot)}$ will be used whenever it is necessary to distinguish a random variable, e.g., $\breve{\Omega}$ and $\breve{\mathbf{o}}$, from its realization, e.g., Ω_k and $\mathbf{o}$.

The decomposition of the *a posteriori class probability* $P_{\breve{\Omega}|\breve{\mathbf{o}}}(\Omega_k \,|\mathbf{o})$ into the *class-conditional likelihood* $p_{\breve{\mathbf{o}}|\breve{\Omega}}(\mathbf{o}\,|\Omega_k)$, the *a priori class probability* $P_{\breve{\Omega}}(\Omega_k)$ and the *marginal likelihood* of the observed feature vector $p_{\breve{\mathbf{o}}}(\mathbf{o})$ according to Bayes' theorem does not change the decision. Neither does dropping $p_{\breve{\mathbf{o}}}(\mathbf{o})$ in (3.3).

Whether or not the decision is based on the a posteriori class probabilities (3.1) or the product (3.3) of the class-conditional likelihood and the a priori class probability is a conceptual question. Some classifiers, e.g., neural networks [35], are able to directly output the a posteriori class probabilities, while others explicitly evaluate the class-conditional likelihoods and the a priori probabilities.

The most widespread approach to automatic speech recognition, speech recognition based on HMMs, falls into the latter category. The details of HMM-based speech recognition – the one also pursued in this work – will be discussed in more detail in this chapter.

Taking the decision rule (3.1) from its generic formulation to the specific problem of speech recognition first of all calls for the identification of *feature* and *classes*.

Though in principle equivalent, the uttered and recorded speech is usually not considered to be characterized by a single feature vector $\mathbf{o}$, e.g., the waveform of the speech signal, but by a sequence of feature vectors $\mathbf{o}_{1:T} := \mathbf{o}_1, \ldots, \mathbf{o}_T$, where the individual feature vectors $\mathbf{o}_t \in \mathbb{R}^{D\times 1}$, $t \in \{1, \ldots, T\}$, are considered to be realizations of a real, vector-valued stochastic process $\breve{\mathbf{o}}_{1:T}$. The dimension D of the feature vectors and the length T of a sequence thereof depend on the applied *feature extraction scheme* and, at least the latter, on the duration of the considered utterance.

Given the sequence of feature vectors $\mathbf{o}_{1:T}$, classification, also referred to as the *decoding* in the context of speech recognition[1], shall eventually output an estimate $\hat{\mathcal{S}}$ of the sentence $\mathcal{S}$ uttered. A sentence $\mathcal{S}$ is understood as an ordered sequence $w_{1:N_w}$ of words w_n stemming from a vocabulary $\Omega = \{\omega_1, \ldots, \omega_\Upsilon\}$ of size Υ. The length of that sequence is denoted by $\pounds\left(w_{1:N_w}\right) = \pounds\left(\mathcal{S}\right) = N_w$.

Since neither the true length of a sentence $\mathcal{S}$ nor the identity of the words within a sentence $\mathcal{S}$ is known, a speech recognizer is facing the challenging task of finding the optimal solution[2] over all word sequences $w_{1:N_w}$ of all possible lengths N_w (ranging from 0 to, theoretically, ∞). Therefore, the sentence $\mathcal{S}$ is considered to be a realization of the stochastic process $\breve{\mathcal{S}}$. The decision rule (3.1) then formally turns into

$$\hat{\mathcal{S}} = \underset{\{\mathcal{S}\}}{\operatorname{argmax}} \left\{ P_{\breve{\mathcal{S}}|\breve{\mathbf{o}}_{1:T}}\left(\mathcal{S} \,|\, \mathbf{o}_{1:T}\right) \right\}. \tag{3.4}$$

By introducing the length N_w of the sentence $\mathcal{S}$ as an additional RV, the above can – applying the *law of total probability* – also be written as

$$\hat{\mathcal{S}} = \underset{\{\mathcal{S}\}}{\operatorname{argmax}} \left\{ \sum_{\{N_w\}} P_{\breve{\mathcal{S}},\breve{N}_w|\breve{\mathbf{o}}_{1:T}}\left(\mathcal{S}, N_w \,|\, \mathbf{o}_{1:T}\right) \right\} \tag{3.5}$$

$$= \underset{\{N_w, \mathcal{S}\in\{\mathcal{S}|\pounds(\mathcal{S})=N_w\}\}}{\operatorname{argmax}} \left\{ P_{\breve{\mathcal{S}},\breve{N}_w|\breve{\mathbf{o}}_{1:T}}\left(\mathcal{S}, N_w \,|\, \mathbf{o}_{1:T}\right) \right\}. \tag{3.6}$$

The last equality holds, since the joint a posteriori ***p**robability **m**ass **f**unction* (PMF) $P_{\breve{\mathcal{S}},\breve{N}_w|\breve{\mathbf{o}}_{1:T}}$ is given by

$$P_{\breve{\mathcal{S}},\breve{N}_w|\breve{\mathbf{o}}_{1:T}}\left(\mathcal{S}, N_w \,|\, \mathbf{o}_{1:T}\right) = \begin{cases} P_{\breve{\mathcal{S}},\breve{N}_w|\breve{\mathbf{o}}_{1:T}}\left(\mathcal{S}, \pounds\left(\mathcal{S}\right) |\, \mathbf{o}_{1:T}\right), & \text{if } N_w = \pounds\left(\mathcal{S}\right) \\ 0, & \text{else} \end{cases} \tag{3.7}$$

and the summation in (3.5) thus encompasses only a single summand that is non-zero. From (3.6) it can now be seen, that the optimization eventually takes into account both sentence identity and length.

Substituting the sentence $\mathcal{S}$ of length $\pounds\left(S\right)$ by the sequence of words $w_{1:N_w}$ of length $\pounds\left(w_{1:N_w}\right) = N_w$ and applying BAYES' rule once more finally results in the decision rule in

[1] According to the *Source-Channel Model of Speech Recognition* [36, p. 9], human speakers are considered to *encode* their thoughts into speech, which is then transmitted over an acoustic channel. Consequently, the recognizer's role is to *decode* the speech – possibly compensating for the acoustic channel – into the uttered sequence of words.

[2] Optimal in terms of the MAP criterion.

its more popular form

$$\hat{w}_{1:\hat{N}_w} = \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ \frac{p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(\mathbf{o}_{1:T}\left|w_{1:N_w},N_w\right.\right) P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w},N_w\right)}{p_{\breve{\mathbf{o}}_{1:T}}\left(\mathbf{o}_{1:T}\right)} \right\} \tag{3.8}$$

$$= \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(\mathbf{o}_{1:T}\left|w_{1:N_w},N_w\right.\right) P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w},N_w\right) \right\} \tag{3.9}$$

$$= \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}\left(\mathbf{o}_{1:T}\left|w_{1:N_w}\right.\right) P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w},N_w\right) \right\}, \tag{3.10}$$

where the maximization has to be carried out over all word sequences $w_{1:N_w}$ of all length N_w. The dependence of the observed sequence of feature vectors $\mathbf{o}_{1:T}$ on the length N_w and the word sequence $w_{1:N_w}$, itself, has thereby been replaced by the latter.

In the context of speech recognition, the conditional PDF $p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}$, from which the likelihood $p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}\left(\mathbf{o}_{1:T}\left|w_{1:N_w}\right.\right)$ of the observed feature vector sequence $\mathbf{o}_{1:T}$ given the hypothesized word sequence $w_{1:N_w}$ is obtained, is denoted by the *acoustic model*. It statistically relates the word sequence $w_{1:N_w}$ to its acoustic realization, the observed feature vector sequence $\mathbf{o}_{1:T}$. The joint PMF $P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}$, from which the a priori probability $P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w},N_w\right)$ of the hypothesized word sequence $w_{1:N_w}$ of length N_w is computed, is termed the (statistical) *language model* since it describes the linguistic content of the speech uttered.

Consequently, the likelihood $p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}\left(\mathbf{o}_{1:T}\left|w_{1:N_w}\right.\right)$ and the probability $P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w},N_w\right)$ are denoted *acoustic score* and *language score*, respectively.

The conceptual architecture of an automatic speech recognition system as it is described above is given in Fig. 3.1. The building blocks, starting with the conversion of the speech signal into the sequence of feature vectors $\mathbf{o}_{1:T}$, will be discussed next. The *lexicon*, which has not been introduced yet, specifies the set of words that may appear in the data to be recognized. However, unlike the vocabulary, it also contains the transcription of each word in terms of *sub-word* units. This modeling detail will be discussed at the end of Section 3.2.

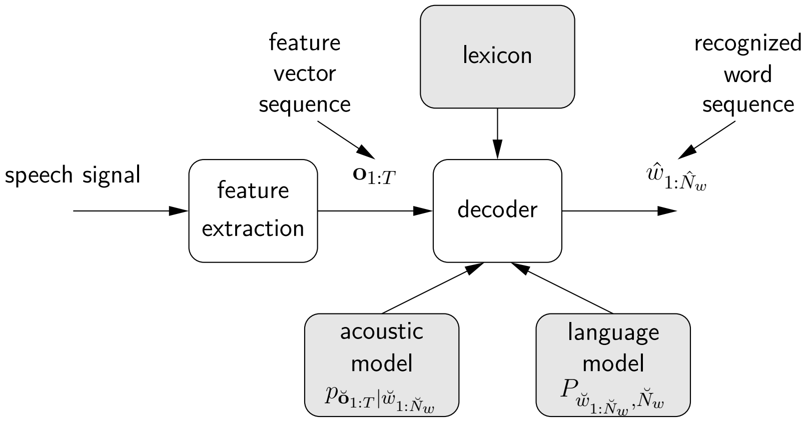


Figure 3.1: *Conceptual architecture of an automatic speech recognition system in a statistical framework.*

3.1 Feature Extraction

Feature extraction, also referred to as the ***f**ront-**e**nd* (FE) of the speech recognition system, is concerned with finding a *compact* representation $\mathbf{o}_{1:T}$ – the sequence of feature vectors – of the (discrete-time) acoustic signal $o^{(\mathsf{MIC})}_{1:\tilde{T}}$ of length $\tilde{T}$ to be used for recognition.

The term *compact* thereby subsumes a variety of desirable properties. The features to be extracted shall, e.g., preserve that part of information that is perceptually important for the distinction of linguistic units while at the same time being insensitive to acoustic variations w.r.t. irrelevant information [37, p. 159]. These irrelevant information also comprise distortions due to the acoustic environment, e.g., reverberation and noise.

In practice, the most successful feature extraction schemes are motivated, at least in part, by the human auditory system. Among those is the extraction employing ***p**erceptual **l**inear **p**rediction* (PLP) analysis of speech [38] (later on ***re**lative **spe**c**tra*** (RASTA)-PLP analysis of speech [39]) and the extraction of the so-called ***m**el **f**requency **c**esptral **c**oefficients* (MFCCs) [40].

The latter is taken as the basis of the considerations following in Sec. 4.3 and hence will be discussed in more detail here. The process of extracting the mel frequency cepstral feature vectors has been standardized by the ***E**uropean **T**elecommunication **S**tandards **I**nstitute* (ETSI) in [41] and is depicted in Fig. 3.2. The input signal captured by the microphone is first converted from its analog to a digital representation by the ***a**nalog-to-**d**igital **c**onversion* (ADC) block. The samples of the discrete-time output sequence are thereby obtained at the sampling rate f_S (equivalent: the sampling period $T_S = f_S^{-1}$) which may take one out of three specified values, i.e., $f_S \in \{8\,\text{kHz}, 11\,\text{kHz}, 16\,\text{kHz}\}$.

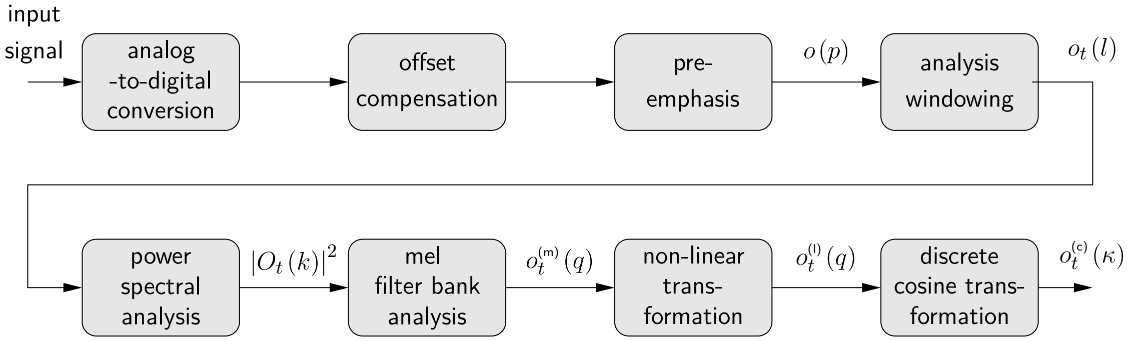


Figure 3.2: *MFCC feature extraction scheme following the ETSI standard [41]*

The offset compensation block then removes any constant offset from the discrete-time input signal, which is further passed to the pre-emphasis block. The pre-emphasis is primarily meant to compensate for the 6 dB/octave attenuation of the lower frequency region in the spectrum of voiced speech [37], however, is applied to the offset-compensated signal irrespective of the voicedness of individual speech segments.

The resulting signal $o(p)$ is than decomposed into a sequence of T overlapping frames by the use of an analysis window $w_{\mathsf{A}}(l) \in \mathbb{R}$ with support on $l \in \{0, L_{w_{\mathsf{A}}} - 1\}$, where the length $L_{w_{\mathsf{A}}} \in \mathbb{N}_{>0}$ of the analysis window (and hence the length of each frame) is chosen such that the signal within a frame may be considered approximately stationary. The shift $B \in \mathbb{N}_{>0}$ of the analysis window is thereby chosen to be lower than $L_{w_{\mathsf{A}}}$ to at least capture a glottal cycle[3] once in the center of a frame. The signal within frame t may thus be written as

$$o_t(l) := w_{\mathsf{A}}(l)\,o(l+tB), \tag{3.11}$$

with index $l \in \{0, L_{w_{\mathsf{A}}}-1\}$.

Each frame is then transformed into the frequency domain by application of the ***d**iscrete* **F**OURIER ***t**ransformation* (DFT), resulting in the ***s**pectral **c**oefficients* (SCs)

$$O_t(k) = \sum_{l=0}^{L_{w_{\mathsf{A}}}-1} o_t(l)\,\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk}, \tag{3.12}$$

where $k \in \{0, K-1\}$ denotes the frequency bin, K the total number of frequency bins and j the imaginary unit. In both ETSI standards [41] and [42], any phase information, which has been found to play only a minor role in human perception of speech [43, p. 52], is discarded by computing either the magnitude spectrum [41] or the power spectrum [42] of the current frame.

Following [42], the ***p**ower **s**pectral **c**oefficients* (PSCs) $|O_t(k)|^2$ are further analyzed in a bank of Q triangular-shaped mel filters $\boldsymbol{\Lambda}_q$, $q \in \{0, Q-1\}$. These filters eventually emulate the human auditory system, since they i) follow the human perception of sound intensity by smoothing the spectrum within critical bands and ii) thereby take into account the non-linear frequency resolution of the human ear by spacing the critical bands linearly on the mel scale. The bandwidth of each individual mel filter in the linear frequency domain is obtained by *warping* adjacent center frequencies in the mel domain back to the linear domain. Denoting the lower and upper cutoff frequency indices for the qth mel filter by $K_q^{(\mathsf{low})}$ and $K_q^{(\mathsf{up})}$, respectively, this bandwidth is given by $K_q^{(\mathsf{up})} - K_q^{(\mathsf{low})} + 1$ and the MPSCs are defined by

$$o_t^{(\mathsf{m})}(q) := \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\,|O_t(k)|^2. \tag{3.13}$$

The human perception of sound intensity also motivates the subsequent non-linear transformation of these coefficients by the natural logarithm, leading to the so-called LMPSCs defined by

$$o_t^{(\mathsf{l})}(q) := \ln\left(o_t^{(\mathsf{m})}(q)\right), \tag{3.14}$$

eventually making up the LMPSC feature vector $\mathbf{o}_t^{(\mathsf{l})} \in \mathbb{R}^{Q\times 1}$

$$\mathbf{o}_t^{(\mathsf{l})} := \left[o_t^{(\mathsf{l})}(0), \ldots, o_t^{(\mathsf{l})}(Q-1)\right]^{\dagger}. \tag{3.15}$$

Thereby $(.)^{\dagger}$ denotes the matrix/vector transpose operator. Due to the overlap of adjacent mel filters, the components (3.14) of the feature vector in the logarithmic mel power

[3] The glottal cycle describes the opening and closing of the vocal cords eventually producing the air pressure variations required for voiced speech.

spectrum are highly correlated with each other. However, the correlation matrix, which approximately exhibits a TOEPLITZ structure, is almost diagonalized by the following ***d**iscrete **c**osine **t**ransformation* (DCT), which is equivalent to the DFT for real-valued, even signals. The transformation from the logarithmic mel power spectrum to the so-called mel *cepstrum* (an anagram of *spectrum*) is given by

$$o_t^{(\mathsf{c})}(\kappa) := \sum_{q=0}^{Q-1} o_t^{(\mathsf{l})}(q) \cos\left(\frac{\pi}{\mathcal{K}}\kappa\left[q+\frac{1}{2}\right]\right), \tag{3.16}$$

with $\kappa \in \{0, \mathcal{K}-1\}$, where the number $\mathcal{K}$ is usually chosen to be smaller than the number Q of LMPSCs (see Tab. 3.1). The corresponding MFCC feature vector $\mathbf{o}_t^{(\mathsf{c})} \in \mathbb{R}^{\mathcal{K}\times 1}$ is defined by

$$\mathbf{o}_t^{(\mathsf{c})} := \left[o_t^{(\mathsf{c})}(0), \ldots, o_t^{(\mathsf{c})}(\mathcal{K}-1)\right]^{\dagger}. \tag{3.17}$$

Note that $\mathbf{o}_t^{(\mathsf{c})}$ may also be written in terms of $\mathbf{o}_t^{(\mathsf{l})}$ by means of the linear transformation

$$\mathbf{o}_t^{(\mathsf{c})} = \mathbf{C}_{\mathsf{DCT}}\mathbf{o}_t^{(\mathsf{l})}, \tag{3.18}$$

where $\mathbf{C}_{\mathsf{DCT}} \in \mathbb{R}^{\mathcal{K}\times Q}$ denotes the DCT matrix with the element in the κth row and qth column given by the cosine term in (3.16). Choosing $\mathcal{K}$ to be lower than Q in (3.16) is motivated by the *source-filter model of speech production* [44, p. 16ff], according to which the speech production process can be characterized by the linear convolution of an excitation signal and a time-variant filter representing the resonance structure of the vocal tract. Compared to the excitation signal, the impulse response of the vocal tract, which is assumed to carry the important information on the currently uttered speech sound, varies only slowly with time. Hence, relevant information w.r.t. the vocal tract can majorly be found at lower cepstral coefficients, while the characteristics of the excitation signal appear at higher cepstral coefficients. Note, that the radiation characteristic of the lips has already been removed, at least in part, by the pre-emphasis filter. The computation of the MFCCs according to (3.16) thus not only reduces the correlation between the components of the resulting MFCC feature vector (3.17) but also separates the information on the vocal tract from that on the excitation.

While the DCT is capable of reducing the correlation of the MFCCs within an MFCC feature vector (referred to as *intra-frame* correlation), those feature vectors themselves are highly correlated over time (referred to as *inter-frame* correlation), which may majorly be attributed to the overlap of adjacent analysis windows.

If the acoustic model would be able to accurately model these inter-frame dependencies, the MFCC feature vectors could directly be passed to the recognizer without any further preparation. However, recognizers employing the acoustic model introduced in Section 3.2, usually benefit from the extension of the feature vector by so-called *dynamic* features. In contrast to features purely providing information on the current frame (so-called *static* features), dynamic features provide information on the context of the current frame. The most prominent dynamic features consist of approximations to the first and second order derivative of the MFCCs, resulting in so-called *delta* (Δ) and *delta-delta* ($\Delta\Delta$) features

[45], defined by

$$\Delta o_t^{(\mathsf{c})}(\kappa) := \frac{\sum\limits_{i=-L_\Delta}^{L_\Delta} i \cdot o_{t+i}^{(\mathsf{c})}(\kappa)}{\sum\limits_{i=-L_\Delta}^{L_\Delta} i^2}, \tag{3.19}$$

$$\Delta\Delta o_t^{(\mathsf{c})}(\kappa) := \frac{\sum\limits_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} j \cdot \Delta o_{t+j}^{(\mathsf{c})}(\kappa)}{\sum\limits_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} j^2} = \frac{\sum\limits_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum\limits_{i=-L_\Delta}^{L_\Delta} ji \cdot o_{t+(j+i)}^{(\mathsf{c})}(\kappa)}{\sum\limits_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum\limits_{i=-L_\Delta}^{L_\Delta} (ji)^2}, \tag{3.20}$$

where $L_\Delta \in \mathbb{N}_{>0}$ and $L_{\Delta\Delta} \in \mathbb{N}_{>0}$ specify the extent to which past and future MFCCs are considered for the computation of $\Delta o_t^{(\mathsf{c})}(\kappa)$ and $\Delta\Delta o_t^{(\mathsf{c})}(\kappa)$, respectively. Eventually, the feature vector $\mathbf{o}_t \in \mathbb{R}^{3\mathcal{K}\times 1}$ subject to be used for speech recognition is composed by concatenating the static MFCCs, the delta MFCCs and the delta-delta MFCCs as

$$\mathbf{o}_t := \left[\left(\mathbf{o}_t^{(\mathsf{c})}\right)^\dagger, \left(\Delta\mathbf{o}_t^{(\mathsf{c})}\right)^\dagger, \left(\Delta\Delta\mathbf{o}_t^{(\mathsf{c})}\right)^\dagger\right]^\dagger, \tag{3.21}$$

with $\Delta\mathbf{o}_t^{(\mathsf{c})} \in \mathbb{R}^{\mathcal{K}\times 1}$ and $\Delta\Delta\mathbf{o}_t^{(\mathsf{c})} \in \mathbb{R}^{\mathcal{K}\times 1}$ defined by

$$\Delta\mathbf{o}_t^{(\mathsf{c})} := \left[\Delta o_t^{(\mathsf{c})}(0), \ldots, \Delta o_t^{(\mathsf{c})}(\mathcal{K}-1)\right]^\dagger, \tag{3.22}$$

$$\Delta\Delta\mathbf{o}_t^{(\mathsf{c})} := \left[\Delta\Delta o_t^{(\mathsf{c})}(0), \ldots, \Delta\Delta o_t^{(\mathsf{c})}(\mathcal{K}-1)\right]^\dagger. \tag{3.23}$$

The formation of the feature vector $\mathbf{o}_t$ can also be considered as the linear transformation of the super-vector consisting of the MFCC feature vectors $\mathbf{o}_{t-(L_\Delta+L_{\Delta\Delta})}^{(\mathsf{c})}, \ldots, \mathbf{o}_{t+(L_\Delta+L_{\Delta\Delta})}^{(\mathsf{c})}$, i.e.,

$$\mathbf{o}_t = \underbrace{\begin{bmatrix} \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta\Delta}} & \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_\Delta} & \mathbf{I}_{\mathcal{K}\times\mathcal{K}} & \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_\Delta} & \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta\Delta}} \\ \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta\Delta}} & \reflectbox{$\mathbf{B}$}_\Delta & \mathbf{0}_{\mathcal{K}\times\mathcal{K}} & -\mathbf{B}_\Delta & \mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta\Delta}} \\ \reflectbox{$\mathbf{D}$}_{\Delta\Delta} & \reflectbox{$\mathbf{C}$}_{\Delta\Delta} & \mathbf{A}_{\Delta\Delta} & \mathbf{C}_{\Delta\Delta} & \mathbf{D}_{\Delta\Delta} \end{bmatrix}}_{:=\mathcal{T}_{L_\Delta,L_{\Delta\Delta}}^{(\mathsf{c})}} \begin{bmatrix} \mathbf{o}_{t+(L_\Delta+L_{\Delta\Delta})}^{(\mathsf{c})} \\ \vdots \\ \mathbf{o}_{t-(L_\Delta+L_{\Delta\Delta})}^{(\mathsf{c})} \end{bmatrix}, \tag{3.24}$$

where

$$\mathbf{B}_\Delta := \begin{bmatrix} \mathbf{I}_{\mathcal{K}\times\mathcal{K}} & 2\mathbf{I}_{\mathcal{K}\times\mathcal{K}} & \cdots & (L_\Delta-1)\,\mathbf{I}_{\mathcal{K}\times\mathcal{K}} & L_\Delta\mathbf{I}_{\mathcal{K}\times\mathcal{K}} \end{bmatrix}, \tag{3.25}$$

$$\mathbf{A}_{\Delta\Delta} := \left(\sum_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum_{\substack{i=-L_\Delta \\ \{i|j+i=0\}}}^{L_\Delta} ji\right) \mathbf{I}_{\mathcal{K}\times\mathcal{K}}, \tag{3.26}$$

$$\mathbf{C}_{\Delta\Delta} := \left[\left(\sum_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum_{\substack{i=-L_\Delta \\ \{i|j+i=-1\}}}^{L_\Delta} ji\right) \mathbf{I}_{\mathcal{K}\times\mathcal{K}} \quad \cdots \quad \left(\sum_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum_{\substack{i=-L_\Delta \\ \{i|j+i=-L_\Delta\}}}^{L_\Delta} ji\right) \mathbf{I}_{\mathcal{K}\times\mathcal{K}}\right], \tag{3.27}$$

$$\mathbf{D}_{\Delta\Delta} := \left[\left(\sum_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum_{\substack{i=-L_\Delta \\ \{i|j+i=-(L_\Delta+1)\}}}^{L_\Delta} ji\right) \mathbf{I}_{\mathcal{K}\times\mathcal{K}} \quad \cdots \quad \left(\sum_{j=-L_{\Delta\Delta}}^{L_{\Delta\Delta}} \sum_{\substack{i=-L_\Delta \\ \{i|j+i=-(L_\Delta+L_{\Delta\Delta})\}}}^{L_\Delta} ji\right) \mathbf{I}_{\mathcal{K}\times\mathcal{K}}\right]. \tag{3.28}$$

The matrix $\reflectbox{$\mathbf{B}$}_{\Delta}$ can be obtained from the matrix $\mathbf{B}_{\Delta}$ by reversing the order of the involved block matrices, i.e.,

$$\reflectbox{$\mathbf{B}$}_{\Delta} := \begin{bmatrix} L_{\Delta}\mathbf{I}_{\mathcal{K}\times\mathcal{K}} & (L_{\Delta}-1)\,\mathbf{I}_{\mathcal{K}\times\mathcal{K}} & \cdots & 2\mathbf{I}_{\mathcal{K}\times\mathcal{K}} & \mathbf{I}_{\mathcal{K}\times\mathcal{K}} \end{bmatrix}. \tag{3.29}$$

The same operation creates the matrices $\reflectbox{$\mathbf{C}$}_{\Delta\Delta}$ and $\reflectbox{$\mathbf{D}$}_{\Delta}$ from the matrices $\mathbf{C}_{\Delta\Delta}$ and $\mathbf{D}_{\Delta\Delta}$, respectively.

Note that the vector $\mathbf{o}_t$ may also directly be obtained from the super-vector composed of the LMPSC feature vectors within the sequence $\mathbf{o}^{(\mathsf{l})}_{t-(L_{\Delta}+L_{\Delta\Delta})},\ldots,\mathbf{o}^{(\mathsf{l})}_{t+(L_{\Delta}+L_{\Delta\Delta})}$ by means of a single linear transformation, i.e.,

$$\mathbf{o}_t = \mathcal{T}^{(\mathsf{l})}_{L_{\Delta},L_{\Delta\Delta}} \begin{bmatrix} \mathbf{o}^{(\mathsf{l})}_{t+(L_{\Delta}+L_{\Delta\Delta})} \\ \vdots \\ \mathbf{o}^{(\mathsf{l})}_{t-(L_{\Delta}+L_{\Delta\Delta})} \end{bmatrix}, \tag{3.30}$$

where the matrix $\mathcal{T}^{(\mathsf{l})}_{L_{\Delta},L_{\Delta\Delta}}$ can easily be obtained from the matrix $\mathcal{T}^{(\mathsf{c})}_{L_{\Delta},L_{\Delta\Delta}}$ defined in (3.24) by replacing the zero matrices $\mathbf{0}_{\mathcal{K}\times\mathcal{K}}$, $\mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta}}$ and $\mathbf{0}_{\mathcal{K}\times\mathcal{K}L_{\Delta\Delta}}$ by the corresponding zero matrices $\mathbf{0}_{Q\times Q}$, $\mathbf{0}_{Q\times QL_{\Delta}}$ and $\mathbf{0}_{Q\times QL_{\Delta\Delta}}$ while also substituting the DCT matrix $\mathbf{C}_{\mathsf{DCT}}$ for the identity matrix $\mathbf{I}_{\mathcal{K}\times\mathcal{K}}$ in (3.24)–(3.28).

Further note that the extension of the static feature vectors by dynamic components is not part of the ETSI standard [41] and only briefly described in [42].

Finally, typical front-end parameters are given in Tab. 3.1 for $T_S^{-1} \in \{8\,\text{kHz}, 16\,\text{kHz}\}$.

Table 3.1: *Parameters for feature extraction according to [41].*

Sampling	Frame	Frame	FFT	#LMPSCs	#MFCCs		
rate T_S^{-1}	shift B	length L_{w_A}	length K	Q	$\mathcal{K}$	L_{Δ}	$L_{\Delta\Delta}$
8 kHz	80	200	256	23	13	3	2
16 kHz	160	400	512	23	13	3	2

3.2 Acoustic Modeling

The acoustic model introduced in (3.10) statistically relates the observed sequence of feature vectors $\mathbf{o}_{1:T}$ to the underlying sequence of words $w_{1:N_w}$. In HMM-based speech recognition, the conditional PDF $p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}$ is obtained by introducing a sequence of underlying (hidden) discrete-valued states $q_{1:T}$ assumed to be realizations of a stochastic process $\breve{q}_{1:T}$, where $q_t \in \{1,\ldots,I\}$ with I denoting the total number of HMM states. Repeatedly applying

BAYES' theorem yields[4]

$$p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_{1:T}|w_{1:N_w}) = \sum_{\{q_{1:T}\}} p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}|\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_{1:T}, q_{1:T}|w_{1:N_w}) \tag{3.31}$$

$$= \sum_{\{q_{1:T}\}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t,\breve{q}_t|\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t-1},\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_t, q_t|\mathbf{o}_{1:t-1}, q_{1:t-1}, w_{1:N_w}) \tag{3.32}$$

$$= \sum_{\{q_{1:T}\}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t},\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_t|\mathbf{o}_{1:t-1}, q_{1:t}, w_{1:N_w}) \\ P_{\breve{q}_t|\breve{q}_{1:t-1},\breve{\mathbf{o}}_{1:t-1},\breve{w}_{1:\breve{N}_w}}(q_t|q_{1:t-1}, \mathbf{o}_{1:t-1}, w_{1:N_w}), \tag{3.33}$$

where the summation has to be carried out over all state sequences $q_{1:T}$.

Given the hypothesized word sequence $w_{1:N_w}$, only a certain subset of state sequences $q_{1:T}$ will have a non-zero probability. Defining this subset by

$$\Phi_{w_{1:N_w}} := \left\{ q_{1:T} \middle| P_{\breve{q}_{1:T}|\breve{w}_{1:\breve{N}_w}}(q_{1:T}|w_{1:N_w}) \neq 0 \right\} \tag{3.34}$$

allows to write (3.33) in its notationally more convenient form of

$$p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_{1:T}|w_{1:N_w}) \tag{3.35}$$

$$= \sum_{\substack{q_{1:T}\in \\ \Phi_{w_{1:N_w}}}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t}}(\mathbf{o}_t|\mathbf{o}_{1:t-1}, q_{1:t}) \, P_{\breve{q}_t|\breve{q}_{1:t-1},\breve{\mathbf{o}}_{1:t-1}}(q_t|q_{1:t-1}, \mathbf{o}_{1:t-1}).$$

W.l.o.g, the last algebraic conversion implies that the sequence of HMM states $q_{1:T}$ uniquely characterizes the word sequence $w_{1:N_w}$, i.e., there is no other word sequence with the same state sequence of length T. For a given sequence of words $w_{1:N_w}$ the statistical dependencies between the involved RVs according to (3.35) are depicted in Fig. 3.3a in terms of a ***dynamic* BAYESIAN *network*** (DBN) [46]. For ease of visualization, the past sequence of states and the past sequence of observations are kept in so-called *history* nodes [47, pp. 309f.].

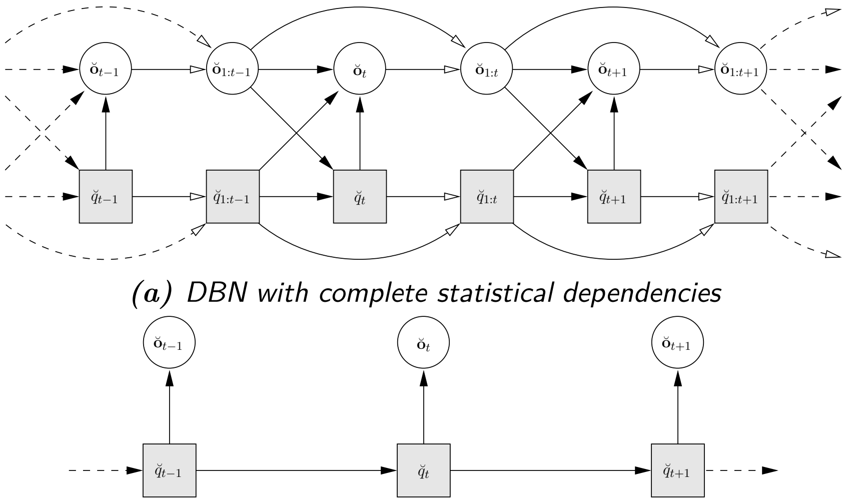
The power of HMM-based speech recognition now lies in two assumptions posed on the involved PDF/PMF [48, p. 322 and references therein]: The first assumption made, known as the *conditional independence* assumption, states that the current observation $\mathbf{o}_t$ is independent of all past observations $\mathbf{o}_{1:t-1}$ and all past states $q_{1:t-1}$ once the current state q_t is given, or, mathematically speaking

$$p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t}}(\mathbf{o}_t|\mathbf{o}_{1:t-1}, q_{1:t}) \approx p_{\breve{\mathbf{o}}_t|\breve{q}_t}(\mathbf{o}_t|q_t). \tag{3.36}$$

The second assumption made models the stochastic process $\breve{q}_{1:T}$ a first-order MARKOV process, i.e.,

$$P_{\breve{q}_t|\breve{q}_{1:t-1},\breve{\mathbf{o}}_{1:t-1}}(q_t|q_{1:t-1}, \mathbf{o}_{1:t-1}) \approx P_{\breve{q}_t|\breve{q}_{t-1}}(q_t|q_{t-1}). \tag{3.37}$$

[4] Here and in the following, the notation $p_{\breve{\mathbf{o}}_{1:T}}(\mathbf{o}_{1:T}) = \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}(\mathbf{o}_t|\mathbf{o}_{1:t-1})$ and alike shall be understood as $p_{\breve{\mathbf{o}}_{1:T}}(\mathbf{o}_{1:T}) = p_{\breve{\mathbf{o}}_1}(\mathbf{o}_1) \prod_{t=2}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}(\mathbf{o}_t|\mathbf{o}_{1:t-1})$ and is introduced for ease of notation.



$\breve{\mathbf{o}}_{t-1}$
$\breve{\mathbf{o}}_{1:t-1}$
$\breve{\mathbf{o}}_{t}$
$\breve{\mathbf{o}}_{1:t}$
$\breve{\mathbf{o}}_{t+1}$
$\breve{\mathbf{o}}_{1:t+1}$
$\breve{q}_{t-1}$
$\breve{q}_{1:t-1}$
$\breve{q}_{t}$
$\breve{q}_{1:t}$
$\breve{q}_{t+1}$
$\breve{q}_{1:t+1}$

(a) *DBN with complete statistical dependencies*

(b) *DBN with statistical dependencies according to the HMM approximations*

Figure 3.3: *The DBNs for speech recognition: Continuous RVs are depicted by circles, discrete ones by squares. The unobservable (hidden) RVs are highlighted in gray. The statistical dependencies between RVs are indicated by black-headed arrows. Deterministic dependencies are indicated by white-headed arrows. The DBN in (a) depicts the complete statistical dependencies after* (3.35)*, the DBN in (b) the statistical dependencies employing the approximations* (3.36) *and* (3.37)*.*

From (3.36) and (3.37) it can be seen, that direct statistical dependencies between consecutive feature vectors within a sequence are modeled only by means of the statistical dependencies between adjacent HMM states. The resulting simplified DBN is given in Fig. 3.3b.

These approximations, and in particular the conditional independence assumption, are commonly cited to be the major limitation of HMMs in ASR and many approaches to overcome these short-comings have been proposed, see for instance [49, 50, 51, 52]. Nevertheless, HMM-based speech recognition employing (3.36) and (3.37) is still the most prominent approach to speech recognition in a statistical framework.

Having said this, Eq. (3.33) turns into

$$p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}(\mathbf{o}_{1:T}|w_{1:N_w}) \approx \sum_{\substack{q_{1:T}\in\\ \Phi_{w_{1:N_w}}}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{q}_t}(\mathbf{o}_t|q_t)\, P_{\breve{q}_t|\breve{q}_{t-1}}(q_t|q_{t-1})\,. \tag{3.38}$$

The conditional PMF $P_{\breve{q}_t|\breve{q}_{t-1}}$ gives the probabilities of, e.g., transiting from HMM state $q_{t-1}=h$ at time instant $t-1$ to HMM state $q_t=i$ at time instant t. These probabilities are denoted as *initial state probabilities* $\pi_i \in [0,1]$ for $t=1$ and as *state transition probabilities* $a_{i|h} \in [0,1]$ for $t>1$, i.e.,

$$\pi_i := P_{\breve{q}_1}(q_1=i)\,, \qquad \text{for } t=1 \tag{3.39}$$

$$a_{i|h} := P_{\breve{q}_t|\breve{q}_{t-1}}(q_t=i|q_{t-1}=h)\,, \qquad \forall t>1 \tag{3.40}$$

subject to $\sum_{i=1}^{I}\pi_i=1$ and $\sum_{i=1}^{I}a_{i|h}=1\ \forall h\in\{1,\ldots,I\}$. As the notation already indicates, all transition probabilities $a_{h,i}$ are considered independent of the time the transition between states actually takes place in the HMM.

The conditional PDF $p_{\breve{\mathbf{o}}_t|\breve{q}_t}$, also termed the *observation density*, is usually modeled as a mixture of elementary PDFs. With an appropriate choice of the shape of the elementary PDF, *mixture densities* are capable of approximating, arbitrarily closely, any continuous-valued PDF [48, p. 350],[53]. Among the elementary densities, the GAUSSIAN density is probably the most widespread one with mixtures thereof commonly known as ***G**AUSSIAN **m**ixture **m**odels* (GMMs) or ***m**ixtures **o**f **G**AUSSIANs* (MOGs). With $q_t = i$, the mixture index $b_t \in \{1,\ldots,J(i)\}$ is formally introduced into the state-conditioned PDF $p_{\breve{\mathbf{o}}_t|\breve{q}_t}$ as the RV $\breve{b}_t$

$$p_{\breve{\mathbf{o}}_t|\breve{q}_t}(\mathbf{o}_t\,|q_t = i) = \sum_{j=1}^{J(i)} P_{\breve{b}_t|\breve{q}_t}(b_t = j\,|q_t = i)\, p_{\breve{\mathbf{o}}_t|\breve{q}_t,\breve{b}_t}(\mathbf{o}_t\,|q_t = i, b_t = j)\,, \tag{3.41}$$

where the total number of mixture components used to model the current state's observation PDF is given by $J(i)$.

The state-conditioned PMF $P_{\breve{b}_t|\breve{q}_t}$ gives the probability of, e.g., being in mixture $b_t = j$ at the current time instant t given the current HMM state $q_t = i$. These *mixture weights* will be denoted by $c_{j|i} \in [0,1]$, i.e.,

$$c_{j|i} := P_{\breve{b}_t|\breve{q}_t}(b_t = j\,|q_t = i)\,, \tag{3.42}$$

satisfying $\sum_{j=1}^{J(i)} c_{j|i} = 1\ \forall k \in \{1,\ldots,I\}$.

For the GAUSSIAN PDF, $p_{\breve{\mathbf{o}}_t|\breve{q}_t,\breve{b}_t}$ is given by

$$p_{\breve{\mathbf{o}}_t|\breve{q}_t,\breve{b}_t}(\mathbf{o}_t\,|q_t = i, b_t = j) = \frac{1}{\sqrt{(2\pi)^D \left|\boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}\right|}} \mathrm{e}^{-\frac{1}{2}\left(\mathbf{o}_t - \boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}\right)^{\dagger} \boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}^{-1}\left(\mathbf{o}_t - \boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}\right)} \tag{3.43}$$

$$=: \mathcal{N}\left(\mathbf{o}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}\right) \tag{3.44}$$

with mean vector $\boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}$. Here, $|\cdot|$ and $(\cdot)^{-1}$ are the determinant of a square matrix and the inverse of a regular matrix, respectively. As with the transition probabilities, mixture weights $c_{j|i}$, means $\boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}$ and covariances $\boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}$ are considered independent of the actual time instant of being in mixture j of state i.

For a given sequence of words $w_{1:N_w}$ the total set of involved RVs and their statistical dependencies in an HMM are depicted in Fig. 3.4 in terms of a DBN [46]. As (3.41) and Fig. 3.4 may already suggest, HMMs with mixture observation densities may also be interpreted as modified HMMs with elementary observation densities [54].

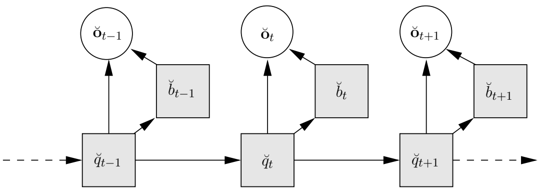


Figure 3.4: *The DBN for continuous mixture HMM-based speech recognition.*

The set of acoustic model parameters, consisting of the initial HMM state probabilities π_i, HMM state transition probabilities $a_{i|h}$ and the parameters of the PDF $p_{\breve{\mathbf{o}}_t|\breve{q}_t}$, namely the GMM weights $c_{j|i}$, mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{o}}|i,j}$ and covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{o}}|i,j}$, will be denoted by $\Theta_{\breve{\mathbf{o}}}^{\mathsf{HMM}}$ and is usually trained by applying the ***e**xpectation **m**aximization* algorithm (EM algorithm) to some *labeled* training data [48].

Since it is impossible to train large HMMs for all possible sequences of words uttered, the HMMs for sequences of words are decomposed into smaller HMMs, each modeling, e.g., one word in the sequence. The single word HMMs are then *glued* together by the language model,which will be described in more detail in Section 3.3. But, even for a moderate vocabulary size Υ, the relative occurrence of each word in a training corpus may be too low to *reliably* train the corresponding HMM parameters. Further decomposition of this *whole-word* HMMs into HMMs for sub-word units, e.g., (context-independent) *monophones* or (context-dependent) *di-* and *triphones* may alleviate this problem to a certain extent (see [48] for a detailed discusssion).

3.3 Language Modeling

The (statistical) language model $P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}$ in (3.10) weights each hypothesized word sequences $w_{1:N_w}$ of length N_w by the probability of its natural occurrence in a language.

Aiming at the practical incorporation of the language model into an automatic speech recognition system, Bayes' theorem is repeatedly applied to the joint PMF $P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}$ to yield

$$P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}(w_{1:N_w}, N_w) = P_{\breve{N}_w|\breve{w}_{1:\breve{N}_w}}(N_w|w_{1:N_w})\, P_{\breve{N}_w}(w_{1:N_w}) \tag{3.45}$$

$$= \delta_{N_w - \pounds(w_{1:N_w})} \prod_{n=1}^{N_w} P_{\breve{w}_n|\breve{w}_{1:n-1}}(w_n|w_{1:n-1})\,. \tag{3.46}$$

where $\delta_{(\cdot)}$ denotes the Kronecker-delta, defined for discrete p by

$$\delta_p = \begin{cases} 1, & \text{if } p = 0 \\ 0, & \text{else} \end{cases}. \tag{3.47}$$

Note that the formulation of the decoding rule (3.10) always ensures $N_w = \pounds(w_{1:N_w})$ and the Kronecker-delta also always evaluates to one and as such might be dropped when incorporating the language model into it.

Eq. (3.46) is then further simplified by introducing the N-*gram* approximation [55]

$$P_{\breve{w}_n|\breve{w}_{1:n-1}}(w_n|w_{1:n-1}) \approx P_{\breve{w}_n|\breve{w}_{n-N+1:n-1}}(w_n|w_{n-N+1:n-1})\,, \tag{3.48}$$

i.e., by taking only $N-1$ past words into account to approximate the PMF $P_{\breve{w}_n|\breve{w}_{1:n-1}}$.

The final language model is thus usually approximated by

$$P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}(w_{1:N_w}, N_w) \approx \delta_{N_w - \pounds(w_{1:N_w})} \left(\prod_{n=1}^{N_w} P_{\breve{w}_n|\breve{w}_{n-N+1:n-1}}(w_n|w_{n-N+1:n-1}) \right)^{\alpha_{\mathsf{LMS}}} \tag{3.49}$$

$$=: \delta_{N_w - \pounds(w_{1:N_w})}\, \tilde{P}^{\alpha_{\mathsf{LMS}}}_{\breve{w}_{1:\breve{N}_w}}(w_{1:N_w}) \tag{3.50}$$

where the *language mode scale* factor $\alpha_{\mathsf{LMS}} \in \mathbb{R}$ is introduced to further compensate for the *locality* of the N-gram approximation [56]. The language model scale factor is usually set to $\alpha_{\mathsf{LMS}} > 1$, eventually increasing the probabilities for more probable word sequences, while decreasing the probability for less probable ones. This eventually amounts to emphasizing the contribution of the language model over the one from the acoustic model in the final decision rule. Optimal values can for instance be determined experimentally by minimizing the word error rate on some development data.

The importance of balancing the contribution of acoustic and linguistic information to the decision process not only by means of *tuning* the aforementioned language model scale factor α_{LMS} but also by accounting for the sentence length N_w is widely accepted though is usually introduced as a heuristic [48, p. 454], [37, p. 206].

A common approach to address this is the so-called ***w**ord **i**nsertion **p**enalty* (WIP), which is typically applied in the logarithmic probability space: By scaling the length N_w of the currently hypothesized word sequence $w_{1:N_w}$ by the so-called word insertion penalty factor $\alpha_{\mathsf{WIP}} \in \mathbb{R}_{<0}$ and adding it to the logarithmic contribution of the acoustic model $p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}$ and the N-gram language model $\tilde{P}^{\alpha_{\mathsf{LMS}}}_{\breve{w}_{1:\breve{N}_w}}$ the rate at which words are either inserted or deleted during recognition can be controlled [48, p. 454], [57, p. 610, Ch. 12.2.2].

3.4 Decoding

The decoder, also referred to as the *back-end* of the speech recognition system, eventually is assigned the task to find the most probable word sequence $\hat{w}_{1:\hat{N}_w}$ out of all possible word sequences given the observed feature vector sequence $\mathbf{o}_{1:T}$. With the approximations (3.38) and (3.49) applied to the acoustic model and the language model, respectively, the decision rule (3.10) turns into[5]

$$\hat{w}_{1:\hat{N}_w} = \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}}\left(\mathbf{o}_{1:T}\,|w_{1:N_w}\right) P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w}, N_w\right) \right\} \tag{3.51}$$

$$\approx \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ \left(\sum_{\substack{q_{1:T}\in \\ \Phi_{w_{1:N_w}}}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{q}_t}\left(\mathbf{o}_t\,|q_t\right) P_{\breve{q}_t|\breve{q}_{t-1}}\left(q_t\,|q_{t-1}\right) \right) \tilde{P}^{\alpha_{\mathsf{LMS}}}_{\breve{w}_{1:\breve{N}_w}}\left(w_{1:N_w}\right) \right\} \tag{3.52}$$

Since the number of possible state sequences $q_{1:T}$ exponentially increases with T, even moderate recognition task do not allow for an *exhaustive* search over all possible hypothesis.

The Viterbi *approximation* circumvents this problem by approximating the sum over all possible state sequences by the most probable state sequence, i.e.

$$\hat{w}_{1:\hat{N}_w} \approx \underset{\{N_w, w_{1:N_w}\}}{\operatorname{argmax}} \left\{ \max_{\substack{q_{1:T}\in \\ \Phi_{w_{1:N_w}}}} \underbrace{\left(\prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{q}_t}\left(\mathbf{o}_t\,|q_t\right) P_{\breve{q}_t|\breve{q}_{t-1}}\left(q_t\,|q_{t-1}\right) \right)}_{\approx p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T}\right)} \tilde{P}^{\alpha_{\mathsf{LMS}}}_{\breve{w}_{1:\breve{N}_w}}\left(w_{1:N_w}\right) \right\}. \tag{3.53}$$

[5] Note that the Kronecker-delta $\delta_{N_w - \pounds\left(w_{1:N_w}\right)}$ has been dropped for ease of readability when incorporating the language model (3.50) into the decoding rule.

The computational efficiency resulting from the VITERBI approximation comes with the fact that the likelihood of the sequence of observation vectors $\mathbf{o}_{1:t}$ along the most probable state sequence ending in HMM state i at time instant t can be computed recursively from the likelihoods of a sub-sequence of observation vectors $\mathbf{o}_{1:t-1}$ along the most probable partial state sequences ending in HMM states $h \in \{1,\ldots,I\}$ at time instant $t-1$.

Repeated application of this *__d__ynamic __p__rogramming* (DP) algorithm – the VITERBI algorithm – eventually returns the desired quantity , i.e., $\max\limits_{q_{1:T}\in\Phi_{w_{1:N_w}}} \Big(p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}(\mathbf{o}_{1:T},q_{1:T})\Big)$.

3.5 Evaluation Metrics

The performance of an ASR system is usually evaluated by comparing the transcription of the sentence to be recognized (the reference transcription) with the transcription given by the recognizer (the recognition result). A string alignment algorithm based on the DP algorithm – the so-called LEVENSHTEIN algorithm [58] – then returns the minimum number of *edit operations* required to transform one transcription into the other. The set of edit operations considered thereby consists of *__sub__stitutions* (SUBs), *__del__etions* (DELs) and *__ins__ertions* (INSs). Denoting the number of substituted, deleted and inserted words by N_{SUB}, N_{DEL} and N_{INS}, respectively, the following two metrics are commonly used:

Word error rate The *__w__ord __e__rror __r__ate* (WER) λ_w is defined as the quotient of the total number of edit operations required to transform the reference transcription $w_{1:N_{w_{\mathsf{ref}}}}$ of length $N_{w_{\mathsf{ref}}}$ into the recognized word sequence $\hat{w}_{1:\hat{N}_w}$ of length $\hat{N}_w$, i.e.,

$$\lambda_{\mathsf{WER}} = \frac{N_{\mathsf{SUB}} + N_{\mathsf{DEL}} + N_{\mathsf{INS}}}{N_{w_{\mathsf{ref}}}} \times 100\,\%. \tag{3.54}$$

Note that, due to a normalization on the length N_w of the reference transcription, the WER may become larger than one. Further note that the WER is independent of the order of the alignment, i.e., the WER of aligning $w_{1:N_{w_{\mathsf{ref}}}}$ with $\hat{w}_{1:\hat{N}_w}$ is the same a the WER of aligning $\hat{w}_{1:\hat{N}_w}$ with $w_{1:N_{w_{\mathsf{ref}}}}$, since changing the order for the alignment only turns insertions into deletions and vice versa.

Word accuracy With the definition of the WER in (3.54), the *word __acc__uracy* (ACC) is simply obtained as

$$\lambda_{\mathsf{ACC}} = 100\,\% - \lambda_{\mathsf{WER}}, \tag{3.55}$$

which, as a consequence of the definition of the WER, may become negative.

3.6 Environmental Robustness

Thus far, it has been assumed that the set of HMM parameters $\Theta_{\breve{\mathbf{o}}}^{\mathsf{HMM}}$ learned from training data via the EM algorithm [59] is capable of accurately modeling the observed data, i.e., that the training data and the observed data are extracted by the same feature extraction scheme and that they exhibit the same statistical properties. The acoustic model and the observed data are said to *match*. Hence, the decoding rule (3.51) may also be written as

$$\hat{w}_{1:\hat{N}_w} = \operatorname*{argmax}_{\{N_w, w_{1:N_w}\}} \left\{ p_{\breve{\mathbf{o}}_{1:T}|\breve{w}_{1:\breve{N}_w}} \left(\mathbf{o}_{1:T} \middle| w_{1:N_w}; \Theta_{\breve{\mathbf{o}}}^{\mathsf{HMM}}\right) P_{\breve{w}_{1:\breve{N}_w},\breve{N}_w}\left(w_{1:N_w}, N_w\right)\right\}, \quad (3.56)$$

where the set of HMM model parameters and the feature vector sequence share the same identifier (here: $\mathbf{o}$) to highlight the *matched condition*.

Operating the recognizer in this matched condition calls for incorporation of knowledge about the statistics of the data to be observed already at the training stage of the HMM or at least proper adaptation of an existing set of HMM parameters to the new condition prior or in parallel to recognition. Failure to do so results in the famous *model mismatch*, particularly reflected by a major decrease in recognition performance when compared to the matched condition.

Hence, if training data representative of the data to be observed are available, *matched training* of an acoustic model is to be favored. If no explicit knowledge about the data to be recognized is given, but merely a vague conjecture, so-called *multi-style* training [60], where (sometimes also artificially generated) data for various expected recognition conditions are used to build the acoustic model, has been found to be effective [61, p. 657].

W.l.o.g, the training data are now assumed to be extracted from a speech signal that is, e.g., neither corrupted by additive noise nor by reverberation and captured with a close-talking microphone. Identifying the feature vectors of this *clean* speech signal by $\mathbf{x}$, the corresponding set of acoustic model parameters is given by $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$. If the data to be recognized and the data underlying the training of the acoustic model are extracted by the same feature extraction scheme and further share the same statistical properties, (3.56) can readily be employed (replacing $\mathbf{o}_{1:T}$ by $\mathbf{x}_{1:T}$ and $\Theta_{\breve{\mathbf{o}}}^{\mathsf{HMM}}$ by $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$). However, if the trained model and the data to be recognized do not match, and neither model retraining nor model adaptation is possible or desired, the set of acoustic model parameters $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$ may still be used for decoding if the mismatch is properly accounted for during computation of the contribution of the acoustic model in the decoding rule.

This may best be seen by introducing the sequence of clean speech feature vectors $\mathbf{x}_{1:T}$ underlying the observed sequence of feature vectors $\mathbf{o}_{1:T}$ as the realization of a (hidden) stochastic process $\breve{\mathbf{x}}_{1:T}$ into the contribution of the acoustic model.[6] This eventually amounts to evaluating the joint likelihood $p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T}, q_{1:T}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)$ of the observed sequence of feature vectors $\mathbf{o}_{1:T}$ and the hypothesized hidden state sequence $q_{1:T}$ employing the parameters $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$ of the *mismatched* model. The sequence of clean speech feature

[6] In general, the data used for HMM training and the data to be observed may be given in different signal domains, i.e., may be extracted by different feature extraction schemes. Nevertheless, for ease of notation, the observations and the training data are assumed to be extracted by the same feature extraction scheme.

vectors $\mathbf{x}_{1:T}$ can now be introduced as[7]

$$\begin{aligned}
&p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&=\int\limits_{\mathbb{R}^{DT}} p_{\breve{\mathbf{o}}_{1:T},\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},\mathbf{x}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_{1:T} && (3.57)\\
&=\int\limits_{\mathbb{R}^{DT}} p_{\breve{\mathbf{o}}_{1:T}|\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T}\left|\mathbf{x}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{x}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_{1:T} && (3.58)\\
&=\int\limits_{\mathbb{R}^{DT}} p_{\breve{\mathbf{o}}_{1:T}|\breve{\mathbf{x}}_{1:T}}\left(\mathbf{o}_{1:T}\left|\mathbf{x}_{1:T}\right.\right)p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{x}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_{1:T}. && (3.59)
\end{aligned}$$

The last algebraic conversion implies that once the underlying sequence of clean speech feature vectors $\mathbf{x}_{1:T}$ is given, the state sequence $q_{1:T}$ with the respective set of model parameters $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$ does not provide any additional information on $\mathbf{o}_{1:T}$. An interesting interpretation of (3.59) can be found by once more applying Bayes' theorem, i.e., by writing

$$\begin{aligned}
&p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\propto\int\limits_{\mathbb{R}^{DT}} p_{\breve{\mathbf{x}}_{1:T}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{1:T}\left|\mathbf{o}_{1:T}\right.\right)\frac{p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{x}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)}{p_{\breve{\mathbf{x}}_{1:T}}\left(\mathbf{x}_{1:T}\right)}\mathrm{d}\mathbf{x}_{1:T} && (3.60)\\
&=E\left[\left.\frac{p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\breve{\mathbf{x}}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)}{p_{\breve{\mathbf{x}}_{1:T}}\left(\breve{\mathbf{x}}_{1:T}\right)}\right|\mathbf{o}_{1:T}\right], && (3.61)
\end{aligned}$$

where the mathematical symbol $\propto$ has been used to indicate that the left-hand and the right-hand side of (3.60) are proportional to each other. The constant of proportionality is given by the likelihood $p_{\breve{\mathbf{o}}_{1:T}}$ of the observed sequence of feature vectors $\mathbf{o}_{1:T}$.

Instead of evaluating $\frac{p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}}{p_{\breve{\mathbf{x}}_{1:T}}}$ for a sequence of clean speech feature vectors $\mathbf{x}_{1:T}$ and an HMM state sequence $q_{1:T}$, as sufficient for a recognition under a matched condition, (3.61) now calls for the computation of its expected value w.r.t. the clean speech feature vectors $\mathbf{x}_{1:T}$, conditioned on the sequence of observations $\mathbf{o}_{1:T}$. Note that while the likelihood $p_{\breve{\mathbf{x}}_{1:T}}$ can be dropped during decoding under matched conditions (compare (3.8)), it cannot be dropped for decoding under mismatched conditions since it is part of the integral (3.60).

3.6.1 Practical Realization

Equation (3.60) represents the theoretically optimal way to compute the likelihood of the sequence of observed feature vectors $\mathbf{o}_{1:T}$ while using an acoustic model with parameter set $\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}$ trained on data potentially exhibiting different statistics than the observed data.

[7] Integrals of the form $\int\limits_{\mathbb{R}^{DT}}(\cdot)\,\mathrm{d}\mathbf{x}_{1:T}$ and alike have to be understood as nested volume integrals, i.e., $\int\limits_{\mathbb{R}^{DT}}(\cdot)\,\mathrm{d}\mathbf{x}_{1:T}=\int\limits_{\mathbb{R}^{D}}\cdots\int\limits_{\mathbb{R}^{D}}(\cdot)\,\mathrm{d}\mathbf{x}_{1}\cdots\mathrm{d}\mathbf{x}_{T}$.

However, (3.60) is of only minor practical use, since a direct computation of the integral turns infeasible for all but the most trivial forms of the involved PDFs, e.g.,

$$p_{\breve{\mathbf{x}}_{1:T}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{1:T}|\mathbf{o}_{1:T}\right) = \delta\left(\mathbf{x}_{1:T} - \hat{\mathbf{x}}_{1:T}\left(\mathbf{o}_{1:T}\right)\right), \tag{3.62}$$

where $\delta(\cdot)$ denotes the DIRAC-delta distribution and $\hat{\mathbf{x}}_{1:T}(\mathbf{o}_{1:T})$ an estimate of the sequence of clean speech feature vectors obtained from the observations $\mathbf{o}_{1:T}$.

The a posteriori PDF (3.62), due to the *sifting property* of the DIRAC-delta distribution w.r.t. integration, turns (3.60) into

$$p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T}, q_{1:T}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \propto p_{\breve{\mathbf{x}}_{1:T},\breve{q}_{1:T}}\left(\hat{\mathbf{x}}_{1:T}\left(\mathbf{o}_{1:T}\right), q_{1:T}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right), \tag{3.63}$$

i.e., treats the estimated sequence $\hat{\mathbf{x}}_{1:T}(\mathbf{o}_{1:T})$ as the true sequence of clean speech feature vectors, which are than *plugged into* the decoding rule (3.53). In contrast to *back-end methods* like acoustic model adaptation or retraining it is usually by far less expensive to infer these estimates in the front-end of a speech recognition system than to adapt or retrain the set of model parameters in the corresponding back-end.

The a posteriori PDF provides optimal estimates for the sequence of clean speech feature vectors $\hat{\mathbf{x}}_{1:T}(\mathbf{o}_{1:T})$ with respect to any criterion; e.g., the mean of the posterior distribution provides the *__m__inimum __m__ean __s__quared __e__rror* (MMSE) estimate and its mode the MAP estimate. Hence, even if the a posteriori PDF may not allow for an analytic solution to the integral (3.60), employing *point estimates* obtained from it may be a convenient, though sub-optimal way to carry out decoding.

However, finding a computationally tractable way of computing (approximate) solutions to (3.60) (or equivalently (3.59)) may eventually allow the decoder to also resolve the uncertainty left in $\mathbf{x}_{1:T}$ after observing $\mathbf{o}_{1:T}$. Approaches adhering to this idea are subsumed under the term UD. A comprehensive collection of robust speech recognition approaches employing uncertainty information is given in [62].

In the following, two practically feasible approximations to (3.59) will be presented. Their derivation starts off by targeting a causal and a non-causal approximation. However, since the non-causal approximation may further be turned into another causal approximation, an additional qualifier is needed to uniquely identify the different approximations. This qualifier is chosen to be the a priori distribution characterizing and distinguishing the two UD rules.

3.6.1.1 Approximation With Predictive Prior

Looking for a causal solution to the aforementioned problem, (3.57) may first be rewritten by repeatedly applying BAYES' theorem and employing the conditional independence assumption (3.36) and the Markov property (3.37), leading to

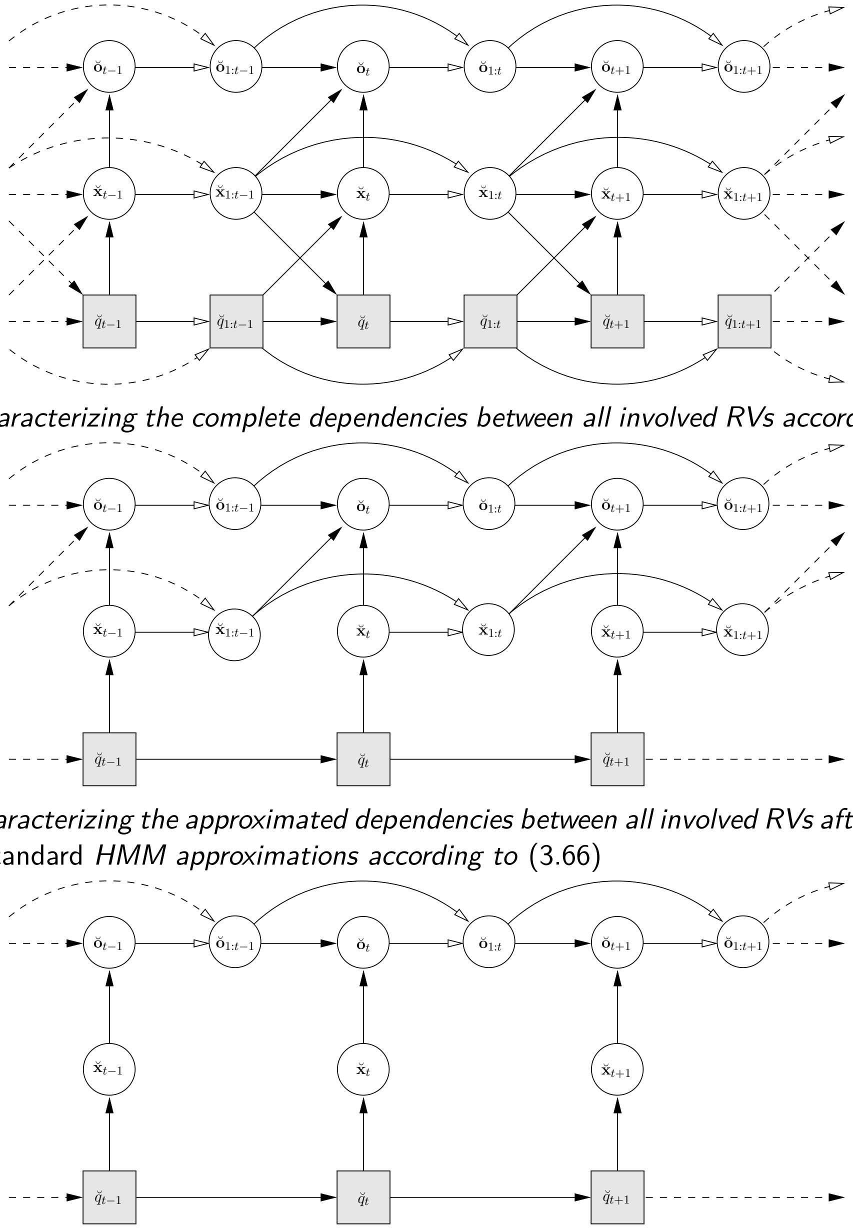
$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T}, q_{1:T}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\quad = \int\limits_{\mathbb{R}^{DT}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_{1:t},\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t}}\left(\mathbf{o}_t \,\middle|\, \mathbf{x}_{1:t}, \mathbf{o}_{1:t-1}, q_{1:t}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\qquad\qquad p_{\breve{\mathbf{x}}_t|\breve{\mathbf{x}}_{1:t-1},\breve{\mathbf{o}}_{1:t-1},\breve{q}_{1:t}}\left(\mathbf{x}_t \,\middle|\, \mathbf{x}_{1:t-1}, \mathbf{o}_{1:t-1}, q_{1:t}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\qquad\qquad P_{\breve{q}_t|\breve{q}_{1:t-1},\breve{\mathbf{x}}_{1:t-1},\breve{\mathbf{o}}_{1:t-1}}\left(q_t \,\middle|\, q_{1:t-1}, \mathbf{x}_{1:t-1}, \mathbf{o}_{1:t-1}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \mathrm{d}\mathbf{x}_{1:T} \qquad (3.64)\\
&\quad = \int\limits_{\mathbb{R}^{DT}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_{1:t},\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,\middle|\, \mathbf{x}_{1:t}, \mathbf{o}_{1:t-1}\right) p_{\breve{\mathbf{x}}_t|\breve{\mathbf{x}}_{1:t-1},\breve{q}_{1:t}}\left(\mathbf{x}_t \,\middle|\, \mathbf{x}_{1:t-1}, q_{1:t}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\qquad\qquad P_{\breve{q}_t|\breve{q}_{1:t-1},\breve{\mathbf{x}}_{1:t-1}}\left(q_t \,\middle|\, q_{1:t-1}, \mathbf{x}_{1:t-1}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \mathrm{d}\mathbf{x}_{1:T} \qquad (3.65)\\
&\quad \approx \int\limits_{\mathbb{R}^{DT}} \prod_{t=1}^{T} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_{1:t},\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,\middle|\, \mathbf{x}_{1:t}, \mathbf{o}_{1:t-1}\right) p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t \,\middle|\, q_t; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\qquad\qquad P_{\breve{q}_t|\breve{q}_{t-1}}\left(q_t \,\middle|\, q_{t-1}; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \mathrm{d}\mathbf{x}_{1:T}. \qquad (3.66)
\end{aligned}
$$

Note that (3.65) is not considered an approximation of (3.64), since i) once all feature vectors $\mathbf{x}_{1:t}$ of the clean speech signal are given, the HMM with state sequence $q_{1:t}$ does not provide additional information on $\mathbf{o}_t$ and ii) once all feature vectors $\mathbf{x}_{1:t-1}$ of the clean speech signal are given, the sequence of observed feature vectors $\mathbf{o}_{1:t-1}$ does not provide additional information on the current clean speech feature vector $\mathbf{x}_t$. The DBN characterizing (3.65) and the one after application of the conditional independence assumption and the Markov property, eventually characterizing (3.66), are shown in Fig. 3.5a and Fig. 3.5b on p. 25, respectively.

The involved PDF $p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_{1:t},\breve{\mathbf{o}}_{1:t-1}}$ may now be approximated by

$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_{1:t},\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,|\, \mathbf{x}_{1:t}, \mathbf{o}_{1:t-1}\right)\\
&\quad = \frac{p_{\breve{\mathbf{x}}_{1:t-1}|\breve{\mathbf{x}}_t,\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_{1:t-1} \,|\, \mathbf{x}_t, \mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_{1:t-1}|\breve{\mathbf{x}}_t,\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_{1:t-1} \,|\, \mathbf{x}_t, \mathbf{o}_{1:t-1}\right)} \frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_t \,|\, \mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_t \,|\, \mathbf{o}_{1:t-1}\right)} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,|\, \mathbf{o}_{1:t-1}\right) \qquad (3.67)\\
&\quad \approx \frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_t \,|\, \mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_t \,|\, \mathbf{o}_{1:t-1}\right)} p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,|\, \mathbf{o}_{1:t-1}\right) \qquad (3.68)\\
&\quad = p_{\breve{\mathbf{o}}_t|\breve{\mathbf{x}}_t,\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t \,|\, \mathbf{x}_t, \mathbf{o}_{1:t-1}\right), \qquad (3.69)
\end{aligned}
$$

i.e., the observation $\mathbf{o}_t$ at time instant t is assumed to not provide additional information on the past clean speech feature vectors, i.e., $\mathbf{x}_{1:t-1}$, once the past observations $\mathbf{o}_{1:t-1}$ and the current clean speech feature vector $\mathbf{x}_t$ are given. The corresponding DBN is given in Fig. 3.5c on p. 25. Though this assumption may be debatable, it allows the DT-dimensional volume integral (3.66) over a product of T real-valued functions to be approximated by the product of T real-valued D-dimensional volume integrals suited for practical realization,


$\check{\mathbf{o}}_{t-1}$
$\check{\mathbf{o}}_{1:t-1}$
$\check{\mathbf{o}}_{t}$
$\check{\mathbf{o}}_{1:t}$
$\check{\mathbf{o}}_{t+1}$
$\check{\mathbf{o}}_{1:t+1}$
$\check{\mathbf{x}}_{t-1}$
$\check{\mathbf{x}}_{1:t-1}$
$\check{\mathbf{x}}_{t}$
$\check{\mathbf{x}}_{1:t}$
$\check{\mathbf{x}}_{t+1}$
$\check{\mathbf{x}}_{1:t+1}$
$\check{q}_{t-1}$
$\check{q}_{1:t-1}$
$\check{q}_{t}$
$\check{q}_{1:t}$
$\check{q}_{t+1}$
$\check{q}_{1:t+1}$

(a) *DBN characterizing the complete dependencies between all involved RVs according to* (3.65)

(b) *DBN characterizing the approximated dependencies between all involved RVs after application of the* standard *HMM approximations according to* (3.66)

(c) *DBN characterizing the dependencies between all involved RVs amenable for practical realization according to* (3.69)

Figure 3.5: *The DBNs for speech recognition accounting for the uncertainty in the observed sequence of feature vectors. Subfigure 3.5a depicts the DBN with the complete dependencies between all involved RVs according to* (3.65)*, Subfig. 3.5a the DBN after employing the standard assumptions of HMM-based speech recognition. Eventually, Subfig. 3.5c depicts the DBN illustrating the dependencies resulting from* (3.69)*, which render the resolution of uncertainty amenable for practical realization.*

i.e.,

$$\begin{aligned} &p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\ &\approx\prod_{t=1}^{T}\left(\int_{\mathbb{R}^D}\frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_t\,|\mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_t\,|\mathbf{o}_{1:t-1}\right)}p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t\,\middle|q_t;\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_t\right)p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{o}_t\,|\mathbf{o}_{1:t-1}\right)\\ &\quad P_{\breve{q}_t|\breve{q}_{t-1}}\left(q_t\,\middle|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right). \end{aligned}\tag{3.70}$$

Since $p_{\breve{\mathbf{o}}_t|\breve{\mathbf{o}}_{1:t-1}}$ equally contributes to all hypothesized sequences of states and words, dropping this term does not change the recognition result and the final decision rule is given by

$$\begin{aligned} \hat{w}_{1:\hat{N}_w}\approx\underset{\{N_w,w_{1:N_w}\}}{\operatorname{argmax}}\Bigg\{&\max_{\substack{q_{1:T}\in\\ \Phi_{w_{1:N_w}}}}\left(\prod_{t=1}^{T}\left[\int_{\mathbb{R}^D}\frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_t\,|\mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_t\,|\mathbf{o}_{1:t-1}\right)}p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t\,\middle|q_t;\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_t\right]\right.\\ &\left.P_{\breve{q}_t|\breve{q}_{t-1}}\left(q_t\,\middle|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\right)\tilde{P}_{\breve{w}_{1:\breve{N}_w}}^{\alpha_{\mathsf{LMS}}}\left(w_{1:N_w}\right)\Bigg\}. \end{aligned}\tag{3.71}$$

The above UD rule will also be denoted as the "causal UD-p" rule, since, besides being causal, it also involves a (p)redictive a priori distribution at time instant t (opposed to a marginal a priori distribution, as will be encountered soon) .

Irrespective of the analytic form of the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ and the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$, (3.71) exhibits some important and desirable properties.

If the current observation $\mathbf{o}_t$ does not provide any information about the current clean speech feature vector $\mathbf{x}_t$ that has not already been present in the sequence of past observations $\mathbf{o}_{1:t-1}$, the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ equals the *predictive* PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ and the integral for the current time instant t reduces to one. Hence, the currently observed feature vector does not contribute to the final decision, which, as a consequence, relies more on the information provided by the state transition probabilities and the contribution of the language model. In the extreme case where all observed feature vectors do not provide any information on the underlying sequence of clean speech features vectors, the speech recognizer decides in favor of the most probable word sequence – the *maximum a posteriori* decision rule in essence turns into a *maximum a priori* decision rule. At this point, the importance of proper language modeling and in particular accurate modeling of the PMF of the length N_w of a word sequence gets apparent.

On the other hand, if the current observation $\mathbf{o}_t$ completely resolves the uncertainty in $\mathbf{x}_t$ that has been left after the observation of the sequence $\mathbf{o}_{1:t-1}$, the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ condenses to a DIRAC-delta distribution centered at the true underlying clean speech feature vector. The sifting property of the DIRAC-delta eventually renders the integration to a simple evaluation of the state-conditioned PDF $p_{\breve{\mathbf{x}}_t|\breve{q}_t}$ at $\hat{\mathbf{x}}_t\left(\mathbf{o}_{1:t}\right)=\mathbf{x}_t$. Note that the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ does not have to be evaluated explicitly in this case, since it equally contributes to all hypothesized word sequences. In the extreme case where the sequence of observations $\mathbf{o}_{1:T}$ is equal to or can unambiguously be mapped to the sequence of clean speech feature vectors $\mathbf{x}_{1:T}$, the decision rule (3.71) reduces to the standard decision rule (3.53).

3.6.1.2 Approximation With Marginal Prior

While the aforementioned approximations are only applicable in the context of a causal processing, a non-causal alternative to (3.71) has been presented in [63]. The presented derivation starts at (3.59) and applies BAYES' theorem, the conditional independence assumption (3.36) and the MARKOV property (3.37) to it to arrive at

$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\propto \int\limits_{\mathbb{R}^{DT}} \frac{p_{\breve{\mathbf{x}}_{1:T}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{1:T}\left|\mathbf{o}_{1:T}\right.\right)}{p_{\breve{\mathbf{x}}_{1:T}}\left(\mathbf{x}_{1:T}\right)} p_{\breve{\mathbf{x}}_{1:T}|\breve{q}_{1:T}}\left(\mathbf{x}_{1:T}\left|q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right) P_{\breve{q}_{1:T}}\left(q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\mathrm{d}\mathbf{x}_{1:T}
\end{aligned}
\tag{3.72}
$$

$$
\begin{aligned}
&= \int\limits_{\mathbb{R}^{DT}} \prod_{t=1}^{T} \frac{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:T},\breve{\mathbf{x}}_{1:t-1}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:T},\mathbf{x}_{1:t-1}\right.\right)}{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{x}}_{1:t-1}}\left(\mathbf{x}_{t}\left|\mathbf{x}_{1:t-1}\right.\right)} p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{x}}_{1:t-1},\breve{q}_{1:T}}\left(\mathbf{x}_{t}\left|\mathbf{x}_{1:t-1},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\\
&\qquad P_{\breve{q}_{t}|\breve{q}_{1:t-1}}\left(q_{t}\left|q_{1:t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\mathrm{d}\mathbf{x}_{1:T}
\end{aligned}
\tag{3.73}
$$

$$
\approx \int\limits_{\mathbb{R}^{DT}} \prod_{t=1}^{T} \frac{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:T}\right.\right)}{p_{\breve{\mathbf{x}}_{t}}\left(\mathbf{x}_{t}\right)} p_{\breve{\mathbf{x}}_{t}|\breve{q}_{t}}\left(\mathbf{x}_{t}\left|q_{t};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right) P_{\breve{q}_{t}|\breve{q}_{t-1}}\left(q_{t}\left|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\mathrm{d}\mathbf{x}_{1:T}
\tag{3.74}
$$

$$
= \prod_{t=1}^{T} P_{\breve{q}_{t}|\breve{q}_{t-1}}\left(q_{t}\left|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right) \int\limits_{\mathbb{R}^{D}} \frac{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:T}\right.\right)}{p_{\breve{\mathbf{x}}_{t}}\left(\mathbf{x}_{t}\right)} p_{\breve{\mathbf{x}}_{t}|\breve{q}_{t}}\left(\mathbf{x}_{t}\left|q_{t};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\mathrm{d}\mathbf{x}_{t}.
\tag{3.75}
$$

Here it is the approximation $p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{x}}_{1:t-1}} \approx p_{\breve{\mathbf{x}}_{t}}$ that allows the DT-dimensional volume integral over a product of T real-valued functions to be approximated by the product of T real-valued D-dimensional volume integrals.

One more step towards a feasible realization involves the approximation

$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_{1:T},\breve{q}_{1:T}}\left(\mathbf{o}_{1:T},q_{1:T};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right)\\
&\approx \prod_{t=1}^{T} P_{\breve{q}_{t}|\breve{q}_{t-1}}\left(q_{t}\left|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right) \int\limits_{\mathbb{R}^{D}} \frac{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:t+\tau}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:t+\tau}\right.\right)}{p_{\breve{\mathbf{x}}_{t}}\left(\mathbf{x}_{t}\right)} p_{\breve{\mathbf{x}}_{t}|\breve{q}_{t}}\left(\mathbf{x}_{t}\left|q_{t};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\mathrm{d}\mathbf{x}_{t}
\end{aligned}
\tag{3.76}
$$

to account for the fact that the context is usually reduced to only a small number of future observations in practice, e.g., $p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:T}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:T}\right.\right) \approx p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:t+\tau}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:t+\tau}\right.\right)$, where $\tau \geq 0$ determines the number of future observations to be taken into account.

This different factorization and the required approximations now lead to the decision rule

$$
\begin{aligned}
\hat{w}_{1:\hat{N}_w} \approx \underset{\{N_w,w_{1:N_w}\}}{\operatorname{argmax}} \Bigg\{ \max_{\substack{q_{1:T}\in\\ \Phi_{w_{1:N_w}}}} \Bigg(\prod_{t=1}^{T} \Bigg[\int\limits_{\mathbb{R}^{D}} \frac{p_{\breve{\mathbf{x}}_{t}|\breve{\mathbf{o}}_{1:t+\tau}}\left(\mathbf{x}_{t}\left|\mathbf{o}_{1:t+\tau}\right.\right)}{p_{\breve{\mathbf{x}}_{t}}\left(\mathbf{x}_{t}\right)} p_{\breve{\mathbf{x}}_{t}|\breve{q}_{t}}\left(\mathbf{x}_{t}\left|q_{t};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\mathrm{d}\mathbf{x}_{t}\Bigg]\\
P_{\breve{q}_{t}|\breve{q}_{t-1}}\left(q_{t}\left|q_{t-1};\Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right.\right)\Bigg) \tilde{P}_{\breve{w}_{1:\breve{N}_w}}^{\alpha_{\mathsf{LMS}}}\left(w_{1:N_w}\right)\Bigg\}.
\end{aligned}
\tag{3.77}
$$

In comparison to its counterpart (3.71), the above UD rule (3.77) now incorporates (possibly all) future observations to determine the numerator PDF. However, the denominator term now is *context-free*, i.e., considers only the (m)arginal distribution of the clean speech feature vector at time instant t. Hence, this UD rule will be denoted as the "non-causal UD-m" rule for $\tau > 0$ and as the "causal UD-m" rule for $\tau = 0$.[8]

3.6.2 Limitations on the Involved PDFs

Though the approximations (3.68) and (3.74) already render the causal and non-causal decision rules (3.71) and (3.77) quite compact, analytic solutions to the involved integrals

$$\mathcal{I}_{i|\mathbf{o}_{1:t}} := \int\limits_{\mathbb{R}^D} \frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}\left(\mathbf{x}_t \middle| \mathbf{o}_{1:t}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\mathbf{x}_t \middle| \mathbf{o}_{1:t-1}\right)} p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t \middle| q_t = i; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \mathrm{d}\mathbf{x}_t \tag{3.78}$$

and

$$\mathcal{I}_{i|\mathbf{o}_{1:t+\tau}} := \int\limits_{\mathbb{R}^D} \frac{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}\left(\mathbf{x}_t \middle| \mathbf{o}_{1:t+\tau}\right)}{p_{\breve{\mathbf{x}}_t}\left(\mathbf{x}_t\right)} p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t \middle| q_t = i; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) \mathrm{d}\mathbf{x}_t \tag{3.79}$$

for a particular state $q_t = i$ are, as already mentioned earlier in this section, subject to restrictions on the involved PDFs.

The state-conditioned PDF $p_{\breve{\mathbf{x}}_t|\breve{q}_t}$ is given by a GMM as in (3.41) and may thus be expressed by

$$p_{\breve{\mathbf{x}}_t|\breve{q}_t}\left(\mathbf{x}_t \middle| q_t = i; \Theta_{\breve{\mathbf{x}}}^{\mathsf{HMM}}\right) = \sum_{j=1}^{J(i)} c_{j|i} \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \tag{3.80}$$

Besides the cited Dirac-delta distribution, the a posteriori PDFs $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ and $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:T}}$ may, in general, be represented by mixture densities. However, targeting an analytic solution to the integrals (3.78) and (3.79), the analytic form of the elementary distributions in the mixture densities may eventually dictate the analytic form the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ or the marginal PDF $p_{\breve{\mathbf{x}}_t}$ should take.

Of the mixture densities, two special cases for the a posteriori PDFs $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ and $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}$ are considered next.

3.6.2.1 Gaussian Mixture Density

As with the state-condition PDFs in the acoustic model, the a posteriori PDFs $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ and $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}$ may be modeled by GMMs.

UD Rule with Predictive Prior By introducing the mixture index $m_t \in \{1, \ldots, M\}$ as the realization of the RV $\breve{m}_t$, the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ required for the causal UD-p

[8] Note that the causal UD-m rule may also be considered an approximation to the causal UD-p rule by simply dropping the context of the a priori PDF in the denominator term of (3.71).

rule may be written as

$$
\begin{aligned}
p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}(\mathbf{x}_t|\mathbf{o}_{1:t}) &= \sum_{m=1}^{M} P_{\breve{m}_t|\breve{\mathbf{o}}_{1:t}}(m_t=m|\mathbf{o}_{1:t})\, p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t},\breve{m}_t}(\mathbf{x}_t|\mathbf{o}_{1:t}, m_t=m) && (3.81)\\
&:= \sum_{m=1}^{M} \gamma_{m|\mathbf{o}_{1:t}} \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right) && (3.82)
\end{aligned}
$$

where $\gamma_{m|\mathbf{o}_{1:t}}$, $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ denote the weight, the mean vector and the covariance matrix of the mth elementary GAUSSIAN PDF. Note that in contrast to (3.80) all parameters of the GMM (3.82) depend on the observation sequence $\mathbf{o}_{1:t}$ and are thus varying with time.

Though the GMM (3.82) may, in theory, approximate any a posteriori PDF arbitrarily close, it calls for the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ to take a specific analytic form to allow the integral (3.78) to be solved analytically. The predictive PDF may, e.g., be given by the GAUSSIAN distribution

$$
p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}(\mathbf{x}_t|\mathbf{o}_{1:t-1}) = \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right), \qquad (3.83)
$$

where the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}$ of the predictive PDF depend on the sequence of past observations $\mathbf{o}_{1:t-1}$. Plugging (3.80), (3.82) and (3.83) into the integral (3.78) while employing (A.1) and (A.9) given in Appendix A.1 to express the quotient and the product of two GAUSSIAN distributions eventually leads to

$$
\mathcal{I}_{i|\mathbf{o}_{1:t}} = \sum_{m=1}^{M} \gamma_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \int_{\mathbb{R}^D} \frac{\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)}{\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)} \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}},|i,j}\right) \mathrm{d}\mathbf{x}_t \qquad (3.84)
$$

$$
= \sum_{m=1}^{M} \gamma^{(\mathsf{eq})}_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \int_{\mathbb{R}^D} \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}, \boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right) \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \mathrm{d}\mathbf{x}_t \qquad (3.85)
$$

$$
= \sum_{m=1}^{M} \gamma^{(\mathsf{eq})}_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \mathcal{N}\left(\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \qquad (3.86)
$$

with

$$
\begin{aligned}
\boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} &= \left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)^{-1} && (3.87)\\
\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} &= \boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} \left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}} \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right) && (3.88)\\
\gamma^{(\mathsf{eq})}_{m|\mathbf{o}_{1:t}} &= \gamma_{m|\mathbf{o}_{1:t}} \frac{\mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)}{\mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}, \boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right) \mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)} && (3.89)
\end{aligned}
$$

Though, strictly speaking, moving from (3.84) to (3.85) is only allowed if the difference $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}} - \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ is non-singular, in practice, the two covariance matrices meet this requirement since the uncertainty about the current clean speech feature vector $\mathbf{x}_t$ after observing $\mathbf{o}_t$ in addition to $\mathbf{o}_{1:t-1}$ is lower than (or at most equal to) the corresponding uncertainty given only $\mathbf{o}_{1:t-1}$.

In (3.86), the covariance matrix of the jth mixture of the ith state-conditioned PDF now has to be increased by the *equivalent covariance matrix* $\mathbf{\Sigma}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ and the mixture finally has to be evaluated at the *equivalent mean vector* $\boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$.

In (3.88) this equivalent mean vector $\boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ is given as a weighted average between the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$ of the mth mixture of the a posteriori distribution and the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}$ of the predictive distribution. The actual weighting of the two is controlled by the *covariance matrix ratio* $\mathbf{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$, which can easier be seen by rewriting (3.88) (see Appendix A.2 for a derivation) as

$$\begin{aligned}\boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} &= \left(\mathbf{I}+\left[\mathbf{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}-\mathbf{I}\right]^{-1}\right)\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\\ &\quad -\left(\mathbf{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}-\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}. \end{aligned} \tag{3.90}$$

Note that the equivalent means (3.87), the equivalent covariance matrices (3.88) and the equivalent weights (3.89) are independent of the current HMM index i and can thus be pre-computed prior to decoding. Further, the contribution (and thus the computation) of the equivalent weight may be dropped if $M=1$.

UD Rule with Marginal Prior Likewise, for the non-causal and causal UD-m rule, the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}$ may be written as

$$p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}\left(\mathbf{x}_t\,|\mathbf{o}_{1:t+\tau}\right) := \sum_{m=1}^{M}\gamma_{m|\mathbf{o}_{1:t+\tau}}\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m},\mathbf{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}\right). \tag{3.91}$$

This time, its weights, mean vectors and covariance matrices also depend on additional τ future observations. The marginal PDF $p_{\breve{\mathbf{x}}_t}$ may now be given by the GAUSSIAN distribution

$$p_{\breve{\mathbf{x}}_t}\left(\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}},\mathbf{\Sigma}_{\breve{\mathbf{x}}}\right), \tag{3.92}$$

whose mean vector and covariance matrix are fixed and, e.g., obtained from the same clean speech training data used to train the HMM acoustic model.

The integral (3.79) then exhibits the closed-form solution

$$\mathcal{I}_{i|\mathbf{o}_{1:t+\tau}} = \sum_{m=1}^{M}\gamma^{(\text{eq})}_{m|\mathbf{o}_{1:t+\tau}}\sum_{j=1}^{J(i)}c_{j|i}\mathcal{N}\left(\boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j},\mathbf{\Sigma}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}+\mathbf{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \tag{3.93}$$

with

$$\mathbf{\Sigma}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m} = \left(\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}-\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}}\right)^{-1} \tag{3.94}$$

$$\boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m} = \mathbf{\Sigma}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}\left(\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}-\mathbf{\Sigma}^{-1}_{\breve{\mathbf{x}}}\boldsymbol{\mu}_{\breve{\mathbf{x}}}\right) \tag{3.95}$$

$$\gamma^{(\text{eq})}_{m|\mathbf{o}_{1:t+\tau}} = \gamma_{m|\mathbf{o}_{1:t+\tau}}\frac{\mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m},\mathbf{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}\right)}{\mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m},\mathbf{\Sigma}^{(\text{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m}\right)\mathcal{N}\left(\mathbf{0}_{D\times 1};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}},\mathbf{\Sigma}_{\breve{\mathbf{x}}}\right)}. \tag{3.96}$$

Comparing the causal and non-causal variants of the UD rules, only the computation of the equivalent weights, mean vectors and covariance matrices can be found to differ.

A special case of the non-causal UD-m rule may be obtained if the uncertainty in the a priori distribution is assumed to be much *larger* than in the a posteriori PDF. In this case the equivalent mean vectors and the equivalent covariance matrices may be approximated by the mean vectors and the covariance matrices of the mixture components of the a posteriori PDF. This, however, is equivalent to completely neglecting the a priori PDF in the decision rule (3.77), as is for instance employed in [64].

3.6.2.2 Dirac-Delta Mixture Density

The a posteriori PDFs $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$ and $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}$ may also be modeled by a mixture of Dirac-delta distributions [65].

UD Rule with Predictive Prior The mixture index m_t is, equivalent to (3.81), introduced as a realization of a random variable $\breve{m}_t$. However, this time the elementary PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t},\breve{m}_t}$ for application of the causal UD-p rule is considered to be a Dirac-delta distribution centered at, e.g., $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$.[9] Hence,

$$p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}(\mathbf{x}_t\,|\mathbf{o}_{1:t}) = \sum_{m=1}^{M} P_{\breve{m}_t|\breve{\mathbf{o}}_{1:t}}(m_t = m\,|\mathbf{o}_{1:t})\, p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t},\breve{m}_t}(\mathbf{x}_t\,|\mathbf{o}_{1:t}, m_t = m) \tag{3.97}$$

$$:= \sum_{m=1}^{M} \gamma_{m|\mathbf{o}_{1:t}}\,\delta\left(\mathbf{x}_t - \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right). \tag{3.98}$$

Plugging (3.98) and (3.80) into (3.78) now leads to

$$\mathcal{I}_{i|\mathbf{o}_{1:t}} = \sum_{m=1}^{M} \gamma_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \int_{\mathbb{R}^D} \frac{\delta\left(\mathbf{x}_t - \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}(\mathbf{x}_t\,|\mathbf{o}_{1:t-1})} \mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \mathrm{d}\mathbf{x}_t \tag{3.99}$$

$$= \sum_{m=1}^{M} \gamma_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \frac{\mathcal{N}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right)}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\,\middle|\mathbf{o}_{1:t-1}\right)} \tag{3.100}$$

$$= \sum_{m=1}^{M} \tilde{\gamma}_{m|\mathbf{o}_{1:t}} \sum_{j=1}^{J(i)} c_{j|i} \mathcal{N}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right), \tag{3.101}$$

with

$$\tilde{\gamma}_{m|\mathbf{o}_{1:t}} = \frac{\gamma_{m|\mathbf{o}_{1:t}}}{p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\,\middle|\mathbf{o}_{1:t-1}\right)}. \tag{3.102}$$

The evaluation of the integral thus does not depend on the actual analytic form of the predictive PDF, which may now, e.g., be modeled by a GMM.

[9] Though the Dirac-delta mixtures may be centered at arbitrary estimates of the clean speech feature vector $\mathbf{x}_t$, the mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$, $m \in \{1, \ldots M\}$ are considered here for better comparison with, e.g., (3.86).

Though each mixture of the state-conditioned PDF is only evaluated at a point estimate of the clean speech feature vector $\mathbf{x}_t$ (e.g., the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$), at least some uncertainty about the clean speech feature vector $\mathbf{x}_t$ is eventually accounted for by means of multiple, weighted hypotheses about the latter in the a posteriori PDF.

Hence, modeling the a posteriori PDF by a weighted sum of Dirac-delta distributions to compute the integral (3.78) may be a viable alternative to modeling the a posteriori PDF by a GMM if i) reliable estimates of the covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}$, $m \in \{1, \ldots, M\}$, are not available, ii) the predictive PDF is clearly non-Gaussian or iii) the computational burden coming along with the computation of the equivalent mean vectors (3.95) and equivalent covariance matrices (3.94) shall be reduced.

UD Rule with Marginal Prior For the non-causal and causal UD rule employing the marginal a priori PDF, very similar equations are obtained. In particular, the integral of interest has the closed-form solution

$$\mathcal{I}_{i|\mathbf{o}_{1:t+\tau}} = \sum_{m=1}^{M} \tilde{\gamma}_{m|\mathbf{o}_{1:t+\tau}} \sum_{j=1}^{J(i)} c_{j|i} \mathcal{N}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|i,j}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|i,j}\right) \tag{3.103}$$

where

$$\tilde{\gamma}_{m|\mathbf{o}_{1:t+\tau}} = \frac{\gamma_{m|\mathbf{o}_{1:t+\tau}}}{p_{\breve{\mathbf{x}}_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}|m}\right)}. \tag{3.104}$$

In practice, the case where $M = 1$ is of particular interest, since the contribution (and thus the evaluation) of the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ to the integral $\mathcal{I}_{i|\mathbf{o}_{1:t}}$ or the contribution of the marginal PDF $p_{\breve{\mathbf{x}}_t}$ to the integral $\mathcal{I}_{i|\mathbf{o}_{1:t+\tau}}$ may be neglected and both the decoding rules (3.71) and (3.77) reduce to the standard decoding rule, however, employing the point estimates $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t}} := \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m=1}$ or $\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau}} := \boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t+\tau},m=1}$ instead of the true clean speech feature vector $\mathbf{x}_t$.

3.6.3 Modeling the Environment

So far, the distortions causing the mismatch between the statistics of the observed feature vectors $\mathbf{o}_{1:T}$ and those of the (clean) feature vectors $\mathbf{x}$ employed for training of the acoustic model have been mentioned only briefly in Sec. 3.1.

In this work, the clean speech signal $x(p)$ is assumed to be corrupted by two kinds of distortions – reverberation and additive (background) noise.

As a direct consequence of using a distant-talking microphone rather than a close-talking one, the speech sound traverses along multiple paths from the speech source to the sensor, along which it suffers different degrees of attenuation and delay. These differently attenuated and delayed version of the speech sound eventually superpose at the microphone and, after ADC, offset-compensation and pre-emphasis, form the *reverberant speech signal* $s(p)$. A system theoretic model for the effect of reverberation is the convolution of the clean speech signal $x(p)$ with the AIR $h(p)$, describing the multi-path propagation of the speech signal from the speech source to the microphone. The reverberant signal $s(p)$ can thus be written as

$$s(p) = (x * h)(p) \tag{3.105}$$

$$= \sum_{p'=0}^{L_h-1} h\left(p'\right) x\left(p-p'\right), \tag{3.106}$$

with $*$ denoting the (linear) convolution operator. Note that (3.106) implicitly assumes the AIR to be causal, time-invariant and of finite length L_h. In practice, the AIR is sensitive to changes within the acoustic environment, e.g., changes in temperature and humidity, speaker movements and other movements within the environment. As a consequence, the AIR is highly time-variant and (3.106) only serves as a simplified model.

Using a distant-talking microphone also increases the chance of capturing signals from other sound sources than the desired speaker. Subsuming these signals, which may also contain speech from interfering speakers, as *background noise* $n(p)$ superposing the reverberant speech signal results in the final model of the environment used throughout this work, given by

$$o(p) = s(p) + n(p) \tag{3.107}$$

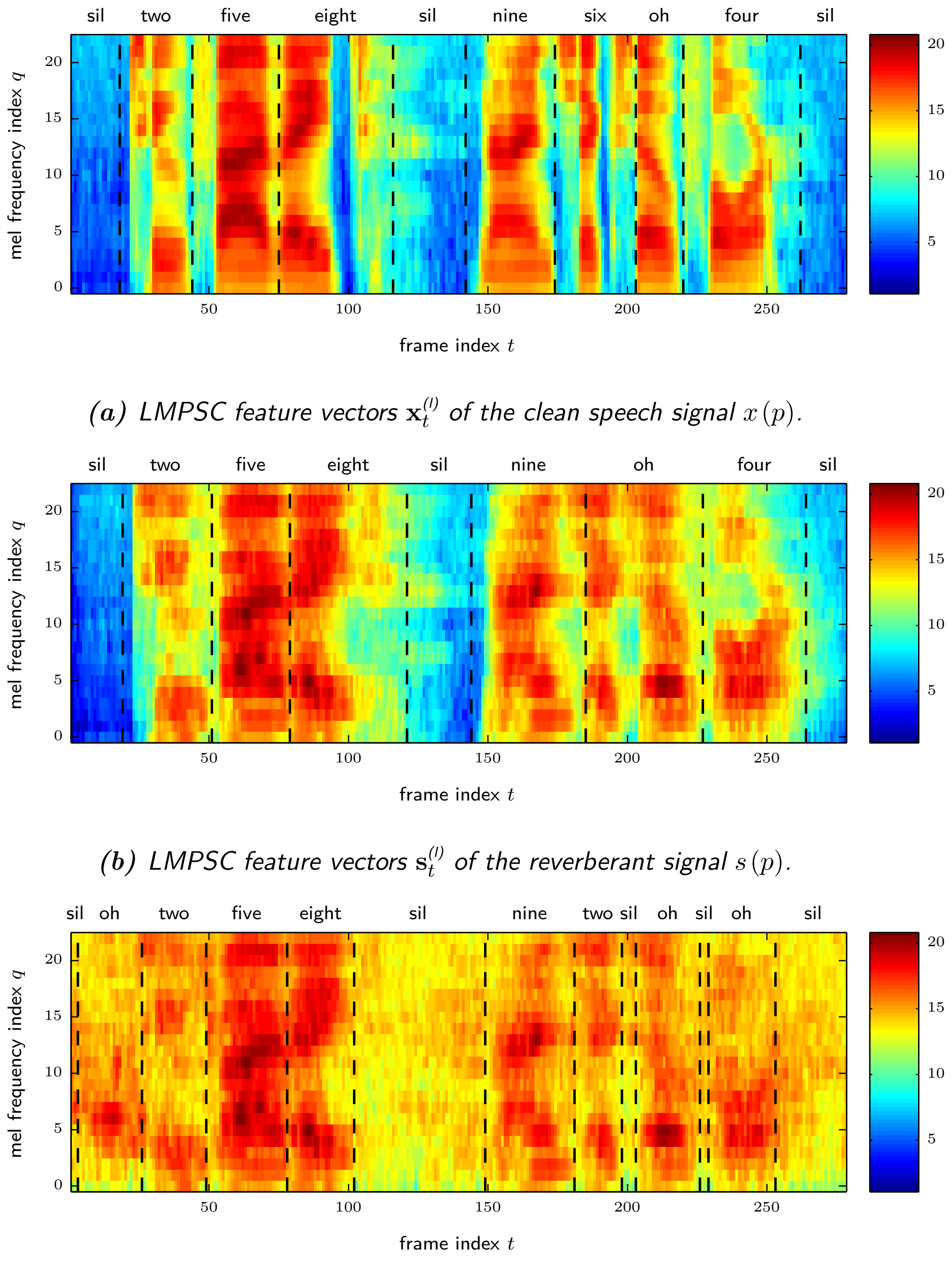
$$= (x * h)(p) + n(p) \tag{3.108}$$

$$= \sum_{p'=0}^{L_h-1} h\left(p'\right) x\left(p-p'\right) + n(p). \tag{3.109}$$

From (3.108) it already becomes apparent that feature vectors $\mathbf{o}_{1:T}$ extracted from the noisy reverberant signal $o(p)$ will be different from $\mathbf{x}_{1:T}$, i.e., those directly extracted from the clean speech signal $x(p)$.

To illustrate this issue, the example utterance "two-five-eight-nine-six-oh-four" taken from the AURORA 5 database (see Sec. 5.3.1 for a detailed description of the database) is considered. With the provided clean speech signal $x(p)$, the reverberant signal $s(p)$ and the noisy reverberant signal $o(p)$, LMPSC feature vectors are extracted according to the described ETSI standard front-end. The LMPSC feature vectors are displayed in Fig. 3.6 together with their transcription as output by a HMM based speech recognizer where the acoustic model trained on clean speech signals has been employed for the three scenarios, i.e., no measures are taken to counteract the detrimental effect of reverberation and noise. Comparing the LMPSC feature vectors of the clean speech signal presented in Fig. 3.6a to the according feature vectors of the reverberant signal displayed in Fig. 3.6b, first, two major differences may be observed.

The first observation is a *coloration* of the logarithmic mel power spectrum in the presence of reverberation. This may majorly be attributed to the *early reflections*, i.e., the relatively sparse first delayed and attenuated version of the speech sound (arriving at the microphone after only one or two reflections at walls, floor, ceiling, ...). The second observation is a *dispersion* of the spectrum along the time-axis. This is due to the *late reverberation*, i.e., the succession of densely populated reflections of diminishing power (arriving at the microphone after multiple reflections). The latter phenomenon may best be seen by comparing the LMPSC feature vectors grouped under the label "eight" in Fig. 3.6a and Fig. 3.6b, respectively. In particular, the dispersion completely masks the stop prior to the voiceless alveolar plosive "/t".



(a) LMPSC feature vectors $\mathbf{x}_t^{(l)}$ of the clean speech signal $x(p)$.

(b) LMPSC feature vectors $\mathbf{s}_t^{(l)}$ of the reverberant signal $s(p)$.

(c) LMPSC feature vectors $\mathbf{o}_t^{(l)}$ of the noisy reverberant signal $o(p)$.

Figure 3.6: *LMPSC feature vectors extracted from the clean speech signal (a), the reverberant signal (b) and the noisy reverberant signal (c) of the utterance "two-five-eight-nine-six-oh-four" taken from the AURORA 5 database. The transcriptions of the signals are provided by a HMM-based speech recognizer with a* clean *acoustic model.*

Due to the aforementioned two phenomena the feature vectors of the reverberant signal will exhibit a different statistical representation than the clean speech feature vectors and, as a consequence, will not match well with the *clean* acoustic model. This mismatch will immediately result in an increased number of recognition errors unless proper countermeasures are taken. In the considered example, the digit "six" is no longer recognized, i.e., a deletion took place.

The additional presence of additive background noise further causes a masking of the LMPSCs of the reverberant signal (see Fig. 3.6c) and renders the mismatch between the *clean* acoustic model and the statistics of the noisy reverberant test data even more severe. In the considered example, this leads to an insertion of the digit "oh" at the beginning of the utterance and two substitutions at its end ("six-oh-four" vs. "two-oh-oh").

The above example illustrates the need for counteracting the detrimental effect of reverberation and noise to arrive at environmentally robust ASR systems. For the derived decoding rules (3.71) and (3.77) to be applicable, knowledge about the statistical relation between the clean speech feature vectors and the observed noisy reverberant feature vectors is required. This statistical relation and its inference will be derived in full detail in the next chapter.

4 Bayesian Estimation of the Speech Feature Posterior

Key to application of the decoding rule (3.71) is the proper determination of the predictive PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t-1}}$ and the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t}}$.

Equivalently, for the decoding rule (3.77), the a posteriori PDF $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+\tau}}$ is required. In this work, the estimation will be approximated to the estimation of $p_{\breve{\mathbf{x}}_t|\breve{\mathbf{o}}_{1:t+L_C-1}}$, i.e., only a lag of $\tau = L_C - 1$ is allowed for the inference of the LMPSC vector of the clean speech signal.

Whether the logarithmic mel power spectral domain is thereby preferred over the cepstral domain or vice versa is a question that cannot be answered unambiguously, however, the logarithmic mel power spectral domain may be favorable, since it provides more detailed information about the acoustic scenery than the cepstral domain. In this work, this domain will be considered not only for that reason, but also for the reason of simplifying the models used for the Bayesian inference.

4.1 Conceptually Optimal Solution

Anticipating the findings in Sec. 4.3, the clean speech LMPSC feature vector $\mathbf{x}_t^{(\mathsf{l})}$ will be extended by $L_C - 1$ past clean speech LMPSC feature vectors and further augmented by the LMPSC feature vector of the noise $\mathbf{n}_t^{(\mathsf{l})}$ to give the *state* vector $\mathbf{z}_t^{(\mathsf{l})} \in \mathbb{R}^{(L_C+1)Q}$, defined as

$$\mathbf{z}_t^{(\mathsf{l})} := \left[\left(\mathbf{x}_t^{(\mathsf{l})}\right)^\dagger \quad \ldots \quad \left(\mathbf{x}_{t-L_C+1}^{(\mathsf{l})}\right)^\dagger \quad \left(\mathbf{n}_t^{(\mathsf{l})}\right)^\dagger \right]^\dagger. \tag{4.1}$$

The parameter $L_C \in \mathbb{N}_{>0}$ thereby determines the number of LMPSC feature vectors of the clean speech signal in the state vector.

Instead of trying to infer the predictive PDF $p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ and the a posteriori PDFs $p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ and $p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t+L_C-1}^{(\mathsf{l})}}$ required for a causal and non-causal decoding, directly, the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ and the a posteriori PDFs $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ and $p_{\breve{\mathbf{z}}_{t+L_C-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t+L_C-1}^{(\mathsf{l})}}$ of the state vector $\breve{\mathbf{z}}_t^{(\mathsf{l})}$ are inferred as an *intermediate* step. The desired PDFs may then be obtained from those PDFs by application of the law of total probability, i.e., via marginalization.

The conditional Bayesian inference now provides a recursive formulation for the estimation of the intermediate PDFs.

Assuming the a posteriori PDF $p_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ at time instant $t-1$ to be given, the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ of the state vector $\breve{\mathbf{z}}_t^{(\mathsf{l})}$ at time instant t is obtained by applying the law of total probability as

$$p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{z}_t^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)$$
$$= \int\limits_{\mathbb{R}^{(L_C+1)Q}} p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{z}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{z}_t^{(l)}\middle|\mathbf{z}_{t-1}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right) p_{\breve{\mathbf{z}}_{t-1}^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{z}_{t-1}^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)\mathrm{d}\mathbf{z}_{t-1}^{(l)}. \tag{4.2}$$

This *state prediction step* is followed by the *state update step*, which employs BAYES' theorem to obtain the a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t}^{(l)}}$ at time instant t from the predictive PDF $p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ obtained from (4.2), i.e.,

$$p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t}^{(l)}}\left(\mathbf{z}_t^{(l)}\middle|\mathbf{o}_{1:t}^{(l)}\right)$$
$$= \frac{p_{\breve{\mathbf{o}}_t^{(l)},\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{o}_t^{(l)},\mathbf{z}_t^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)}{p_{\breve{\mathbf{o}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{o}_t^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)} \tag{4.3}$$
$$= \frac{p_{\breve{\mathbf{o}}_t^{(l)}|\breve{\mathbf{z}}_t^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{o}_t^{(l)}\middle|\mathbf{z}_t^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right) p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{z}_t^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)}{\int\limits_{\mathbb{R}^{(L_C+1)Q}} p_{\breve{\mathbf{o}}_t^{(l)}|\breve{\mathbf{z}}_t^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{o}_t^{(l)}\middle|\mathbf{z}_t^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right) p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{z}_t^{(l)}\middle|\mathbf{o}_{1:t-1}^{(l)}\right)\mathrm{d}\mathbf{z}_t^{(l)}}. \tag{4.4}$$

The statistical dependencies of the involved random variables are depicted in the DBN of Fig. 4.1. The two steps of predicting and updating the distribution of the state vector $\breve{\mathbf{z}}_t^{(l)}$ require the additional knowledge of two distributions, namely $p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{z}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ and $p_{\breve{\mathbf{o}}_t^{(l)}|\breve{\mathbf{z}}_t^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$.

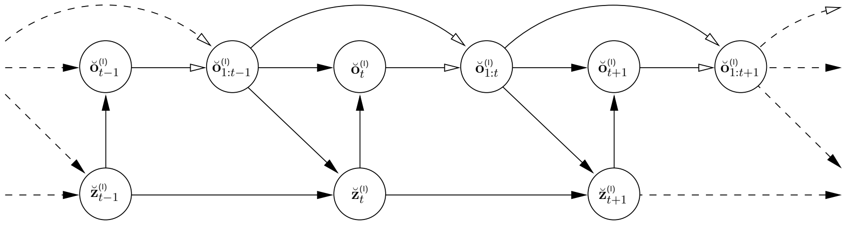

$\breve{\mathbf{o}}_{t-1}^{(l)}$
$\breve{\mathbf{o}}_{1:t-1}^{(l)}$
$\breve{\mathbf{o}}_{t}^{(l)}$
$\breve{\mathbf{o}}_{1:t}^{(l)}$
$\breve{\mathbf{o}}_{t+1}^{(l)}$
$\breve{\mathbf{o}}_{1:t+1}^{(l)}$
$\breve{\mathbf{z}}_{t-1}^{(l)}$
$\breve{\mathbf{z}}_{t}^{(l)}$
$\breve{\mathbf{z}}_{t+1}^{(l)}$

Figure 4.1: *DBN characterizing the dependencies between all involved RVs according to* (4.4) *and* (4.2)*. Again, deterministic dependencies between RVs are indicated by white-headed arrows, statistical dependencies by black-headed arrows.*

The first PDF, $p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{z}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$, statistically describes the evolution of the state vector $\mathbf{z}_t^{(l)}$ at time instant t, given the state vector $\mathbf{z}_{t-1}^{(l)}$ and all past observation $\mathbf{o}_{1:t-1}^{(l)}$. Since no information about the current observation $\mathbf{o}_t^{(l)}$ is incorporated, this statistical model is called *the a priori model*.

The second PDF, $p_{\breve{\mathbf{o}}_t^{(l)}|\breve{\mathbf{z}}_t^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$, statistically relates the state vector $\mathbf{z}_t^{(l)}$ at time instant t and the sequence of past observation vectors $\mathbf{o}_{1:t-1}^{(l)}$ to the current observation vector $\mathbf{o}_t^{(l)}$ and is called *the observation model*.

The recursive inference of the predictive PDF and the a posteriori PDF according to (4.2) and (4.4) is depicted in Fig. 4.2. However, though quite appealing, (4.2) and (4.4) are optimal only in theory. In practice, due to the recursive formulation, the quality of

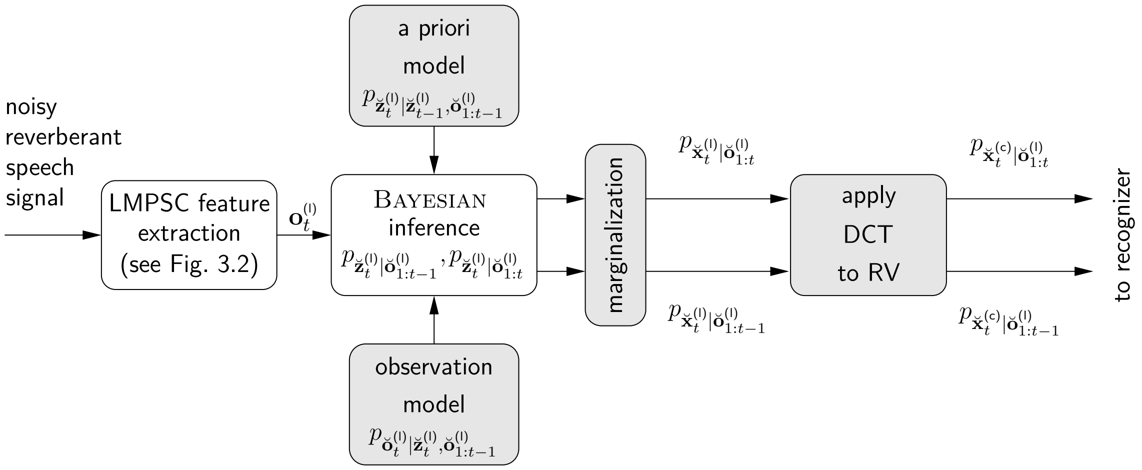

a priori
model
$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$
noisy
reverberant
speech
signal
LMPSC feature
extraction
(see Fig. 3.2)
$\mathbf{o}_t^{(\mathsf{l})}$
BAYESIAN
inference
$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}, p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$
marginalization
$p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$
$p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$
apply
DCT
to RV
$p_{\breve{\mathbf{x}}_t^{(\mathsf{c})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$
$p_{\breve{\mathbf{x}}_t^{(\mathsf{c})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$
to recognizer
observation
model
$p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$

Figure 4.2: *Conceptually optimal solution to the inference of the predictive PDF and the a posteriori PDF in a* BAYESIAN *framework (here: only causal inference is considered).*

the prediction step will depend on the quality of the update step and so forth. Further, both the prediction and the update step, i.e., the integrals over and the product of the involved PDFs, should be analytically and computationally tractable. Hence, the quality of the inference will be highly sensitive to both the chosen a priori model and the employed observation model. A detailed discussion and derivation thereof is given in Sec. 4.2 and Sec. 4.3, respectively. A computationally tractable approach resulting from the proposed a priori model will be outlined in Sec. 4.8.

4.2 A Priori Model

The role of the a priori model $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ is to statistically predict the state vector $\mathbf{z}_t^{(\mathsf{l})}$ at time instant t, given the state vector $\mathbf{z}_{t-1}^{(\mathsf{l})}$ at time instant $t-1$ and all past observations $\mathbf{o}_{1:t-1}^{(\mathsf{l})}$. Since the state vector $\mathbf{z}_t^{(\mathsf{l})}$ consists of the sequence of L_C (current and past) clean speech LMPSC feature vectors and the current noise LMPSC feature vector (compare (4.1)), the a priori model may also be written as

$$\begin{aligned}
&p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{z}_{t-1}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)\\
&\quad=p_{\breve{\mathbf{x}}_{t-L_C+1:t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-1:t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{x}_{t-L_C+1:t}^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},\mathbf{n}_{t-1:t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)\\
&\qquad p_{\breve{\mathbf{n}}_t^{(\mathsf{l})}|\breve{\mathbf{n}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{n}_t^{(\mathsf{l})}\middle|\mathbf{n}_{t-1}^{(\mathsf{l})},\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right) \qquad (4.5)\\
&\quad=p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-1:t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},\mathbf{n}_{t-1:t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)\\
&\qquad p_{\breve{\mathbf{x}}_{t-L_C+1:t-1}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-1:t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{x}_{t-L_C+1:t-1}^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},\mathbf{n}_{t-1:t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)\\
&\qquad p_{\breve{\mathbf{n}}_t^{(\mathsf{l})}|\breve{\mathbf{n}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{n}_t^{(\mathsf{l})}\middle|\mathbf{n}_{t-1}^{(\mathsf{l})},\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right). \qquad (4.6)
\end{aligned}$$

Noting that the second term in (4.6) is nothing but a DIRAC-delta distribution centered at $\mathbf{x}_{t-L_C+1:t-1}^{(\mathsf{l})}$, the a priori model can be found to be completely described by the two terms $p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{n}}_{t-1:t}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ and $p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$. Applying the assumption of (conditional) independence between clean speech and noise LMPSC feature vectors to these PDFs, i.e.,

$$\begin{aligned}
&p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{n}}_{t-1:t}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{x}_t^{(l)}\middle|\mathbf{x}_{t-L_C:t-1}^{(l)},\mathbf{n}_{t-1:t}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right)\\
&\quad\approx p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{x}_t^{(l)}\middle|\mathbf{x}_{t-L_C:t-1}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right), \qquad (4.7)\\
&p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{n}_t^{(l)}\middle|\mathbf{n}_{t-1}^{(l)},\mathbf{x}_{t-L_C:t-1}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right)\\
&\quad\approx p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{n}_t^{(l)}\middle|\mathbf{n}_{t-1}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right), \qquad (4.8)
\end{aligned}$$

eventually allows the a priori model $p_{\breve{\mathbf{z}}_t^{(l)}|\breve{\mathbf{z}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the state vector to be decomposed into an a priori model $p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the clean speech LMPSC feature vector and an a priori model $p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the LMPSC feature vector of the noise. The DBN depicting the statistical dependencies in the inference after application of the above approximations is given in Fig. 4.3

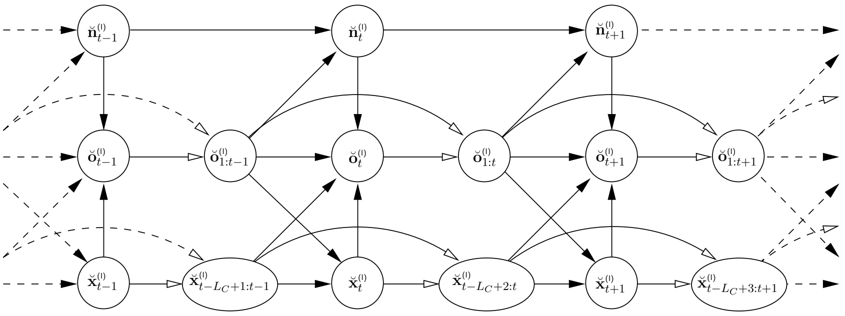

$\breve{\mathbf{n}}_{t-1}^{(l)}$
$\breve{\mathbf{n}}_{t}^{(l)}$
$\breve{\mathbf{n}}_{t+1}^{(l)}$
$\breve{\mathbf{o}}_{t-1}^{(l)}$
$\breve{\mathbf{o}}_{1:t-1}^{(l)}$
$\breve{\mathbf{o}}_{t}^{(l)}$
$\breve{\mathbf{o}}_{1:t}^{(l)}$
$\breve{\mathbf{o}}_{t+1}^{(l)}$
$\breve{\mathbf{o}}_{1:t+1}^{(l)}$
$\breve{\mathbf{x}}_{t-1}^{(l)}$
$\breve{\mathbf{x}}_{t-L_C+1:t-1}^{(l)}$
$\breve{\mathbf{x}}_{t}^{(l)}$
$\breve{\mathbf{x}}_{t-L_C+2:t}^{(l)}$
$\breve{\mathbf{x}}_{t+1}^{(l)}$
$\breve{\mathbf{x}}_{t-L_C+3:t+1}^{(l)}$

Figure 4.3: *DBN characterizing the dependencies between all involved RVs according to* (4.4) *and* (4.2) *after application of the conditional independence assumptions* (4.7) *and* (4.8).

4.2.1 A Priori Model for Speech

The a priori model $p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the clean speech LMPSC feature vector $\mathbf{x}_t^{(l)}$ shall incorporate knowledge about the past LMPSC feature vectors $\mathbf{x}_{t-L_C:t-1}^{(l)}$ and also take into account the information provided by the sequence of observed LMPSC feature vectors $\mathbf{o}_{1:t-1}^{(l)}$. To decouple the information provided by the past clean speech LMPSC feature vectors from that provided by the past observations, an underlying (hidden) sequence of discrete-valued *dynamic states* $m_{1:T}$ is introduced. The state sequence $m_{1:T}$ with $m_t \in \{1,\ldots,M\}$ is assumed to be a realization of the stochastic process $\breve{m}_{1:T}$. Thereby M denotes the total number of states the *dynamic model* may take.

The a priori model may thus formally be written in terms of the underlying state sequence as

$$\begin{aligned}
&p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right)\\
&\quad=\sum_{\{m_{1:T}\}} p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:T}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1},m_{1:T}\right)\\
&\qquad\qquad P_{\breve{m}_{1:T}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{1:T}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right) && (4.9)\\
&\quad=\sum_{\{m_{1:T}\}} p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:T}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}m_{1:T}\right)\\
&\qquad\qquad P_{\breve{m}_{t+1:T}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:t}}\left(m_{t+1:T}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1},m_{1:t}\right)\\
&\qquad\qquad P_{\breve{m}_{t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:t-1}}\left(m_{t}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1},m_{1:t-1}\right)\\
&\qquad\qquad P_{\breve{m}_{1:t-1}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{1:t-1}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right) && (4.10)
\end{aligned}$$

The decoupling of the information provided by the sequence of past clean speech LMPSC feature vectors and the past observations now comes with two assumptions.

The first assumption, a conditional independence assumption similar to (3.36), states that once the current state m_t and the past clean speech LMPSC feature vectors $\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1}$ are given, the clean speech LMPSC feature vector $\mathbf{x}_t^{(\mathsf{l})}$ is independent of the past observations $\mathbf{o}^{(\mathsf{l})}_{1:t-1}$ and all but the current state, i.e.,

$$\begin{aligned}
&p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:T}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1},m_{1:T}\right)\\
&\qquad\approx p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{m}_t}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},m_t\right) && (4.11)
\end{aligned}$$

The second assumption made models the stochastic process $\breve{m}_{1:T}$ a first-order MARKOV process, i.e.,

$$P_{\breve{m}_{t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{1:t-1}}\left(m_{t}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1},m_{1:t-1}\right)\approx P_{\breve{m}_t|\breve{m}_{t-1}}\left(m_t\,|m_{t-1}\right). \quad (4.12)$$

The a priori model may thus finally be approximated by

$$\begin{aligned}
&p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right)\\
&\quad\approx\sum_{\{m_t\}} p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{m}_t}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},m_t\right)\\
&\qquad\quad\sum_{\{m_{t-1}\}} P_{\breve{m}_t|\breve{m}_{t-1}}\left(m_t\,|m_{t-1}\right)P_{\breve{m}_{t-1}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{t-1}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right)\\
&&& (4.13)
\end{aligned}$$

where the last sum can be considered an approximation to the predictive state PMF $P_{\breve{m}_t|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$. For $t=1$ this sum simply gives the *a priori probability*, which will be denoted by $\pi_i\in[0,1]$, i.e.,

$$\pi_i := P_{\breve{m}_1}\left(m_1 = i\right), \quad (4.14)$$

satisfying $\sum\limits_{i=1}^{M} \pi_i = 1$. For all other time instants, this sum consists of the a posteriori state PMF $P_{\breve{m}_{t-1}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ and the PMF $P_{\breve{m}_t|\breve{m}_{t-1}}$. The latter denotes the probability of, e.g., switching from sub-model $m_{t-1} = j$ at time instant $t-1$ to sub-model $m_t = i$ at time instant t. These *switching probabilities* will be assumed to be time-invariant and will be denoted by $a_{i|j} \in [0,1]$, i.e.,

$$a_{i|j} := P_{\breve{m}_t|\breve{m}_{t-1}}\left(m_t = i\,|m_{t-1} = j\,\right), \tag{4.15}$$

satisfying $\sum\limits_{i=1}^{M} a_{i|j} = 1, \forall j \in \{1, \dots M\}$.

The sub-model specific predictive PDFs $p_{\breve{\mathbf{x}}^{(\mathsf{l})}_t|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{m}_t}$ will now be approximated by GAUSSIAN linear ***auto**regressive* (AR) models. In practice, the order to be chosen to model the AR process will depend on the correlation inherent to the underlying samples of the speech signal and on the correlation introduced by the feature extraction scheme, i.e., the frame size and the frame shift. In [66, p. 29] AR models of orders up to one have been found to describe the data with a reasonable accuracy. Similar findings have been presented in [67], where the modeling power of AR models of order one has been found to be comparable to that of AR models of higher order, however, at a lower computational cost. Following these findings, the current work focuses on AR models of order zero (denoted by AR-0 model) and of order one (denoted by AR-1 model).

AR-0 model For the AR-0 model, the sub-model specific predictive PDF for state $m_t = i$ is approximated by a single GAUSSIAN distribution with time-invariant mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}$ and $\mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ and time-invariant covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}$ and $\mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ for the first and all other time instants, respectively, according to

$$\begin{aligned} & p_{\breve{\mathbf{x}}^{(\mathsf{l})}_t|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C:t-1},\breve{m}_t}\left(\mathbf{x}^{(\mathsf{l})}_t \middle| \mathbf{x}^{(\mathsf{l})}_{t-L_C:t-1}, m_t = i\right) \\ & \approx p_{\breve{\mathbf{x}}^{(\mathsf{l})}_t|\breve{m}_t}\left(\mathbf{x}^{(\mathsf{l})}_t \middle| m_t = i\right) \end{aligned} \tag{4.16}$$

$$:= \begin{cases} \mathcal{N}\left(\mathbf{x}^{(\mathsf{l})}_t; \boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}\right), & \text{for } t = 1 \\ \mathcal{N}\left(\mathbf{x}^{(\mathsf{l})}_t; \mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}, \mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right), & \text{for } 1 < t \leq T \end{cases}. \tag{4.17}$$

Note that though the sub-model specific statistics of the clean speech LMPSC vector $\breve{\mathbf{x}}^{(\mathsf{l})}_t$ are time-invariant, the resulting a priori model (4.13) remains time-variant. Since it bears resemblance to a GMM, however, with time-variant weights, this kind of a priori model is denoted as **M**ARKOV *switching* **G**AUSSIAN ***m**ixture **m**odel* (MSGMM). The corresponding DBN is given in Fig. 4.4. The MSGMM reduces to a standard GMM if the switching probabilities $a_{i|j}$ in (4.15) are made independent of the predecessor state and if no special treatment is applied to the first frame of each sequence. Further note that the observation $\breve{\mathbf{o}}^{(\mathsf{l})}_t$ still depends on the complete state vector $\breve{\mathbf{z}}^{(\mathsf{l})}_t$, i.e., the clean speech LMPSC vector $\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C+1:t}$ and the LMPSC vector of the noise $\breve{\mathbf{n}}^{(\mathsf{l})}_t$.

AR-1 model In contrast to the AR-0 model, the AR-1 model now explicitly takes the correlation between the current and the past clean speech LMPSC feature vectors $\breve{\mathbf{x}}^{(\mathsf{l})}_t$

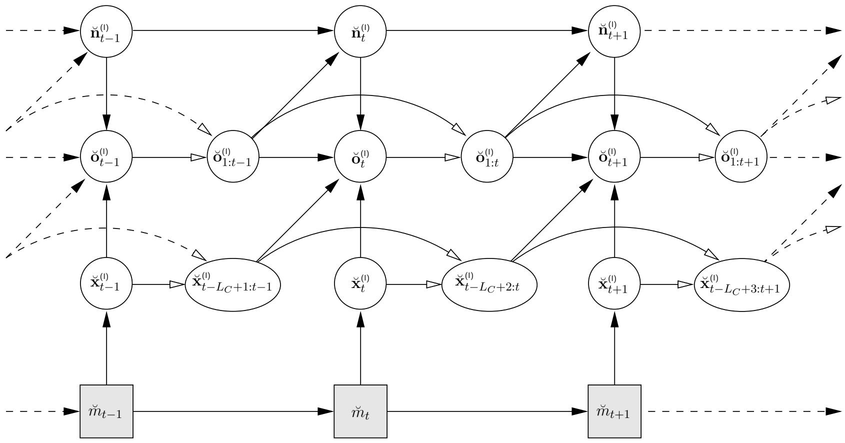

$\breve{\mathbf{n}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{n}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{n}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t+1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+1:t-1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+2:t}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+3:t+1}^{(\mathsf{l})}$
$\breve{m}_{t-1}$
$\breve{m}_{t}$
$\breve{m}_{t+1}$

Figure 4.4: *DBN characterizing the dependencies between all involved RVs according to* (4.4) *and* (4.2) *after application of the approximations* (4.7) *and* (4.8). *Further, the a priori model for the clean speech LMPSC feature vector trajectory is composed of AR-0 sub-models, resulting in an MSGMM.*

and $\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})}$ into account by modeling the mean of the Gaussian PDF $p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})},\breve{m}_t}$ a linear function of the clean speech LMPSC feature vector $\mathbf{x}_{t-1}^{(\mathsf{l})}$, i.e.,

$$\begin{aligned} & p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t-1}^{(\mathsf{l})},m_t=i\right) \\ &\approx p_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{x}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-1}^{(\mathsf{l})},m_t=i\right) \qquad (4.18) \\ &:= \begin{cases} \mathcal{N}\left(\mathbf{x}_t^{(\mathsf{l})};\boldsymbol{\mu}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i},\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}\right), & \text{for } t=1 \\ \mathcal{N}\left(\mathbf{x}_t^{(\mathsf{l})};\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})}+\mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i},\mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right), & \text{for } 1<t\leq T \end{cases}, \qquad (4.19) \end{aligned}$$

where $\boldsymbol{\mu}_{\mathbf{x}_1^{(\mathsf{l})}|i}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_1^{(\mathsf{l})}|i}$ are the mean vector and the covariance matrix of the ith sub-model specific distribution for the first time instant and $\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$, $\mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ and $\mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ are the *transition* matrix, the *prediction bias* vector and the *prediction error covariance* matrix.

The resulting a priori model (4.13) is termed a **M**arkov ***s****witching* ***l****inear* ***d****ynamic* ***m****odel* (MSLDM). The corresponding DBN is given in Fig. 4.5. It exploits the information present in the sequence of observations $\mathbf{o}_{1:t-1}^{(\mathsf{l})}$ by means of the a posteriori state PMF $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ and also considers the dynamic of the clean speech LMPSC feature vector trajectory by taking the correlation of two consecutive clean speech LMPSC feature vectors into account.

Note that the MSGMM (4.17) may be considered a special case of the MSLDM (4.19) where the transition matrices are all zero, i.e. $\mathbf{A}_i := \mathbf{0}_{Q\times Q}, \forall i \in \{1,\ldots,M\}$.

Model training Since the sequence of dynamic states is not observable, the set of AR-0 model parameters

$$\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}^{\mathsf{MSGMM}} := \left\{\pi_i, a_{i|j}, \boldsymbol{\mu}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}, \mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}, \mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} \middle| i,j \in \{1,\ldots,M\}\right\} \qquad (4.20)$$

characterizing the MSGMM and the set of AR-1 model parameters

$$\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}^{\mathsf{MSLDM}} := \left\{\pi_i, a_{i|j}, \boldsymbol{\mu}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}, \mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}, \mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}, \mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} \middle| i,j \in \{1,\ldots,M\}\right\} \qquad (4.21)$$

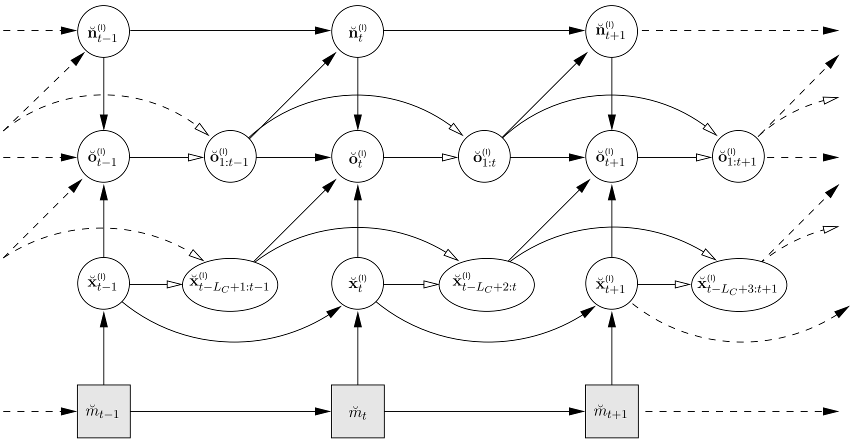

$\breve{\mathbf{n}}^{(\mathsf{l})}_{t-1}$
$\breve{\mathbf{n}}^{(\mathsf{l})}_{t}$
$\breve{\mathbf{n}}^{(\mathsf{l})}_{t+1}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{t}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{t+1}$
$\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t+1}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t-1}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C+1:t-1}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C+2:t}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t+1}$
$\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C+3:t+1}$
$\breve{m}_{t-1}$
$\breve{m}_{t}$
$\breve{m}_{t+1}$

Figure 4.5: *DBN characterizing the dependencies between all involved RVs according to* (4.4) *and* (4.2) *after application of the approximations* (4.7) *and* (4.8). *Further, the a priori model for the clean speech LMPSC feature vector trajectory is composed of AR-1 sub-models, resulting in an MSLDM.*

characterizing the MSLDM are trained in an unsupervised manner by application of the EM algorithm [59]. The EM algorithm is an iterative procedure to find an approximate ***m**aximum **l**ikelihood* (ML) solution to a set of model parameters $\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ (which may here be either $\Theta^{\mathsf{MSGMM}}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ or $\Theta^{\mathsf{MSLDM}}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$).

Assuming a set $X_{1:U}$ of U sequences of clean speech LMPSC training vectors $\mathbf{x}^{(\mathsf{l})}_{1:T(u),u}, u \in \{1,\ldots,U\}$ defined by

$$X_{1:U} := \left\{ \mathbf{x}^{(\mathsf{l})}_{1:T(u),u} \middle| u \in \{1,\ldots,U\} \right\} \tag{4.22}$$

and an initial set of model parameters, denoted by $\Theta^{[0]}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$, to be given, the EM algorithm iteratively maximizes a lower bound on the likelihood function

$$\mathcal{L}\left(\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right) := p_{\breve{X}_{1:U}}\left(X_{1:U}; \Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right). \tag{4.23}$$

The lower bound on the likelihood function (4.23) is obtained by maximizing the so-called *auxiliary function*

$$Q\left(\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}; \Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right) := E\left[\ln\left(p_{\breve{X}_{1:U}, \breve{S}_{1:U}}\left(X_{1:U}, S_{1:U}; \Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right)\right) \middle| X_{1:U}; \Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}} \right] \tag{4.24}$$

w.r.t. the model parameters $\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}$. Thereby

$$S_{1:U} := \left\{ m_{1:T(u),u} \middle| u \in \{1,\ldots,U\} \right\} \tag{4.25}$$

denotes the set of state sequences underlying the set of clean speech LMPSC sequences $X_{1:U}$ and $\Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ is the set of model parameters obtained after the ιth iteration of the EM algorithm. The new set of model parameters $\Theta^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ is then obtained as

$$\Theta^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}} = \underset{\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}}{\operatorname{argmax}} \left\{ Q\left(\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}; \Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right)\right\}. \tag{4.26}$$

Equation (4.24) and (4.26) make up the eponymous E- and M-step of the EM algorithm and are iterated until some convergence criteria, e.g., w.r.t. changes of the likelihood function and the number of EM iterations, are met.

With the a posteriori probabilities of a single state and a pair of successive states defined by

$$\gamma_{t,u}^{(\iota)}(i) := P_{\breve{m}_{t,u}|\breve{\mathbf{x}}^{(\mathsf{l})}_{1:T(u),u}}\left(m_{t,u}=i\,\middle|\,\mathbf{x}^{(\mathsf{l})}_{1:T(u),u}.\Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right), \tag{4.27}$$

$$\eta_{t,u}^{(\iota)}(j,i) := P_{\breve{m}_{t-1:t,u}|\breve{\mathbf{x}}^{(\mathsf{l})}_{1:T(u),u}}\left(m_{t-1,u}=j, m_{t,u}=i\,\middle|\,\mathbf{x}^{(\mathsf{l})}_{1:T(u),u},\Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right), \tag{4.28}$$

where $\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ has to be substituted by either $\Theta^{\mathsf{MSGMM}}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ or $\Theta^{\mathsf{MSLDM}}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ for the AR-0 and AR-1 model, respectively, the sub-model specific parameters obtained after the $(\iota+1)$th iteration of the EM algorithm are given by [68, 69]

$$\boldsymbol{\mu}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i} := \frac{\sum\limits_{u=1}^{U}\gamma_{1,u}^{(\iota)}(i)\,\mathbf{x}^{(\mathsf{l})}_{1,u}}{\sum\limits_{u=1}^{U}\gamma_{1,u}^{(\iota)}(i)}, \tag{4.29}$$

$$\boldsymbol{\Sigma}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i} := \frac{\sum\limits_{u=1}^{U}\gamma_{1,u}^{(\iota)}(i)\left(\mathbf{x}^{(\mathsf{l})}_{1,u}-\boldsymbol{\mu}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}\right)\left(\mathbf{x}^{(\mathsf{l})}_{1,u}-\boldsymbol{\mu}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}_1|i}\right)^{\dagger}}{\sum\limits_{u=1}^{U}\gamma_{1,u}^{(\iota)}(i)}, \tag{4.30}$$

$$\mathbf{V}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} := \frac{\sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\left(\mathbf{x}^{(\mathsf{l})}_{t,u}-\mathbf{A}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\mathbf{x}^{(\mathsf{l})}_{t-1,u}-\mathbf{b}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right)\left(\mathbf{x}^{(\mathsf{l})}_{t,u}-\mathbf{A}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\mathbf{x}^{(\mathsf{l})}_{t-1,u}-\mathbf{b}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right)^{\dagger}}{\sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)}, \tag{4.31}$$

where for the AR-1 model $\mathbf{A}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ and $\mathbf{b}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ are the solutions to the set of linear equations

$$\begin{aligned}&\begin{bmatrix}\sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\,\mathbf{x}^{(\mathsf{l})}_{t-1,u}\left(\mathbf{x}^{(\mathsf{l})}_{t-1,u}\right)^{\dagger} & \sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\,\mathbf{x}^{(\mathsf{l})}_{t-1,u}\\ \sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\left(\mathbf{x}^{(\mathsf{l})}_{t-1,u}\right)^{\dagger} & \sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\end{bmatrix}\begin{bmatrix}\left(\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right)^{\dagger}\\ \left(\mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}\right)^{\dagger}\end{bmatrix}\\ &=\begin{bmatrix}\sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\,\mathbf{x}^{(\mathsf{l})}_{t-1,u}\left(\mathbf{x}^{(\mathsf{l})}_{t,u}\right)^{\dagger}\\ \sum\limits_{u=1}^{U}\sum\limits_{t=2}^{T(u)}\gamma_{t,u}^{(\iota)}(i)\left(\mathbf{x}^{(\mathsf{l})}_{t,u}\right)^{\dagger}\end{bmatrix}\end{aligned} \tag{4.32}$$

For the AR-0 model, the transition matrix $\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ is fixed , i.e., $\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}=\mathbf{0}_{Q\times Q}$, and thus not subject to the M-step of the EM algorithm. However, for (4.31) to be applicable, formally $\mathbf{A}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ is set to

$$\mathbf{A}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} := \mathbf{A}^{(0)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} = \mathbf{0}_{Q\times Q}, \quad \forall \iota\in\mathbb{N}_{\geq 0}. \tag{4.33}$$

The solution to $\mathbf{b}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i+1}$ can then directly be given as

$$\mathbf{b}^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} := \frac{\sum\limits_{u=1}^{U} \sum\limits_{t=2}^{T(u)} \gamma^{(\iota)}_{t,u}(i)\,\mathbf{x}^{(\mathsf{l})}_{t,u}}{\sum\limits_{u=1}^{U} \sum\limits_{t=2}^{T(u)} \gamma^{(\iota)}_{t,u}(i)} \tag{4.34}$$

The initial state probabilities $\pi_i^{(\iota+1)}$ and the state transition probabilities $a_{i|j}^{(\iota+1)}$ are finally obtained as

$$\pi_i^{(\iota+1)} := \frac{\sum\limits_{u=1}^{U} \gamma^{(\iota)}_{1,u}(i)}{\sum\limits_{u=1}^{U} 1} \tag{4.35}$$

$$a_{i|j}^{(\iota+1)} := \frac{\sum\limits_{u=1}^{U} \sum\limits_{t=2}^{T(u)} \eta^{(\iota)}_{t,u}(j,i)}{\sum\limits_{u=1}^{U} \sum\limits_{t=2}^{T(u)} \gamma^{(\iota)}_{t-1,u}(j)}. \tag{4.36}$$

In comparison to an ML training, the EM algorithm thus compensates for the unknown state sequences by using the a posteriori probabilities (4.27) and (4.28) as *soft weights* in (4.29)–(4.34).

Though repeated application of the EM algorithm ensures a monotonic increase of the likelihood function (4.23), i.e.,

$$\mathcal{L}\left(\Theta^{(\iota+1)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right) \geq \mathcal{L}\left(\Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}\right), \quad \forall \iota \in \mathbb{N}_{\geq 0}, \tag{4.37}$$

the model parameters $\Theta^{(\iota)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ are only guaranteed to converge to a local maximum of the likelihood function $\mathcal{L}(\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}})$ [59].

This, however, makes the algorithm and the practical use of the obtained model parameters quite sensitive to the choice of the initial set of model parameters $\Theta^{(0)}_{\breve{\mathbf{x}}^{(\mathsf{l})}}$ – a problem commonly known as *seeding*.

The most sophisticated seeding algorithms thereby borrow ideas from the *k-means++* algorithm [70] as done in [71], the *fuzzy-k-means* clustering [72] as done in [73] or the *model splitting* carried out during the training of GMMs [74]. While the first two approaches try to directly infer the model parameters for the targeted M sub-models from the training data, the latter divides the seeding problem for M sub-models into the problem of seeding the parameters for $\tilde{M} < M$ sub-models and specifying a (usually deterministic) rule to obtain the parameters for the targeted M sub-models from the trained parameters of the $\tilde{M}$ sub-models. Repeated application of this rule, starting with the ML estimates of the model parameters for an a priori model consisting of just one sub-model, eventually yields the desired set of model parameters.

The model splitting approach also differs from the other two approaches by another fact: Due to the stochastic nature of the algorithms, the initial set of model parameters obtained from the *k-means++*-like algorithm and the one motivated by the *fuzzy-k-means* clustering are random and so is the final set of model parameters after application of the EM algorithm. The splitting approach, initialized with the ML estimates of the parameters for an a priori model with just one sub-model, however, is entirely deterministic.[1] Thus, without an evaluation of the models obtained by the three aforementioned approaches on some development data, neither of the three models can a priori be said to outperform the other two.

In this work, majorly the model splitting approach will be followed. However, to highlight the sensitivity of the proposed inference schemes w.r.t. the choice of the a priori model for speech, control experiments will also utilize previously employed *k-means++*-like initialized AR-1 a priori speech models.

4.2.2 A Priori Model for Noise

Due to the diversity of noise signals that may be encountered when operating the automatic speech recognizer in a real-word scenario, finding an appropriate a priori model $p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the LMPSC feature vector of the noise $\breve{\mathbf{n}}_t^{(l)}$ is much more challenging than finding an a priori model $p_{\breve{\mathbf{x}}_t^{(l)}|\breve{\mathbf{x}}_{t-L_C:t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}$ for the clean speech LMPSC feature vector $\breve{\mathbf{x}}_t^{(l)}$. Though, in theory, a Markov switching model may also be considered to characterize the a priori model for a specific noise type, in practice, accounting for all types of noises turns infeasible since it would i) require a huge amount of training data and ii) result in a large number of sub-models. A viable alternative to model all possible noise types in the a priori model at once is to assume the current utterance to be only affected by a certain type of noise. By further assuming the statistics of the noise LMPSC vector $\breve{\mathbf{n}}_t^{(l)}$ to be independent of the observation process $\breve{\mathbf{o}}_{1:t-1}^{(l)}$ once the realization $\mathbf{n}_{t-1}^{(l)}$ of the random variable $\breve{\mathbf{n}}_{t-1}^{(l)}$ is given, e.g.,

$$p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)},\breve{\mathbf{o}}_{1:t-1}^{(l)}}\left(\mathbf{n}_t^{(l)}\middle|\mathbf{n}_{t-1}^{(l)},\mathbf{o}_{1:t-1}^{(l)}\right) \approx p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)}}\left(\mathbf{n}_t^{(l)}\middle|\mathbf{n}_{t-1}^{(l)}\right), \tag{4.38}$$

training of the a priori model may be carried out on sequences of *noise-only* LMPSC feature vectors identified by some ***v**oice **a**ctivity **d**etection* (VAD). As a further simplification, only the AR-0 process is considered in this work. The involved parameters are thereby again assumed to be time-invariant. For the AR-0 process, the a priori model is thus given by

$$p_{\breve{\mathbf{n}}_t^{(l)}|\breve{\mathbf{n}}_{t-1}^{(l)}}\left(\mathbf{n}_t^{(l)}\middle|\mathbf{n}_{t-1}^{(l)}\right) \approx p_{\breve{\mathbf{n}}_t^{(l)}}\left(\mathbf{n}_t^{(l)}\right) \tag{4.39}$$

$$:= \begin{cases} \mathcal{N}\left(\mathbf{n}_t^{(l)};\boldsymbol{\mu}_{\breve{\mathbf{n}}_1^{(l)}},\boldsymbol{\Sigma}_{\breve{\mathbf{n}}_1^{(l)}}\right), & \text{for } t=1 \\ \mathcal{N}\left(\mathbf{n}_t^{(l)};\mathbf{b}_{\breve{\mathbf{n}}^{(l)}},\mathbf{V}_{\breve{\mathbf{n}}^{(l)}}\right), & \text{for } 1<t\leq T \end{cases}, \tag{4.40}$$

where $\boldsymbol{\mu}_{\breve{\mathbf{n}}_1^{(l)}}$ and $\mathbf{b}_{\breve{\mathbf{n}}^{(l)}}$ are the means for the first and all other time instants, respectively, and $\boldsymbol{\Sigma}_{\breve{\mathbf{n}}_1^{(l)}}$ and $\mathbf{V}_{\breve{\mathbf{n}}^{(l)}}$ denote the corresponding covariance matrices.

The parameters are essentially given by (4.29)-(4.31) and (4.34). However, an utterance is now considered to be a sequences of consecutive LMPSC feature vectors marked as noise-only by the VAD. Further, the noise LMPSC feature vectors identified by the VAD

[1] Provided that the splitting rule is deterministic.

have to be substituted for the clean speech LMPSC feature vectors and the a posteriori probabilities have to be set to one.

The final DBN characterizing the dependencies between all involved variables in the inference is exemplarily given in Fig. 4.6 for the combination of an MSLDM for the clean speech LMPSC feature vector trajectory and an AR-0 model for the trajectory of the LMPSC feature vector of the noise.

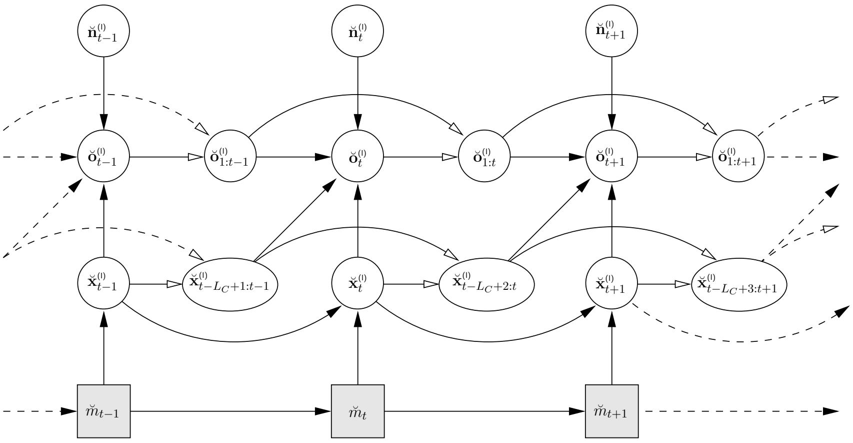

$\breve{\mathbf{n}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{n}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{n}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{o}}_{1:t+1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+1:t-1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+2:t}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t+1}^{(\mathsf{l})}$
$\breve{\mathbf{x}}_{t-L_C+3:t+1}^{(\mathsf{l})}$
$\breve{m}_{t-1}$
$\breve{m}_{t}$
$\breve{m}_{t+1}$

Figure 4.6: *DBN characterizing the dependencies between all involved RVs according to* (4.4) *and* (4.2) *after application of the approximations* (4.7) *and* (4.8). *Further, the a priori model for the clean speech LMPSC feature vector trajectory is composed of AR-1 sub-models, resulting in an MSLDM. The a priori model for the noise LMPSC feature vector trajectory is assumed to be an AR-0 model.*

4.3 Observation Models

The a priori model $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ discussed in the previous section is the core component of the prediction step. Given the a posteriori PDF $p_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ at time instant $t-1$, the integral in (4.2) eventually characterizes the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$.

The observation model $p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ now builds the core component of the update step. Given the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$, the a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ can be obtained from (4.2).

Employing the definition of the state vector $\mathbf{z}_t^{(\mathsf{l})}$ in (4.1), the observation model may also be written as

$$p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{z}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)=p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C+1:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right). \tag{4.41}$$

The observation model thus calls for a characterization of the statistical dependency between the current observation $\breve{\mathbf{o}}_t^{(\mathsf{l})}$ and the (current and past) L_C clean speech LMPSCs $\breve{\mathbf{x}}_{t-L_C+1:t}^{(\mathsf{l})}$, the current LMPSC of the noise $\breve{\mathbf{n}}_t^{(\mathsf{l})}$ and all past observed LMPSC vectors $\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}$.

The following subsections gradually lay out the derivation of tractable, Gaussian approximations to the observation models in i) the presence of reverberation and the absence of noise (Sec. 4.3.2, Eq. (4.96)), ii) the presence of reverberation and background noise (Sec. 4.3.3, Eq. (4.130)) and iii) the absence of reverberation and the presence of background noise (Sec. 4.3.4, Eq. (4.143)).

4.3.1 From a Deterministic Relation in the Short-Time Discrete-Time Fourier Domain to a Stochastic Relation in the Logarithmic Mel Power Spectral Domain

All derivations start from a deterministic relation between the SC $O_t(k)$ of the noisy reverberant speech signal $o(p)$ and the SCs of the underlying clean speech signal $x(p)$ and the noise signal $n(p)$, denoted by $X_t(k)$ and $N_t(k)$, respectively. Though this deterministic relation has already been published in [75] it will be repeated here (in a slightly different and more condensed form) not only for convenience, but also to introduce the required notation.

Due to a loss of information inherent to the signal model and the employed feature extraction scheme (see Sec. 3.1), this deterministic relation cannot be held up when actually transforming the SCs to the logarithmic mel power spectral domain. However, the relation between the LMPSC $\mathbf{o}_t^{(\mathsf{l})}$ and the LMPSCs $\mathbf{x}_{t-L_C+1:t}^{(\mathsf{l})}$, $\mathbf{n}_t^{(\mathsf{l})}$ and $\mathbf{o}_{1:t-1}^{(\mathsf{l})}$ may be described in a stochastic rather than deterministic way. The stochastic component in the resulting observation model describes the remaining uncertainty about the observation by means of an observation error, which eventually accounts for all approximations necessary to come up with a tractable observation model.

4.3.1.1 Deterministic Relation in the Short-Time Discrete-Time Fourier Domain

The SC of the noisy reverberant speech signal is first, due to the linearity of the ***d**iscrete-**t**ime **s**hort-**t**ime* **Fourier** ***t**ransformation* (DTSTFT), written as

$$O_t(k) = S_t(k) + N_t(k). \tag{4.42}$$

According to [75], the SC $S_t(k)$ of the reverberant speech signal $s(p)$ may now be expressed in terms of the SCs of the underlying clean speech signal $x(p)$ by means of the following steps.

First, the signal model (3.109) is plugged into the definition of the SC as used in (3.12). The SC of the reverberant speech signal is thus given by

$$\begin{aligned} S_t(k) &= \sum_{l=0}^{L_{w_\mathsf{A}}-1} s_t(l)\,\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk} && (4.43)\\ &= \sum_{l=0}^{L_{w_\mathsf{A}}-1} w_\mathsf{A}(l)\,s(l+tB)\,\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk} && (4.44)\\ &= \sum_{l=0}^{L_{w_\mathsf{A}}-1} w_\mathsf{A}(l)\left(\sum_{p'=0}^{L_h-1} h\left(p'\right)x\left(l+tB-p'\right)\right)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk}. && (4.45)\end{aligned}$$

Next, the clean speech signal $x(p)$ is considered to be synthesized from its SCs according to [76, ch. 3, p. 49]

$$x(p) = \sum_{t'=-\infty}^{\infty} \sum_{k'=0}^{K-1} w_{\mathsf{S}}\left(p - t'B\right) X_{t'}\left(k'\right) \mathrm{e}^{\mathrm{j}\frac{2\pi}{K}(p-t'B)k'}, \tag{4.46}$$

where $w_{\mathsf{S}}(l) \in \mathbb{R}$ denotes a synthesis window of length $L_{w_{\mathsf{S}}} \in \mathbb{N}_{>0}$ with support on $l \in \{0, L_{w_{\mathsf{S}}} - 1\}$. For the speech signal $x(p)$ to be exactly recovered from its SCs by (4.46), the analysis and synthesis window have to meet the so-called *completeness condition* [77], which, after choosing equal lengths of the analysis and synthesis window, i.e., $L_{w_{\mathsf{A}}} = L_{w_{\mathsf{S}}} = L_w \leq K$, is given by [67, Ch. A2, Eq. (A.80)]

$$\sum_{t'=0}^{\lceil \frac{L_w}{B} \rceil} w_{\mathsf{A}}\left(p - t'B\right) w_{\mathsf{S}}\left(p - t'B\right) = \frac{1}{K}, \quad \forall p \in \{0, B-1\}. \tag{4.47}$$

Assuming the analysis and synthesis window to meet the completeness condition (4.47), the SC of the reverberant speech signal given by (4.45) may eventually be written as

$$S_t(k) = \sum_{t'=-\infty}^{\infty} \sum_{k'=0}^{K-1} X_{t'}\left(k'\right) h_{t-t'}\left(k, k'\right) \tag{4.48}$$

$$= \sum_{k'=0}^{K-1} \sum_{t'=-\infty}^{\infty} X_{t-t'}\left(k'\right) h_{t'}\left(k, k'\right) \tag{4.49}$$

where

$$h_t\left(k, k'\right) := \sum_{p'=0}^{L_h-1} h\left(p'\right) \sum_{l=0}^{L_w-1} w_{\mathsf{A}}(l)\, w_{\mathsf{S}}\left(l + tB - p'\right) \mathrm{e}^{\mathrm{j}\frac{2\pi}{K}(l+tB-p')k'} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk} \tag{4.50}$$

are the so-called *cross-band filters* for $k \neq k'$ and *band-to-band filters* for $k = k'$ [75]. In [75] it has also been shown that the energy of a cross-band/band-to-band filter $h_t(k, k')$ decreases as $|k - k'| \bmod K$ increases.

The SC $S_t(k)$ of the reverberant signal $s(p)$ given in (4.49) may be interpreted as a sum of convolutions of the SCs $X_t(k')$ of the clean speech signal $x(p)$ with the cross-band/band-to-band filter $h_t(k, k')$, where the summation has to be carried out over all frequency bands k'. Eq. (4.50) may now be written as

$$h_t\left(k, k'\right) = \sum_{p'=0}^{L_h-1} h\left(p'\right) \mathit{\Phi}_{tB-p'}\left(k, k'\right) \tag{4.51}$$

$$= \sum_{p''=tB-L_h+1}^{tB} h\left(tB - p''\right) \mathit{\Phi}_{p''}\left(k, k'\right) \tag{4.52}$$

with the auxiliary function

$$\mathit{\Phi}_p\left(k, k'\right) := \sum_{l=0}^{L_w-1} w_{\mathsf{A}}(l)\, w_{\mathsf{S}}(l + p)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{K}(l+p)k'} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}lk}. \tag{4.53}$$

Noting that $\Phi_p(k,k')$ is, due to the limited support of the two window functions $w_{\mathsf{A}}(l)$ and $w_{\mathsf{S}}(l)$, only non-zero for $-L_w+1 \leq p \leq L_w-1$, the cross-band/band-to-band filter $h_t(k,k')$ may finally be written as

$$h_t\left(k,k'\right) = \sum_{p''=\mathcal{L}(t)}^{\mathcal{U}(t)} h\left(tB-p''\right)\Phi_{p''}\left(k,k'\right), \tag{4.54}$$

where

$$\mathcal{L}(t) = \max\{tB-L_h+1, -L_w+1\} \tag{4.55}$$
$$\mathcal{U}(t) = \min\{tB, L_w-1\}. \tag{4.56}$$
$$\tag{4.57}$$

Since the AIR $h(p)$ is assumed to be causal and of finite length L_h, the computation of SC $S_t(k)$ by a summation over an infinite number of elements according to (4.49) eventually simplifies to a summation over a finite number of elements as

$$S_t(k) = \sum_{t'=-L_{H,\ell}}^{L_H}\sum_{k'=0}^{K-1} X_{t-t'}\left(k'\right)h_{t'}\left(k,k'\right). \tag{4.58}$$

The lower limit

$$L_{H,\ell} = \left\lfloor \frac{L_w-1}{B} \right\rfloor, \tag{4.59}$$

where $\lfloor\cdot\rfloor$ denotes the flooring operator, is thereby characterized by that t', for which $t'B < -L_w+1$ Similarly, the upper limit

$$L_H = \left\lfloor \frac{L_w+L_h-2}{B} \right\rfloor \tag{4.60}$$

is characterized by that t', for which $t'B-L_h+1 > L_w-1$.

Employing the found representation (4.58) of the reverberant signal in the SC domain, the SC $O_t(k)$ of the noisy reverberant speech signal $o(p)$ is thus eventually given by

$$O_t(k) = S_t(k) + N_t(k) \tag{4.61}$$
$$= \sum_{t'=-L_{H,\ell}}^{L_H}\sum_{k'=0}^{K-1} X_{t-t'}\left(k'\right)h_{t'}\left(k,k'\right) + N_t(k). \tag{4.62}$$

4.3.1.2 Stochastic Relation in the Logarithmic Mel Power Spectral Domain

Equation (4.62) expresses the SC $O_t(k)$ of the noisy reverberant speech signal $o(p)$ in terms of the underlying SCs of the clean speech signal $x(p)$ and the signal of the noise $n(p)$. This expression is deterministic, i.e., given the SCs of the clean speech signal $x(p)$ and the signal of the noise $n(p)$, the SC $O_t(k)$ can uniquely be recovered from them.

Looking for an equivalent deterministic relation in the power spectral domain, where

$$
\begin{aligned}
|O_t(k)|^2 &= (S_t(k) + N_t(k))\,(S_t(k) + N_t(k))^* && (4.63)\\
&= S_t(k)\,S_t^*(k) + 2\,\mathrm{Re}\{(S_t(k)\,N_t^*(k)\} + N_t(k))\,N_t^*(k) && (4.64)\\
&= \sum_{t'=-L_{H,\ell}}^{L_H} \sum_{k'=0}^{K-1} \left|X_{t-t'}\left(k'\right)\right|^2 \left|h_{t'}\left(k,k'\right)\right|^2 \\
&\quad + \sum_{t'=-L_{H,\ell}}^{L_H} \sum_{\substack{k',k''=0\\ k'\neq k''}}^{K-1} X_{t-t'}\left(k'\right) X_{t-t'}^*\left(k''\right) h_{t'}\left(k,k'\right) h_{t'}^*\left(k,k''\right) \\
&\quad + \sum_{\substack{t',t''=-L_{H,\ell}\\ t'\neq t''}}^{L_H} \sum_{k',k''=0}^{K-1} X_{t-t'}\left(k'\right) X_{t-t''}^*\left(k''\right) h_{t'}\left(k,k'\right) h_{t''}^*\left(k,k''\right) \\
&\quad + 2\,\mathrm{Re}\left\{\left(\sum_{t'=-L_{H,\ell}}^{L_H} \sum_{k'=0}^{K-1} X_{t-t'}\left(k'\right) h_{t'}\left(k,k'\right)\right) N_t^*(k)\right\} + |N_t(k)|^2 && (4.65)
\end{aligned}
$$

with $\mathrm{Re}\{\cdot\}$ denoting the real part operator, it becomes apparent that, given only the PSCs of the clean speech signal $x(p)$ and the signal of the noise $n(p)$, the PSC of the noisy reverberant speech signal cannot be recovered uniquely. Hence, the relation among the PSCs has to be formulated in a stochastic rather than deterministic manner. The stochastic description in the power spectral domain will then automatically lead to a stochastic description in the mel power spectral domain and finally in the targeted logarithmic mel power spectral domain.

For ease of derivation, the stochastic model for the LMPSCs of the noisy reverberant speech signal will be developed from the stochastic model for the LMPSCs of the reverberant but noise-free speech signal. The latter one has been derived in full detail in [67]. However, since it builds the basis for the stochastic model for the LMPSCs of the noisy reverberant speech signal, its derivation will be repeated here. Until stated otherwise, the AIR $h(p)$ is thereby assumed to be a known, deterministic quantity. Later, in Sec. 4.4, this assumption will be dropped and a stochastic model for the AIR $\breve{h}(p)$ will be introduced, instead.

4.3.2 Presence of Reverberation and Absence of Background Noise

In the absence of background noise, the observable SCs $O_t(k)$ are equal to the ones of the reverberant speech signal $s(p)$. The same holds for the observed LMPSC feature vector $\mathbf{o}_t^{(\mathsf{l})}$ and the reverberant LMPSC feature vector $\mathbf{s}_t^{(\mathsf{l})}$. For the conditional PDF (4.41) it thus holds

$$
p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right) \rightarrow p_{\breve{\mathbf{s}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_C:t}^{(\mathsf{l})},\breve{\mathbf{s}}_{1:t}^{(\mathsf{l})}}\left(\mathbf{s}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_C:t}^{(\mathsf{l})},\mathbf{s}_{1:t-1}^{(\mathsf{l})}\right). \tag{4.66}
$$

Starting in the power spectral domain, the derivation of a tractable analytic form for (4.66) will be presented in the following.

From (4.65) the PSC $S_t(k)$ of the reverberant speech signal can be found to be

$$|S_t(k)|^2 = \sum_{t',t''=-L_{H,\ell}}^{L_H} \sum_{k',k''=0}^{K-1} X_{t-t'}\left(k'\right) X^*_{t-t''}\left(k''\right) h_{t'}\left(k,k'\right) h^*_{t''}\left(k,k''\right) \tag{4.67}$$

$$\begin{aligned} &= \sum_{t'=-L_{H,\ell}}^{L_H} \sum_{k'=0}^{K-1} \left|X_{t-t'}\left(k'\right)\right|^2 \left|h_{t'}\left(k,k'\right)\right|^2 \\ &+ \sum_{t'=-L_{H,\ell}}^{L_H} \sum_{\substack{k',k''=0 \\ k'\neq k''}}^{K-1} X_{t-t'}\left(k'\right) X^*_{t-t'}\left(k''\right) h_{t'}\left(k,k'\right) h^*_{t'}\left(k,k''\right) \\ &+ \sum_{\substack{t',t''=-L_{H,\ell} \\ t'\neq t''}}^{L_H} \sum_{k',k''=0}^{K-1} X_{t-t'}\left(k'\right) X^*_{t-t''}\left(k''\right) h_{t'}\left(k,k'\right) h^*_{t''}\left(k,k''\right). \end{aligned} \tag{4.68}$$

Following [67] the PSC of the reverberant speech signal will now be written as

$$|S_t(k)|^2 = C_P(k) \sum_{t'=0}^{L_H} |X_{t-t'}(k)|^2 |h_{t'}(k,k)|^2 + E_t(k). \tag{4.69}$$

The introduced *error term* $E_t(k)$ and the so-called frequency dependent *power compensation constant* $C_P(k) \in \mathbb{R}_{>0}$ thereby account for all terms in (4.67) where $t' \neq t''$ (third line in (4.68)), $t' = t''$ but $k' \neq k''$ (second line in (4.68)) and the non-causal terms where $t' < 0$ and $k' \neq 0$ (first line in (4.68)).

The error term is now assumed to be a realization of a real-valued random variable $\breve{E}_t(k)$ and the power compensation constant $C_P(k)$ is assumed to be a deterministic quantity chosen such that the error term is zero-mean, i.e.,

$$E\left[\breve{E}_t(k)\right] \overset{!}{=} 0. \tag{4.70}$$

$$\Leftrightarrow$$

$$E\left[\left|\breve{S}_t(k)\right|^2\right] \overset{!}{=} C_P(k)\, E\left[\sum_{t'=0}^{L_H} \left|\breve{X}_{t-t'}(k)\right|^2 |h_{t'}(k,k)|^2\right]$$

$$\Leftrightarrow$$

$$C_P(k) \overset{!}{=} \frac{E\left[\left|\breve{S}_t(k)\right|^2\right]}{E\left[\sum_{t'=0}^{L_H} \left|\breve{X}_{t-t'}(k)\right|^2 |h_{t'}(k,k)|^2\right]} \tag{4.71}$$

$$= \frac{E\left[\left|\sum_{t'=-L_{H,\ell}}^{L_H} \sum_{k'=0}^{K-1} \breve{X}_{t-t'}(k')\, h_{t'}(k,k')\right|^2\right]}{E\left[\sum_{t'=0}^{L_H} \left|\breve{X}_{t-t'}(k)\right|^2 |h_{t'}(k,k)|^2\right]}. \tag{4.72}$$

The required expectations can only be computed if the statistics of the LMPSCs computed from the clean signal $x(p)$, or, equivalently, those of the clean speech signal itself, are known. Since this is rarely the case, a tractable solution may be obtained if, e.g., the clean signal $x(p)$ is assumed to be a realization of a real-valued white GAUSSIAN random process. The resulting constant is then given by (see Appendix A.6)

$$C_P(k) := \frac{C_N(k)}{C_D(k)}, \tag{4.73}$$

where

$$C_N(k) := \sum_{t',t''=-L_{H,\ell}}^{L_H} \sum_{l=0}^{L_{w_\mathsf{A}}-1} w_\mathsf{A}(l)\, w_\mathsf{A}\left(l + \left(t''-t'\right)B\right) \\ \sum_{k',k''=0}^{K-1} h_{t'}\left(k,k'\right) h^*_{t''}\left(k,k''\right) \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk' - \left(l+\left[t''-t'\right]B\right)k''\right)} \tag{4.74}$$

and

$$C_D(k) := \left(\sum_{l=0}^{L_{w_\mathsf{A}}-1} w^2_\mathsf{A}(l)\right) \sum_{t'=0}^{L_H} \left|h_{t'}(k,k)\right|^2. \tag{4.75}$$

Matters get even more involved when taking (4.69) to the mel power spectral domain. The MPSC of the reverberant speech signal is given by

$$\begin{aligned} s^{(\mathsf{m})}_t(q) &= \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|S_t(k)\right|^2 & (4.76)\\ &= \sum_{t'=0}^{L_H} \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|X_{t-t'}(k)\right|^2 C_P(k) \left|h_{t'}(k,k)\right|^2 + \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, E_t(k) & (4.77)\\ &= \sum_{t'=0}^{L_H} \mathcal{H}^X_{t,t'}(q) \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|X_{t-t'}(k)\right|^2 + \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, E_t(k) & (4.78)\\ &= \sum_{t'=0}^{L_H} \mathcal{H}^X_{t,t'}(q)\, x^{(\mathsf{m})}_{t-t'}(q) + e^{(\mathsf{m})}_t(q) & (4.79) \end{aligned}$$

where $e^{(\mathsf{m})}_t(q)$ denotes the MPSC of the error term and where

$$\mathcal{H}^X_{t,t'}(q) := \frac{\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|X_{t-t'}(k)\right|^2 C_P(k) \left|h_{t'}(k,k)\right|^2}{\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|X_{t-t'}(k)\right|^2} \tag{4.80}$$

denotes the AIR representation in the mel power spectral domain. Obviously, (4.79) can only be expressed in terms of the MPSCs $x^{(\mathsf{m})}_t(q)$ of the clean speech signal without introducing any additional uncertainty if $\mathcal{H}^X_{t,t'}(q)$ does not depend on the PSCs of the clean speech signal. Looking at (4.80), this, however, may only hold for arbitrary clean speech signals if the product of the frequency dependent power compensation constant and the power of the band-to-band filters is constant over the frequency bins covered by the current mel filter q.

In [67], where a frequency independent power compensation constant has been employed, it has thus been proposed to approximate the power of the band-to-band filters in the qth mel band by their arithmetic mean. However here, in line with the computation of the frequency dependent power compensation constant $C_P(k)$, the AIR representation in the mel power spectral domain $\mathcal{H}^X_{t,t'}(q)$ will be approximated by $\mathcal{H}_{t'}(q)$, where $\mathcal{H}_{t'}(q)$ is chosen to satisfy

$$\begin{aligned} &E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, C_P(k) \left|\breve{X}_{t-t'}(k)\right|^2 \left|h_{t'}(k,k)\right|^2\right] \\ &\overset{!}{=} \mathcal{H}_{t'}(q)\, E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|\breve{X}_{t-t'}(k)\right|^2\right]. \end{aligned} \tag{4.81}$$

This eventually leads to

$$\begin{aligned} s_t^{(\mathsf{m})}(q) &= \sum_{t'=0}^{L_H} \left(\mathcal{H}_{t'}(q)\, x_{t-t'}^{(\mathsf{m})}(q) + \epsilon_{t,t'}^{(\mathsf{m})}(q)\right) + e_t^{(\mathsf{m})}(q) && (4.82)\\ &= \sum_{t'=0}^{L_H} \mathcal{H}_{t'}(q)\, x_{t-t'}^{(\mathsf{m})}(q) + \sum_{t'=0}^{L_H} \epsilon_{t,t'}^{(\mathsf{m})}(q) + e_t^{(\mathsf{m})}(q) && (4.83)\\ &= \sum_{t'=0}^{L_H} \mathcal{H}_{t'}(q)\, x_{t-t'}^{(\mathsf{m})}(q) + \bar{\epsilon}_t^{(\mathsf{m})}(q) + e_t^{(\mathsf{m})}(q). && (4.84) \end{aligned}$$

The new error terms $\epsilon_{t,t'}^{(\mathsf{m})}(q)$ thereby account for the approximations of $\mathcal{H}^X_{t,t'}(q)$ by $\mathcal{H}_{t'}(q)$. Choosing $\mathcal{H}_{t'}(q)$ to satisfy (4.81) thus ensures $\epsilon_{t,t'}^{(\mathsf{m})}(q)$ to be of zero mean and thus also their sum $\bar{\epsilon}_t^{(\mathsf{m})}(q)$. Since the power compensation constant $C_P(k)$ is chosen such that $E_t(k)$ and thus also $e_t^{(\mathsf{m})}(q)$ is of zero mean, too, the final approximation error, i.e., $\bar{\epsilon}_t^{(\mathsf{m})}(q) + e_t^{(\mathsf{m})}(q)$ also exhibits a zero mean.

For a practical realization, the clean signal $x(p)$ may again be assumed to be a realization of a real-valued white GAUSSIAN random process eventually resulting in (see Appendix A.7)

$$\mathcal{H}^X_{t,t'}(q) \approx \mathcal{H}_{t'}(q) := \frac{\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, C_P(k) \left|h_{t'}(k,k)\right|^2}{\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)}. \tag{4.85}$$

Finally, by applying the natural logarithm to both sides of (4.82), the LMPSC of the reverberant speech signal $s(p)$ is given by

$$\begin{aligned} s_t^{(\mathrm{l})}(q) &= \ln\left(s_t^{(\mathrm{m})}(q)\right) && (4.86)\\ &= \ln\left(\sum_{t'=0}^{L_H} \mathcal{H}_{t'}(q)\, x_{t-t'}^{(\mathrm{m})}(q) + \bar{\epsilon}_t^{(\mathrm{m})}(q) + e_t^{(\mathrm{m})}(q)\right) && (4.87)\\ &= \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{x_{t-t'}^{(\mathrm{l})}(q) + \bar{h}_{t'}^{(\mathrm{l})}(q)}\right) + v_{s_t}^{(\mathrm{l})}(q), && (4.88)\end{aligned}$$

with

$$\bar{h}_{t'}^{(\mathrm{l})}(q) := \ln\left(\mathcal{H}_{t'}(q)\right) \qquad (4.89)$$

denoting the approximate representation of the AIR in the logarithmic mel power spectral domain and

$$v_{s_t}^{(\mathrm{l})}(q) := \ln\left(1 + \frac{\bar{\epsilon}_t^{(\mathrm{m})}(q) + e_t^{(\mathrm{m})}(q)}{\sum\limits_{t'=0}^{L_H} \mathcal{H}_{t'}(q)\, x_{t-t'}^{(\mathrm{m})}(q)}\right) \qquad (4.90)$$

characterizing the error in that very domain.

Given the LMPSCs of the clean speech signal, this error term is the only stochastic component when describing the LMPSC of the observation in the noise-free case by (4.88) and is thus denoted as the *observation error*.

The corresponding LMPSC feature vector $\mathbf{s}_t^{(\mathrm{l})}$ may thus be written in terms of the clean speech LMPSC feature vectors $\mathbf{x}_{t-L_H:t}^{(\mathrm{l})}$ and the LMPSC vector of the error $\mathbf{v}_{s_t}^{(\mathrm{l})}$ as

$$\begin{aligned} \mathbf{s}_t^{(\mathrm{l})} &= \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{l})} + \bar{\mathbf{h}}_{t'}^{(\mathrm{l})}}\right) + \mathbf{v}_{s_t}^{(\mathrm{l})} && (4.91)\\ &= f_s^{(\mathrm{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathrm{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathrm{l})}\right) + \mathbf{v}_{s_t}^{(\mathrm{l})} && (4.92)\end{aligned}$$

where

$$f_s^{(\mathrm{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathrm{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathrm{l})}\right) := \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{l})} + \bar{\mathbf{h}}_{t'}^{(\mathrm{l})}}\right) \qquad (4.93)$$

will in the following be denoted as the *observation mapping*. All mathematical operations thereby have to be understood to be applied to the vectors component-wise.

The error vector $\mathbf{v}_{s_t}^{(\mathrm{l})}$ will now be assumed to be a realization of a RV $\breve{\mathbf{v}}_{s_t}^{(\mathrm{l})}$ with conditional PDF $p_{\breve{\mathbf{v}}_{s_t}^{(\mathrm{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathrm{l})}, \breve{\mathbf{s}}_{1:t-1}^{(\mathrm{l})}}$. The stochastic observation model for reverberant speech thus turns into

$$\begin{aligned} & p_{\breve{\mathbf{s}}_t^{(\mathrm{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathrm{l})}, \breve{\mathbf{s}}_{1:t-1}^{(\mathrm{l})}}\left(\mathbf{s}_t^{(\mathrm{l})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathrm{l})}, \mathbf{s}_{1:t-1}^{(\mathrm{l})}\right)\\ &= \; p_{\breve{\mathbf{v}}_{s_t}^{(\mathrm{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathrm{l})}, \breve{\mathbf{s}}_{1:t-1}^{(\mathrm{l})}}\left(\mathbf{s}_t^{(\mathrm{l})} - f_s^{(\mathrm{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathrm{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathrm{l})}\right) \middle| \mathbf{x}_{t-L_H:t}^{(\mathrm{l})}, \mathbf{s}_{1:t-1}^{(\mathrm{l})}\right). \qquad (4.94)\end{aligned}$$

The observation model in the absence of background noise is thus completely determined by the conditional PDF $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{s}}^{(\mathsf{l})}_{1:t-1}}$ of the observation error $\mathbf{v}^{(\mathsf{l})}_{s_t}$.

In [67], the observation error $\mathbf{v}^{(\mathsf{l})}_{s_t}$ has been been considered as a realization of a real-valued white, stationary and ergodic Gaussian process that is independent of the most recent L_H+1 LMPSC vectors of the clean speech signal and all past LMPSC vectors of the reverberant speech signal, i.e.,

$$p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{s}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{v}^{(\mathsf{l})}_{s_t}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_H:t},\mathbf{s}^{(\mathsf{l})}_{1:t-1}\right)\approx p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}}\left(\mathbf{v}^{(\mathsf{l})}_{s_t}\right):=\mathcal{N}\left(\mathbf{v}^{(\mathsf{l})}_{s_t};\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s},\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}\right), \tag{4.95}$$

which is employed $\forall t\in\{1,\cdots,T\}$. The mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}$ have thereby been obtained from artificially reverberated training data and the underlying clean data. A detailed analysis of the observation error in the presence of reverberation and the absence of noise will follow in Sec. 4.7.1.

The observation model in the absence of background noise will thus finally be approximated by

$$p_{\breve{\mathbf{s}}^{(\mathsf{l})}_t|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{s}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{s}^{(\mathsf{l})}_t\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_H:t},\mathbf{s}^{(\mathsf{l})}_{1:t-1}\right)\approx\mathcal{N}\left(\mathbf{s}^{(\mathsf{l})}_t;\ln\left(\sum_{t'=0}^{L_H}\mathrm{e}^{\mathbf{x}^{(\mathsf{l})}_{t-t'}+\bar{\mathbf{h}}^{(\mathsf{l})}_{t'}}\right)+\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s},\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}\right). \tag{4.96}$$

4.3.3 Presence of Reverberation and Background Noise

In the presence of background noise, the noisy reverberant PSCs $|O_t(k)|^2$ is given by (4.64), which could also be written as

$$\begin{aligned}|O_t(k)|^2&=|S_t(k)|^2+2\,\mathrm{Re}\{S_t(k)\,N^*_t(k)\}+|N_t(k)|^2 &(4.97)\\&=|S_t(k)|^2+2\,|S_t(k)|\,|N_t(k)|\cos\left(\varphi_{S_t(k),N_t(k)}\right)+|N_t(k)|^2, &(4.98)\end{aligned}$$

where $\varphi_{S_t(k),N_t(k)}$ denotes the phase difference between the complex-valued SCs of the reverberant speech signal $s(p)$ and the noise signal $n(p)$ at time instant t and frequency bin k. However, since only the superposition of the reverberant speech signal and the noise can be observed, any information about the phase between $S_t(k)$ and $N_t(k)$ is irretrievably lost. The observation model for noisy reverberant speech thus has to account for the uncertainty about this phase, however, not in the power spectral but in the logarithmic mel power spectral domain.

In the mel power spectral domain, the MPSC of the observed noisy reverberant speech signal is given by

$$\begin{aligned}o^{(\mathsf{m})}_t(q)=&\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}}\Lambda_q(k)\,|S_t(k)|^2+2\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}}\Lambda_q(k)\,|S_t(k)|\,|N_t(k)|\cos\left(\varphi_{S_t(k),N_t(k)}\right)\\&+\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}}\Lambda_q(k)\,|N_t(k)|^2 &(4.99)\\=&\,s^{(\mathsf{m})}_t(q)+2\alpha_t(q)\sqrt{s^{(\mathsf{m})}_t(q)}\sqrt{n^{(\mathsf{m})}_t(q)}+n^{(\mathsf{m})}_t(q), &(4.100)\end{aligned}$$

where the introduced *phase factor* $\alpha_t(q) \in [-1,+1]$ is defined by[2]

$$\alpha_t(q) := \frac{\sum\limits_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} \Lambda_q(k)\,|S_t(k)|\,|N_t(k)| \cos\left(\varphi_{S_t(k),N_t(k)}\right)}{\sqrt{\sum\limits_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} \Lambda_q(k)\,|S_t(k)|^2}\sqrt{\sum\limits_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} \Lambda_q(k)\,|N_t(k)|^2}}. \tag{4.101}$$

The phase factor thus accounts for all terms in the spectral domain that are not directly accessible in the mel power spectral domain, i.e., the magnitudes of the SCs $S_t(k)$, $N_t(k)$ and their respective phase differences $\varphi_{S_t(k),N_t(k)}$. However, the magnitudes of the SCs in (4.101) are normalized by the square root of their average powers in the current mel band which eventually, as will soon be discussed in Sec. 4.6, allows the phase factors $\alpha_t(q)$, $t \in \{1,\dots,T\}$ to be considered as realizations of ***i**ndependent and **i**dentically **d**istributed* (i.i.d.) RVs $\breve{\alpha}_t(q)$:

Taking the natural logarithm on both sides of (4.100) results in the LMPSC $o_t^{(\text{l})}(q)$ of the noisy reverberant speech signal to be given by

$$o_t^{(\text{l})}(q) = \ln\left(\mathrm{e}^{s_t^{(\text{l})}(q)} + 2\alpha_t(q)\,\mathrm{e}^{\frac{s_t^{(\text{l})}(q)+n_t^{(\text{l})}(q)}{2}} + \mathrm{e}^{n_t^{(\text{l})}(q)}\right), \tag{4.102}$$

which, after substituting (4.88) for $s_t^{(\text{l})}(q)$, eventually provides the desired relationship of the LMPSC of the noisy reverberant speech signal $s(p)$ and those of the underlying LMPSCs $x_{t-L_H:t}^{(\text{l})}(q)$ and $n_t^{(\text{l})}(q)$ of the clean speech signal and the noise, respectively. It is given by

$$\begin{aligned} o_t^{(\text{l})}(q) = \ln\Bigg(& \mathrm{e}^{v_{s_t}^{(\text{l})}(q)} \sum_{t'=0}^{L_H} \mathrm{e}^{x_{t-t'}^{(\text{l})}(q)+\bar{h}_{t'}^{(\text{l})}(q)} \\ & + 2\alpha_t(q)\,\mathrm{e}^{\frac{v_{s_t}^{(\text{l})}(q)}{2}}\,\mathrm{e}^{\frac{\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{x_{t-t'}^{(\text{l})}(q)+\bar{h}_{t'}^{(\text{l})}(q)}\right)+n_t^{(\text{l})}(q)}{2}} + \mathrm{e}^{n_t^{(\text{l})}(q)}\Bigg) \end{aligned} \tag{4.103}$$

and may be rewritten as

$$o_t^{(\text{l})}(q) = \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{x_{t-t'}^{(\text{l})}(q)+\bar{h}_{t'}^{(\text{l})}(q)} + \mathrm{e}^{n_t^{(\text{l})}(q)}\right) + v_{o_t}^{(\text{l})}(q) \tag{4.104}$$

where

$$v_{o_t}^{(\text{l})}(q) := \ln\left(1 + \left(\mathrm{e}^{v_{s_t}^{(\text{l})}(q)} - 1\right)\xi\left(r_t^{(\text{l})}(q)\right) + 2\alpha_t(q)\,\mathrm{e}^{\frac{v_{s_t}^{(\text{l})}(q)}{2}}\zeta\left(r_t^{(\text{l})}(q)\right)\right). \tag{4.105}$$

[2] The term *phase factor* has previously been introduced in [78, 79] for an observation model of noisy speech. However, the phase factor presented here can be considered a generalization thereof with equality holding only if the clean speech signal is not corrupted by reverberation, i.e., $S_t(k) \to X_t(k)$.

with the two auxiliary functions

$$\xi\left(r_t^{(\mathrm{l})}(q)\right) := \frac{1}{1+\mathrm{e}^{-\frac{\ln(10)}{10} r_t^{(\mathrm{l})}(q)}} = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\ln(10)}{20} r_t^{(\mathrm{l})}(q)\right) \tag{4.106}$$

$$\zeta\left(r_t^{(\mathrm{l})}(q)\right) := \frac{\mathrm{e}^{-\frac{\ln(10)}{20} r_t^{(\mathrm{l})}(q)}}{1+\mathrm{e}^{-\frac{\ln(10)}{10} r_t^{(\mathrm{l})}(q)}} = \frac{1}{2}\operatorname{sech}\left(\frac{\ln(10)}{20} r_t^{(\mathrm{l})}(q)\right) \tag{4.107}$$

is the final observation error in the logarithmic mel power spectral domain depending on the IRNR , defined by

$$r_t^{(\mathrm{l})}(q) := \frac{10}{\ln(10)}\left[\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{x_{t-t'}^{(\mathrm{l})}(q)+\bar{h}_{t'}^{(\mathrm{l})}(q)}\right) - n_t^{(\mathrm{l})}(q)\right]. \tag{4.108}$$

The IRNR may be interpreted as measure of the frame and mel band specific ratio of the reverberant speech power to noise power.

The two auxiliary functions $\xi\left(r_t^{(\mathrm{l})}(q)\right)$ and $\zeta\left(r_t^{(\mathrm{l})}(q)\right)$ are sketched in Fig. 4.7, from which the following properties of the observation error in the presence of noise may already be deduced:

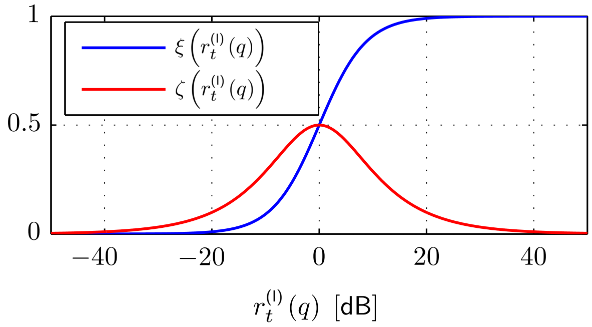


Figure 4.7: *IRNR-dependent auxiliary functions $\xi\left(r_t^{(\mathrm{l})}(q)\right)$ and $\zeta\left(r_t^{(\mathrm{l})}(q)\right)$.*

- $r_t^{(\mathrm{l})}(q) \leq -40\,\mathrm{dB}$: If the value of the IRNR $r_t^{(\mathrm{l})}(q)$ is very small, both auxiliary functions $\xi\left(r_t^{(\mathrm{l})}(q)\right)$ and $\zeta\left(r_t^{(\mathrm{l})}(q)\right)$ tend to zero. The error $v_{\mathrm{O}t}^{(\mathrm{l})}(q)$ in the presence of reverberation and noise thus approaches zero, too.

  The noise LMPSC $n_t^{(\mathrm{l})}(q)$ can directly be observed.

- $-40\,\mathrm{dB} < r_t^{(\mathrm{l})}(q) \leq 0\,\mathrm{dB}$: With increasing IRNR $r_t^{(\mathrm{l})}(q)$, both auxiliary functions $\xi\left(r_t^{(\mathrm{l})}(q)\right)$ and $\zeta\left(r_t^{(\mathrm{l})}(q)\right)$ monotonically increase. Hence, both summands contribute to the observation error $v_{\mathrm{O}t}^{(\mathrm{l})}(q)$. The contribution of the phase factor related term thereby reaches its maximum at $r_t^{(\mathrm{l})}(q) = 0\,\mathrm{dB}$.

- $0\,\mathrm{dB} < r_t^{(\mathrm{l})}(q) \leq 40\,\mathrm{dB}$: With further increasing IRNR $r_t^{(\mathrm{l})}(q)$, the monotonically increase in $\xi\left(r_t^{(\mathrm{l})}(q)\right)$ continues, while $\zeta\left(r_t^{(\mathrm{l})}(q)\right)$ now monotonically decreases. The influence of the phase factor related term on the observation error slowly diminishes.

- $40\,\mathrm{dB} < r_t^{(\mathsf{l})}(q)$: If the value of the IRNR $r_t^{(\mathsf{l})}(q)$ becomes very large, auxiliary function $\xi\left(r_t^{(\mathsf{l})}(q)\right)$ tends to one and auxiliary function $\zeta\left(r_t^{(\mathsf{l})}(q)\right)$ to zero. The error $v_{o_t}^{(\mathsf{l})}(q)$ in the presence of reverberation and noise thus approaches the error $v_{s_t}^{(\mathsf{l})}(q)$ in the presence of reverberation but absence of noise.

  The reverberant LMPSC $s_t^{(\mathsf{l})}(q)$ can directly be observed and the observation model for noisy reverberant speech coincides with the observation model for reverberant-only speech.

The related, more detailed discussion on the distribution of the observation error and its functional dependency on the IRNR (4.108) will follow in Sec. 4.7.2.

By employing the vector notation introduced for the observation model in the absence of noise, the LMPSC vector of the observed noisy reverberant speech signal may be expressed as

$$\mathbf{o}_t^{(\mathsf{l})} = \ln\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} \circ \sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})} + \bar{\mathbf{h}}_{t'}^{(\mathsf{l})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{l})}} + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}} \circ \mathrm{e}^{\frac{\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})} + \bar{\mathbf{h}}_{t'}^{(\mathsf{l})}}\right) + \mathbf{n}_t^{(\mathsf{l})}}{2}}\right) \tag{4.109}$$

$$= f_o^{(\mathsf{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathsf{l})}\right) + \mathbf{v}_{o_t}^{(\mathsf{l})}, \tag{4.110}$$

with the observation mapping defined by

$$f_o^{(\mathsf{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathsf{l})}\right) := \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})} + \bar{\mathbf{h}}_{t'}^{(\mathsf{l})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{l})}}\right). \tag{4.111}$$

Thereby $\circ$ denotes the Schur/Hadamard product, i.e., element-wise multiplication of the involved vectors.[3] The observation error vector is thus given by

$$\mathbf{v}_{o_t}^{(\mathsf{l})} := \ln\left(1 + \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} - 1\right) \circ \xi\left(\mathbf{r}_t^{(\mathsf{l})}\right) + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}} \circ \zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right) \tag{4.112}$$

with the IRNR vector

$$\mathbf{r}_t^{(\mathsf{l})} := \frac{10}{\ln(10)}\left[\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})} + \bar{\mathbf{h}}_{t'}^{(\mathsf{l})}}\right) - \mathbf{n}_t^{(\mathsf{l})}\right]. \tag{4.113}$$

Assuming the error vector $\mathbf{v}_{o_t}^{(\mathsf{l})}$ for a given IRNR vector $\mathbf{r}_t^{(\mathsf{l})}$, i.e., given the LMPSC vectors $\mathbf{x}_{t-L_H:t}^{(\mathsf{l})}$, $\mathbf{n}_t^{(\mathsf{l})}$, and the sequence of past observations $\mathbf{o}_{1:t-1}^{(\mathsf{l})}$ to be a realization of the RV $\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}$ with conditional PDF $p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ gives the stochastic observation model for noisy reverberant speech, i.e.,

$$p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}, \mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)$$

$$= p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})} - f_o^{(\mathsf{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathsf{l})}\right) \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}, \mathbf{o}_{1:t-1}^{(\mathsf{l})}\right) \tag{4.114}$$

[3] In general, whenever a Schur/Hadamard product is encountered in an equation all other operations, e.g., division or exponentiation, also have to be understood as being applied to the vectors componentwise.

Looking at (4.105), the RV $\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}$ can further be found to be a function of the RV $\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}$ and the RV $\breve{\boldsymbol{\alpha}}_t$. If the approximation (4.95) is applied, the conditional PDF turns into (see Appendix A.3)

$$\begin{aligned} & p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}} \left(\mathbf{v}_{o_t}^{(\mathsf{l})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}, \mathbf{o}_{1:t-1}^{(\mathsf{l})} \right) \\ & \approx \int\limits_{\mathbb{R}^Q} \left(\prod_{q=0}^{Q-1} \frac{\mathrm{e}^{v_{o_t}^{(\mathsf{l})}(q)}}{2\zeta\left(r_t^{(\mathsf{l})}(q)\right)} \right) p_{\breve{\boldsymbol{\alpha}}_t} \left(\frac{\mathrm{e}^{\mathbf{v}_{o_t}^{(\mathsf{l})}} - 1 - \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} - 1 \right) \circ \xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}} \circ \zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)} \right) p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}} \left(\mathbf{v}_{s_t}^{(\mathsf{l})} \right) \mathrm{d}\mathbf{v}_{s_t}^{(\mathsf{l})}. \end{aligned} \quad (4.115)$$

It is worth looking at the *limiting* cases of the IRNR, i.e., where $\mathbf{r}_t^{(\mathsf{l})} \to +\infty$ (component-wise) and $\mathbf{r}_t^{(\mathsf{l})} \to -\infty$ (component-wise). While the *factor* inside of the integral goes to infinity for both of the two scenarios, the argument of $p_{\breve{\boldsymbol{\alpha}}_t}$ goes to infinity and the value of the PDF to zero for all but two special cases.

If $\mathbf{r}_t^{(\mathsf{l})} \to +\infty$, the value $\mathbf{v}_{o_t}^{(\mathsf{l})} = \mathbf{v}_{s_t}^{(\mathsf{l})}$ renders the argument to take a value of zero, too. ("$\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ converges *faster* towards one than $\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ converges towards zero for $\mathbf{r}_t^{(\mathsf{l})} \to +\infty$"; see Fig. 4.7). Hence,

$$\left(\prod_{q=0}^{Q-1} \frac{\mathrm{e}^{v_{o_t}^{(\mathsf{l})}(q)}}{2\zeta\left(r_t^{(\mathsf{l})}(q)\right)} \right) p_{\breve{\boldsymbol{\alpha}}_t} \left(\frac{\mathrm{e}^{\mathbf{v}_{o_t}^{(\mathsf{l})}} - 1 - \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} - 1 \right) \circ \xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}} \circ \zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)} \right) \to \delta\left(\mathbf{v}_{o_t}^{(\mathsf{l})} - \mathbf{v}_{s_t}^{(\mathsf{l})} \right) \quad (4.116)$$

as $\mathbf{r}_t^{(\mathsf{l})} \to +\infty$ and thus

$$p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}} \left(\mathbf{v}_{o_t}^{(\mathsf{l})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}, \mathbf{o}_{1:t-1}^{(\mathsf{l})} \right) \to p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}} \left(\mathbf{v}_{o_t}^{(\mathsf{l})} \right) \text{ as } \mathbf{r}_t^{(\mathsf{l})} \to +\infty. \quad (4.117)$$

The observation error in the noisy reverberant scenario thus has the distribution of the observation error in the reverberant scenario as its first *limiting* distribution.

If $\mathbf{r}_t^{(\mathsf{l})} \to -\infty$, the value $\mathbf{v}_{o_t}^{(\mathsf{l})} = 0$ renders the argument to take a value of zero, too. ("$\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ converges *faster* towards zero than does $\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ for $\mathbf{r}_t^{(\mathsf{l})} \to -\infty$"; see Fig. 4.7). Hence,

$$\left(\prod_{q=0}^{Q-1} \frac{\mathrm{e}^{v_{o_t}^{(\mathsf{l})}(q)}}{2\zeta\left(r_t^{(\mathsf{l})}(q)\right)} \right) p_{\breve{\boldsymbol{\alpha}}_t} \left(\frac{\mathrm{e}^{\mathbf{v}_{o_t}^{(\mathsf{l})}} - 1 - \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} - 1 \right) \circ \xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}} \circ \zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)} \right) \to \delta\left(\mathbf{v}_{o_t}^{(\mathsf{l})} \right) \quad (4.118)$$

as $\mathbf{r}_t^{(\mathsf{l})} \to -\infty$ and so does $p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$, i.e.,

$$p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})} | \breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})}, \breve{\mathbf{n}}_t^{(\mathsf{l})}, \breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}} \left(\mathbf{v}_{o_t}^{(\mathsf{l})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}, \mathbf{o}_{1:t-1}^{(\mathsf{l})} \right) \to \delta\left(\mathbf{v}_{o_t}^{(\mathsf{l})} \right) \text{ as } \mathbf{r}_t^{(\mathsf{l})} \to -\infty. \quad (4.119)$$

The observation error in the noisy reverberant scenario thus has the DIRAC-delta centered at zero as its second *limiting* distribution.

As already indicated by the argument of the PDF of the phase factor, (4.115) may not be solved analytically for all but the most trivial forms of $p_{\breve{\boldsymbol{\alpha}}_t}$. However, it represents a very convenient and attractive way to compute the conditional PDF of the observation error by means of numerical integration. With the PDF $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}}$ given by (4.95) – only the analytic form of $p_{\breve{\boldsymbol{\alpha}}_t}$ has to be known. A tractable analytic form thereof will be presented in Sec. 4.6.3.

The conditional PDF $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{n}}^{(\mathsf{l})}_t,\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ of the observation error in the presence of noise is thus completely characterized by the PDF $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}}$ of the observation error in the absence of noise and the PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ of the vector of phase factors.

In the following, $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{n}}^{(\mathsf{l})}_t,\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ will be approximated by a Gaussian distribution as

$$p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{n}}^{(\mathsf{l})}_t,\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{v}^{(\mathsf{l})}_{o_t}\left|\mathbf{x}^{(\mathsf{l})}_{t-L_H:t},\mathbf{n}^{(\mathsf{l})}_t,\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right.\right)=p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}|\breve{\mathbf{r}}^{(\mathsf{l})}_t}\left(\mathbf{v}^{(\mathsf{l})}_{o_t}\left|\mathbf{r}^{(\mathsf{l})}_t\right.\right) \tag{4.120}$$

$$\approx\mathcal{N}\left(\mathbf{v}^{(\mathsf{l})}_{o_t};\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right),\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right), \tag{4.121}$$

where the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)$ are functions of the IRNR $\mathbf{r}^{(\mathsf{l})}_t$ and thus varying with time.

Since the Gaussian approximation of the PDF $p_{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{n}}^{(\mathsf{l})}_t,\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ given in (4.120) is equivalent to approximating the PDF $p_{\breve{\mathbf{v}}^{(\mathsf{m})}_{o_t}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t},\breve{\mathbf{n}}^{(\mathsf{l})}_t,\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ of the RV $\breve{\mathbf{v}}^{(\mathsf{m})}_{o_t}:=\mathrm{e}^{\breve{\mathbf{v}}^{(\mathsf{l})}_{o_t}}$ by a Log-Normal distribution, these quantities are given by[4]

$$\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)=\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right)-\frac{1}{2}\mathrm{diag}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right), \tag{4.122}$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)=\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right)^{\dagger}+\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right)-\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)\right)^{\dagger}\right). \tag{4.123}$$

Note that the logarithm has to be applied to the occurring vectors and matrices componentwise. Thereby $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{m})}_o}\left(\mathbf{r}^{(\mathsf{l})}_t\right)$ denote the conditional mean vector and the conditional covariance matrix of the (consequently) log-normally distributed RV $\breve{\mathbf{v}}^{(\mathsf{m})}_{o_t}$.

The components of these quantities are related to the components of the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}$ and the components of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l})}_s}$ of the observation error $\breve{\mathbf{v}}^{(\mathsf{l})}_{s_t}$ in the noise-free case and the components of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ by (see Appendix A.5 for more details)

$$\mu_{\breve{v}^{(\mathsf{m})}_o(q)}\left(\mathbf{r}^{(\mathsf{l})}_t\right):=1+\left(E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l})}_{s_t}(q)}\right]-1\right)\xi\left(r^{(\mathsf{l})}_t(q)\right), \tag{4.124}$$

$$\begin{aligned}\sigma_{\breve{v}^{(\mathsf{m})}_o(q),\breve{v}^{(\mathsf{m})}_o(q')}\left(\mathbf{r}^{(\mathsf{l})}_t\right):=&\left(E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l})}_{s_t}(q)+\breve{v}^{(\mathsf{l})}_s(q')}\right]-E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l})}_s(q)}\right]E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l})}_{s_t}(q')}\right]\right)\xi\left(r^{(\mathsf{l})}_t(q)\right)\xi\left(r^{(\mathsf{l})}_t\left(q'\right)\right)\\&+4\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}^{(\mathsf{l})}_{s_t}(q)+\breve{v}^{(\mathsf{l})}_{s_t}(q')\right)}\right]\zeta\left(r^{(\mathsf{l})}_t(q)\right)\zeta\left(r^{(\mathsf{l})}_t\left(q'\right)\right),\end{aligned} \tag{4.125}$$

[4] Equations (4.122) and (4.123) build the multivariate extension to the relation of the mean and the variance of a univariate, log-normally distributed variable to those of the underlying univariate, normally distributed variable given in [80]. A detailed derivation thereof is given in Appendix A.11.

with

$$E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)}\right] = \mathrm{e}^{\mu_{\breve{v}_s^{(\mathsf{l})}(q)} + \frac{1}{2}\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)}}, \tag{4.126}$$

$$E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q) + \breve{v}_s^{(\mathsf{l})}(q')}\right] = \mathrm{e}^{\mu_{\breve{v}_s^{(\mathsf{l})}(q)} + \mu_{\breve{v}_s^{(\mathsf{l})}(q')} + \frac{1}{2}\left(\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)} + \sigma^2_{\breve{v}_s^{(\mathsf{l})}(q')} + 2\sigma_{\breve{v}_s^{(\mathsf{l})}(q),\breve{v}_s^{(\mathsf{l})}(q')}\right)}, \tag{4.127}$$

$$E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}_{s_t}^{(\mathsf{l})}(q) + \breve{v}_{s_t}^{(\mathsf{l})}(q')\right)}\right] = \mathrm{e}^{\frac{1}{2}\left(\mu_{\breve{v}_s^{(\mathsf{l})}(q)} + \mu_{\breve{v}_s^{(\mathsf{l})}(q')}\right) + \frac{1}{8}\left(\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)} + \sigma^2_{\breve{v}_s^{(\mathsf{l})}(q')} + 2\sigma_{\breve{v}_s^{(\mathsf{l})}(q),\breve{v}_s^{(\mathsf{l})}(q')}\right)}. \tag{4.128}$$

As can be seen, the vector of phase factors contributes (only) in terms of its first two central moment (see Sec. 4.6.4). Note that $\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ computed by (4.122) and $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ computed by (4.122) converge to $\boldsymbol{\mu}_{\breve{\mathbf{v}}_s^{(\mathsf{l})}}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_s^{(\mathsf{l})}}$, i.e., the mean vector and the covariance matrix of the observation error in the absence of noise, respectively, as $\mathbf{r}_t^{(\mathsf{l})} \to +\infty$ (component-wise). Further, as $\mathbf{r}_t^{(\mathsf{l})} \to -\infty$ (component-wise), i.e., in the absence of reverberant speech, both moments converge to zero.

With (4.114), the final (approximate) observation model is thus given by

$$\begin{aligned}
& p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right) \\
& \approx \left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{o_t^{(\mathsf{l})}(q) - f_o^{(\mathsf{l})}\left(x_{t-L_H:t}^{(\mathsf{l})}(q), n_t^{(\mathsf{l})}(q); \bar{h}_{0:L_H}^{(\mathsf{l})}(q)\right)}}{2\zeta\left(r_t^{(\mathsf{l})}(q)\right)}\right) \\
& \quad \int\limits_{\mathbb{R}^Q} p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{o}_t^{(\mathsf{l})} - f_o^{(\mathsf{l})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{l})}, \mathbf{n}_t^{(\mathsf{l})}; \bar{\mathbf{h}}_{0:L_H}^{(\mathsf{l})}\right)} - 1 - \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}} - 1\right)\circ\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}}\circ\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)}\right) p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\mathbf{v}_{s_t}^{(\mathsf{l})}\right)\mathrm{d}\mathbf{v}_{s_t}^{(\mathsf{l})}
\end{aligned} \tag{4.129}$$

$$\approx \mathcal{N}\left(\mathbf{o}_t^{(\mathsf{l})}; \ln\left(\sum_{t'=0}^{L_H}\mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})} + \bar{\mathbf{h}}_{t'}^{(\mathsf{l})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{l})}}\right) + \boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right), \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right). \tag{4.130}$$

An analysis of this Gaussian approximation will be given in Sec. 4.7.2.

4.3.4 Absence of Reverberation and Presence of Background Noise

The observation model in the absence of reverberation and the presence of noise can be deduced from the observation model in the presence of both reverberation and noise by choosing

$$h(p) = \delta_p, \tag{4.131}$$

where $\delta_{(\cdot)}$ denotes previously introduced the Kronecker-delta. Under this AIR, the SCs $S_t(k)$ of the reverberant speech signal are equal to the SCs $X_t(k)$ of the clean speech signal. Substituting $S_t(k)$ for $X_t(k)$ in the definition of the phase factor (4.101) and all consecutive feature extraction steps eventually results in the LMPSC vector of the noisy speech signal, which will be denoted by $\mathbf{y}_t^{(\mathsf{I})}$ to ease distinction between the different scenarios, to be given by [78]

$$\mathbf{y}_t^{(\mathsf{I})} = \ln\left(\mathrm{e}^{\mathbf{x}_t^{(\mathsf{I})}} + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{x}_t^{(\mathsf{I})}+\mathbf{n}_t^{(\mathsf{I})}}{2}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{I})}}\right) \tag{4.132}$$

$$= f_y^{(\mathsf{I})}\left(\mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}\right) + \mathbf{v}_{y_t}^{(\mathsf{I})}, \tag{4.133}$$

with the observation mapping

$$f_y^{(\mathsf{I})}\left(\mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}\right) := \ln\left(\mathrm{e}^{\mathbf{x}_t^{(\mathsf{I})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{I})}}\right) \tag{4.134}$$

and the associated observation error

$$\mathbf{v}_{y_t}^{(\mathsf{I})} := \ln\left(1 + 2\boldsymbol{\alpha}_t \circ \frac{\mathrm{e}^{\frac{\mathbf{x}_t^{(\mathsf{I})}-\mathbf{n}_t^{(\mathsf{I})}}{2}}}{1+\mathrm{e}^{\mathbf{x}_t^{(\mathsf{I})}-\mathbf{n}_t^{(\mathsf{I})}}}\right) \tag{4.135}$$

$$= \ln\left(1 + 2\boldsymbol{\alpha}_t \circ \zeta\left(\mathbf{r}_t^{(\mathsf{I})}\right)\right). \tag{4.136}$$

The auxiliary function $\zeta(\cdot)$ introduced in (4.107) thereby has to be evaluated at the IRNR vector in the absence of reverberation, which reduces to the definition of the ISNR – a frame and mel band specific ratio of the instantaneous speech signal power to noise power. It is given by

$$\mathbf{r}_t^{(\mathsf{I})} = \frac{10}{\ln(10)}\left(\mathbf{x}_t^{(\mathsf{I})} - \mathbf{n}_t^{(\mathsf{I})}\right). \tag{4.137}$$

Looking at Fig. 4.7, it immediately gets apparent that the contribution of the phase factor component $\alpha_t(q)$ first monotonically increases with increasing ISNR $r_t^{(\mathsf{I})}(q)$, reaches its maximum at $r_t^{(\mathsf{I})}(q) = 0\,\mathrm{dB}$ and eventually monotonically decreases towards zero. The stochastic observation model for noisy speech thus turns into

$$\begin{aligned} & p_{\breve{\mathbf{y}}_t^{(\mathsf{I})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{I})},\breve{\mathbf{n}}_t^{(\mathsf{I})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{I})}}\left(\mathbf{y}_t^{(\mathsf{I})} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}, \mathbf{y}_{1:t-1}^{(\mathsf{I})}\right) \\ = \; & p_{\breve{\mathbf{v}}_{y_t}^{(\mathsf{I})}|\breve{\mathbf{x}}_t^{(\mathsf{I})},\breve{\mathbf{n}}_t^{(\mathsf{I})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{I})}}\left(\mathbf{y}_t^{(\mathsf{I})} - f_y^{(\mathsf{I})}\left(\mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}\right) \middle| \mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}, \mathbf{y}_{1:t-1}^{(\mathsf{I})}\right) \end{aligned} \tag{4.138}$$

and can be found to be completely characterized by the conditional PDF $p_{\breve{\mathbf{v}}_{y_t}^{(\mathsf{I})}|\breve{\mathbf{x}}_t^{(\mathsf{I})},\breve{\mathbf{n}}_t^{(\mathsf{I})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{I})}}$ of the RV $\breve{\mathbf{v}}_{y_t}^{(\mathsf{I})}$, which may be expressed as (see Appendix A.4)

$$\begin{aligned} & p_{\breve{\mathbf{v}}_{y_t}^{(\mathsf{I})}|\breve{\mathbf{x}}_t^{(\mathsf{I})},\breve{\mathbf{n}}_t^{(\mathsf{I})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{I})}}\left(\mathbf{v}_{y_t}^{(\mathsf{I})} \middle| \mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}, \mathbf{y}_{1:t-1}^{(\mathsf{I})}\right) \\ \approx & \left(\prod_{q=0}^{Q-1} \frac{\mathrm{e}^{v_{y_t}^{(\mathsf{I})}(q)}}{2\zeta\left(r_t^{(\mathsf{I})}(q)\right)}\right) p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{v}_{y_t}^{(\mathsf{I})}} - 1}{2\zeta\left(\mathbf{r}_t^{(\mathsf{I})}\right)}\right) \end{aligned} \tag{4.139}$$

and as such is completely characterized by the PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ of the RV $\breve{\boldsymbol{\alpha}}_t$ of the vector of phase factors. Note that the distribution of the observation error is only non-zero if $\ln\left(1 - 2\zeta\left(r_t^{(\mathsf{I})}(q)\right)\right) \leq v_{y_t}^{(\mathsf{I})}(q) \leq \ln\left(1 + 2\zeta\left(r_t^{(\mathsf{I})}(q)\right)\right) \leq \ln(2)$ holds.

Combining (4.138) and (4.139) while employing (4.137) in the definition (4.107) eventually returns the desired stochastic observation model in the absence of reverberation and the presence of noise to be

$$\begin{aligned}
&p_{\breve{\mathbf{y}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{y}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{y}_{1:t-1}^{(\mathsf{l})}\right)\\
&\approx\left(\frac{1}{2^Q}\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{y_t^{(\mathsf{l})}(q)}}{\mathrm{e}^{\frac{x_t^{(\mathsf{l})}(q)+n_t^{(\mathsf{l})}(q)}{2}}}\right)p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{y}_t^{(\mathsf{l})}}-\mathrm{e}^{\mathbf{x}_t^{(\mathsf{l})}}-\mathrm{e}^{\mathbf{n}_t^{(\mathsf{l})}}}{2\mathrm{e}^{\frac{\mathbf{x}_t^{(\mathsf{l})}+\mathbf{n}_t^{(\mathsf{l})}}{2}}}\right) &(4.140)\\
&=\frac{1}{2^Q}\mathrm{e}^{-\frac{1}{2}\left(\mathbf{x}_t^{(\mathsf{l})}+\mathbf{n}_t^{(\mathsf{l})}-2\mathbf{y}_t^{(\mathsf{l})}\right)^\dagger\mathbf{1}_{Q\times1}}p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{y}_t^{(\mathsf{l})}}-\mathrm{e}^{\mathbf{x}_t^{(\mathsf{l})}}-\mathrm{e}^{\mathbf{n}_t^{(\mathsf{l})}}}{2\mathrm{e}^{\frac{\mathbf{x}_t^{(\mathsf{l})}+\mathbf{n}_t^{(\mathsf{l})}}{2}}}\right), &(4.141)
\end{aligned}$$

where $\mathbf{1}_{Q\times1}$ is a column vector of all ones with Q elements.

If (4.141) is to be approximated by a GAUSSIAN distribution with ISNR-dependent moments as

$$\begin{aligned}
&p_{\breve{\mathbf{y}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{y}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{y}_{1:t-1}^{(\mathsf{l})}\right)\\
&\approx\mathcal{N}\left(\mathbf{y}_t^{(\mathsf{l})};f_y^{(\mathsf{l})}\left(\mathbf{x}_t^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})}\right)+\boldsymbol{\mu}_{\breve{\mathbf{y}}^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right),\boldsymbol{\Sigma}_{\breve{\mathbf{y}}^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right), &(4.142)
\end{aligned}$$

which is equivalent to approximating the conditional PDF $p_{\breve{\mathbf{v}}_{y_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_t^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{l})}}$ of the observation error $\breve{\mathbf{v}}_{y_t}^{(\mathsf{l})}$ by a GAUSSIAN as

$$p_{\breve{\mathbf{v}}_{y_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_t^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{y}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{v}_{y_t}^{(\mathsf{l})}\middle|\mathbf{x}_t^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{y}_{1:t-1}^{(\mathsf{l})}\right)\approx\mathcal{N}\left(\mathbf{y}_t^{(\mathsf{l})};\boldsymbol{\mu}_{\breve{\mathbf{y}}^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right),\boldsymbol{\Sigma}_{\breve{\mathbf{y}}^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right), \qquad (4.143)$$

then the elements of the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{y}}^{(\mathsf{l})}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{y}}^{(\mathsf{l})}}$ may be estimated by again assuming the observation error in the MPSC domain $\breve{\mathbf{v}}_{y_t}^{(\mathsf{m})}:=\mathrm{e}^{\breve{\mathbf{v}}_{y_t}^{(\mathsf{l})}}$ to be log-normally distributed. Equivalent to (4.122) and (4.123), these moments are given by

$$\begin{aligned}
\boldsymbol{\mu}_{\breve{\mathbf{v}}_y^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)&=-\frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_y^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right), &(4.144)\\
\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_y^{(\mathsf{l})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)&=\ln\left(\mathbf{1}+\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_y^{(\mathsf{m})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right), &(4.145)
\end{aligned}$$

where the components of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_y^{(\mathsf{m})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)$ are given by

$$\sigma_{\breve{v}_y^{(\mathsf{m})}(q),\breve{v}_y^{(\mathsf{m})}(q')}\left(\mathbf{r}_t^{(\mathsf{l})}\right):=4\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\zeta\left(r_t^{(\mathsf{l})}(q)\right)\zeta\left(r_t^{(\mathsf{l})}\left(q'\right)\right), \qquad (4.146)$$

The simplified form of (4.144) and (4.145) over (4.122) and (4.123) comes with the fact that under the Log-Normal assumption, $\mu_{\breve{\mathbf{v}}_y^{(\mathsf{m})}}\left(\mathbf{r}_t^{(\mathsf{l})}\right)=\mathbf{1}_{Q\times1}$.

For $\mathbf{r}_t^{(\mathsf{l})}\rightarrow\pm\infty$, both mean vector and covariance matrix converge to zero.

4.3.5 Overview of Observation Models

Though the derivations of the different observation models start with the observation model in the presence of reverberation and the absence of background noise, the observation model in the presence of both reverberation and background noise presented in Sec. 4.3.3 can be considered as a *generalized observation model* from which the observation model in the presence of reverberation and the absence of noise presented in Sec. 4.3.2 and the observation in the absence of reverberation and the presence of background noise presented in Sec. 4.3.4 can be deduced – as illustrated in Fig. 4.8.

Figure 4.8 also highlights that the stochastic observation models are completely characterized by the PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ and the PDF $p_{\breve{\mathbf{v}}_{st}^{(\mathrm{l})}}$ of the observation error in the presence of reverberation and the absence of noise $\breve{\mathbf{v}}_{st}^{(\mathrm{l})}$. Both PDFs will be looked at in full detail in the following sections.

4.4 AIR Model

For the evaluation of the observation models (4.96) and (4.130) in the presence of reverberation the logarithmic mel power spectral representation of the AIR $\bar{\mathbf{h}}_{0:L_H}^{(\mathrm{l})}$ is required. Since the underlying AIR, from which $\bar{\mathbf{h}}_{t'}^{(\mathrm{l})}$ may be computed via (4.85) by employing (4.54) with (4.53), is usually unknown in practice, a stochastic model for the AIR, as proposed in [28], will be employed in this work. Thus, a sensitive blind estimation of the AIR may be avoided.

The stochastic model of the AIR employed here has been introduced in [31] and characterizes the AIR as a realization of a real-valued, stationary stochastic process according to

$$\breve{h}(p) = \sigma_{\breve{h}} \chi_{L_h}(p)\, \breve{\nu}_h(p)\, \mathrm{e}^{-\frac{p}{\tau_h}}. \tag{4.147}$$

Thereby, $\breve{\nu}_h(p)$ denotes a real-valued, zero-mean white GAUSSIAN stochastic process with auto-correlation function

$$E\left[\breve{\nu}_h(p)\, \breve{\nu}_h\left(p'\right)\right] = \delta\left(p - p'\right) \tag{4.148}$$

and $\chi_{L_h}(p)$ the binary indicator function defined by

$$\chi_{L_h}(p) = \begin{cases} 1, & 0 \le p \le L_h - 1 \\ 0, & \text{else.} \end{cases} \tag{4.149}$$

The binary indicator function ensures the AIR to be causal and of finite length L_h. Figure 4.9 shows a measured AIR of a room with $T_{60} \approx 300\,\mathrm{ms}$ (green line) and a possible sample realization of the AIR according to the model (4.147) (blue line).[5]. The deficiencies to properly model the direct sound and the early reflections become apparent. Strictly speaking, the model (4.147) is thus only approximately valid for the late reverberation. A detailed discussion on the deficiencies of the AIR model may, e.g., be found in [28]. Due to the fact that the model is in essence characterized by only two parameters, namely the decay constant τ_h and the AIR energy $\sigma_{\breve{h}}^2$, (4.147) is preferred here over more accurate (but also more complicated) models. The auto-correlation function of the stochastic process

[5] Note that the measured AIR has been shifted such that the direct sound occurs at $p = 0$. Further, only $0 \le p \le 2L_h$ is displayed, although the measured AIR extends much more along the x-axis.

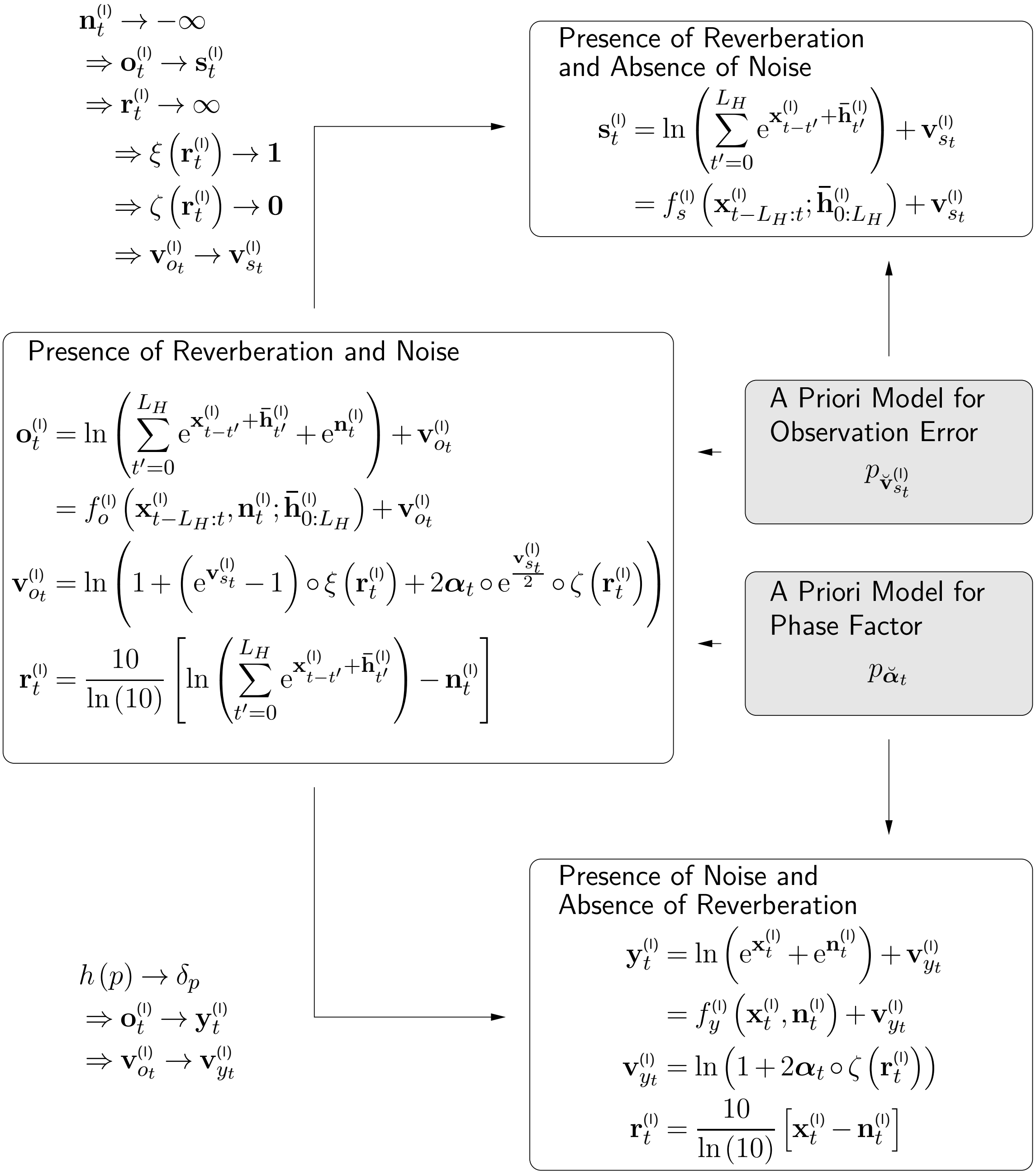

$\mathbf{n}_t^{(l)} \rightarrow -\infty$
$\Rightarrow \mathbf{o}_t^{(l)} \rightarrow \mathbf{s}_t^{(l)}$
$\Rightarrow \mathbf{r}_t^{(l)} \rightarrow \infty$
$\Rightarrow \xi\left(\mathbf{r}_t^{(l)}\right) \rightarrow \mathbf{1}$
$\Rightarrow \zeta\left(\mathbf{r}_t^{(l)}\right) \rightarrow \mathbf{0}$
$\Rightarrow \mathbf{v}_{o_t}^{(l)} \rightarrow \mathbf{v}_{s_t}^{(l)}$
Presence of Reverberation and Absence of Noise
$\mathbf{s}_t^{(l)} = \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(l)}+\bar{\mathbf{h}}_{t'}^{(l)}}\right) + \mathbf{v}_{s_t}^{(l)}$
$= f_s^{(l)}\left(\mathbf{x}_{t-L_H:t}^{(l)};\bar{\mathbf{h}}_{0:L_H}^{(l)}\right) + \mathbf{v}_{s_t}^{(l)}$
Presence of Reverberation and Noise
$\mathbf{o}_t^{(l)} = \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(l)}+\bar{\mathbf{h}}_{t'}^{(l)}} + \mathrm{e}^{\mathbf{n}_t^{(l)}}\right) + \mathbf{v}_{o_t}^{(l)}$
$= f_o^{(l)}\left(\mathbf{x}_{t-L_H:t}^{(l)},\mathbf{n}_t^{(l)};\bar{\mathbf{h}}_{0:L_H}^{(l)}\right) + \mathbf{v}_{o_t}^{(l)}$
$\mathbf{v}_{o_t}^{(l)} = \ln\left(1 + \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(l)}} - 1\right) \circ \xi\left(\mathbf{r}_t^{(l)}\right) + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(l)}}{2}} \circ \zeta\left(\mathbf{r}_t^{(l)}\right)\right)$
$\mathbf{r}_t^{(l)} = \frac{10}{\ln(10)}\left[\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(l)}+\bar{\mathbf{h}}_{t'}^{(l)}}\right) - \mathbf{n}_t^{(l)}\right]$
A Priori Model for Observation Error
$p_{\breve{\mathbf{v}}_{s_t}^{(l)}}$
A Priori Model for Phase Factor
$p_{\breve{\boldsymbol{\alpha}}_t}$
$h(p) \rightarrow \delta_p$
$\Rightarrow \mathbf{o}_t^{(l)} \rightarrow \mathbf{y}_t^{(l)}$
$\Rightarrow \mathbf{v}_{o_t}^{(l)} \rightarrow \mathbf{v}_{y_t}^{(l)}$
Presence of Noise and Absence of Reverberation
$\mathbf{y}_t^{(l)} = \ln\left(\mathrm{e}^{\mathbf{x}_t^{(l)}} + \mathrm{e}^{\mathbf{n}_t^{(l)}}\right) + \mathbf{v}_{y_t}^{(l)}$
$= f_y^{(l)}\left(\mathbf{x}_t^{(l)},\mathbf{n}_t^{(l)}\right) + \mathbf{v}_{y_t}^{(l)}$
$\mathbf{v}_{y_t}^{(l)} = \ln\left(1 + 2\boldsymbol{\alpha}_t \circ \zeta\left(\mathbf{r}_t^{(l)}\right)\right)$
$\mathbf{r}_t^{(l)} = \frac{10}{\ln(10)}\left[\mathbf{x}_t^{(l)} - \mathbf{n}_t^{(l)}\right]$

Figure 4.8: *Overview of observation models: the observation model for the presence of both reverberation and noise as a generalization of the observation models for the presence of either reverberation or noise.*

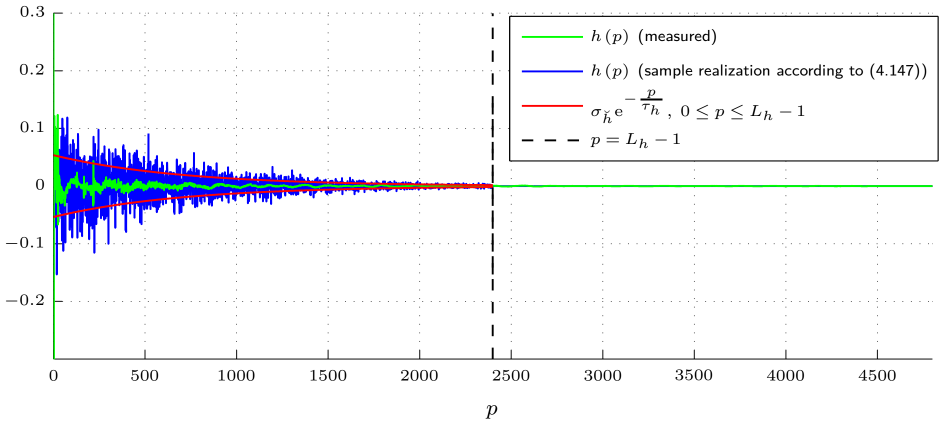


Figure 4.9: *Measured AIR of a room with $T_{60} \approx 300\,\mathrm{ms}$ (green line, $T_S^{-1} = 16\,\mathrm{kHz}$), a sample realization of the AIR according to the model* (4.147) *(blue line) and the exponentially decaying envelop (red line). The parameters of the model have been determined to be $\tau_h \approx 695$, $L_h = 2400$ and $\sigma_{\breve{h}} = 5.36 \cdot 10^{-2}$ according to* (4.159), (4.157) *and* (4.154).

$\breve{h}(p)$ is given by

$$E\left[\breve{h}(p)\,\breve{h}\left(p'\right)\right] = \sigma_{\breve{h}}^2 \chi_{L_h}(p)\,\chi_{L_h}\left(p'\right) E\left[\breve{\nu}_h(p)\,\breve{\nu}_h\left(p'\right)\right] e^{-\frac{p+p'}{\tau_h}} \tag{4.150}$$

$$= \sigma_{\breve{h}}^2 \chi_{L_h}(p)\, e^{-\frac{2p}{\tau_h}} \delta\left(p - p'\right). \tag{4.151}$$

Higher order moments of $\breve{h}(p)$ may be expressed by means of ISSERLIS theorem [81]. In particular,

$$E\left[\breve{h}(p)\,\breve{h}\left(p'\right)\breve{h}\left(p''\right)\breve{h}\left(p''\right)\right]$$

$$= \sigma_{\breve{h}}^4 \chi_{L_h}(p)\,\chi_{L_h}\left(p'\right)\chi_{L_h}\left(p''\right)\chi_{L_h}\left(p'''\right) E\left[\breve{\nu}_h(p)\,\breve{\nu}_h\left(p'\right)\breve{\nu}_h\left(p''\right)\breve{\nu}_h\left(p'''\right)\right] e^{-\frac{p+p'+p''+p'''}{\tau_h}} \tag{4.152}$$

$$= \sigma_{\breve{h}}^4 \chi_{L_h}(p)\,\chi_{L_h}\left(p'\right)\chi_{L_h}\left(p''\right)\chi_{L_h}\left(p'''\right) e^{-\frac{p+p'+p''+p'''}{\tau_h}} \left[\delta\left(p-p'\right)\delta\left(p''-p'''\right) + \delta\left(p-p''\right)\delta\left(p'-p'''\right) + \delta\left(p-p'''\right)\delta\left(p'-p''\right)\right] \tag{4.153}$$

will be utilized later on.

4.4.1 Model Parameters

The required two parameters $\sigma_{\breve{h}}^2$ and τ_h denote the energy of the AIR and the decay constant, respectively. The decay constant is related to the reverberation time T_{60} and the sampling period T_S through [67, p. 142, A.2.3.]

$$\tau_h := \frac{T_{60}}{3 \ln(10)\, T_S}. \tag{4.154}$$

The energy parameter $\sigma_{\breve{h}}$ may, e.g., be estimated by assuming the speech signal $x(p)$ and the noise signal $n(p)$ to be realizations of uncorrelated stationary stochastic processes $\breve{x}(p)$ and $\breve{n}(p)$. Denoting their powers by $\sigma_{\breve{x}}^2 := E\left[\breve{x}^2(p)\right]$ and $\sigma_{\breve{n}}^2 := E\left[\breve{n}^2(p)\right]$ and the power of the noisy reverberant speech signal by $\sigma_{\breve{o}}^2 := E\left[\breve{o}^2(p)\right]$, $\sigma_{\breve{h}}$ may be found by requiring

$$E\left[\breve{o}^2(p)\right] = E\left[(\breve{s}(p) + \breve{n}(p))^2\right] \tag{4.155}$$

$$\overset{!}{=} E\left[\breve{s}^2(p)\right] + E\left[\breve{n}^2(p)\right]. \tag{4.156}$$

Employing the signal model (3.106) relating the reverberant speech signal to the clean speech signal and the AIR and the stochastic model of the AIR given in (4.147) eventually yields

$$\sigma_{\breve{h}} := \sqrt{\frac{\sigma_{\breve{o}}^2 - \sigma_{\breve{n}}^2}{\sigma_{\breve{x}}^2}} \cdot \tilde{\sigma}_{\breve{h}} \tag{4.157}$$

where

$$\tilde{\sigma}_{\breve{h}} := \sqrt{\frac{1 - \mathrm{e}^{-\frac{2}{\tau_h}}}{1 - \mathrm{e}^{-\frac{2L_h}{\tau_h}}}}. \tag{4.158}$$

Note that $\sigma_{\breve{o}}^2 - \sigma_{\breve{n}}^2$ may be considered an estimate of the power of the reverberant speech signal $\breve{s}(p)$. Thus, in the noise free case, $\sigma_{\breve{o}}^2 - \sigma_{\breve{n}}^2$ has to be replaced by $\sigma_{\breve{s}}^2$. In artificially created databases usually $\sigma_{\breve{s}}^2 = \sigma_{\breve{x}}^2$ and thus $\sigma_{\breve{h}} = \tilde{\sigma}_{\breve{h}}$.

With a reasonable length L_h given by [28]

$$L_h := \left\lceil -\frac{\tau_h}{2} \ln(\epsilon_h) \right\rceil, \tag{4.159}$$

which arises as the outcome of minimizing the AIR length under the constraint that the relative energy of the remaining tail of the AIR is smaller than some ϵ_h (in this work, $\epsilon_h = 10^{-3}$), all parameters of the stochastic model of the AIR may be obtained from an estimate of the reverberation time T_{60} and estimates of the power of the involved random processes, i.e., $\sigma_{\breve{o}}^2$, $\sigma_{\breve{n}}^2$ and $\sigma_{\breve{x}}^2$. In practice, these estimates may, e.g., be obtained from available clean speech training data and noisy reverberant test data.

The AIR parameters for different reverberation times for $T_S^{-1} = 8\,\text{kHz}$ and $\epsilon_h = 10^{-3}$ are displayed in Tab. 4.1. Further, the number $L_H + 1$ of speech feature vectors considered in the evaluation of (4.96) and (4.130), which is related to L_h via (repeated here from (4.60) for convenience)

$$L_H = \left\lfloor \frac{L_w + L_h - 2}{B} \right\rfloor, \tag{4.160}$$

may be inferred from the last column.

Table 4.1: *Parameters of the AIR model for different reverberation times and* $T_S^{-1} = 8\,\mathrm{kHz}$ *(*$\epsilon_h = 10^{-3}$*). For the computation of* L_H *the front-end parameters* $L_w = 200$ *and* $B = 80$ *are taken from Tab. 3.1.*

Reverberation	Decay	AIR Energy	AIR length	
time T_{60} [ms]	Constant τ_h	parameter $\tilde{\sigma}_{\breve{h}}$	L_h	L_H
200	232	$9.28 \cdot 10^{-2}$	800	12
300	347	$7.58 \cdot 10^{-2}$	1200	17
400	463	$6.57 \cdot 10^{-2}$	1600	22
500	579	$5.87 \cdot 10^{-2}$	2000	27
600	695	$5.36 \cdot 10^{-2}$	2400	32
700	811	$4.97 \cdot 10^{-2}$	2800	37

4.4.2 Representation of the AIR in the Logarithmic Mel Domain

With the stochastic model of the AIR given in (4.147), the approximate representation of the AIR in the logarithmic mel domain, denoted by $\bar{\mathbf{h}}_{t'}^{(\mathrm{l})}$ (see (4.89)), may now be replaced by its expected value, i.e.,

$$\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathrm{l})}} := E\left[\breve{\bar{\mathbf{h}}}_{t'}^{(\mathrm{l})}\right]. \tag{4.161}$$

Since $\bar{\mathbf{h}}_{t'}^{(\mathrm{l})}$ may be considered to be approximately GAUSSIAN distributed [67], (4.161) may be expressed in terms of the mean vector $\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathcal{H}}_{t'}}$ of the, consequently approximately log-normally distributed, RV $\breve{\mathcal{H}}_{t'}$ by

$$\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathrm{l})}} = \ln\left(\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}\right) - \frac{1}{2}\mathrm{diag}\left(\boldsymbol{\Sigma}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathrm{l})}}\right), \tag{4.162}$$

where

$$\boldsymbol{\Sigma}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathrm{l})}} = \ln\left(\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}\left(\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}\right)^{\dagger} + \boldsymbol{\Sigma}_{\breve{\mathcal{H}}_{t'}}\right) - \ln\left(\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}\left(\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}\right)^{\dagger}\right). \tag{4.163}$$

Note that the logarithm has to be applied to the occurring vectors and matrices componentwise. Equations (4.162) and (4.163) build the multivariate extension to the relation of the mean and the variance of a univariate, log-normally distributed variable to those of the underlying univariate, normally distributed variable given in [80]. A detailed derivation thereof is given in Appendix A.11.

The required elements of the mean vector $\boldsymbol{\mu}_{\breve{\mathcal{H}}_{t'}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathcal{H}}_{t'}}$ are, employing the stochastic model of the AIR (4.147), given by

$$\mu_{\breve{\mathcal{H}}_{t'}(q)} := E\left[\breve{\mathcal{H}}_{t'}(q)\right] \tag{4.164}$$

$$= C_P \sigma_{\breve{h}}^2 \mathcal{V}_{t'}(0) \tag{4.165}$$

and

$$\sigma_{\breve{\mathcal{H}}_{t'}(q),\breve{\mathcal{H}}_{t'}(q')} := E\left[\left(\breve{\mathcal{H}}_{t'}(q) - E\left[\breve{\mathcal{H}}_{t'}(q)\right]\right)\left(\breve{\mathcal{H}}_{t'}\left(q'\right) - E\left[\breve{\mathcal{H}}_{t'}\left(q'\right)\right]\right)\right] \tag{4.166}$$

$$= \sigma_{\breve{h}}^4 C_P^2 \frac{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \sum\limits_{k'=K_{q'}^{(\mathsf{low})}}^{K_{q'}^{(\mathsf{up})}} \Lambda_q(k)\,\Lambda_{q'}(k')\left[\left|\mathcal{V}_{t'}(k+k')\right|^2 + \left|\mathcal{V}_{t'}(k-k')\right|^2\right]}{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \sum\limits_{k'=K_{q'}^{(\mathsf{low})}}^{K_{q'}^{(\mathsf{up})}} \Lambda_q(k)\,\Lambda_{q'}(k')} \tag{4.167}$$

with

$$\mathcal{V}_{t'}(k) = \sum_{p=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p\right)}{\tau_h}} w^2(p)\,\mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}pk} \tag{4.168}$$

and the window function

$$w\left(p'\right) := \sum_{l'=0}^{L_w-1} w_{\mathsf{A}}\left(l'\right) w_{\mathsf{S}}\left(l'+p'\right) \tag{4.169}$$

with support on $p' \in \{-L_w+1,\ldots,L_w-1\}$. The summation limits $\mathcal{L}(t')$ and $\mathcal{U}(t')$ are defined in (4.55) and (4.56), respectively. A detailed derivation is given in Appendix A.7 (according to [28], however, with the modification (4.81) applied). Note that the covariance matrices (4.163) are independent of the energy parameter $\sigma_{\breve{h}}^2$ which only contributes to the components of the mean vectors (4.162) in terms of the additive term $\ln\left(\sigma_{\breve{h}}^2\right)$.

Further note that the frequency dependent power compensation constant $C_P(k)$ defined in (4.73) has been replaced by a frequency independent power compensation constant C_P given by

$$C_P := \frac{E\left[\breve{C}_N(k)\right]}{E\left[\breve{C}_D(k)\right]} \tag{4.170}$$

$$= \frac{C_N}{C_D} \tag{4.171}$$

with (see Appendix A.6.1)

$$C_N := K^2 \sum_{t',t''=-L_{H,\ell}}^{L_H} \sum_{l=0}^{L_w-1} w_{\mathsf{A}}(l)\,w_{\mathsf{S}}(l)\,w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right) w_{\mathsf{S}}\left(l+\left(t''-t'\right)B\right) \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w_{\mathsf{A}}^2\left(l-p'\right) \tag{4.172}$$

$$C_D := \left(\sum_{l=0}^{L_w-1} w_{\mathsf{A}}^2(l)\right) \sum_{t'=0}^{L_H} \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w\left(p'\right). \tag{4.173}$$

The frequency independence of the power compensation constant and its independence of the AIR energy $\sigma^2_{\check{h}}$ is a direct consequence of applying the stochastic model of the AIR (4.147) already at the power spectral domain to ensure an unbiased prediction of the PSC of the reverberant speech signal (compare (4.69)). It can further be seen that the constant depends only on the parameters employed for the feature extraction, i.e. the analysis and synthesis windows, the number of frequency bins etc., and, via the decay constant τ_h, on the reverberation time T_{60}.

The frequency independent power compensation constant obtained with the parameter values of to the ETSI standard front-end for $T_S^{-1} = 8\,\text{kHz}$ (listed in Tab. 3.1) is given in Fig. 4.10 for different reverberation times. Its value clearly is upper bounded and further about 8 for a large range of practically relevant reverberation times.

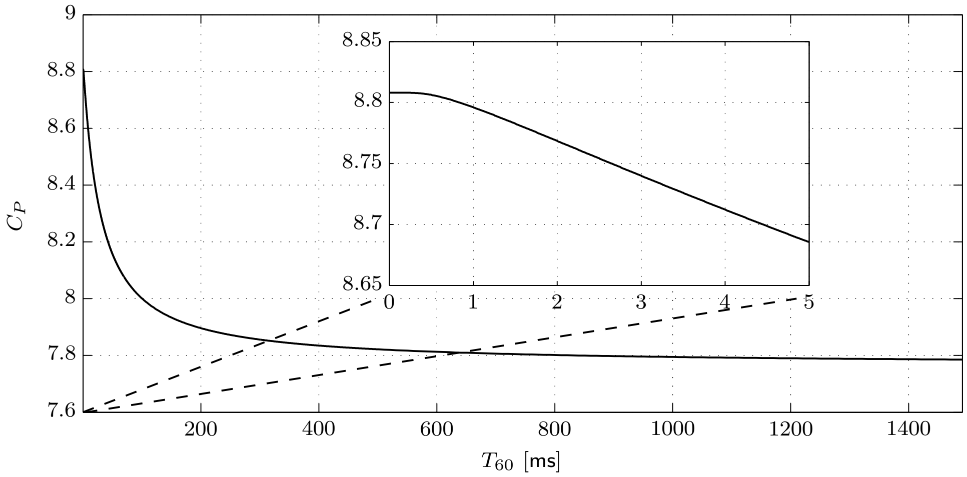


Figure 4.10: Frequency independent power compensation constant C_P for different reverberation times and $T_S^{-1} = 8\,\text{kHz}$ ($\epsilon_h = 10^{-3}$). As analysis window w_A a HAMMING *window has been chosen. Its length $L_w = 200$ and the shift $B = 80$ are taken from Tab. 3.1. The synthesis window w_S for the given analysis window w_A is computed as the least squares solution to the set of linear equations given in* (4.47).

4.5 Observation Models – AIR Model Applied

The AIR model (4.147) allows the approximate representation of the AIR in the logarithmic mel domain $\bar{\mathbf{h}}^{(l)}_{0:L_H}$ to be replaced by its expected value $\boldsymbol{\mu}_{\check{\mathbf{h}}^{(l)}_{0:L_H}}$ given by (4.162). In doing so, the observation error in the observation model for the presence of reverberation and the absence of noise (which also drives the observation error in the presence of both reverberation and noise) may partly compensate for this approximation.

In addition, the AIR model (4.147) also allows the observation models (4.92) and (4.110) to be reformulated in terms of recursive observation mappings relating the LMPSC feature vectors of the reverberant/noisy reverberant observation at time instant t to a reduced number of $L_R \ll L_H + 1$ LMPSC feature vectors of the clean speech signal and the reverberant/noisy reverberant observation at time instant $t - L_R$ [67, 82]. Since L_H linearly depends on the length of the AIR, the number of clean speech LMPSC feature vectors to be considered in the observation mappings becomes large even in the presence of *moderate* reverberation (compare Tab. 4.1). A recursive formulation thus may significantly reduce the computational effort coming along with the evaluation of the observation mappings $f_s^{(\mathsf{l})}(\cdot)$ and $f_o^{(\mathsf{l})}(\cdot)$ in (4.92) and (4.110). These observation models will, opposed to the *non-recursive* ones dealt with so far, in the following be denoted by *recursive* observation models.

Starting with an overview of the non-recursive observation models employing the stochastic AIR model (4.147) in Sec. 4.5.1, a recursive observation model for the presence of reverberation and the absence of noise will be derived in Sec. 4.5.2. The findings will then be employed to also derive a recursive observation model for the presence of both reverberation and noise in Sec. 4.5.3. An overview of the recursive observation models will finally be given in Sec. 4.5.4.

4.5.1 Overview of Non-Recursive Observation Models

Figure 4.11 gives an overview of the non-recursive observation models if the AIR is modeled as a stochastic process according to (4.147) (compare Fig. 4.8 for the case where the AIR is known rather than modeled.). Note that due to the change of the AIR representation in the LMPSC domain to its expected value under the AIR model (4.147), the observation error, now denoted by $\mathbf{v}_{ot}^{(\mathsf{l},\mathsf{N})}$, also has to compensate for this approximation. Further, the IRNR vector will now denoted by $\mathbf{r}_t^{(\mathsf{l},\mathsf{N})}$, since it has to account for the changed AIR representation in the LMPSC domain, too. Though in the absence of reverberation $\mathbf{r}_t^{(\mathsf{l},\mathsf{N})} = \mathbf{r}_t^{(\mathsf{l})}$ and thus $\mathbf{v}_{yt}^{(\mathsf{l},\mathsf{N})} = \mathbf{v}_{yt}^{(\mathsf{l})}$, a unified notation has been chosen here to ease a comparison of the different observation models.

4.5.2 Recursive Observation Model in the Presence of Reverberation and the Absence of Background Noise

The recursive observation model in the presence of reverberation and the absence of noise is based on a recursive formulation of the expected value of the power of the band-to-band filters $\breve{h}_{t'}(k,k)$ at time instant t'.

With the expected value of the power of the band-to-band filters $\breve{h}_{t'}(k,k)$ at time instant t' given by (see (A.105) in Appendix A.6.1)

$$E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] = \sigma_{\breve{h}}^2 \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{4.174}$$

the power of the band-to-band filters $\breve{h}_{t'+L_R}(k,k)$ at time instant $t'+L_R$ may be expressed

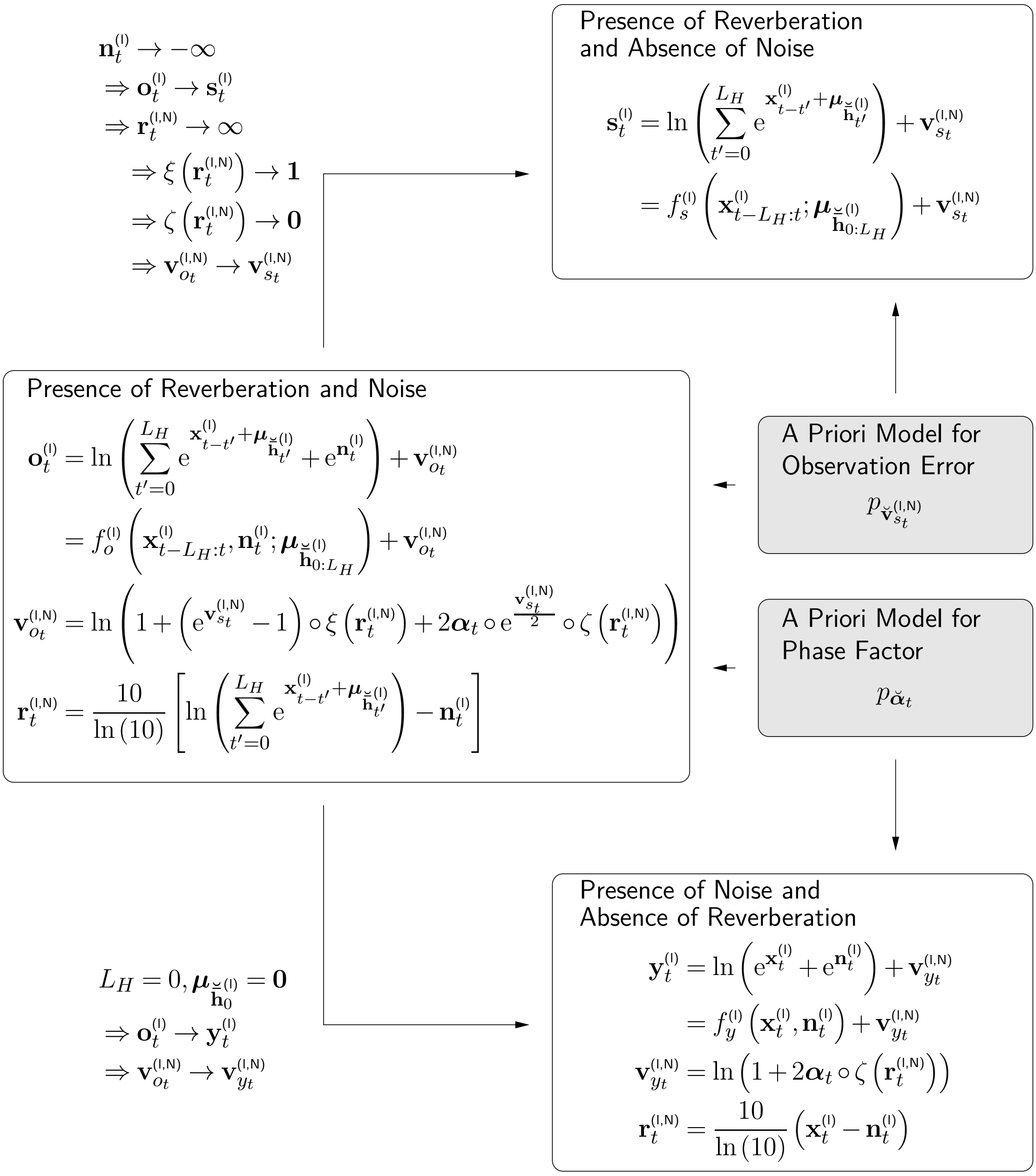

$\mathbf{n}_t^{(\mathsf{I})} \rightarrow -\infty$
$\Rightarrow \mathbf{o}_t^{(\mathsf{I})} \rightarrow \mathbf{s}_t^{(\mathsf{I})}$
$\Rightarrow \mathbf{r}_t^{(\mathsf{I,N})} \rightarrow \infty$
$\Rightarrow \xi\left(\mathbf{r}_t^{(\mathsf{I,N})}\right) \rightarrow \mathbf{1}$
$\Rightarrow \zeta\left(\mathbf{r}_t^{(\mathsf{I,N})}\right) \rightarrow \mathbf{0}$
$\Rightarrow \mathbf{v}_{o_t}^{(\mathsf{I,N})} \rightarrow \mathbf{v}_{s_t}^{(\mathsf{I,N})}$
Presence of Reverberation and Absence of Noise
$\mathbf{s}_t^{(\mathsf{I})} = \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathsf{I})}}}\right) + \mathbf{v}_{s_t}^{(\mathsf{I,N})} = f_s^{(\mathsf{I})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{I})}; \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_H}^{(\mathsf{I})}}\right) + \mathbf{v}_{s_t}^{(\mathsf{I,N})}$
Presence of Reverberation and Noise
$\mathbf{o}_t^{(\mathsf{I})} = \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathsf{I})}}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{I})}}\right) + \mathbf{v}_{o_t}^{(\mathsf{I,N})} = f_o^{(\mathsf{I})}\left(\mathbf{x}_{t-L_H:t}^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}; \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_H}^{(\mathsf{I})}}\right) + \mathbf{v}_{o_t}^{(\mathsf{I,N})}$
$\mathbf{v}_{o_t}^{(\mathsf{I,N})} = \ln\left(1 + \left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{I,N})}} - 1\right) \circ \xi\left(\mathbf{r}_t^{(\mathsf{I,N})}\right) + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{I,N})}}{2}} \circ \zeta\left(\mathbf{r}_t^{(\mathsf{I,N})}\right)\right)$
$\mathbf{r}_t^{(\mathsf{I,N})} = \frac{10}{\ln(10)}\left[\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathsf{I})}}}\right) - \mathbf{n}_t^{(\mathsf{I})}\right]$
A Priori Model for Observation Error
$p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{I,N})}}$
A Priori Model for Phase Factor
$p_{\breve{\boldsymbol{\alpha}}_t}$
$L_H = 0, \boldsymbol{\mu}_{\breve{\mathbf{h}}_0^{(\mathsf{I})}} = \mathbf{0}$
$\Rightarrow \mathbf{o}_t^{(\mathsf{I})} \rightarrow \mathbf{y}_t^{(\mathsf{I})}$
$\Rightarrow \mathbf{v}_{o_t}^{(\mathsf{I,N})} \rightarrow \mathbf{v}_{y_t}^{(\mathsf{I,N})}$
Presence of Noise and Absence of Reverberation
$\mathbf{y}_t^{(\mathsf{I})} = \ln\left(\mathrm{e}^{\mathbf{x}_t^{(\mathsf{I})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathsf{I})}}\right) + \mathbf{v}_{y_t}^{(\mathsf{I,N})} = f_y^{(\mathsf{I})}\left(\mathbf{x}_t^{(\mathsf{I})}, \mathbf{n}_t^{(\mathsf{I})}\right) + \mathbf{v}_{y_t}^{(\mathsf{I,N})}$
$\mathbf{v}_{y_t}^{(\mathsf{I,N})} = \ln\left(1 + 2\boldsymbol{\alpha}_t \circ \zeta\left(\mathbf{r}_t^{(\mathsf{I,N})}\right)\right)$
$\mathbf{r}_t^{(\mathsf{I,N})} = \frac{10}{\ln(10)}\left(\mathbf{x}_t^{(\mathsf{I})} - \mathbf{n}_t^{(\mathsf{I})}\right)$

Figure 4.11: *Overview of the non-recursive observation models: the AIR representation in the LMPSC domain has been replaced by its expected value under the AIR model* (4.147)*. The observation model for the presence of both reverberation and noise may again be considered as a generalization of the observation models for the presence of either reverberation or noise. Note the changed definition for the IRNR vector* $\mathbf{r}_t^{(\mathsf{I,N})}$ *and the observation errors* $\mathbf{v}_{s_t}^{(\mathsf{I,N})}$ *in the presence of reverberation. The notation for the absence of reverberation has only been adapted to ease comparison between the observation models.*

as

$$E\left[\left|\breve{h}_{t'+L_R}(k,k)\right|^2\right] = \sigma^2_{\breve{h}} \sum_{p'=\mathcal{L}(t'+L_R)}^{\mathcal{U}(t'+L_R)} \mathrm{e}^{-\frac{2\left(\left(t'+L_R\right)B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{4.175}$$

$$= \sigma^2_{\breve{h}} \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \sum_{p'=\mathcal{L}(t'+L_R)}^{\mathcal{U}(t'+L_R)} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{4.176}$$

$$\approx \sigma^2_{\breve{h}} \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{4.177}$$

$$= \mathrm{e}^{-\frac{2L_RB}{\tau_h}} E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right]. \tag{4.178}$$

Since $\mathcal{L}(t'+L_R) = \mathcal{L}(t')$ for $t'+L_R \geq \frac{L_h-L_w}{B}$ and $\mathcal{U}(t'+L_R) = \mathcal{U}(t')$ for $t'+L_R \leq \frac{L_w-1}{B}$, (4.178) is exact for $\frac{L_w-1}{B} \leq t'+L_R \leq \frac{L_h-L_w}{B}$. Eq. (4.178) is on the outset of the recursive observation models presented in the following.

By replacing the the power $\left|\breve{h}_{t'}(k,k)\right|^2$ of the band-to-band filters $\breve{h}_{t'}(k,k)$ in (4.69) by their expected value under the AIR model given in (4.174), the PSC of the reverberant speech signal turns into

$$|S_t(k)|^2 = C_P \sum_{t'=0}^{L_H} |X_{t-t'}(k)|^2 E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] + \tilde{E}_t(k) \tag{4.179}$$

$$= C_P \sum_{t'=0}^{L_R-1} |X_{t-t'}(k)|^2 E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] + \sum_{t'=0}^{L_H-L_R} \left|X_{t-(t'+L_R))}(k)\right|^2 E\left[\left|\breve{h}_{t'+L_R}(k,k)\right|^2\right] + \tilde{E}_t(k). \tag{4.180}$$

Note that by replacing the frequency dependent power compensation constant $C_P(k)$ by the frequency independent power compensation constant C_P defined in (4.170), the new error term $\tilde{E}_t(k)$ still is approximately of zero mean.

Employing the approximation (4.178) then yields

$$|S_t(k)|^2 = C_P \sum_{t'=0}^{L_R-1} |X_{t-t'}(k)|^2 E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] + \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \sum_{t'=0}^{L_H} \left|X_{t-(t'+L_R))}(k)\right|^2 E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] + \tilde{\tilde{E}}_t(k). \tag{4.181}$$

The additional errors introduced by the approximation (4.178) are captured by the new error term $\tilde{\tilde{E}}_t(k)$. Note that since the AIR is assumed to be of finite length, the implicit change of the upper summation limit of the second sum from $L_H - L_R$ to L_H does not introduce additional errors.

Finally, the similarity of the last sum to $\left|S_{t-L_R}(k)\right|^2$ may be employed to obtain

$$|S_t(k)|^2 = C_P \sum_{t'=0}^{L_R-1} |X_{t-t'}(k)|^2 E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] + \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \left|S_{t-L_R}(k)\right|^2 + \tilde{\tilde{\tilde{E}}}_t(k). \tag{4.182}$$

Since the last sum is not exactly $\left|S_{t-L_R}(k)\right|^2$ but rather an estimate thereof, the error due to this approximation is again compensated for by means of the new error term $\tilde{\tilde{E}}_t(k)$.

Taking (4.182) to the LMPSC domain finally yields

$$\mathbf{s}_t^{(\mathsf{l})} = \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})}+\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathsf{l})}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}}\mathrm{e}^{\mathbf{s}_{t-L_R}^{(\mathsf{l})}}\right) + \mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})} \tag{4.183}$$

$$= f_{s,L_R}^{(\mathsf{l,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{t-L_R}^{(\mathsf{l})}; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right) + \mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}, \tag{4.184}$$

with the observation mapping

$$f_{s,L_R}^{(\mathsf{l,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{t-L_R}^{(\mathsf{l})}; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right) := \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})}+\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{(\mathsf{l})}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}}\mathrm{e}^{\mathbf{s}_{t-L_R}^{(\mathsf{l})}}\right) \tag{4.185}$$

and the observation error $\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}$. Note that the frequency independent power compensation constant C_P has been absorbed by the expected AIR representation $\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R}^{(\mathsf{l})}}$.

Following the considerations in Sec. 4.3.2, the conditional PDF of the LMPSC vector of the reverberant speech signal is thus given by

$$\begin{aligned} & p_{\breve{\mathbf{s}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{s}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{s}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{1:t-1}^{(\mathsf{l})}\right) \\ & = p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{s}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{s}_t^{(\mathsf{l})} - f_{s,L_R}^{(\mathsf{l,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{t-L_R}^{(\mathsf{l})}; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right)\middle|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{1:t-1}^{(\mathsf{l})}\right), \end{aligned} \tag{4.186}$$

i.e., completely characterized by the conditional PDF of the observation error. In [67], the observation error $\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}$ has been been considered as a realization of a real-valued white, stationary and ergodic GAUSSIAN process that is independent of the most recent L_R LMPSC vectors of the clean speech signal and all past LMPSC vectors of the reverberant speech signal, i.e.,

$$p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{s}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}\middle|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{1:t-1}^{(\mathsf{l})}\right) \approx p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}}\left(\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}\right) \tag{4.187}$$

$$:= \mathcal{N}\left(\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}; \boldsymbol{\mu}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}, \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}\right), \tag{4.188}$$

which is applied $\forall t \in \{1, \cdots, T\}$. The mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}$ may thereby again be obtained from artificially reverberated training data and the underlying clean data. A detailed analysis of the observation error in the recursive observation model in the presence of reverberation and the absence of noise will follow in Sec. 4.7.1.

The recursive observation model in the presence of reverberation and the absence of noise will thus finally be approximated by

$$\begin{aligned} & p_{\breve{\mathbf{s}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{s}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{s}_t^{(\mathsf{l})}\middle|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{1:t-1}^{(\mathsf{l})}\right) \\ & \approx \mathcal{N}\left(\mathbf{s}_t^{(\mathsf{l})} - f_{s,L_R}^{(\mathsf{l,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})}, \mathbf{s}_{t-L_R}^{(\mathsf{l})}; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right); \boldsymbol{\mu}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}, \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{l,R})}}\right). \end{aligned} \tag{4.189}$$

4.5.3 Recursive Observation Model in the Presence of Reverberation and Background Noise

With the recursive observation mapping in the absence of noise and the resulting observation model given in (4.185) and (4.189), respectively, the derivation of a recursive observation model in the presence of both reverberation and noise may, in theory, be carried out as done in Sec. 4.3.3. There, the non-recursive observation model in the presence of both reverberation and noise has been derived from the non-recursive observation model in the presence of reverberation but the absence of noise.

However, the LMPSC feature vector $\mathbf{s}_{t-L_R}^{(\mathsf{l})}$ of the reverberant speech signal at time instant $t-L_R$ required by the recursive observation mapping (4.185) is not part of the observable sequence of feature vectors $\mathbf{o}_{1:t-1}^{(\mathsf{l})}$ in the additional presence of noise. Since $\mathbf{s}_{t-L_R}^{(\mathsf{l})}$ is linked to $\mathbf{n}_{t-L_R}^{(\mathsf{l})}$ via (4.102), the LMPSC feature vector of the noise $\mathbf{n}_{t-L_R}^{(\mathsf{l})}$ will temporarily be included in the state vector. Hence, for the derivation of the targeted observation PDF, i.e., $p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{n}}_{t-L_R}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$, it may now be assumed that the LMPSC feature vector of the noise at time instant $t-L_R$ is known.

In the following, $\mathrm{e}^{\mathbf{s}_{t-L_R}^{(\mathsf{l})}} = \mathbf{s}_{t-L_R}^{(\mathsf{m})}$ (the essentially required quantity) will thus be replaced by an approximate MMSE estimate at time instant $t-L_R$ defined by (see Appendix A.8 for a detailed derivation)

$$\hat{s}_{t-L_R}^{(\mathsf{m,R})}(q) := E\left[\breve{s}_{t-L_R}^{(\mathsf{m})}(q)\middle|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{n}}_{t-L_R}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}\right] \tag{4.190}$$

$$\approx \begin{cases} \left(2\sigma_{\breve{\alpha}_q}^2-1\right) n_{t-L_R}^{(\mathsf{m})}(q)+o_{t-L_R}^{(\mathsf{m})}(q), & \text{if } o_{t-L_R}^{(\mathsf{m})}(q) \geq n_{t-L_R}^{(\mathsf{m})}(q) \\ \left(2\sigma_{\breve{\alpha}_q}^2-1\right) o_{t-L_R}^{(\mathsf{m})}(q)+n_{t-L_R}^{(\mathsf{m})}(q), & \text{if } o_{t-L_R}^{(\mathsf{m})}(q) < n_{t-L_R}^{(\mathsf{m})}(q) \end{cases}. \tag{4.191}$$

where $\sigma_{\breve{\alpha}_q}^2$ denotes the variance of the phase factor at mel frequency index q. Note that approximation (4.191) is exact for $o_{t-L_R}^{(\mathsf{m})}(q) \geq n_{t-L_R}^{(\mathsf{m})}(q)$.

Employing vector notation and expressing (4.191) in terms of the corresponding LMPSC vectors further allows the MMSE estimate to be written as

$$\mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{(\mathsf{l,R})}} := \max\left\{\left(2\boldsymbol{\sigma}_{\breve{\alpha}}^2-\mathbf{1}\right)\circ\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathsf{l})}}+\mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathsf{l})}},\ \left(2\boldsymbol{\sigma}_{\breve{\alpha}}^2-\mathbf{1}\right)\circ\mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathsf{l})}}+\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathsf{l})}}\right\} \tag{4.192}$$

$$= \left(2\boldsymbol{\sigma}_{\breve{\alpha}}^2-\mathbf{1}\right)\circ\min\left\{\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathsf{l})}},\ \mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathsf{l})}}\right\}+\max\left\{\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathsf{l})}},\ \mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathsf{l})}}\right\} \tag{4.193}$$

where $\boldsymbol{\sigma}_{\breve{\alpha}}^2$ now denotes the variance vector associated with the vector of phase factors. Note that the minimum/maximum operation has to be carried out on the vectors componentwise.

With the estimate $\hat{\mathbf{s}}_{t-L_R}^{(\mathsf{m,R})}$ of $\mathbf{s}_{t-L_R}^{(\mathsf{m})}$ given by (4.193), the LMPSC feature vector of the reverberant speech signal $\mathbf{s}_t^{(\mathsf{l})}$ may be written as (from (4.183))

$$\mathbf{s}_t^{(\mathsf{l})} = \ln\left(\sum_{t'=0}^{L_R-1}\mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathsf{l})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathsf{l})}}}+\mathrm{e}^{-\frac{2L_RB}{\tau_h}}\mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{(\mathsf{l,R})}}\right)+\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}+\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})} \tag{4.194}$$

with

$$\mathbf{w}^{(\mathrm{l,R})}_{ot,L_R} := \ln\left(\frac{\sum\limits_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\mathbf{s}^{(\mathrm{l})}_{t-L_R}}}{\sum\limits_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}}}\right) \tag{4.195}$$

$$= \ln\left(\mathbf{1} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}} \frac{\mathrm{e}^{\mathbf{s}^{(\mathrm{l})}_{t-L_R}} - \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}}}{\sum\limits_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}}}\right) \tag{4.196}$$

denoting the error introduced by replacing the LMPSC feature vector $\mathbf{s}^{(\mathrm{l})}_{t-L_R}$ of the reverberant speech signal by the estimate (4.190).

Following the calculus carried out in (4.102)-(4.111), the LMPSC feature vector $\mathbf{o}^{(\mathrm{l})}_t$ of the noisy reverberant observation is then characterized by

$$\begin{aligned}\mathbf{o}^{(\mathrm{l})}_t = \ln\Bigg(& \mathrm{e}^{\mathbf{w}^{(\mathrm{l,R})}_{ot,L_R}+\mathbf{v}^{(\mathrm{l,R})}_{st,L_R}} \circ \left[\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}}\right] + \mathrm{e}^{\mathbf{n}^{(\mathrm{l})}_t} \\ & +2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{w}^{(\mathrm{l,R})}_{ot,L_R}+\mathbf{v}^{(\mathrm{l,R})}_{st,L_R}}{2}} \circ \mathrm{e}^{\frac{\ln\left(\sum\limits_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}}\right)+\mathbf{n}^{(\mathrm{l})}_t}{2}}\Bigg)\end{aligned} \tag{4.197}$$

$$= f^{(\mathrm{l,R})}_{o,L_R}\left(\mathbf{x}^{(\mathrm{l})}_{t-L_R+1:t}, \mathbf{n}^{(\mathrm{l})}_{t-L_R,t}, \mathbf{o}^{(\mathrm{l})}_{t-L_R}; \boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{0:L_R-1}}\right) + \mathbf{v}^{(\mathrm{l,R})}_{ot,L_R} \tag{4.198}$$

where, for ease of reading, the short-hand notation $\breve{\mathbf{n}}^{(\mathrm{l})}_{t-L_R,t} = \breve{\mathbf{n}}^{(\mathrm{l})}_{t-L_R}, \breve{\mathbf{n}}^{(\mathrm{l})}_t$ has been introduced. The observation mapping $f^{(\mathrm{l,R})}_{o,L_R}(\cdot)$ and the corresponding error $\mathbf{v}^{(\mathrm{l,R})}_{ot,L_R}$ are defined as

$$\begin{aligned} & f^{(\mathrm{l,R})}_{o,L_R}\left(\mathbf{x}^{(\mathrm{l})}_{t-L_R+1:t}, \mathbf{n}^{(\mathrm{l})}_{t-L_R,t}, \mathbf{o}^{(\mathrm{l})}_{t-L_R}; \boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{0:L_R-1}}\right) \\ & := \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathrm{l})}_{t-t'}+\boldsymbol{\mu}_{\breve{\mathbf{h}}^{(\mathrm{l})}_{t'}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\, \mathrm{e}^{\hat{\mathbf{s}}^{(\mathrm{l,R})}_{t-L_R}} + \mathrm{e}^{\mathbf{n}^{(\mathrm{l})}_t}\right)\end{aligned} \tag{4.199}$$

and

$$\mathbf{v}^{(\mathrm{l,R})}_{ot,L_R} := \ln\left(1 + \left(\mathrm{e}^{\mathbf{v}^{(\mathrm{l,R})}_{st,L_R}+\mathbf{w}^{(\mathrm{l,R})}_{ot,L_R}} - 1\right) \circ \xi\left(\mathbf{r}^{(\mathrm{l,R})}_{t,L_R}\right) + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}^{(\mathrm{l,R})}_{st,L_R}+\mathbf{w}^{(\mathrm{l,R})}_{ot,L_R}}{2}} \circ \zeta\left(\mathbf{r}^{(\mathrm{l,R})}_{t,L_R}\right)\right), \tag{4.200}$$

respectively. The IRNR is now defined as

$$\mathbf{r}^{(\mathsf{I,R})}_{t,L_R} := \frac{\ln(10)}{10}\left[\ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}^{(\mathsf{I})}_{t-t'}+\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{I})}_{t'}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}}\,\mathrm{e}^{\hat{\mathbf{s}}^{(\mathsf{I,R})}_{t-L_R}}\right) - \mathbf{n}^{(\mathsf{I})}_t\right]. \tag{4.201}$$

Eq. (4.200) thus may be considered the recursive equivalent to the observation error (4.105) in the non-recursive observation model with the only differences in the definition of the IRNR (4.200) and the additional error term $\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}$. Consequently, the conditional PDF of the observation error $\breve{\mathbf{v}}^{(\mathsf{I,R})}_{o_t,L_R}$ is given by (compare (4.115))

$$\begin{aligned}
&p_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{o_t,L_R}|\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}}\left(\mathbf{v}^{(\mathsf{I,R})}_{o_t,L_R}\middle|\mathbf{x}^{(\mathsf{I})}_{t-L_R+1:t},\mathbf{n}^{(\mathsf{I})}_{t-L_R,t},\mathbf{o}^{(\mathsf{I})}_{1:t-1}\right)\\
&\approx\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v^{(\mathsf{I,R})}_{o_t,L_R}(q)}}{2\zeta\left(r^{(\mathsf{I,R})}_{t,L_R}(q)\right)}\right)\\
&\quad\int\limits_{\mathbb{R}^Q}\int\limits_{\mathbb{R}^Q} p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{v}^{(\mathsf{I,R})}_{o_t,L_R}}-1-\left(\mathrm{e}^{\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}+\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}}-1\right)\circ\xi\left(\mathbf{r}^{(\mathsf{I,R})}_{t,L_R}\right)}{2\mathrm{e}^{\frac{\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}+\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}}{2}}\circ\zeta\left(\mathbf{r}^{(\mathsf{I,R})}_{t,L_R}\right)}\right)\\
&\qquad\qquad p_{\breve{\mathbf{w}}^{(\mathsf{I,R})}_{o_t,L_R}|\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}\left(\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}\middle|\mathbf{x}^{(\mathsf{I})}_{t-L_R+1:t},\mathbf{n}^{(\mathsf{I})}_{t-L_R,t},\mathbf{o}^{(\mathsf{I})}_{1:t-1},\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}\right)\\
&\qquad\qquad p_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}\left(\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}\right)\mathrm{d}\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}\mathrm{d}\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}.
\end{aligned}\tag{4.202}$$

Due to the exponentially decaying factor in $\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}$ defined in (4.195), $\mathbf{w}^{(\mathsf{I,R})}_{o_t,L_R}$ may be considered to be approximately zero for large recursion lengths L_R. Intuitively, at larger recursion length L_R more of the reverberant observation $\mathrm{s}^{(\mathsf{I})}_t$ may be *explained away* by the accumulated contribution of the present and past speech feature vectors in (4.194), as the upper sum limit is also growing with L_R. Hence, (4.202) may be approximated by

$$\begin{aligned}
&p_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{o_t,L_R}|\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}}\left(\mathbf{v}^{(\mathsf{I,R})}_{o_t,L_R}\middle|\mathbf{x}^{(\mathsf{I})}_{t-L_R+1:t},\mathbf{n}^{(\mathsf{I})}_{t-L_R,t},\mathbf{o}^{(\mathsf{I})}_{1:t-1}\right)\\
&\overset{L_R\gg}{\approx}\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v^{(\mathsf{I,R})}_{o_t,L_R}(q)}}{2\zeta\left(r^{(\mathsf{I,R})}_{t,L_R}(q)\right)}\right)\\
&\quad\int\limits_{\mathbb{R}^Q} p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{v}^{(\mathsf{I,R})}_{o_t,L_R}}-1-\left(\mathrm{e}^{\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}}-1\right)\circ\xi\left(\mathbf{r}^{(\mathsf{I,R})}_{t,L_R}\right)}{2\mathrm{e}^{\frac{\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}}{2}}\circ\zeta\left(\mathbf{r}^{(\mathsf{I,R})}_{t,L_R}\right)}\right)p_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}\left(\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}\right)\mathrm{d}\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}.
\end{aligned}\tag{4.203}$$

With the same reasoning, the stochastic observation model $p_{\breve{\mathbf{o}}^{(\mathsf{I})}_t|\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}}$ may directly be inferred from (4.129) without the tedious math involved to be

$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-L_R,t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\left|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})},\mathbf{n}_{t-L_R,t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right.\right)\\
&\approx\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{o_t^{(\mathsf{l})}(q)-f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\mu_{\breve{\bar{h}}_{0:L_R-1}^{(\mathsf{l})}}(q)\right)}}{2\zeta\left(r_{t,L_R}^{(\mathsf{l,R})}(q)\right)}\right)\\
&\qquad\int\limits_{\mathbb{R}^Q}\int\limits_{\mathbb{R}^Q}p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{o}_t^{(\mathsf{l})}-f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right)}-1-\left(\mathrm{e}^{\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}+\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}}-1\right)\circ\xi\left(\mathbf{r}_{t,L_R}^{(\mathsf{l,R})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}+\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}}{2}}\circ\zeta\left(\mathbf{r}_{t,L_R}^{(\mathsf{l,R})}\right)}\right)\\
&\qquad\quad p_{\breve{\mathbf{w}}_{o_t,L_R}^{(\mathsf{l,R})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-L_R,t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}}\left(\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}\left|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})},\mathbf{n}_{t-L_R,t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}\right.\right)\\
&\qquad\quad p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}}\left(\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}\right)\mathrm{d}\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}\mathrm{d}\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})},
\end{aligned}
\tag{4.204}
$$

where the short-hand notations

$$
f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right):=f_{o,L_R}^{(\mathsf{l,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})},\mathbf{n}_{t-L_R,t}^{(\mathsf{l})},\mathbf{o}_{t-L_R}^{(\mathsf{l})};\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right)\tag{4.205}
$$

$$
f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\mu_{\breve{\bar{h}}_{0:L_R-1}^{(\mathsf{l})}}(q)\right):=f_{o,L_R}^{(\mathsf{l,R})}\left(x_{t-L_R+1:t}^{(\mathsf{l})}(q),n_{t-L_R,t}^{(\mathsf{l})}(q),o_{t-L_R}^{(\mathsf{l})}(q);\mu_{\breve{\bar{h}}_{0:L_R-1}^{(\mathsf{l})}}(q)\right)\tag{4.206}
$$

have been introduced for ease of readability.

For large recursion lengths, (4.204) is approximately given by

$$
\begin{aligned}
&p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_R+1:t}^{(\mathsf{l})},\breve{\mathbf{n}}_{t-L_R,t}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\left|\mathbf{x}_{t-L_R+1:t}^{(\mathsf{l})},\mathbf{n}_{t-L_R,t}^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right.\right)\\
&\overset{L_R\gg}{\approx}\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{o_t^{(\mathsf{l})}(q)-f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\mu_{\breve{\bar{h}}_{0:L_R-1}^{(\mathsf{l})}}(q)\right)}}{2\zeta\left(r_{t,L_R}^{(\mathsf{l,R})}(q)\right)}\right)\\
&\qquad\int\limits_{\mathbb{R}^Q}p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{o}_t^{(\mathsf{l})}-f_{o,L_R}^{(\mathsf{l,R})}\left(\cdots;\boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{(\mathsf{l})}}\right)}-1-\left(\mathrm{e}^{\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}}-1\right)\circ\xi\left(\mathbf{r}_{t,L_R}^{(\mathsf{l,R})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}}{2}}\circ\zeta\left(\mathbf{r}_{t,L_R}^{(\mathsf{l,R})}\right)}\right)\\
&\qquad\quad p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathsf{l,R})}}\left(\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}\right)\mathrm{d}\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}.
\end{aligned}
\tag{4.207}
$$

Since (4.203) and thus also (4.207) are rather bulky and only of minor practical use,

$p_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o_t,L_R}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{l})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ will in the following be approximated by a GAUSSIAN distribution as

$$\begin{aligned} &p_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o_t,L_R}|\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{l})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(\mathbf{v}^{(\mathsf{l,R})}_{o_t,L_R}\middle|\mathbf{x}^{(\mathsf{l})}_{t-L_R+1:t},\mathbf{n}^{(\mathsf{l})}_{t-L_R,t},\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right)\\ &:= p_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o_t,L_R}|\breve{\mathbf{r}}^{(\mathsf{l,R})}_{t,L_R}}\left(\mathbf{v}^{(\mathsf{l,R})}_{o_t,L_R}\middle|\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)\\ &\approx\mathcal{N}\left(\mathbf{v}^{(\mathsf{l,R})}_{o_t,L_R};\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right),\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)\right), \end{aligned} \tag{4.208}$$

where the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{o,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)$ are functions of the IRNR $\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}$ and thus varying with time. They can be related to the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l,R})}_{t,L_R}\right)$ of the observation error $\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s_t,L_R}$ in the recursive observation model in the absence of noise in equivalence to (4.122)–(4.123), i.e.,

$$\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)=\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\right)-\frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\right) \tag{4.209}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)&=\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\right)^{\dagger}+\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\right)\\ &\quad-\ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)\right)^{\dagger}\right). \end{aligned} \tag{4.210}$$

The logarithm again has to be applied to the occurring vectors and matrices componentwise. Thereby $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s,L_R}}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right)$ denote the conditional mean vector and the conditional covariance matrix of the (consequently) log-normally distributed RV $\breve{\mathbf{v}}^{(\mathsf{m,R})}_{s_t,L_R} := \mathrm{e}^{\breve{\mathbf{v}}^{(\mathsf{l,R})}_{s_t,L_R}}$. The required expectation values are given by

$$\mu_{\breve{v}^{(\mathsf{m,R})}_{s,L_R}(q)}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right):=1+\left(E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l,R})}_{s_t,L_R}(q)}\right]-1\right)\xi\left(r^{(\mathsf{l})}_{t}(q)\right) \tag{4.211}$$

$$\begin{aligned} \sigma_{\breve{v}^{(\mathsf{m,R})}_{s,L_R}(q),\breve{v}^{(\mathsf{m})}_{o}(q')}\left(\mathbf{r}^{(\mathsf{l})}_{t}\right):=&\left(E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l,R})}_{s_t,L_R}(q)+\breve{v}^{(\mathsf{l})}_{s}(q')}\right]-E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l,R})}_{s,L_R}(q)}\right]E\left[\mathrm{e}^{\breve{v}^{(\mathsf{l,R})}_{s_t,L_R}(q')}\right]\right)\\ &\cdot\xi\left(r^{(\mathsf{l})}_{t}(q)\right)\xi\left(r^{(\mathsf{l})}_{t}\left(q'\right)\right)\\ &+4\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}^{(\mathsf{l,R})}_{s_t,L_R}(q)+\breve{v}^{(\mathsf{l,R})}_{s_t,L_R}(q')\right)}\right]\zeta\left(r^{(\mathsf{l})}_{t}(q)\right)\zeta\left(r^{(\mathsf{l})}_{t}\left(q'\right)\right), \end{aligned} \tag{4.212}$$

with

$$E\left[\mathrm{e}^{\breve{v}^{\text{(l,R)}}_{s_t,L_R}(q)}\right] = \mathrm{e}^{\mu_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q)} + \frac{1}{2}\sigma^2_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q)}} \tag{4.213}$$

$$E\left[\mathrm{e}^{\breve{v}^{\text{(l,R)}}_{s_t,L_R}(q)+\breve{v}^{\text{(l,R)}}_{s,L_R}\left(q'\right)}\right] = \mathrm{e}^{\mu_{\breve{v}^{\text{(l)}}_{s}(q)} + \mu_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q')} + \frac{1}{2}\left(\sigma^2_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q)} + \sigma^2_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q')} + 2\sigma_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q),\breve{v}^{\text{(l,R)}}_{s,L_R}(q')}\right)} \tag{4.214}$$

$$\begin{aligned} E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}^{\text{(l,R)}}_{s_t,L_R}(q)+\breve{v}^{\text{(l,R)}}_{s_t,L_R}\left(q'\right)\right)}\right] &= \mathrm{e}^{\frac{1}{2}\left(\mu_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q)} + \mu_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q')}\right)} \\ &\quad \cdot \mathrm{e}^{\frac{1}{8}\left(\sigma^2_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q)} + \sigma^2_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q')} + 2\sigma_{\breve{v}^{\text{(l,R)}}_{s,L_R}(q),\breve{v}^{\text{(l,R)}}_{s,L_R}(q')}\right)}. \end{aligned} \tag{4.215}$$

Again, the vector of phase factors can be found to contribute (only) in terms of its first two central moments (see Sec. 4.6.4). The final approximate observation model is thus given by

$$\begin{aligned} &p_{\breve{\mathbf{o}}^{\text{(l)}}_t|\breve{\mathbf{x}}^{\text{(l)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\text{(l)}}_{t-L_R,t},\breve{\mathbf{o}}^{\text{(l)}}_{1:t-1}}\left(\mathbf{o}^{\text{(l)}}_t \middle| \mathbf{x}^{\text{(l)}}_{t-L_R+1:t}, \mathbf{n}^{\text{(l)}}_{t-L_R,t}, \mathbf{o}^{\text{(l)}}_{1:t-1}\right) \approx \\ &\mathcal{N}\left(\mathbf{o}^{\text{(l)}}_t; f^{\text{(l,R)}}_{o,L_R}\left(\mathbf{x}^{\text{(l)}}_{t-L_R+1:t}, \mathbf{n}^{\text{(l)}}_{t-L_R,t}, \mathbf{o}^{\text{(l)}}_{t-L_R}; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{\text{(l)}}_{0:L_R-1}}\right) + \boldsymbol{\mu}_{\breve{\mathbf{v}}^{\text{(l,R)}}_{o,L_R}}\left(\mathbf{r}^{\text{(l,R)}}_{t,L_R}\right), \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{\text{(l,R)}}_{o,L_R}}\left(\mathbf{r}^{\text{(l,R)}}_{t,L_R}\right)\right). \end{aligned} \tag{4.216}$$

4.5.4 Overview of Recursive Observation Models

Figure 4.12 now gives an overview of the recursive observation models. The observation model in the presence of both reverberation and background noise may again be considered a generalization of the observation model in the presence of reverberation and the absence of background noise. The observation model in the absence of reverberation may, due to the recursive formulation, not be derived from the former any more.

4.6 Vector of Phase Factors

Looking at Fig. 4.8, Fig. 4.11 or Fig. 4.12, the PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ can be found to be an integral part of the stochastic observation models in the presence of noise. In fact, in the absence of reverberation, as $\mathbf{o}^{\text{(l)}}_t \rightarrow \mathbf{y}^{\text{(l)}}_t$, the phase factor is the only source of uncertainty about the observation $\mathbf{y}^{\text{(l)}}_t$ if the LMPSC feature vectors $\mathbf{x}^{\text{(l)}}_{t-L_H:t}$ of the clean speech signal and that of the noise, i.e., $\mathbf{n}^{\text{(l)}}_t$, are given.

For the analysis following, majorly the phase factor in the presence of both reverberation and noise will be considered. However, the presented results will be contrasted with the ones obtained for the phase factor in the absence of reverberation and the presence of noise to eventually prove the key findings to be valid irrespective of the absence or presence of reverberation.

Moreover, since the vector of phase factors is only accessible in a *supervised* scenario, all empirical studies will be carried out on the AURORA 5 database for the *noisy reverberant*

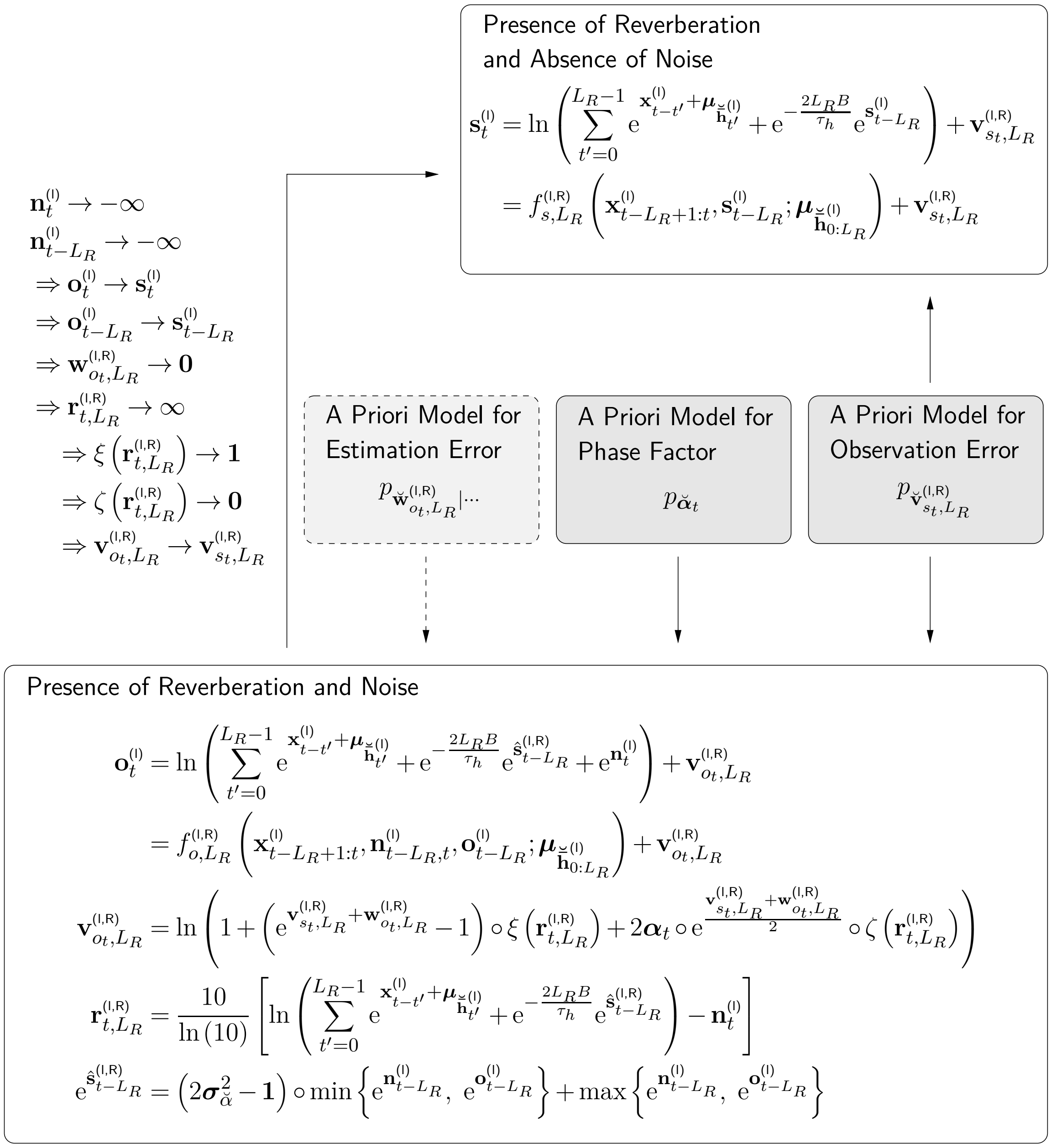

Presence of Reverberation and Absence of Noise
$\mathbf{s}_t^{(\mathrm{I})} = \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \mathrm{e}^{\mathbf{s}_{t-L_R}^{(\mathrm{I})}}\right) + \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}$
$= f_{s,L_R}^{(\mathrm{I,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathrm{I})}, \mathbf{s}_{t-L_R}^{(\mathrm{I})}; \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_R}^{(\mathrm{I})}}\right) + \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}$
$\mathbf{n}_t^{(\mathrm{I})} \to -\infty$
$\mathbf{n}_{t-L_R}^{(\mathrm{I})} \to -\infty$
$\Rightarrow \mathbf{o}_t^{(\mathrm{I})} \to \mathbf{s}_t^{(\mathrm{I})}$
$\Rightarrow \mathbf{o}_{t-L_R}^{(\mathrm{I})} \to \mathbf{s}_{t-L_R}^{(\mathrm{I})}$
$\Rightarrow \mathbf{w}_{o_t,L_R}^{(\mathrm{I,R})} \to \mathbf{0}$
$\Rightarrow \mathbf{r}_{t,L_R}^{(\mathrm{I,R})} \to \infty$
$\Rightarrow \xi\left(\mathbf{r}_{t,L_R}^{(\mathrm{I,R})}\right) \to \mathbf{1}$
$\Rightarrow \zeta\left(\mathbf{r}_{t,L_R}^{(\mathrm{I,R})}\right) \to \mathbf{0}$
$\Rightarrow \mathbf{v}_{o_t,L_R}^{(\mathrm{I,R})} \to \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}$
A Priori Model for Estimation Error
$p_{\breve{\mathbf{w}}_{o_t,L_R}^{(\mathrm{I,R})}|\cdots}$
A Priori Model for Phase Factor
$p_{\breve{\boldsymbol{\alpha}}_t}$
A Priori Model for Observation Error
$p_{\breve{\mathbf{v}}_{s_t,L_R}^{(\mathrm{I,R})}}$
Presence of Reverberation and Noise
$\mathbf{o}_t^{(\mathrm{I})} = \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{(\mathrm{I,R})}} + \mathrm{e}^{\mathbf{n}_t^{(\mathrm{I})}}\right) + \mathbf{v}_{o_t,L_R}^{(\mathrm{I,R})}$
$= f_{o,L_R}^{(\mathrm{I,R})}\left(\mathbf{x}_{t-L_R+1:t}^{(\mathrm{I})}, \mathbf{n}_{t-L_R,t}^{(\mathrm{I})}, \mathbf{o}_{t-L_R}^{(\mathrm{I})}; \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_R}^{(\mathrm{I})}}\right) + \mathbf{v}_{o_t,L_R}^{(\mathrm{I,R})}$
$\mathbf{v}_{o_t,L_R}^{(\mathrm{I,R})} = \ln\left(1 + \left(\mathrm{e}^{\mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}+\mathbf{w}_{o_t,L_R}^{(\mathrm{I,R})}} - 1\right) \circ \xi\left(\mathbf{r}_{t,L_R}^{(\mathrm{I,R})}\right) + 2\boldsymbol{\alpha}_t \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}+\mathbf{w}_{o_t,L_R}^{(\mathrm{I,R})}}{2}} \circ \zeta\left(\mathbf{r}_{t,L_R}^{(\mathrm{I,R})}\right)\right)$
$\mathbf{r}_{t,L_R}^{(\mathrm{I,R})} = \frac{10}{\ln(10)}\left[\ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}} + \mathrm{e}^{-\frac{2L_RB}{\tau_h}} \mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{(\mathrm{I,R})}}\right) - \mathbf{n}_t^{(\mathrm{I})}\right]$
$\mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{(\mathrm{I,R})}} = \left(2\boldsymbol{\sigma}_{\breve{\alpha}}^2 - \mathbf{1}\right) \circ \min\left\{\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathrm{I})}},\ \mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathrm{I})}}\right\} + \max\left\{\mathrm{e}^{\mathbf{n}_{t-L_R}^{(\mathrm{I})}},\ \mathrm{e}^{\mathbf{o}_{t-L_R}^{(\mathrm{I})}}\right\}$

Figure 4.12: *Overview of the recursive observation models: the AIR representation in the LMPSC domain has been replaced by its expected value under the AIR model (4.147). Note that the a priori model for the observation error $\breve{\mathbf{w}}_{o_t,L_R}^{(I,R)}$ approximately turns into a* Dirac*-delta distribution as L_R becomes large.*

and the *noisy* case (see Ch. 5 for more details on this and all other databases employed in this work). With the individual MPSC feature vectors of the reverberant signal, the noise and the noisy reverberant signal assumed to be available (4.100) may then be employed to obtain realizations $\boldsymbol{\alpha}_t$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$. These realizations will further be assumed to stem from a vector-valued, stationary and ergodic process whose RVs are mutually independent and identically distributed.

4.6.1 General Properties

The phase factor $\alpha_t(q)$ at time instant t and mel frequency bin index q given by (4.101) may also be written as

$$\alpha_t(q) = \mathrm{Re}\left\{ \frac{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, S_t(k)\, N_t^*(k)}{\sqrt{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, S_t(k)\, S_t^*(k)}\sqrt{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, N_t(k)\, N_t^*(k)}} \right\} \tag{4.217}$$

and may be interpreted as the real part of a (mel) weighted sample correlation coefficient.

Defining the complex-valued column vectors $\mathbf{S}_{t,q}, \mathbf{N}_{t,q} \in \mathbb{C}^{\left(K_q^{(\mathsf{up})}-K_q^{(\mathsf{low})}+1\right)\times 1}$ and the real-valued diagonal matrix $\boldsymbol{\Lambda}_q^{\frac{1}{2}} \in \mathbb{R}^{\left(K_q^{(\mathsf{up})}-K_q^{(\mathsf{low})}+1\right)\times\left(K_q^{(\mathsf{up})}-K_q^{(\mathsf{low})}+1\right)}$ as

$$\mathbf{S}_{t,q} := \left[S_t\left(K_q^{(\mathsf{low})}\right), \ldots, S_t\left(K_q^{(\mathsf{up})}\right)\right]^\dagger, \tag{4.218}$$

$$\mathbf{N}_{t,q} := \left[N_t\left(K_q^{(\mathsf{low})}\right), \ldots, N_t\left(K_q^{(\mathsf{up})}\right)\right]^\dagger, \tag{4.219}$$

$$\boldsymbol{\Lambda}_q^{\frac{1}{2}} := \mathrm{diag}\left(\left[\sqrt{\Lambda_q\left(K_q^{(\mathsf{low})}\right)}, \ldots, \sqrt{\Lambda_q\left(K_q^{(\mathsf{up})}\right)}\right]\right) \tag{4.220}$$

allows the phase factor (4.217) to also be written as the real part of the inner product of two complex-valued vectors of unit length, e.g.,

$$\alpha_t(q) = \mathrm{Re}\left\{ \frac{\left(\boldsymbol{\Lambda}_q^{\frac{1}{2}}\mathbf{S}_{t,q}\right)^H}{\left|\boldsymbol{\Lambda}_q^{\frac{1}{2}}\mathbf{S}_{t,q}\right|} \frac{\boldsymbol{\Lambda}_q^{\frac{1}{2}}\mathbf{N}_{t,q}}{\left|\boldsymbol{\Lambda}_q^{\frac{1}{2}}\mathbf{N}_{t,q}\right|} \right\}. \tag{4.221}$$

However, the results of the inner product of two complex-valued vectors of unit length always lies on the complex unit circle. Consequently, the elements $\alpha_t(q)$ of the vector of phase factor $\boldsymbol{\alpha}_t$ are bounded by

$$-1 \leq \alpha_t(q) \leq +1 \tag{4.222}$$

and the vector itself lies within the $\mathbb{R}^Q$ hypercube of edge length 2 centered at the origin, e.g.,

$$\boldsymbol{\alpha}_t \in [-1, +1]^Q. \tag{4.223}$$

Hence, the PDF of the vector of phase factors is non-zero only within this $\mathbb{R}^Q$ hypercube.

Looking at (4.221) again also reveals that the phase factor is not only insensitive to the global broadband ***r**everberant-to-**n**oise **r**atio* (RNR) the reverberant speech signal and the noise signal mix at, but also insensitive to the frame specific, i.e., local broadband RNR[6]. The normalization of the vectors $\mathbf{S}_{t,q}$ and $\mathbf{N}_{t,q}$ in (4.221) to unit length compensates for any (constant) scaling of the two involved signals.

Assuming the PSCs of the reverberant speech signal and that of the noise to be approximately constant for those frequency bins covered by a particular mel filter even removes the sensitivity of the phase factor to the power of the involved signals at all. Looking at (4.101), this approximation just leaves the contribution of the phase differences $\varphi_{S_t(k),N_t(k)}$ to the phase factors which then approximates to

$$\alpha_t(q) \approx \sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} c_q(k) \cos\left(\varphi_{S_t(k),N_t(k)}\right), \tag{4.224}$$

where

$$c_q(k) := \frac{\Lambda_q(k)}{\sum\limits_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} \Lambda_q(k)}, \quad k \in \left\{K_q^{(\text{low})}, \ldots, K_q^{(\text{up})}\right\}, \tag{4.225}$$

are the normalized coefficients of the q-th mel filter.

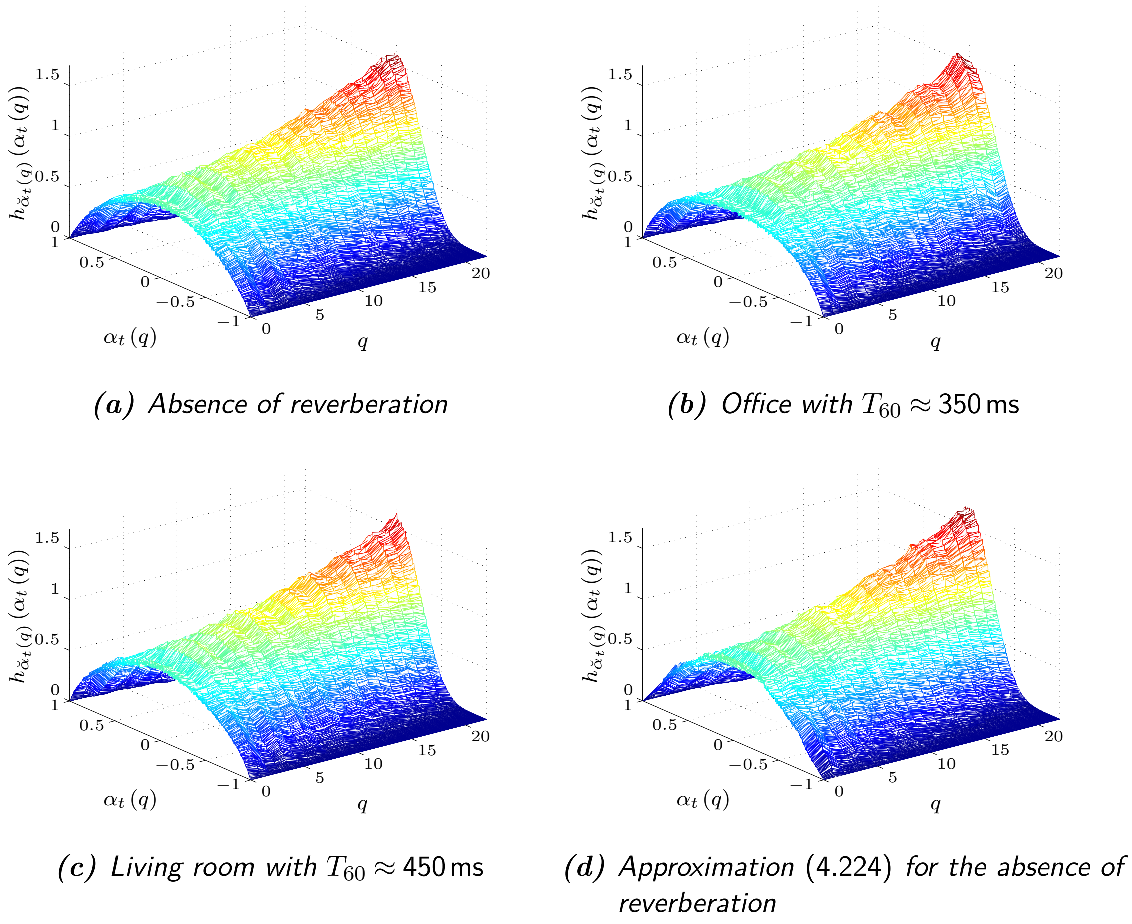
Approximation (4.224) is also on the outset of (4.115) (and in an analogous way of (4.139)). Thus, although the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ formally depends on $\breve{\mathbf{n}}_t^{(\text{l})}$ and via (4.92) on $\breve{\mathbf{x}}_{t-L_H:t}^{(\text{l})}$ and $\breve{\mathbf{v}}_{s_t}^{(\text{l})}$, it is reasonable to assume $\breve{\boldsymbol{\alpha}}_t$ to be independent of these variates.

4.6.2 Empirical Distribution

Figure 4.13 shows the histograms $h_{\breve{\alpha}_t(q)}(\alpha_t(q))$ of the phase factors $\breve{\alpha}_t(q), q \in \{0, \ldots, Q-1\}$ at a global broadband RNR of 10 dB for the absence of reverberation (Subfig. 4.13a), the presence of reverberation as typical for a small-sized "office" with $T_{60} \approx$ 350 ms (Subfig. 4.13b) and the presence of reverberation as typical for a mid-sized "living room" with $T_{60} \approx$ 450 ms (Subfig. 4.13c).

Finally, Subfig. 4.13d shows the histograms $h_{\breve{\alpha}_t(q)}(\alpha_t(q))$ in the absence of reverberation if the phase factors are computed according to approximation (4.224), e.g., by just taking the phase differences $\varphi_{S_t(k),N_t(k)}$ of the complex valued STDFT coefficients of the reverberant speech signal and that of the noise signal into account. Since the histograms in Subfigs. 4.13a-4.13c only slightly differ (in particular only visible at low and high mel indices), it may be concluded here that the marginal distributions of the phase factors are approximately identical, irrespective of the presence and extent of reverberation. The marginal distributions of the phase factors have further been found to be independent of the type of noise [83].

[6] The global broadband RNR is defined as the ratio of the average power of the reverberant speech signal to that of the noise signal. The local broadband RNR averages the powers of the signals only within an analysis window.



(a) Absence of reverberation

(b) Office with $T_{60} \approx 350\,\mathrm{ms}$

(c) Living room with $T_{60} \approx 450\,\mathrm{ms}$

(d) Approximation (4.224) for the absence of reverberation

Figure 4.13: *Histograms $h_{\breve{\alpha}_t(q)}(\alpha_t(q))$ of the phase factors $\breve{\alpha}_t(q), q \in \{0, \ldots, Q-1\}$ at a global broadband RNR of 10 dB for the absence of reverberation (a), the presence of reverberation at $T_{60} \approx 350\,\mathrm{ms}$ (b) and at $T_{60} \approx 450\,\mathrm{ms}$ (c) and the histograms of the phase factors if computed according to approximation (4.224) in the absence of reverberation (d).*

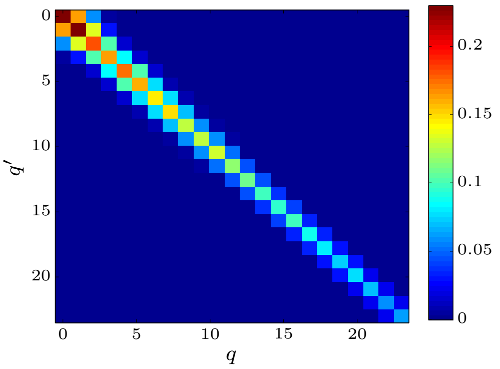


Figure 4.14: *Sample covariance matrix* $\hat{\mathbf{\Sigma}}_{\breve{\alpha}}$ *of the vector of phase factors* $\breve{\boldsymbol{\alpha}}_t$ *at a reverberation time of* $T_{60} \approx 350\,\mathrm{ms}$ *at a global broadband RNR of* $10\,\mathrm{dB}$.

Moreover, a comparison of Subfigs. 4.13a-4.13c with Subfig. 4.13d shows the approximation (4.224) to be fairly decent.

In general, the histograms of the phase factors approach a Gaussian-like shape with increasing mel indices but are clearly non-Gaussian at lower mel indices. Thus, even though the marginal distribution is sometimes modeled as a zero-mean Gaussian in literature [79, 84], this approximation can be considered to be approximately valid only for high mel frequency bin indices. And although the the zero-mean assumption can be found to hold for all mel frequency bin indices, the distribution will, due to the limited range of the phase factors $\breve{\alpha}_t(q)$, formally never be Gaussian.

Thus far, only the marginal PDFs $p_{\breve{\alpha}_t(q)}(\alpha_t(q))$ of the phase factors $\breve{\alpha}_t(q)$ have been considered in terms of their histograms. Since the PDF $p_{\breve{\boldsymbol{\alpha}}_t}(\boldsymbol{\alpha}_t)$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ is not easily accessible in an equivalent manner, a closer look is taken at the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\breve{\alpha}}$ displayed in Fig. 4.14 for the presence of reverberation at $T_{60} \approx 350\,\mathrm{ms}$. From Fig. 4.14 the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\breve{\alpha}}$ can be found to be dominated by its main diagonal entries. Substantial correlations between phase factors $\breve{\alpha}_t(q)$ and $\breve{\alpha}_t(q')$, which are strictly positive, are majorly limited to a few adjacent mel frequency bin indices and rapidly decrease with increasing $|q - q'|$ and also with increasing q. Besides the correlation inherent to the involved reverberant speech signal and that of the noise, this correlation structure may majorly be attributed to the overlap of adjacent mel filters. For a simplified modeling, the phase factors $\breve{\alpha}_t(q)$ at different mel frequency indices may, however, be assumed to be uncorrelated.

4.6.3 Parametric Approximation to its Distribution

As outlined in Sec. 4.6.2, a common approximation to the distribution of the phase factors $\breve{\alpha}_t(q)$ is to assume them to be Gaussian distributed. By further assuming them to be jointly Gaussian, the observation made in the previous section, i.e., that components of the vector of phase factors are approximately uncorrelated, may then motivate the assumption of their statistically independence resulting in the joint PDFs $p_{\breve{\boldsymbol{\alpha}}_t}(\boldsymbol{\alpha}_t)$ turning into the product of the marginal PDFs $p_{\breve{\alpha}_t(q)}(\alpha_t(q))$.

For reasons also explained in Sec. 4.6.2, the Gaussian approximation to the PDFs of the phase factors $\breve{\alpha}_t(q)$ is, however, by no means very accurate. A more accurate approximation may be derived from approximation (4.224). The only RVs involved in (4.224) are the phase differences $\varphi_{S_t(k),N_t(k)}$ of the complex-valued STDFT coefficients of the reverberant speech signal and that of the noise. These phase differences may now be assumed to be realizations of i.i.d. uniformly distributed RVs $\breve{\varphi}_{S_t(k),N_t(k)}$ with

$$p_{\breve{\varphi}_{S_t(k),N_t(k)}}\left(\varphi_{S_t(k),N_t(k)}\right) = \begin{cases} \frac{1}{2\pi} & \text{, if } -\pi < \varphi_{S_t(k),N_t(k)} \leq \pi \\ 0 & \text{, else} \end{cases}. \tag{4.226}$$

The distribution of the mel weighted cosine of $\breve{\varphi}_{S_t(k),N_t(k)}$, i.e.,

$$\breve{\nu}_{t,q}(k) := c_q(k) \cos\left(\breve{\varphi}_{S_t(k),N_t(k)}\right) \tag{4.227}$$

may then be obtained by application of the transformation rule [85, p. 201, Eq. (6-115)] as

$$p_{\breve{\nu}_{t,q}(k)}\left(\nu_{t,q}(k)\right) = \begin{cases} \frac{1}{\pi}\frac{1}{\sqrt{c_q(k)^2-\nu_{t,q}^2(k)}} & \text{, if } -c_q(k) \leq \nu_{t,q}(k) \leq c_q(k) \\ 0 & \text{, else} \end{cases}. \tag{4.228}$$

Note that the $\breve{\nu}_{t,q}(k)$ are still independent but no longer identically distributed. The distribution of the sum of independent RVs, as given by (4.224), is now given by the convolution of the individual RVs' distributions [85, p. 182, Eq. (6-43)], i.e.,

$$p_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right) \approx \left(p_{\breve{\nu}_{t,q}\left(K_q^{(\text{low})}\right)} * \ldots * p_{\breve{\nu}_{t,q}\left(K_q^{(\text{up})}\right)}\right)\left(\alpha_t(q)\right), \tag{4.229}$$

for which, unfortunately, no tractable parametric form can be specified. Such a form may however be found, if a transformation

$$\breve{\gamma}_t(q) := g\left(\breve{\alpha}_t(q)\right) \tag{4.230}$$

of the RV $\breve{\alpha}_t(q)$ can be specified that turns the PDF $p_{\breve{\gamma}_t(q)}$ of the transformed RV $\breve{\gamma}_t(q)$ into a PDF with a known parametric form.

Denoting the **c**umulativ **d**istribution **f**unctions (CDFs) of the RVs $\breve{\alpha}_t(q)$ and $\breve{\gamma}_t(q)$ by

$$F_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right) := P_{\breve{\alpha}_t(q)}\left(\breve{\alpha}_t(q) \leq \alpha_t(q)\right) = \int\limits_{-\infty}^{\alpha_t(q)} p_{\breve{\alpha}_t(q)}(\tau)\,\mathrm{d}\tau \tag{4.231}$$

$$F_{\breve{\gamma}_t(q)}\left(\gamma_t(q)\right) := P_{\breve{\gamma}_t(q)}\left(\breve{\gamma}_t(q) \leq \gamma_t(q)\right) = \int\limits_{-\infty}^{\gamma_t(q)} p_{\breve{\gamma}_t(q)}(\tau)\,\mathrm{d}\tau, \tag{4.232}$$

respectively, this transformation is given by [86, p. 11, Theorem 4.1]

$$g\left(\breve{\alpha}_t(q)\right) = F_{\breve{\gamma}_t(q)}^{-1}\left(F_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right)\right), \tag{4.233}$$

where $F_{\breve{\gamma}_t(q)}^{-1}(\cdot)$ denotes the inverse function of the CDF $F_{\breve{\gamma}_t(q)}(\cdot)$. Equivalently, the inverse transformation is given by

$$g^{-1}\left(\breve{\gamma}_t(q)\right) = F_{\breve{\alpha}_t(q)}^{-1}\left(F_{\breve{\gamma}_t(q)}\left(\gamma_t(q)\right)\right). \tag{4.234}$$

Targeting a GAUSSIAN distribution for the RV $\breve{\gamma}_t(q)$ with $E[\breve{\gamma}_t(q)] = 0$ and $E\left[\breve{\gamma}_t^2(q)\right] = 1$, for which the CDF is given by

$$F_{\breve{\gamma}_t(q)}(\gamma_t(q)) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\gamma_t(q)}{\sqrt{2}}\right)\right), \tag{4.235}$$

with $\operatorname{erf}(\cdot)$ denoting the error function, this inverse transformation may be approximated by a *scaled* error function, i.e.,

$$g^{-1}(\breve{\gamma}_t(q)) \approx \operatorname{erf}\left(\sigma_{\breve{\gamma}_q}\breve{\gamma}_t(q)\right) \tag{4.236}$$

with scaling factor $\sigma_{\breve{\gamma}_q}$. This approximate identity is illustrated in Fig. 4.15, where the CDFs of the phase factors $\breve{\alpha}_t(q)$ at different mel frequency bin indices (Subfig. 4.15a), the CDF of a standard normally distributed RV $\breve{\gamma}_t(q)$ (Subfig. 4.15b) and the resulting inverse transformation functions $g^{-1}(\breve{\gamma}_t(q))$ (Subfig. 4.15c) are displayed. That the inverse transformation functions $g^{-1}(\breve{\gamma}_t(q))$ are indeed *scaled* error functions is further illustrated in Subfig. 4.15d, where the scale factors $\sigma_{\breve{\gamma}_q}$, which can be deduced from the slope of $g^{-1}(\gamma_t(q))$ at $\gamma_t(q) = 0$ as

$$\left.\frac{\mathrm{d}\, g^{-1}(\gamma_t(q))}{\mathrm{d}\, \gamma_t(q)}\right|_{\gamma_t(q)=0} \overset{!}{=} \left.\frac{\mathrm{d}\, \operatorname{erf}\left(\sigma_{\breve{\gamma}_q}\gamma_t(q)\right)}{\mathrm{d}\, \gamma_t(q)}\right|_{\gamma_t(q)=0} \tag{4.237}$$

$$= \sigma_{\breve{\gamma}_q}\frac{2}{\sqrt{\pi}}\left.\mathrm{e}^{-\left(\sigma^2_{\breve{\gamma}_q}\gamma_t(q)\right)^2}\right|_{\gamma_t(q)=0} \tag{4.238}$$

$$= \sigma_{\breve{\gamma}_q}\frac{2}{\sqrt{\pi}}, \tag{4.239}$$

are compensated for. The displayed *scale-normalized* inverse transformation function is thereby defined as

$$\tilde{g}^{-1}(\breve{\gamma}_t(q)) := \operatorname{erf}\left(\frac{\operatorname{erf}^{-1}\left(g^{-1}(\breve{\gamma}_t(q))\right)}{\sigma_{\breve{\gamma}_q}}\right), \tag{4.240}$$

where $\operatorname{erf}^{-1}(\cdot)$ is the inverse error function, and can be found to match the error function (adumbrated by the black crosses in Subfig. 4.15d) pretty well. The scaling factor $\sigma_{\breve{\gamma}_q}$ may alternatively also be considered to first transform the zero-mean standard normally distributed RV $\breve{\gamma}_t(q)$ into a zero-mean normally distributed RV $\breve{\gamma}_t(q)$ with variance $E\left[\breve{\gamma}_t^2(q)\right] = \sigma^2_{\breve{\gamma}_q}$.

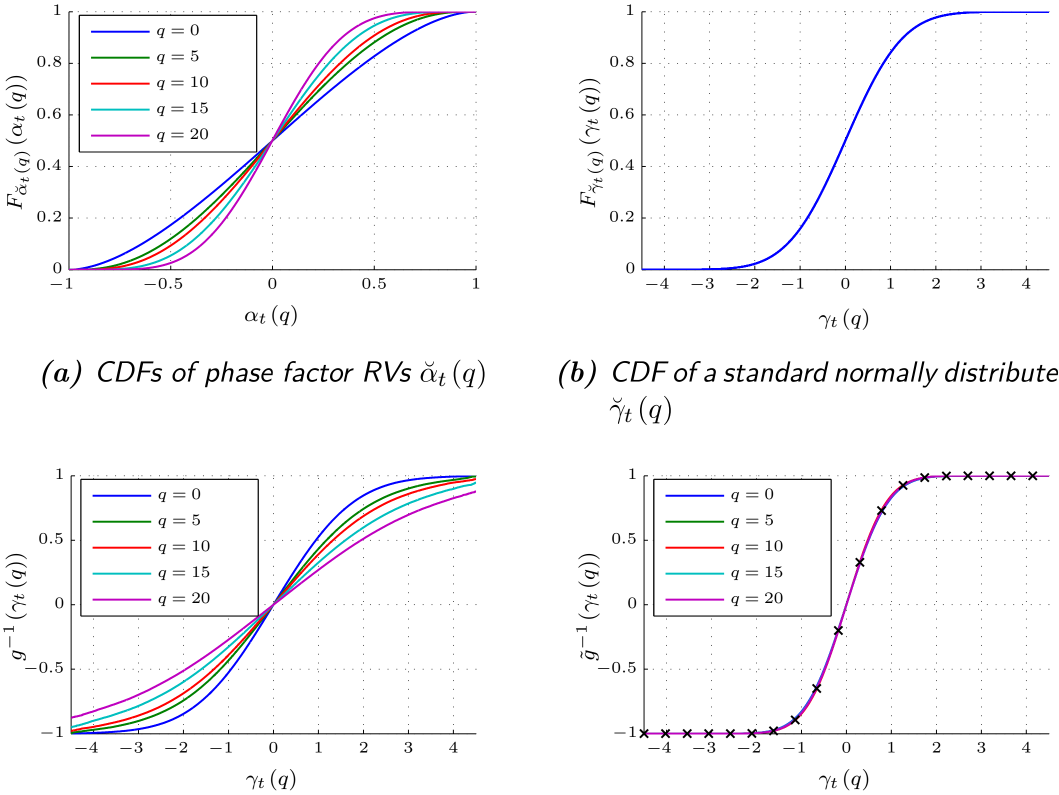
Assuming the RVs $\breve{\gamma}_t(q)$ making up the vector $\breve{\boldsymbol{\gamma}}_t$ to be jointly GAUSSIAN distributed it may then be concluded that the transformation

$$\breve{\boldsymbol{\gamma}}_t = \operatorname{erf}^{-1}(\breve{\boldsymbol{\alpha}}_t) \tag{4.241}$$

transforms the RV $\breve{\boldsymbol{\alpha}}_t$ with PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ into the RV $\breve{\boldsymbol{\gamma}}_t$ with PDF $p_{\breve{\boldsymbol{\gamma}}_t}$ given by[7]

$$p_{\breve{\boldsymbol{\gamma}}_t}(\boldsymbol{\gamma}_t) = \mathcal{N}\left(\boldsymbol{\gamma}_t; \mathbf{0}, \boldsymbol{\Sigma}_{\breve{\gamma}}\right). \tag{4.242}$$

[7] At this point, it should be mentioned that the found transformation bears close resemblance to FISHER's



(a) *CDFs of phase factor RVs $\breve{\alpha}_t(q)$*

(b) *CDF of a standard normally distributed RV $\breve{\gamma}_t(q)$*

(c) *Transformation functions $g^{-1}(\breve{\gamma}_t(q))$*

(d) *Scale-normalized transformation functions $\tilde{g}^{-1}(\breve{\gamma}_t(q))$.*

Figure 4.15: *CDFs of the phase factors $\breve{\alpha}_t(q)$ at different mel frequency bin indices (a) at the presence of reverberation at $T_{60} \approx 350\,\text{ms}$ and a global broadband RNR of $10\,\text{dB}$, the CDF of a standard normally distributed RV $\breve{\gamma}_t(q)$ (b) and the resulting inverse transformation functions $g^{-1}(\gamma_t(q))$ (c) transforming the standard normally distributed RVs $\breve{\gamma}_t(q)$ into RVs with PDF $p_{\breve{\alpha}_t(q)}$. The scale-normalized inverse transformation functions $\tilde{g}^{-1}(\gamma_t(q))$ in (d) match the error function $\operatorname{erf}(\gamma_t(q))$, which is indicated by the black crosses.*

The covariance matrix $\mathbf{\Sigma}_{\breve{\gamma}}$ of the RV $\breve{\boldsymbol{\gamma}}_t$ may now elegantly be determined from the covariance matrix $\mathbf{\Sigma}_{\breve{\alpha}}$ of the RV $\breve{\boldsymbol{\alpha}}_t$ by requiring the transformation of the RV $\breve{\boldsymbol{\gamma}}_t$ by the inverse transformation of (4.241), given by

$$\breve{\boldsymbol{\alpha}}_t = \operatorname{erf}(\breve{\boldsymbol{\gamma}}_t), \tag{4.243}$$

z-transformation [87], i.e., $z = \tanh^{-1}(r)$, where $\tanh^{-1}(\cdot)$ denotes the inverse hyperbolic tangent, which arises as a *variance stabilization* transformation in the context of the sample correlation coefficient r. If the sample correlation coefficient is computed from realizations of correlated, i.i.d. bivariate normally distributed RVs, the distribution of the sample correlation coefficient may be specified in closed form [88, p. 219, Eq. (8.67)]. Though the phase factor RV $\breve{\alpha}_t(q)$ may be considered as the real part of a complex-valued weighted sample correlation coefficient, the weighting, however, renders the considerations made inapplicable to the problem at hand.

to match the covariance $\boldsymbol{\Sigma}_{\breve{\alpha}}$ of the (zero-mean) RV $\breve{\boldsymbol{\alpha}}_t$, i.e.,

$$\boldsymbol{\Sigma}_{\breve{\alpha}} \overset{!}{=} \int\limits_{\mathbb{R}^Q} \operatorname{erf}(\breve{\boldsymbol{\gamma}}_t)\left(\operatorname{erf}(\breve{\boldsymbol{\gamma}}_t)\right)^{\dagger} p_{\breve{\boldsymbol{\gamma}}_t}(\boldsymbol{\gamma}_t)\,\mathrm{d}\boldsymbol{\gamma}_t \tag{4.244}$$

$$= \int\limits_{\mathbb{R}^Q} \operatorname{erf}(\breve{\boldsymbol{\gamma}}_t)\left(\operatorname{erf}(\breve{\boldsymbol{\gamma}}_t)\right)^{\dagger} \mathcal{N}\left(\boldsymbol{\gamma}_t;\mathbf{0},\boldsymbol{\Sigma}_{\breve{\gamma}}\right)\mathrm{d}\boldsymbol{\gamma}_t. \tag{4.245}$$

A closed-form solution to the required integrals is given in Appendix A.10, where the diagonal and off-diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$, denoted by $\sigma^2_{\breve{\gamma}_q}$ and $\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}$ respectively, are given in terms of the diagonal and off-diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$, denoted by $\sigma^2_{\breve{\alpha}_q}$ and $\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}$ respectively, as

$$\sigma^2_{\breve{\gamma}_q} = \frac{\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2},\ \forall q \in \{0,\dots Q-1\} \tag{4.246}$$

$$\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \frac{1}{2}\tan\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)\sqrt{\frac{\left(1+2\sigma^2_{\breve{\gamma}_q}\right)\left(1+2\sigma^2_{\breve{\gamma}_{q'}}\right)}{1+\tan^2\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)}},\ \forall q,q' \in \left\{0,\dots,Q-1 | q \neq q'\right\}. \tag{4.247}$$

Plugging (4.246) into (4.247) then results in the compact formulation

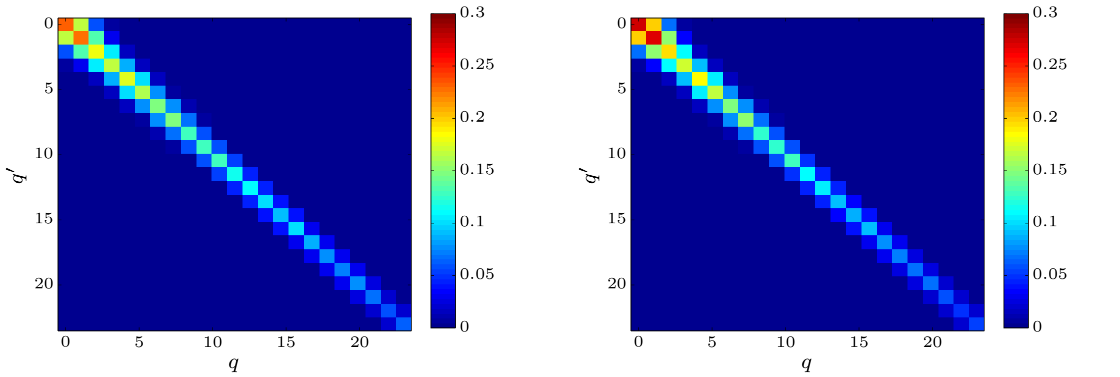
$$\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \frac{\tan\left(\frac{\pi}{4}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)}{1+\tan^2\left(\frac{\pi}{4}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)} \cdot \frac{\sqrt{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}}{1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)} \cdot \frac{\sqrt{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_{q'}}\right)}}{1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_{q'}}\right)} \tag{4.248}$$

which is now applicable $\forall q,q' \in \{0,\dots,Q-1\}$ and where $\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}} = \sigma^2_{\breve{\alpha}_q}$ and $\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \sigma^2_{\breve{\gamma}_q}$ for $q = q'$.[8] The resulting covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$ and the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\breve{\alpha}}$ from which it is computed are displayed in Fig. 4.16. Note that the same scale has been chosen for Subfig. 4.16a ($\hat{\boldsymbol{\Sigma}}_{\breve{\alpha}}$) and Subfig. 4.16b ($\boldsymbol{\Sigma}_{\breve{\gamma}}$) to ease a comparison between the two covariance matrices.

With the (diagonal) JACOBIAN matrix of $\operatorname{erf}(\breve{\boldsymbol{\gamma}}_t)$ given by

$$J_{\operatorname{erf}(\breve{\boldsymbol{\gamma}}_t),\boldsymbol{\gamma}_t} = \operatorname{diag}\left(\left[\tfrac{2}{\sqrt{\pi}}\mathrm{e}^{-\gamma_t^2(0)} \quad \cdots \quad \tfrac{2}{\sqrt{\pi}}\mathrm{e}^{-\gamma_t^2(Q-1)}\right]\right) \tag{4.249}$$

[8] Equivalently, the elements of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$ may be computed from those of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$ by applying (A.245) and (A.277) given in Appendix A.10.



(a) *Sample covariance matrix* $\hat{\boldsymbol{\Sigma}}_{\breve{\alpha}}$ *of RV* $\breve{\boldsymbol{\alpha}}_t$

(b) *Covariance matrix* $\boldsymbol{\Sigma}_{\breve{\gamma}}$ *of RV* $\breve{\boldsymbol{\gamma}}_t$ *according to* (4.248)

Figure 4.16: *Sample covariance matrix* $\hat{\boldsymbol{\Sigma}}_{\breve{\alpha}}$ *of the vector of phase factors* $\breve{\boldsymbol{\alpha}}_t$ *at a reverberation time of* $T_{60} \approx 350\,\mathrm{ms}$ *and a global broadband RNR of* 10 dB *(a) and the covariance matrix* $\boldsymbol{\Sigma}_{\breve{\gamma}}$ *of the transformed vector of phase factors* $\breve{\boldsymbol{\gamma}}_t$ *computed from* $\boldsymbol{\Sigma}_{\breve{\alpha}}$ *according to* (4.248)*.*

the PDF $p_{\breve{\boldsymbol{\alpha}}_t}$ of the RV $\breve{\boldsymbol{\alpha}}_t$ may then eventually be approximated by [85, p. 244, Eq. (7-8)]

$$p_{\breve{\boldsymbol{\alpha}}_t}(\boldsymbol{\alpha}_t) \approx \frac{p_{\breve{\boldsymbol{\gamma}}_t}\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)\right)}{\left|J_{\mathrm{erf}(\breve{\boldsymbol{\gamma}}_t),\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)}\right|} \tag{4.250}$$

$$= \left(\prod_{q=0}^{Q-1} \frac{1}{\frac{2}{\sqrt{\pi}}\mathrm{e}^{-\left(\mathrm{erf}^{-1}(\alpha_t(q))\right)^2}}\right) \mathcal{N}\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t);\mathbf{0},\boldsymbol{\Sigma}_{\breve{\gamma}}\right) \tag{4.251}$$

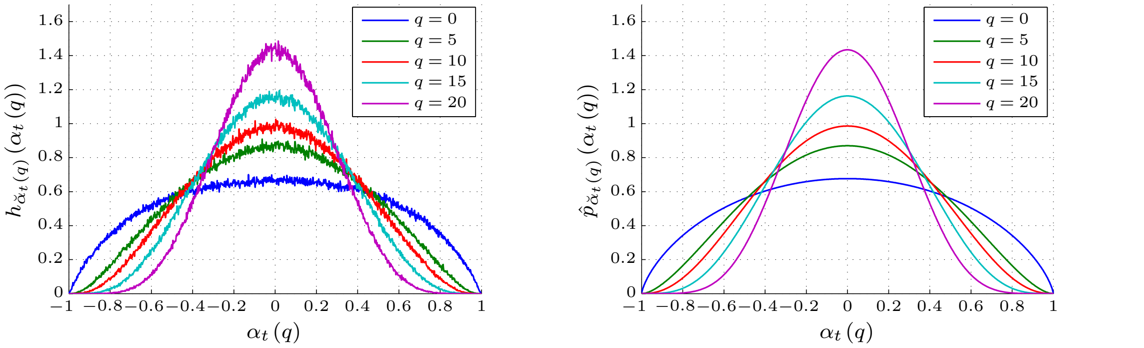
$$= \left(\prod_{q=0}^{Q-1} \frac{1}{\frac{2}{\sqrt{\pi}}\mathrm{e}^{-\left(\mathrm{erf}^{-1}(\alpha_t(q))\right)^2}}\right) \frac{1}{\sqrt{(2\pi)^Q\left|\boldsymbol{\Sigma}_{\breve{\gamma}}\right|}}\mathrm{e}^{-\frac{1}{2}\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)\right)^{\dagger}\boldsymbol{\Sigma}_{\breve{\gamma}}{}^{-1}\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)\right)} \tag{4.252}$$

$$= \frac{1}{\sqrt{8^Q\left|\boldsymbol{\Sigma}_{\breve{\gamma}}\right|}}\mathrm{e}^{-\frac{1}{2}\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\breve{\gamma}}{}^{-1}-2\mathbf{I}_{Q\times Q}\right)\left(\mathrm{erf}^{-1}(\boldsymbol{\alpha}_t)\right)}. \tag{4.253}$$

The marginal distributions are then given by

$$p_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right) \approx \frac{1}{\sqrt{8}\sigma_{\breve{\gamma}_q}}\mathrm{e}^{-\frac{1-2\sigma_{\breve{\gamma}_q}^2}{2\sigma_{\breve{\gamma}_q}^2}\left(\mathrm{erf}^{-1}(\alpha_t(q))\right)^2} \tag{4.254}$$

and displayed in Fig. 4.17 together with the histograms computed from the samples of the vector of phase factors at a reverberation time of $T_{60} \approx 350\,\mathrm{ms}$. The difference between the *true* distribution (Subfig. 4.17a) and the approximated one (Subfig. 4.17b) can be found to be almost imperceptible. This observation is also supported by the **K**ULBACK-**L**EIBLER



(a) *Histograms of the phase factors*

(b) *Approximation of PDFs according to* (4.254)

Figure 4.17: *Histogram approximation (a) to the PDFs of the phase factors at the presence of reverberation at* $T_{60} \approx 350\,\mathrm{ms}$ *and a global broadband RNR of* 10 dB *and the approximation by a transformed* GAUSSIAN *(b) according to* (4.254)*.*

divergence (KL divergence)

$$D_{\mathsf{KL}}\left(p_{\breve{\alpha}_t(q)} \middle\| \hat{p}_{\breve{\alpha}_t(q)}\right) := \int\limits_{[-1,+1]^Q} p_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right) \ln\left(\frac{p_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right)}{\hat{p}_{\breve{\alpha}_t(q)}\left(\alpha_t(q)\right)}\right) \mathrm{d}\alpha_t(q) \tag{4.255}$$

between the *true* distribution $p_{\breve{\alpha}_t(q)}$ and its approximation $\hat{p}_{\breve{\alpha}_t(q)}$.

The KL divergence $D_{\mathsf{KL}}\left(p_{\breve{\alpha}_t(q)} \middle\| \hat{p}_{\breve{\alpha}_t(q)}\right)$, displayed in Fig. 4.18, where the true distribution $p_{\breve{\alpha}_t(q)}$ has been approximated by the histogram $h_{\breve{\alpha}_t(q)}$, is very small for all mel frequency bin indices q and further decreases with increasing q. The approximation of the PDFs of the phase factors by the transformed GAUSSIAN thus becomes more accurate with increasing mel frequency bin index q. Moreover, the properties of the true distribution of being symmetric around $\alpha_t(q) = 0$ and non-zero only for $\alpha_t(q) \in \{-1, \ldots +1\}$ are preserved by the approximation.

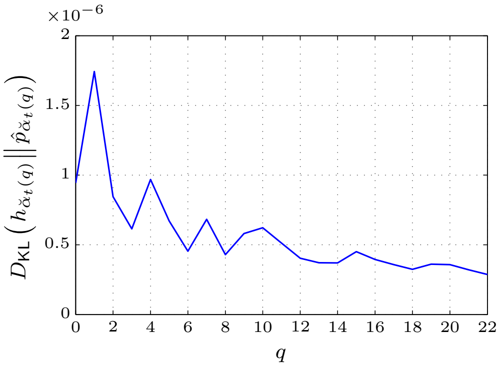


Figure 4.18: *KL divergence between the histogram* $h_{\breve{\alpha}_t(q)}$ *and the transformed* GAUSSIAN *approximation* $\hat{p}_{\breve{\alpha}_t(q)}$ *for mel frequency bin indices* $q \in \{0, \ldots, Q-1\}$*.*

4.6.4 Analytic Solution to its Central Moments

The transformation (4.243) not only allows for finding the parametric model (4.253) approximating the distribution $p_{\breve{\boldsymbol{\alpha}}_t}$ of the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ but also yields a very convenient way to draw samples from it by first drawing samples from the multivariate GAUSSIAN (4.242) (for which efficient methods exist) and then applying (4.243). However, both sampling from and application of the model (4.253) require the covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$, which is linked with the covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$ of the vector of phase factors via (4.248), to be given. Though it may, as done previously, be obtained from stereo data in terms of a sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\breve{\alpha}}$, a completely analytic solution may be obtained from (4.229) if statistical independence of the phase factors is assumed.

Although the convolution of the marginal PDFs (4.228) of the mel weighted cosines $\breve{\nu}_{t,q}(k)$ could not be employed to find a tractable parametric form of the marginal PDF $p_{\breve{\alpha}_t(q)}$, the corresponding characteristic functions $\Phi_{\breve{\nu}_{t,q}(k)}(\tau)$, which are defined as

$$\Phi_{\breve{\nu}_{t,q}(k)}(\tau) := E\left[\mathrm{e}^{\,\mathrm{j}\tau\breve{\nu}_{t,q}(k)}\right] \tag{4.256}$$

$$= \int_{\mathbb{R}} p_{\breve{\nu}_{t,q}(k)}(\nu_{t,q}(k))\,\mathrm{e}^{\,\mathrm{j}\tau\nu_{t,q}(k)}\mathrm{d}\nu_{t,q}(k) \tag{4.257}$$

$$= 2\pi\mathcal{F}^{-1}\left\{p_{\breve{\nu}_{t,q}(k)}(\nu_{t,q}(k))\right\}, \tag{4.258}$$

with $\mathcal{F}^{-1}\{\cdot\}$ denoting the inverse FOURIER transformation operator, may i) be utilized to obtain the characteristic function $\Phi_{\breve{\alpha}_t(q)}(\tau)$ of the phase factor RV $\breve{\alpha}_t(q)$ and thus ii) be employed to express the moments of $\breve{\alpha}_t(q)$ in a purely analytic way.

Since the characteristic function of a RV formally equals the inverse FOURIER transform of the associated PDF, properties of the FOURIER transformation and tables of well-known FOURIER transforms may be employed to first find [89]

$$\Phi_{\breve{\nu}_{t,q}(k)}(\tau) = \mathrm{J}_0(c_q(k)\tau), \quad \forall\tau\in\mathbb{R}, \tag{4.259}$$

with $\mathrm{J}_0(\cdot)$ denoting the BESSEL function of order zero, and finally, since the RVs $\breve{\nu}_{t,q}(k)$ are assumed to be statistically independent,

$$\Phi_{\breve{\alpha}_t(q)}(\tau) = \prod_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Phi_{\breve{\nu}_{t,q}(k)}(\tau) \tag{4.260}$$

$$= \prod_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \mathrm{J}_0(c_q(k)\tau). \tag{4.261}$$

The *raw* moments of order n of the RV $\breve{\alpha}_t(q)$ may now be obtained by n-fold differentiation of (4.260) w.r.t. τ and evaluation of the result at $\tau=0$, i.e.,

$$E\left[\breve{\alpha}_t^n(q)\right] = (-\mathrm{j})^n \left.\frac{\mathrm{d}^n\Phi_{\alpha_t(q)}(\tau)}{\mathrm{d}\tau^n}\right|_{\tau=0}, \quad \forall n\in\mathbb{N}. \tag{4.262}$$

Following this approach, the derivations and proofs carried out in Appendix A.9 show these raw moments to be given by

$$E\left[\breve{\alpha}_t^{2m-1}(q)\right] = 0, \tag{4.263}$$

$$E\left[\breve{\alpha}_t^{2m}(q)\right] = \sum_{l=1}^{m} \binom{2m-1}{2l-1} \left(\sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} \left(\frac{c_q(k)}{2}\right)^{2l} \right) \varsigma_l E\left[\breve{\alpha}_t^{2(m-l)}(q)\right], \tag{4.264}$$

where $m \in \mathbb{N}_{>0}$ and where ς_l may recursively be computed by

$$\varsigma_l = \binom{2l}{l} - \sum_{i=1}^{l-1} \binom{2l-1}{2i-1} \binom{2(l-i)}{l-i} \varsigma_i. \tag{4.265}$$

While the zero-mean property of the RV $\breve{\alpha}_t(q)$ and the symmetry of its distribution w.r.t. $\alpha_t(q) = 0$ is reflected by (4.263) (stating that all moments of uneven order are zero), (4.264) provides a recursive way of obtaining all moments of even order and with

$$E\left[\breve{\alpha}_t^{0}(q)\right] = 1 \tag{4.266}$$

$$E\left[\breve{\alpha}_t^{2}(q)\right] = \sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} \left(\frac{c_q(k)}{2}\right)^{2} 2E\left[\breve{\alpha}_t^{0}(q)\right] = \frac{1}{2} \sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} c_q^2(k) \tag{4.267}$$

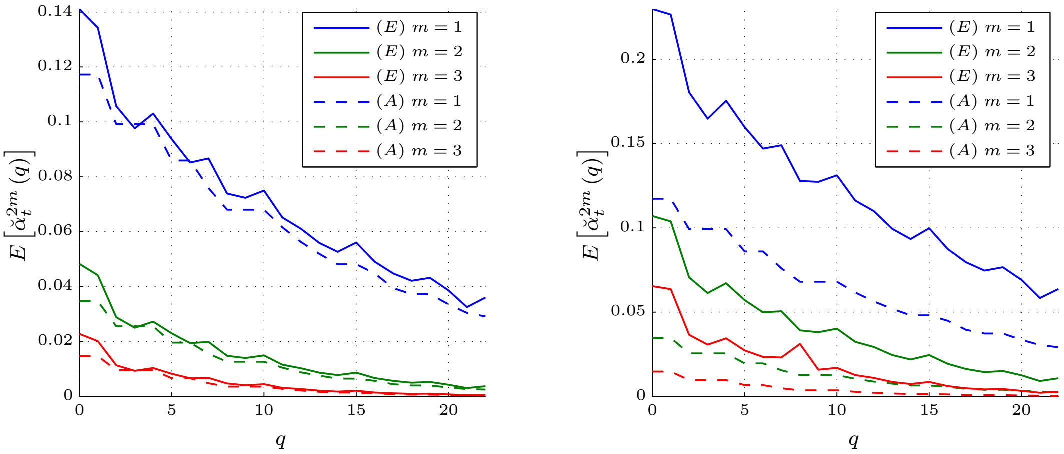
$$E\left[\breve{\alpha}_t^{4}(q)\right] = 3 \underbrace{\sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} \left(\frac{c_q(k)}{2}\right)^{2} 2}_{E\left[\breve{\alpha}_t^{2}(q)\right]} E\left[\breve{\alpha}_t^{2}(q)\right] - \sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} \left(\frac{c_q(k)}{2}\right)^{4} 6E\left[\breve{\alpha}_t^{0}(q)\right] \tag{4.268}$$

$$= 3\left(E\left[\breve{\alpha}_t^{2}(q)\right]\right)^2 - \frac{3}{8} \sum_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}} c_q^4(k) < 3\left(E\left[\breve{\alpha}_t^{2}(q)\right]\right)^2 \tag{4.269}$$

also imposingly points out the *sub*-Gaussian nature of the phase factor RV $\breve{\alpha}_t(q)$.[9]

When comparing the even moments computed from (4.264) with the sample moments it may, however, be observed that the assumption of independence posed on the phase differences $\varphi_{S_t(k),N_t(k)}$ is approximately valid only for a rectangular analysis window. This observation is illustrated in Fig. 4.19, where the sample moments of the phase factors $\breve{\alpha}_t(q)$, obtained at a reverberation time of $T_{60} \approx 350\,\mathrm{ms}$, are compared with the analytically found solution (4.264) in case the front-end analysis employs a rectangular window (Subfig. 4.19a) and a Hamming window (Subfig. 4.19b). For a rectangular analysis window, the even moments can be found to be reasonably well approximated by the solution given in (4.264). However for the Hamming window the empirically determined and the analytically found even moments differ significantly. Looking at the second moment given by (4.267) more

[9] For a Gaussian distributed RV the fourth central moment is equal to three times the square of the second central moment.



(a) Rectangular window

(b) HAMMING *window*

Figure 4.19: *Comparison of the (E)mpirically determined even moments of the phase factors at the presence of reverberation at* $T_{60} \approx 350\,\mathrm{ms}$ *and a global broadband RNR of* 10 dB *with the (A)nalytically determined ones according to* (4.264) *employing a rectangular window (a) and a* HAMMING *window (b) in the front-end analysis.*

closely, the approximation

$$E\left[\breve{\alpha}_t^2(q)\right] = \sum_{k,k'=K_q^{(\text{low})}}^{K_q^{(\text{up})}} E\left[\breve{\nu}_{t,q}(k)\,\breve{\nu}_{t,q}\left(k'\right)\right] \tag{4.270}$$

$$= \sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} E\left[\breve{\nu}_{t,q}^2(k)\right] + 2\sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}-1} \sum_{k'=k+1}^{K_q^{(\text{up})}} E\left[\breve{\nu}_{t,q}(k)\,\breve{\nu}_{t,q}\left(k'\right)\right] \tag{4.271}$$

$$\approx \sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} E\left[\breve{\nu}_{t,q}^2(k)\right], \tag{4.272}$$

where $E\left[\breve{\nu}_{t,q}^2(k)\right] = \frac{1}{2}c_q^2(k)$, can be found to neglect any correlation between the phase differences at different frequency bins. If the analysis window now induces additional correlations between the phase differences $\breve{\varphi}_{S_t(k),N_t(k)}$, these neglected correlations have to be compensated for.

Though the correlations appear as an additive term in (4.271), the compensation may also be achieved by a multiplicative term. Such a multiplicative factor, denoted by F_{w_A}, is given by [90, p. 150, Theorem 5.6.4]

$$F_{w_\mathsf{A}} := L_{w_\mathsf{A}} \frac{\sum_{l=0}^{L_{w_\mathsf{A}}-1} w_\mathsf{A}^4(l)}{\left(\sum_{l=0}^{L_{w_\mathsf{A}}-1} w_\mathsf{A}^2(l)\right)^2}. \tag{4.273}$$

For the Hamming window of length $L_{w_\mathsf{A}} = 200$, $F_{w_\mathsf{A}} = 1.8257$. Since $F_{w_\mathsf{A}} = 1$ for the rectangular analysis window, the second central moment of the phase factor RV $\breve{\alpha}_t(q)$ may eventually be written as

$$E\left[\breve{\alpha}_t^2(q)\right] \approx F_{w_\mathsf{A}} \frac{1}{2} \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} c_q^2(k)\,, \tag{4.274}$$

irrespective of the applied analysis window. The correction may even be used to approximate the fourth central moment as

$$E\left[\breve{\alpha}_t^4(q)\right] \approx F_{w_\mathsf{A}}^2 \left(\frac{3}{4} \left(\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} c_q^2(k) \right)^2 - \frac{3}{8} \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} c_q^4(k) \right). \tag{4.275}$$

The corrected second and fourth central moment are displayed in Fig. 4.20 together with the empirically determined ones for the Hamming window. With (4.274) and (4.275), approximate values for the second and fourth central moment of the phase factor RV $\alpha_t(q)$ may thus be obtained in a purely analytic manner.

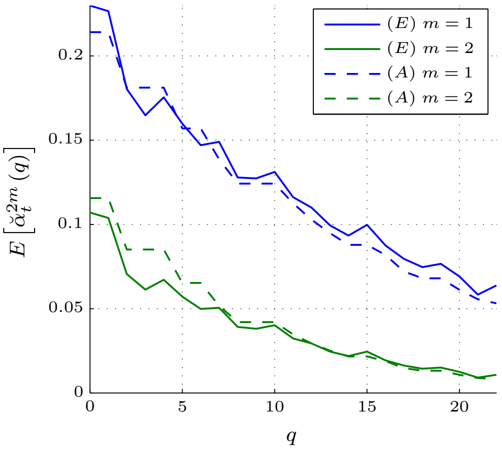


Figure 4.20: *Comparison of the (E)mpirically determined second and fourth central moments of the phase factors at the presence of reverberation at* $T_{60} \approx 350$ ms *and a global broadband RNR of* 10 dB *with the corrected (A)nalytically determined ones according to* (4.274) *and* (4.275) *employing a* Hamming *window in the front-end analysis.*

4.7 Observation Errors

In Secs. 4.3.2, 4.3.3 and 4.3.4 it has been outlined that the stochastic observation models in the reverberant, noisy reverberant and noisy case are completely characterized by the conditional PDFs of the respective observation errors.

In the following, a closer look will be taken at the different observation errors in terms of their marginal distributions conditioned on the IRNR (which in the absence of reverberation reduces to the ISNR). While in the absence of reverberation, the parametric approximation to the distribution of the vector of phase factors introduced in Sec. 4.6.3 allows for a completely analytic solution to the desired PDFs, the PDFs in the presence of reverberation cannot be given in analytic forms and will thus be approximated by GAUSSIAN PDFs whose mean vectors and covariance matrices are modeled as functions of the IRNR. The goodness of this approximation will primarily be illustrated for the non-recursive observation model and only briefly be discussed for the recursive observation model.

Since only an artificially composed database provides reliable access to the IRNR underlying the noisy reverberant and noisy LMPSC feature vector of the observation, all experiments are again based on the AURORA 5 database. The provided clean speech signals and the noise signals were used together with a set of 100 artificially created AIRs, generated by the image method [91], to first create the reverberant speech signals by convolving the clean speech signal with a randomly chosen AIR and later on the noisy and noisy reverberant signals by adding noise at a desired global broadband ***s**ignal-to-**n**oise **r**atio* (SNR)/RNR to the clean/reverberated speech signals.

The reverberation time $\hat{T}_{60}$ of each of the 100 created AIRs thereby lies in the range ±50 ms around a given (average) reverberation time T_{60}.[10] The image method thereby employs a cubic virtual room measuring 5 m × 6 m × 3 m (width×depth×height), which corresponds to an average-sized living room. Further, the position of the speaker has been chosen to lie within one half of the room and the position of the microphone within the other half of the room by random. Both microphone and speaker position were fixed to a height of 1.5 m having a minimum distance of 0.5 m from the walls.

4.7.1 Presence of Reverberation and Absence of Background Noise

Key assumption on the observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ of the non-recursive observation model in the presence of reverberation and absence of noise is its independence on the past LMPSC feature vectors $\breve{\mathbf{s}}_{1:t-1}^{(\mathsf{I})}$ of the reverberant speech signal and the most recent LMPSC feature vectors $\breve{\mathbf{x}}_{t-L_H+1:t}^{(\mathsf{I})}$ of the clean speech signal (compare (4.95)). Though informal correlation tests have already shown this assumption to be fairly optimistic and thus quite debatable, no attempt is made here to model these correlations. The same approximation is made for the observation error $\breve{\mathbf{v}}_{st,L_R}^{(\mathsf{I,R})}$ of the recursive observation model in the presence of reverberation and the absence of noise.

Both observation errors will thus be modeled by GAUSSIAN PDFs with mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ for the non-recursive observation model and mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{I,R})}}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathsf{I,R})}}$ for the recursive observation model, respectively.

Note that the analyses presented in this section merely subsume the more detailed ones presented in [67].

[10] Such an accuracy may be expected from state-of-the-art reverberation time estimators [92]. The presented analyses thus already take actually occurring estimation errors into account.

4.7.1.1 The Non-Recursive Observation Model

The left column of Fig. 4.21 shows the sample covariances matrices $\hat{\mathbf{\Sigma}}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ of the observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ at a reverberation time of $T_{60} = 350\,\mathrm{ms}$ (Subfig. 4.21a) and a reverberation time of $T_{60} = 550\,\mathrm{ms}$ (Subfig. 4.21c).

It can clearly be seen that the sample covariance matrices $\hat{\mathbf{\Sigma}}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ are dominated by their diagonal elements and that the correlations between the observation error components are strictly positive. Substantial elements on the secondary diagonals are mainly due to the overlap of adjacent mel frequency bands. For a simplified modeling it may thus be reasonable to assume the individual observation error components to be uncorrelated. Moreover, the variance of the observation error is, in general, slightly larger for $T_{60} = 550\,\mathrm{ms}$ than for $T_{60} = 350\,\mathrm{ms}$ and decreases with increasing mel frequency index q. The former already indicates that it will be easier to predict the reverberant observation at lower reverberation times than at higher reverberation times.

In case of a Gaussian approximation, the individual observation error components may like-wise be assumed to be independent. The right column of Fig. 4.21 shows the (normalized) histogram approximation to the PDFs $p_{\breve{v}_{st}^{(\mathsf{I,N})}(q)}$ of the observation error components $\breve{v}_{st}^{(\mathsf{I,N})}(q)$ and the Gaussian fit to it for a reverberation time of $T_{60} = 350\,\mathrm{ms}$ (Subfig. 4.21b) and a reverberation time of $T_{60} = 550\,\mathrm{ms}$ (Subfig. 4.21d). The mean and the variance of the Gaussian approximation have thereby been chosen to match the sample mean and the sample variance, respectively.

The Gaussian approximations can be found to match the histograms quite well and further become more accurate with increasing reverberation time. The analysis of the *excess kurtosis*

$$\xi_{\breve{v}_s^{(\mathsf{I,N})}(q)} := \frac{E\left[\left(\breve{v}_{st}^{(\mathsf{I,N})}(q) - E\left[\breve{v}_{st}^{(\mathsf{I,N})}(q)\right]\right)^4\right]}{\left(E\left[\left(\breve{v}_{st}^{(\mathsf{I,N})}(q) - E\left[\breve{v}_{st}^{(\mathsf{I,N})}(q)\right]\right)^2\right]\right)^2} - 3 \tag{4.276}$$

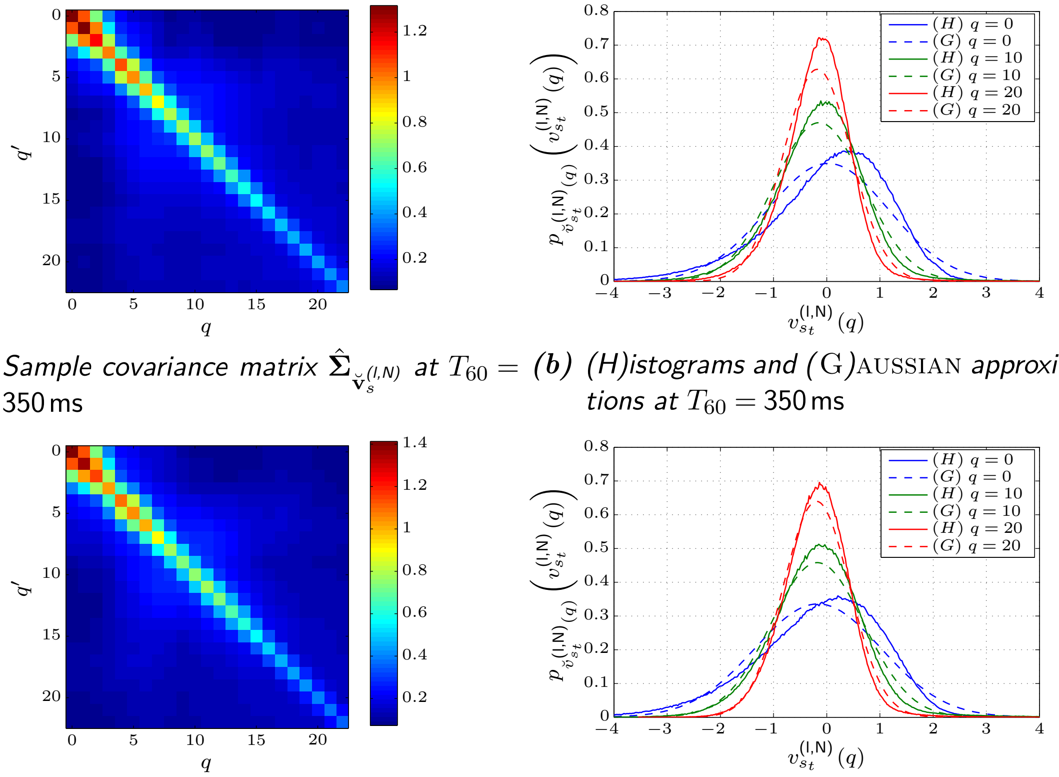
and the *skewness*

$$\nu_{\breve{v}_s^{(\mathsf{I,N})}(q)} := \frac{E\left[\left(\breve{v}_{st}^{(\mathsf{I,N})}(q) - E\left[\breve{v}_{st}^{(\mathsf{I,N})}(q)\right]\right)^3\right]}{\left(E\left[\left(\breve{v}_{st}^{(\mathsf{I,N})}(q) - E\left[\breve{v}_{st}^{(\mathsf{I,N})}(q)\right]\right)^2\right]\right)^{\frac{3}{2}}} \tag{4.277}$$

of the respective random variables presented in Fig. 4.22 supports these findings. While the skewness is mostly about 0 and between -0.5 and $+0.5$ for all mel frequency indices, irrespective of the reverberation time, the excess kurtosis increases from about 1 at low mel frequency indices to about 2 at higher mel frequency indices. Although the skewness and the excess kurtosis of a Gaussian distributed RV are both zero, the illustrated deviations from it do not seem to be too large and the approximation of the marginal PDFs by Gaussian distributions thus quite reasonable.

4.7.1.2 The Recursive Observation Model

The results obtained for the recursive observation mapping look quite similar except for some subtle differences. For lower mel frequency indices q and low recursion lengths L_R



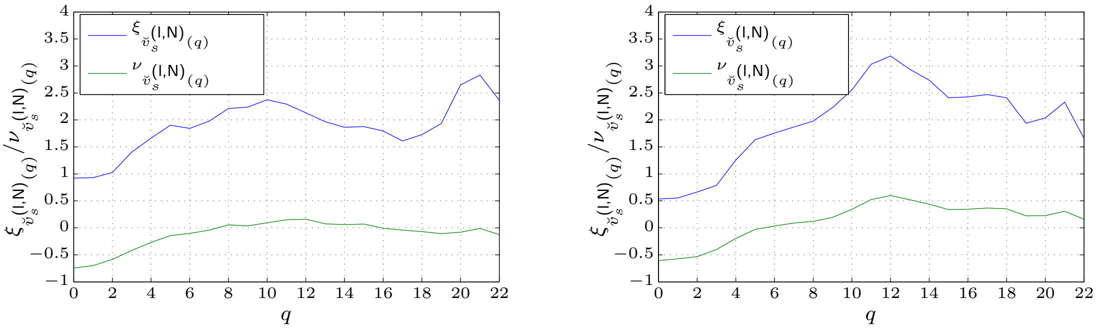
(a) Sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ at $T_{60} =$ 350 ms

(b) (H)istograms and (G)AUSSIAN approximations at $T_{60} =$ 350 ms

(c) Sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ at $T_{60} =$ 550 ms

(d) (H)istograms and (G)AUSSIAN approximations at $T_{60} =$ 550 ms

Figure 4.21: *Non-recursive observation model: Sample covariance matrices $\hat{\boldsymbol{\Sigma}}_{\breve{\mathbf{v}}_s^{(\mathsf{I,N})}}$ of the observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ (left column) for $T_{60} \in \{350\,\mathrm{ms}, 550\,\mathrm{ms}\}$ with the corresponding (H)istogram approximations and (G)AUSSIAN approximations to the marginal PDFs $p_{\breve{v}_{st}^{(\mathsf{I,N})}(q)}$ (right column) for $q \in \{0, 10, 20\}$.*



(a) Excess kurtosis and skewness at $T_{60} =$ 350 ms

(b) Excess kurtosis and skewness at $T_{60} =$ 550 ms

Figure 4.22: *Non-recursive observation model: Excess kurtosis $\xi_{\breve{v}_s^{(\mathsf{I,N})}(q)}$ and skewness $\nu_{\breve{v}_s^{(\mathsf{I,N})}(q)}$ of the observation error $\breve{v}_{st}^{(\mathsf{I,N})}(q)$ for a reverberation time of $T_{60} =$ 350 ms (a) and $T_{60} =$ 550 ms (b).*

the leptokurtotic property of the histograms of the observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ is even more pronounced than for the observation error in the non-recursive observation model. This can best be inferred from figure Fig. 4.23, where the histogram approximations and the GAUSSIAN approximations to the PDFs $p_{\breve{v}_{st,L_R}^{(\mathsf{I,R})}(q)}$ of the observation error components $\breve{v}_{st,L_R}^{(\mathsf{I,R})}(q)$ are displayed for different recursion length $L_R \in \{1,3,6,9\}$ at reverberation times of $T_{60} \in \{350\,\mathrm{ms}, 550\,\mathrm{ms}\}$. However, the GAUSSIAN approximation still seems quite reasonable and can further be found to become more accurate with increasing recursion length L_R and increasing reverberation time T_{60}. For $L_R = 9$, the histograms and the GAUSSIAN approximations to the PDF of the observation error in the recursive observation model given in Fig. 4.23g and Fig. 4.23h in fact are almost indistinguishable from the ones given in Fig. 4.21b and Fig. 4.21d for the non-recursive observation model.

In summary, the observation errors of the non-recursive and the recursive observation model may reasonably well be modeled by GAUSSIAN distributions. Their mean vectors and (diagonal) covariance matrices may thereby be obtained from artificially reverberated training data.

4.7.2 Presence of Reverberation and Background Noise

In the additional presence of background noise, as, e.g., highlighted by (4.112) in Sec. 4.3.3, the observation error $\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}$ of the non-recursive observation model can completely be described in terms of the previously discussed observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ in the presence of reverberation and absence of background noise, the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ and the IRNR $\breve{\mathbf{r}}_t^{(\mathsf{I,N})}$.

A similar formulation can be found for the observation error $\breve{\mathbf{v}}_{ot,L_R}^{(\mathsf{I,R})}$ of the recursive observation model. However, the uncertainty about the MPSC of the reverberant speech signal at time instant $t - L_R$ (which is replaced by the MMSE estimate (4.193)) introduces an additional error term $\breve{\mathbf{w}}_{ot,L_R}^{(\mathsf{I,R})}$.

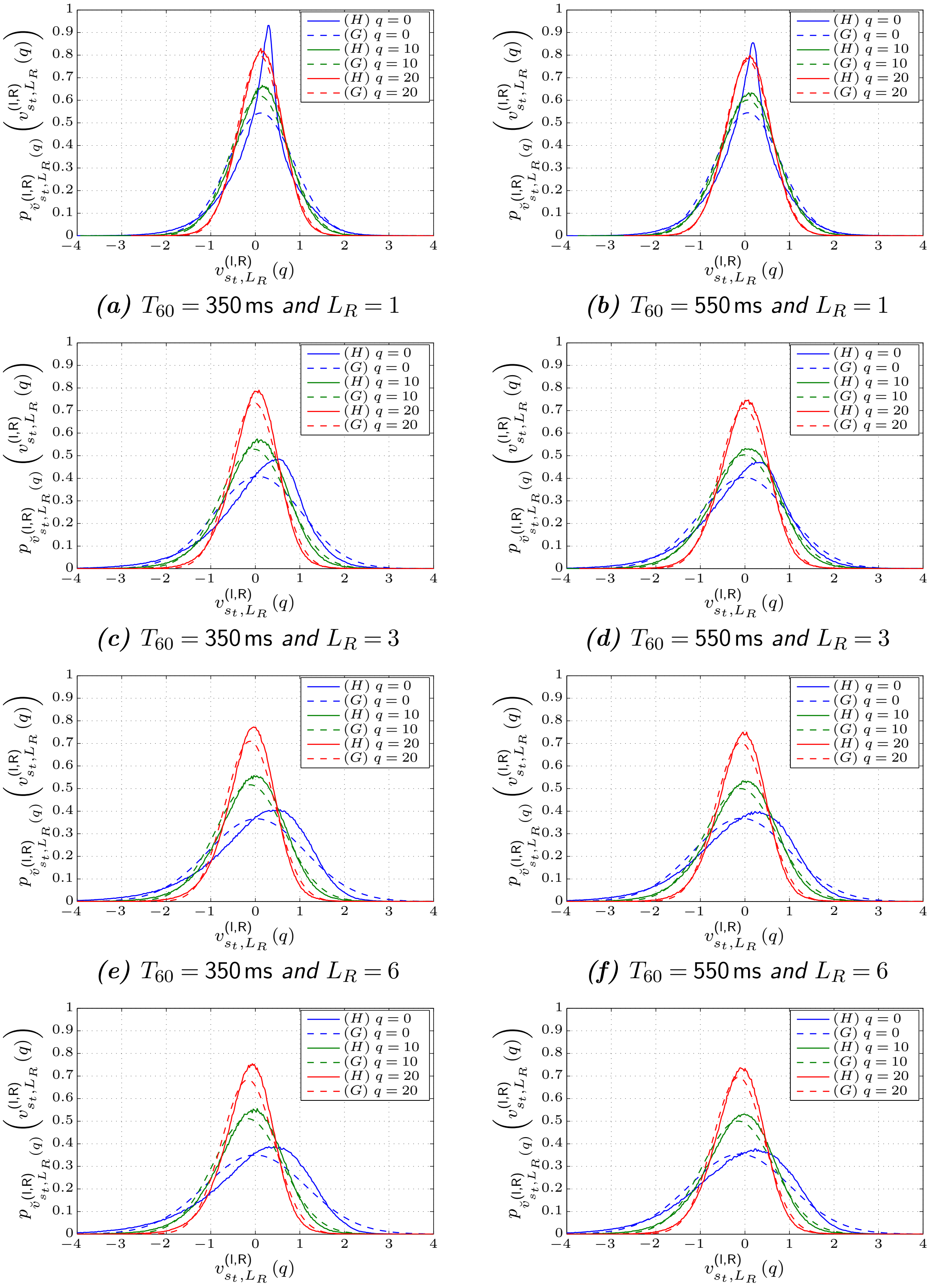
Since the PDFs of both observation errors will be approximated by GAUSSIAN distributions whose mean vectors and covariance matrices are made dependent on the IRNR $\breve{\mathbf{r}}_t^{(\mathsf{I,N})}$ and $\breve{\mathbf{r}}_{t,L_R}^{(\mathsf{I,R})}$ for the non-recursive and recursive observation model, respectively, the following analyses aims at i) illustrating the sensitivity of the PDFs of the observation error on the IRNR and the phase factor and ii) assessing the quality of the GAUSSIAN approximations.

4.7.2.1 The Non-Recursive Observation Model

The (normalized) histogram approximations to the conditional PDFs $p_{\breve{v}_{ot}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ of the observation error components $\breve{v}_{ot}^{(\mathsf{I,N})}(q)$ and the GAUSSIAN approximations to it are illustrated in the left and right column of Fig. 4.24, respectively, at a reverberation time of $T_{60} = 450\,\mathrm{ms}$ and a global broadband RNR of $10\,\mathrm{dB}$. Each subfigure's row thereby represents the distribution of the observation error $\breve{v}_{ot}^{(\mathsf{I,N})}(q)$ for a given IRNR $\breve{r}_t^{(\mathsf{I,N})}(q)$ (to be read off the y-axis).[11]

The means and variances of the GAUSSIAN approximation to the PDFs of the observation error $\breve{v}_{ot}^{(\mathsf{I,N})}(q)$ have thereby been obtained from the means and variances of the observation

[11] Note that the actually values of the approximations to $p_{\breve{v}_{ot}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ are compressed by the 4-th root for visualization purposes.



(a) $T_{60} = 350\,\mathrm{ms}$ *and* $L_R = 1$

(b) $T_{60} = 550\,\mathrm{ms}$ *and* $L_R = 1$

(c) $T_{60} = 350\,\mathrm{ms}$ *and* $L_R = 3$

(d) $T_{60} = 550\,\mathrm{ms}$ *and* $L_R = 3$

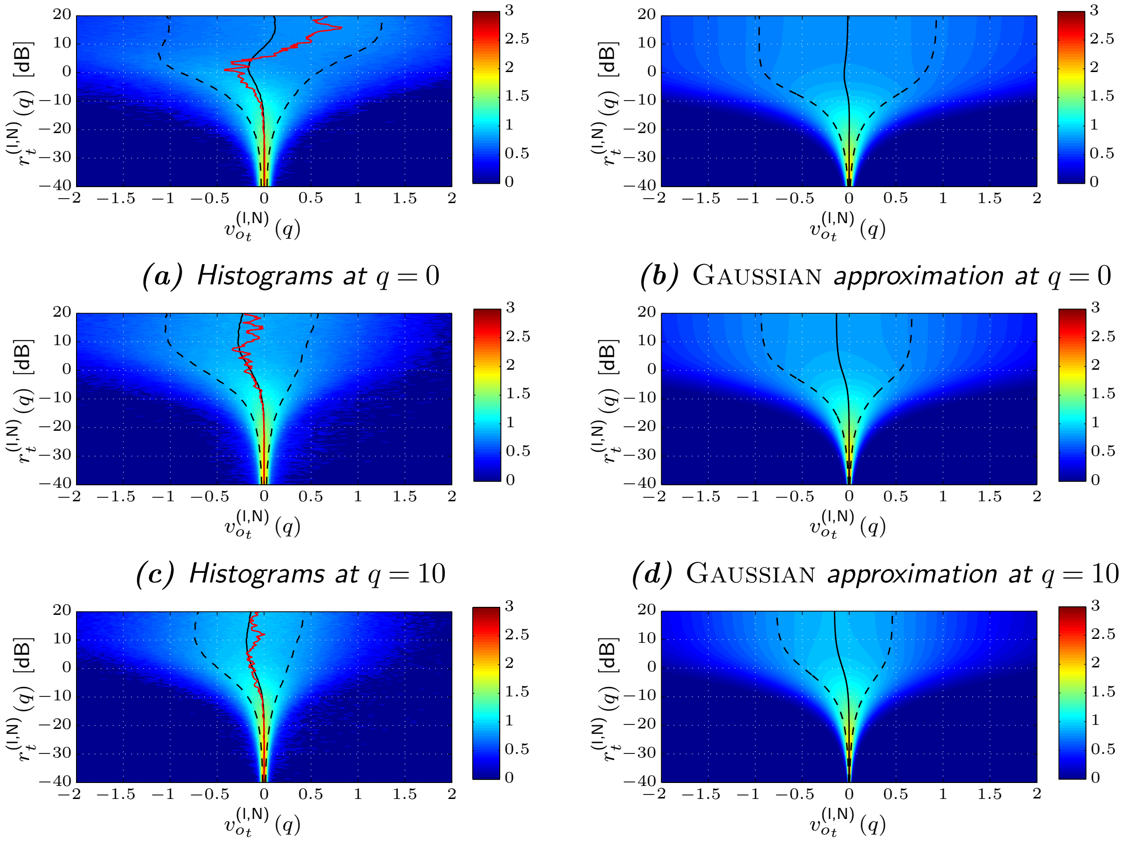
(e) $T_{60} = 350\,\mathrm{ms}$ *and* $L_R = 6$

(f) $T_{60} = 550\,\mathrm{ms}$ *and* $L_R = 6$

(g) $T_{60} = 350\,\mathrm{ms}$ *and* $L_R = 9$

(h) $T_{60} = 550\,\mathrm{ms}$ *and* $L_R = 9$

Figure 4.23: *Recursive observation model: (H)istogram approximations and (*G*)*AUSSIAN *approximations to the marginal PDFs* $p_{\breve{v}^{(I,R)}_{st,L_R}(q)}$ *for* $T_{60} = 350\,\mathrm{ms}$ *(right column) and* $T_{60} = 550\,\mathrm{ms}$ *(left column) for* $L_R \in \{1,3,6,9\}$ *and* $q \in \{0,10,20\}$.



(a) *Histograms at $q = 0$*

(b) Gaussian *approximation at $q = 0$*

(c) *Histograms at $q = 10$*

(d) Gaussian *approximation at $q = 10$*

(e) *Histograms at $q = 20$*

(f) Gaussian *approximation at $q = 20$*

Figure 4.24: *Non-recursive observation model: Histogram approximations to the conditional PDFs $p_{\breve{v}_{ot}^{(I,N)}(q)|\breve{r}_t^{(I,N)}(q)}$ (left column) and the* Gaussian *approximation according to* (4.121) *(right column) at a global broadband RNR of* 10 dB *for a reverberation time of $T_{60} =$* 450 ms *and mel frequency indices $q \in \{0, 10, 20\}$. The solid and dashed black lines indicate the mean and the standard deviation contours of $\breve{v}_{ot}^{(I,N)}(q)$ for a given $\breve{r}_t^{(I,N)}(q)$, respectively. The red solid lines indicate the mode of the conditional PDFs (only for the histograms).*

error $\breve{v}_{st}^{(\text{I,N})}(q)$ by application of (4.122) and (4.123).

Note that the global broadband RNR only influences the a priori probabilities of the IRNR $\breve{r}_t^{(\text{I,N})}(q)$ and thus does not change the conditional PDFs. The shape of both histograms and Gaussian approximations can be found to consistently follow the findings of the discussion on the IRNR-dependent auxiliary functions $\xi\left(r_t^{(\text{I,N})}(q)\right)$ and $\zeta\left(r_t^{(\text{I,N})}(q)\right)$ carried out on p. 59 in Sec. 4.3.3 (see also Fig. 4.7).

For large values of the IRNR $r_t^{(\text{I,N})}(q)$ (here limited to 20 dB for illustration purposes) the conditional PDF of the observation error $\breve{v}_{ot}^{(\text{I,N})}(q)$ approaches the marginal PDF of the observation error $\breve{v}_{st}^{(\text{I,N})}(q)$ in the absence of noise and so do its mean and variance.

With decreasing IRNR $r_t^{(\text{I,N})}(q)$, auxiliary function $\xi\left(r_t^{(\text{I,N})}(q)\right)$ monotonically decreases while $\zeta\left(r_t^{(\text{I,N})}(q)\right)$ first monotonically increases until $r_t^{(\text{I,N})}(q) = 0$ dB. At this point, the influence of the phase factor related term reaches its maximum.

With further decreasing IRNR $r_t^{(\text{I,N})}(q)$, both auxiliary functions now monotonically decrease towards zero. As a consequence, the observation error $\breve{v}_{ot}^{(\mathsf{I,N})}(q)$ more and more concentrates around zero. Thereby, the observation error's variance decreases considerably as the IRNR $r_t^{(\mathsf{I,N})}(q)$ decreases. The influence of the IRNR on the mean, which is approximately zero all along the line, however, can be found to be almost negligible. Both findings can best be seen by looking at the supporting solid and dashed black lines indicating the means and the standard deviation contours, respectively.

The approximately zero-mean for all IRNR values $r_t^{(\mathsf{I,N})}(q)$ can directly be linked to employing the frequency independent power compensation constant C_P computed by (4.170).

The Relevance of the Power Compensation Constant The choice of the compensation constant C_P has immediate consequences on the conditional distribution of the observation error in the presence of both reverberation and noise. The sensitivity of the distribution w.r.t. the power compensation constant C_P shall be illustrated by the considerations following.

Denoting the power compensation constant computed by (4.170) as the *optimal* power compensation constant $C_P^{(\mathsf{opt})}$, any actually chosen value of the power compensation constant C_P may be expressed as

$$C_P := \varpi C_P^{(\mathsf{opt})}. \tag{4.278}$$

The parameter $\varpi \in \mathbb{R}_{>0}$ thus specifies the deviation of the chosen power compensation constant C_P from the optimal value $C_P^{(\mathsf{opt})}$. The observation error components in the absence of noise may then be considered a function of the power compensation constant. In particular, since the frequency independent power compensation constant C_P can be taken out of the sum in (4.85), the observation error may be expressed in terms of the one obtained using the optimal power compensation constant and the parameter ϖ as

$$v_{st}^{(\mathsf{I,N})}(q;\varpi) := v_{st}^{(\mathsf{I,N})}(q;\varpi=1) - \ln(\varpi). \tag{4.279}$$

It may thus immediately be seen, that the parameter $\ln(\varpi)$ only influences the mean of the observation error $v_{st}^{(\mathsf{I,N})}(q;\varpi)$. The observation error $v_{ot}^{(\mathsf{I,N})}(q;\varpi)$ in the presence of both reverberation and noise may then be expressed in terms of $v_{st}^{(\mathsf{I,N})}(q;\varpi=1)$ as

$$v_{ot}^{(\mathsf{I,N})}(q;\varpi) = \ln\left(1 + \left(\mathrm{e}^{v_{st}^{(\mathsf{I,N})}(q;\varpi)} - 1\right)\xi\left(r_t^{(\mathsf{I,N})}(q;\varpi)\right) + 2\alpha_t(q)\,\mathrm{e}^{\frac{v_{st}^{(\mathsf{I,N})}(q;\varpi)}{2}}\zeta\left(r_t^{(\mathsf{I,N})}(q;\varpi)\right)\right) \tag{4.280}$$

$$\begin{aligned} &= \ln\Bigg(1 + \left(\mathrm{e}^{v_{st}^{(\mathsf{I,N})}(q;\varpi=1)-\ln(\varpi)} - 1\right)\xi\left(r_t^{(\mathsf{I,N})}(q;\varpi=1) + \frac{10}{\ln(10)}\ln(\varpi)\right) \\ &\qquad + 2\alpha_t(q)\,\mathrm{e}^{\frac{v_{st}^{(\mathsf{I,N})}(q;\varpi=1)-\ln(\varpi)}{2}}\zeta\left(r_t^{(\mathsf{I,N})}(q;\varpi=1) + \frac{10}{\ln(10)}\ln(\varpi)\right)\Bigg) \end{aligned} \tag{4.281}$$

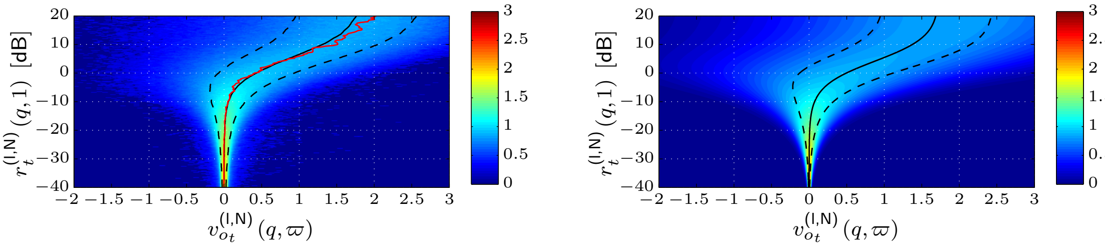
For the last equality, the IRNR for an arbitrary parameter ϖ is written as a function of the IRNR for the optimal power compensation constant, i.e., $\varpi = 1$, as

$$r_t^{(\mathsf{I,N})}(q;\varpi) := r_t^{(\mathsf{I,N})}(q;\varpi=1) + \frac{10}{\ln(10)}\ln(\varpi). \tag{4.282}$$

Clearly, the parameter ϖ causes a shift of the IRNR compared to the IRNR $r_t^{(\mathsf{I,N})}(q;\varpi=1)$ under the optimal power compensation constant.

From (4.280) it can be seen that a deviation from the optimal value $C_P^{(\mathsf{opt})}$ will, due to the non-linearity, affect all moments of the observation error and also shifts the corresponding PDF with respect to the IRNR. Nevertheless, the observation error components may again be modeled by Gaussian distributions whose means and variances are determined by application of (4.122) and (4.123) while employing the means and the variances of the observation error in the absence of noise with the sub-optimally chosen power compensation constant.

To illustrate the influence of the power compensation constant on the conditional PDFs of the observation error $v_{ot}^{(\mathsf{I,N})}(q;\varpi)$, the histogram approximations and the Gaussian approximations to it will be considered for $\varpi^{-1}=C_P^{(\mathsf{opt})}$, i.e., $C_P=1$. The results are depicted in Fig. 4.25 for a reverberation time of $T_{60}=450\,\text{ms}$ at $q=10$ and at a global broadband RNR of 10 dB. To ease a comparison with the conditional PDFs presented in Fig. 4.24 for the optimal power compensation constant, the IRNR $r_t^{(\mathsf{I,N})}(q;1)$ is used on the y-axis, i.e., the shift of the distribution with respect to the IRNR is compensated for. For low values of the IRNR $r_t^{(\mathsf{I,N})}(q;1)$, the sub-optimally chosen power compensation constant $C_P=1$ ($\varpi^{-1}=C_P^{(\mathsf{opt})}$) only has a minor effect on the distribution of the observation error. The observation error still concentrates around zero for low values of the IRNR. However, for increasing values of the IRNR a shift of the histograms' means towards $\mu_{\breve{v}_{st}^{(\mathsf{I,N})}(q;1)}-\ln(\varpi)$ can be observed.

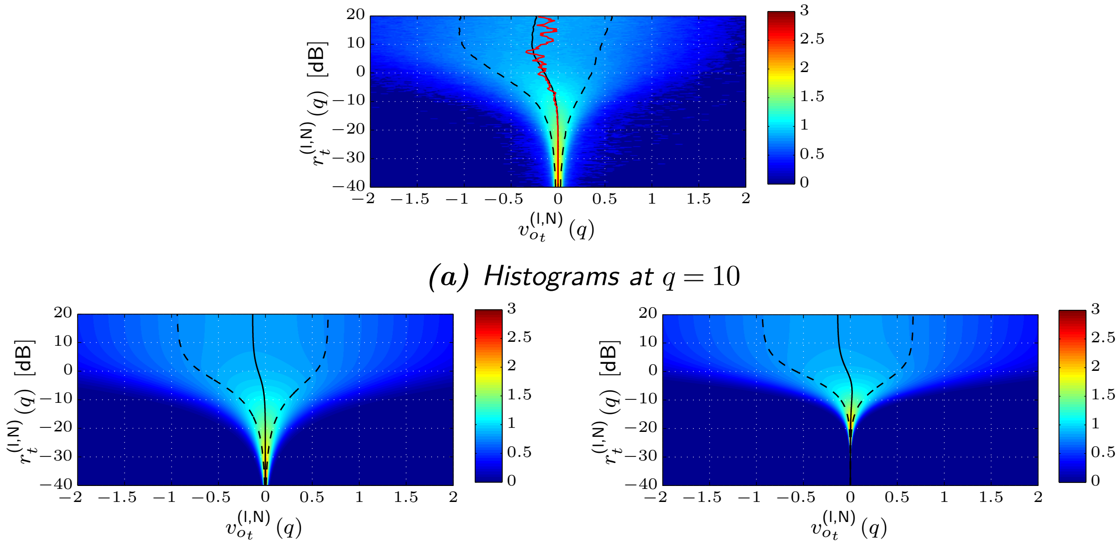


(a) Histograms at $q=10$

(b) Gaussian approximation at $q=10$

Figure 4.25: *Non-recursive observation model: Histogram approximations to the conditional PDFs $p_{\breve{v}_{ot}^{(I,N)}(q,\varpi)|\breve{r}_t^{(I,N)}(q,1)}$ (a) and the Gaussian approximation according to (4.121) (b) at a global broadband RNR of 10 dB for a reverberation time of $T_{60}=450\,\text{ms}$ and mel frequency index $q=10$. The parameter $\varpi^{-1}=C_P^{(opt)}$ has been chosen to force the power compensation constant to be $C_P=1$. The solid and dashed black lines indicate the mean and the standard deviation contours of $\breve{v}_{ot}^{(I,N)}(q,\varpi)$ for a given $\breve{r}_t^{(I,N)}(q,1)$, respectively. The red solid lines indicate the mode of the conditional PDFs (only for the histograms).*

The Influence of the Vector of Phase Factors Finally, Fig. 4.26 illustrates the influence of the phase factor $\breve{\alpha}_t(q)$ on the Gaussian approximation of the observation error's conditional PDF (now again with the optimally chosen power compensation constant). Without considering the phase factor, which amounts to setting $\alpha_t(q)=0$ in (4.112) and consequently $\sigma^2_{\breve{\alpha}_q}=0$ in (4.125), a severe mismatch of the Gaussian approximations



(a) Histograms at $q = 10$

(b) Gaussian *approximation with phase factor consideration at* $q = 10$

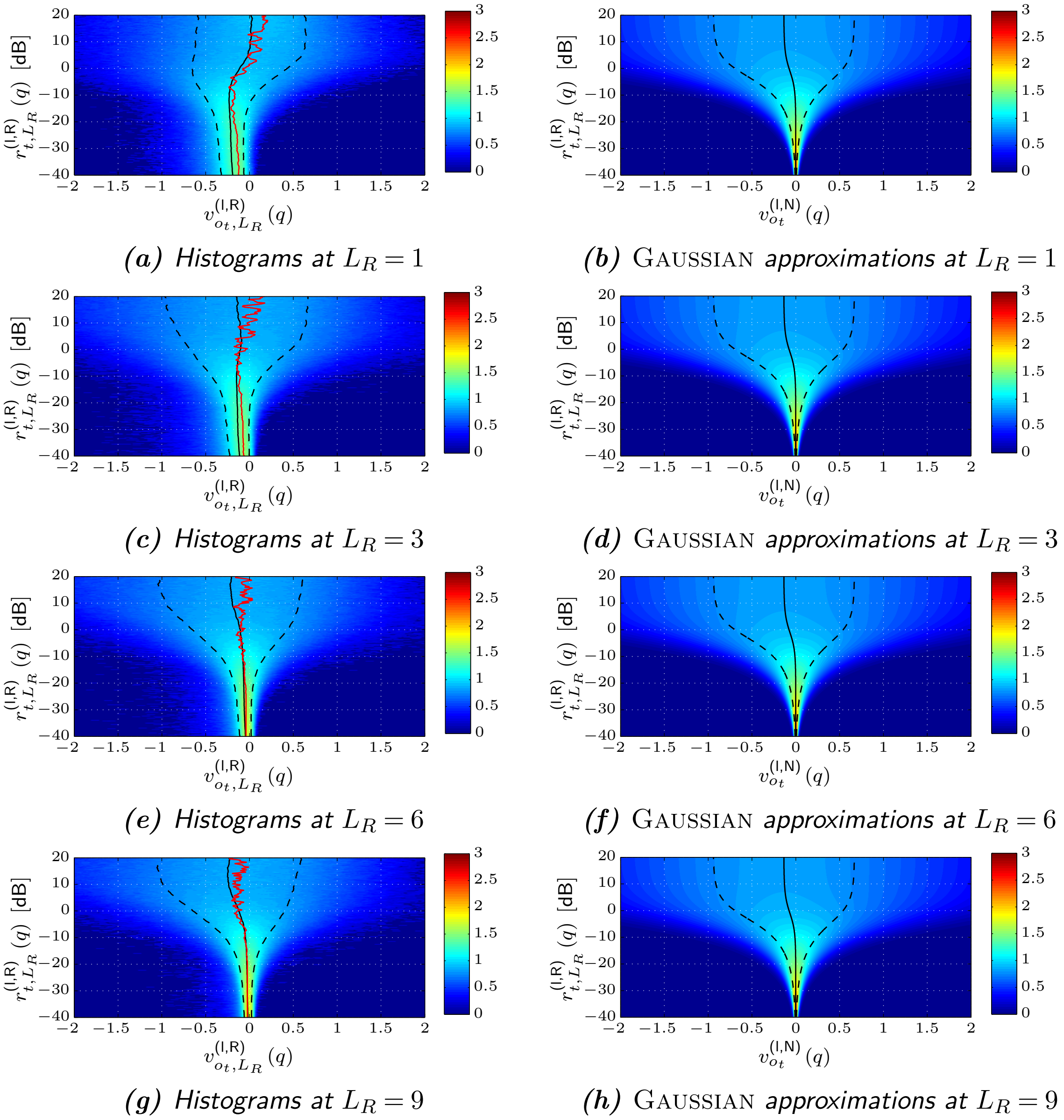
(c) Gaussian *approximation without phase factor consideration at* $q = 10$

Figure 4.26: *Non-recursive observation model: Histogram approximations to the conditional PDFs* $p_{\breve{v}_{o_t}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ *(a) and the* Gaussian *approximations with (b) and without (c) consideration of the phase factor* $\breve{\alpha}_t(q)$ *to the observation error. at a global broadband RNR of* 10 dB, *a reverberation time of* $T_{60} = 450$ ms *and mel frequency index* $q = 10$.

(Subfig. 4.26b) to the histograms (Subfig. 4.26a) can be observed, especially at low IRNR values.

4.7.2.2 The Recursive Observation Model

Figure 4.27 now shows the histogram approximations (left column) and the Gaussian approximations (right column) to the conditional PDFs $p_{\breve{v}_{o_t,L_R}^{(\mathsf{I,R})}(q)|\breve{r}_{t,L_R}^{(\mathsf{I,R})}(q)}$ of the observation error $\breve{v}_{o_t,L_R}^{(\mathsf{I,R})}(q)$ of the recursive observation model for different recursion length L_R at a global broadband RNR of 10 dB, a reverberation time of $T_{60} = 450$ ms and mel frequency index $q = 10$. As expected, the uncertainty in the estimation of the MPSC of the reverberant speech signal at time instant $t - L_R$ causes the histograms to significantly differ from the ones obtained from the non-recursive observation model (compare Fig. 4.24). In particular, the histograms exhibit negative skewness' at mid and low levels of the IRNR. However, with an increasing recursion length L_R, the influence of the estimation error becomes negligible and the histograms of the observation error approach those of the non-recursive model. Hence, for sufficiently large L_R, the derived Gaussian approximation may still be applied.



(a) Histograms at $L_R = 1$

(b) GAUSSIAN approximations at $L_R = 1$

(c) Histograms at $L_R = 3$

(d) GAUSSIAN approximations at $L_R = 3$

(e) Histograms at $L_R = 6$

(f) GAUSSIAN approximations at $L_R = 6$

(g) Histograms at $L_R = 9$

(h) GAUSSIAN approximations at $L_R = 9$

Figure 4.27: *Recursive observation model: Histogram approximations to the conditional PDFs $p_{\breve{v}^{(I,R)}_{ot,L_R}(q)|\breve{r}^{(I,R)}_{t,L_R}(q)}$ (left column) and the* GAUSSIAN *approximation according to* (4.208) *(right column) at a global broadband RNR of* 10 dB *for different recursion length $L_R \in \{1,3,6,9\}$ at a reverberation time of $T_{60} = 450$ ms and mel frequency index $q = 10$. The solid and dashed black lines indicate the mean and the standard deviation contours of $\breve{v}^{(I,R)}_{ot,L_R}(q)$ for a given $\breve{r}^{(I,R)}_{t,L_R}(q)$, respectively. The red solid lines indicate the mode of the conditional PDFs (only for the histograms).*

4.7.3 Absence of Reverberation and Presence of Background Noise

In the absence of reverberation, the conditional PDF $p_{\breve{v}_{yt}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ of the observation error $\breve{v}_{yt}^{(\mathsf{I,N})}(q)$ for a given ISNR $\breve{r}_t^{(\mathsf{I,N})}(q)$ is completely characterized by the PDF $p_{\breve{\alpha}_t(q)}$ of the phase factor $\breve{\alpha}_t(q)$.

With the parametric approximation to the PDF of the phase factor given by (4.254) and the analytic solution to the required second central moment given by (4.274), the conditional PDF $p_{\breve{v}_{yt}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ of the observation error $\breve{v}_{yt}^{(\mathsf{I})}(q)$ may be approximated by the parametric distribution

$$p_{\breve{v}_{yt}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}\left(v_{yt}^{(\mathsf{I,N})}(q)\,\middle|\,r_t^{(\mathsf{I,N})}(q)\right)=\frac{\mathrm{e}^{v_{yt}^{(\mathsf{I,N})}(q)}}{2\zeta\left(r_t^{(\mathsf{I,N})}(q)\right)}p_{\breve{\alpha}_t(q)}\left(\frac{\mathrm{e}^{v_{yt}^{(\mathsf{I,N})}(q)}-1}{2\zeta\left(r_t^{(\mathsf{I,N})}(q)\right)}\right) \qquad (4.283)$$

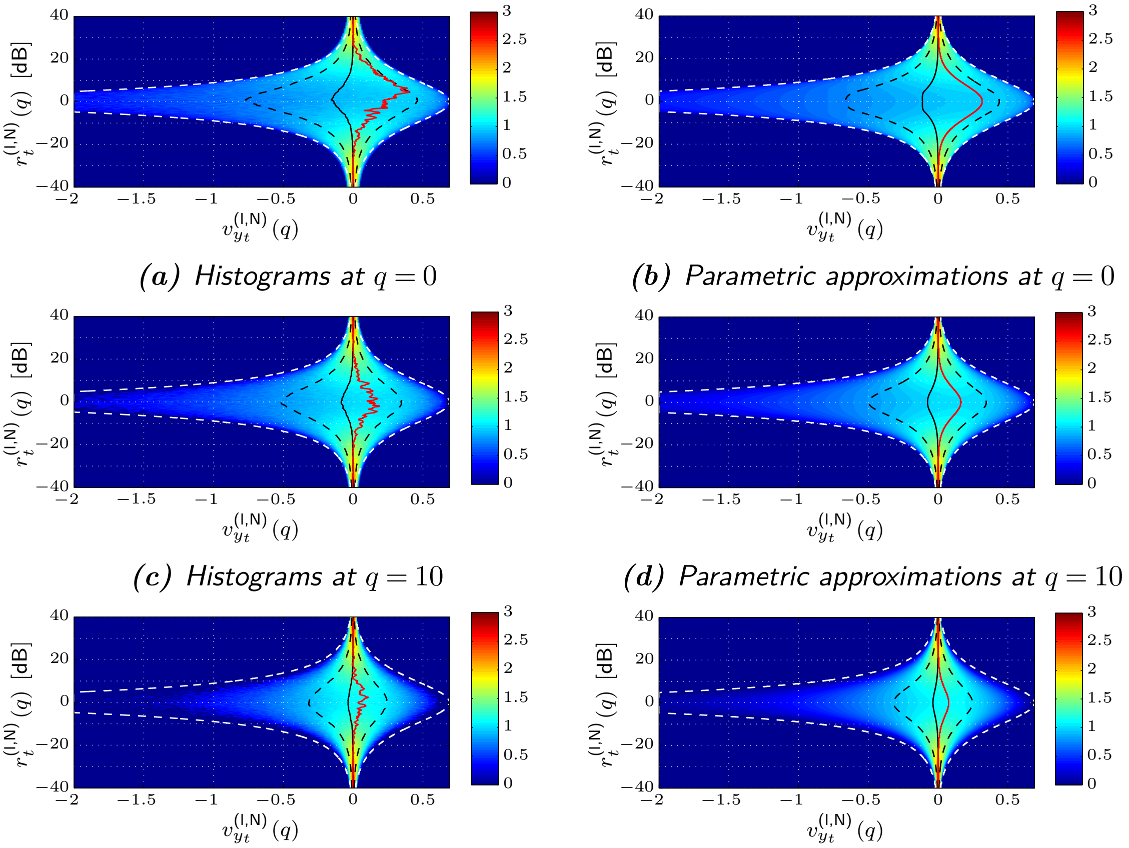
$$\approx\frac{\mathrm{e}^{v_{yt}^{(\mathsf{I})}(q)}}{\sqrt{32}\zeta\left(r_t^{(\mathsf{I})}(q)\right)\sigma_{\breve{\gamma}_q}}\mathrm{e}^{-\frac{1-2\sigma_{\breve{\gamma}_q}^2}{2\sigma_{\breve{\gamma}_q}^2}\left(\mathrm{erf}^{-1}\left(\frac{\mathrm{e}^{v_{yt}^{(\mathsf{I})}(q)}-1}{2\zeta\left(r_t^{(\mathsf{I})}(q)\right)}\right)\right)^2}, \qquad (4.284)$$

which has been obtained by plugging the parametric approximation to the PDF of the phase factor given in (4.254) into (4.283). Since the phase factor $\alpha_t(q)$ always lies in the interval $[-1,+1]$, (4.135) may be employed to find the conditional PDF $p_{\breve{v}_{yt}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ of the observation error $\breve{v}_{yt}^{(\mathsf{I,N})}(q)$ to be non-zero only for $\ln\left(1-2\zeta\left(r_t^{(\mathsf{I})}(q)\right)\right)\leq v_{yt}^{(\mathsf{I,N})}(q)\leq\ln\left(1+2\zeta\left(r_t^{(\mathsf{I})}(q)\right)\right)\leq\ln(2)$. Since the parametric approximation to the PDF of the phase factor obeys the bounds on the phase factor, i.e., is zero outside the interval $[-1,+1]$, this property is also preserved by the parametric approximation (4.284).

Figure 4.28 now shows the histogram approximation to the conditional PDF $p_{\breve{v}_{yt}^{(\mathsf{I,N})}(q)|\breve{r}_t^{(\mathsf{I,N})}(q)}$ of the observation error $\breve{v}_{yt}^{(\mathsf{I,N})}(q)$ (left column) as well as the parametric approximation to it (right column) given in (4.284) for a global broadband SNR of 10 dB and different mel frequency indices $q\in\{0,10,20\}$. The regions where both histogram and parametric approximations are non-zero and zero, respectively, are separated by the dashed white lines. Clearly, these regions are the same, irrespective of the mel frequency index q. However, the conditional PDFs vary with the mel frequency index q. This can best be seen by looking at the evolution of the mean and the variance (illustrated in terms of the standard deviation contours) with decreasing ISNR.

The overall characteristic is common to all mel frequency indices: the mean is monotonically decreasing and the variance monotonically increasing until reaching an ISNR of $r_t^{(\mathsf{I,N})}(q)=0\,\mathrm{dB}$; beyond this point, the mean now monotonically increases while the variance monotonically decreases. However, the observation error $\breve{v}_{yt}^{(\mathsf{I})}(q)$ at lower mel frequency indices exhibits a larger variance than the observation error at upper mel frequency indices.. This can be attributed to the property of the phase factors and in particular their variances, which show the same characteristic (compare Fig. 4.20).

From Fig. 4.28 it also becomes apparent that neglecting the phase factor contribution to the observation error results in fairly rough approximations to the conditional PDF of the



(a) Histograms at $q = 0$

(b) Parametric approximations at $q = 0$

(c) Histograms at $q = 10$

(d) Parametric approximations at $q = 10$

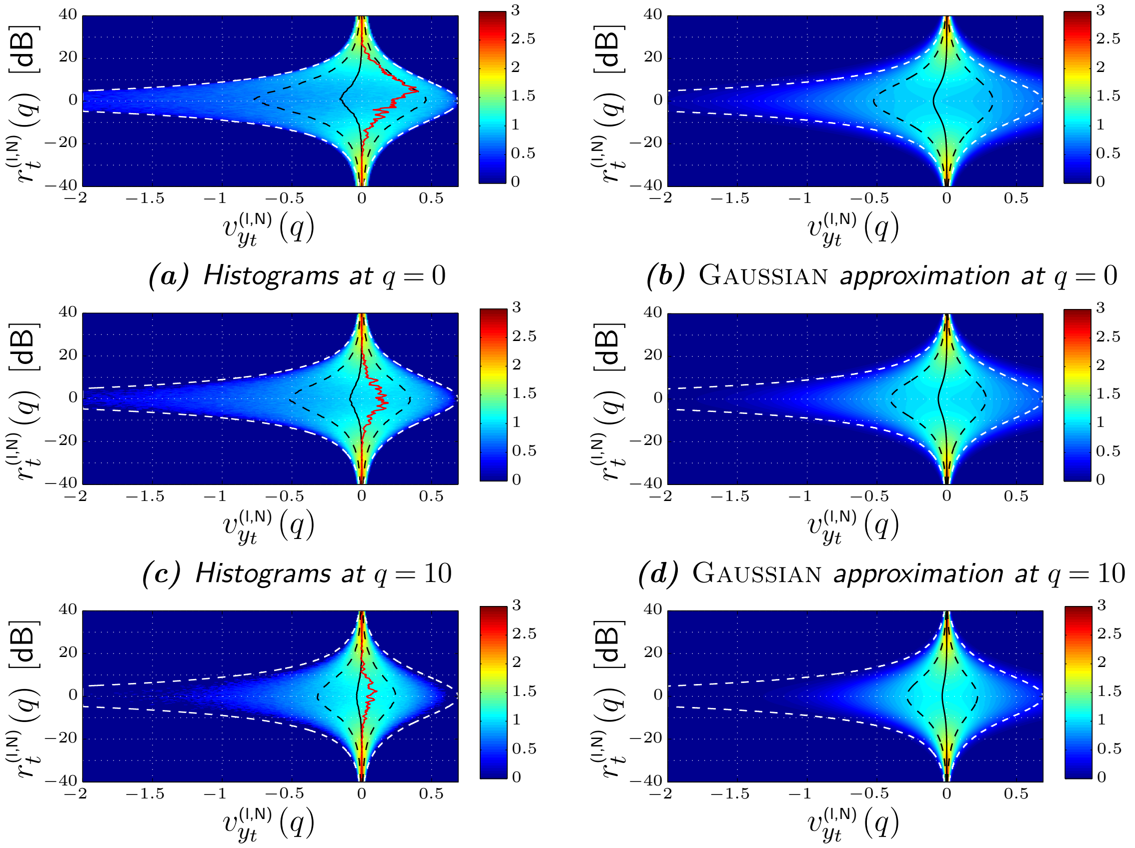
(e) Histograms at $q = 20$

(f) Parametric approximations at $q = 20$

Figure 4.28: *Histogram approximations to the conditional PDFs* $p_{\breve{v}_{yt}^{(l)}(q)|\breve{r}_t^{(l,N)}(q)}$ *(left column) and the parametric approximation according to* (4.141) *with* (4.254) *(right column) at a global broadband SNR of* 10 dB *for mel frequency indices* $q \in \{0, 10, 20\}$*. The dashed white lines mark the beginning of the regions where the conditional PDFs are zero. The solid and dashed black lines indicate the mean and the standard deviation contours of the* $\breve{v}_{yt}^{(l)}(q)$ *for a given* $\breve{r}_t^{(l,N)}(q)$*, respectively. The red solid lines indicate the mode of the conditional PDFs.*

observation error. Neglecting the phase factor's contribution to the observation error would result in a Dirac-Delta distribution of the observation error centered at 0, independent of the ISNR. However, this approximation is true only for an ISNR of $r_t^{(\mathrm{l,N})}(q) = \pm\infty$dB, i.e., either in the absence of noise or the absence of speech, and in particular untenable at, e.g., $r_t^{(\mathrm{l,N})}(q) = 0$ dB.

Finally, Fig. 4.29 illustrates the fit of the Gaussian approximation to the conditional PDFs $p_{\breve{v}_{yt}^{(\mathrm{l})}(q)|r_t^{(\mathrm{l,N})}(q)}$ of the observation error $\breve{v}_{yt}^{(\mathrm{l})}(q)$ according to (4.143). The Gaussian approximations can be found to extend beyond the zero/non-zero bounds of the conditional PDFs of the observation error and further fail to model the mode of the PDFs correctly for mid-level ISNR values around 0 dB. However, the means and variances, indicated by the solid and dashed black lines, respectively, can be found to match those under the histograms quite well.



(a) Histograms at $q=0$

(b) Gaussian *approximation at $q=0$*

(c) Histograms at $q=10$

(d) Gaussian *approximation at $q=10$*

(e) Histograms at $q=20$

(f) Gaussian *approximation at $q=20$*

Figure 4.29: *Histogram approximations to the conditional PDFs $p_{\breve{v}_{yt}^{(l)}(q)|\breve{r}_t^{(l,N)}(q)}$ (left column, repeated from Figs. 4.28a, 4.28c, 4.28e for convenience) and the* Gaussian *approximation according to* (4.143) *(right column) at a global broadband SNR of* 10 dB *for mel frequency indices $q \in \{0, 10, 20\}$. The dashed white lines mark the beginning of the regions where the conditional PDFs are/should be zero. The solid and dashed black lines indicate the mean and the standard deviation contours of the $\breve{v}_{yt}^{(l)}(q)$ for a given $\breve{r}_t^{(l,N)}(q)$, respectively. The red solid lines indicate the mode of the conditional PDFs (only for the histograms).*

4.8 Inference

With the a priori models and the (approximate) observation models given (see Sec. 4.2 and Sec. 4.3/Sec. 4.5, respectively), the conceptually optimal solution to the inference of the a posteriori PDF of the state vector $\breve{\mathbf{z}}_t^{(\mathsf{l})}$ is given in terms of the prediction step (4.2) and the update step (4.3)/(4.4).

However, an exact inference under the derived observation models is not possible, since the a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ as the normalized product of the observation PDF $p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ and the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$, as to be calculated by (4.4), can i) not be given in closed form and ii) does not build a conjugate distribution to the predictive PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$.

Even if the a posteriori and the predictive PDF were conjugate distributions, the multimodality of the a priori model $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ would not allow for a computationally tractable solution: the number of mixtures in the a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ would increase exponentially over time. At the first time instant it would consist of M mixture components, at the second time instant of M^2 mixture components, at the third time instant of M^3 mixture components ...

Any inference scheme thus has to cope with the aforementioned two issue. Since the state update step can separately be carried out on each mixture of the predictive PDF, the first issue is addressed by the so-called *model-specific* inference. The second issue is then addressed by the so-called *multi-model* inference discussed first.

4.8.1 Approximate Multi-Model Inference

In a multi-model inference scheme, the a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ at time instant t may be written in terms of the model state sequence $\breve{m}_{1:t}$ as

$$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})}\right)=\sum_{\{m_1\}}\cdots\sum_{\{m_t\}}p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_{1:t}}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})},m_{1:t}\right)P_{\breve{m}_{1:t}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},}\left(m_{1:t}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})}\right). \tag{4.285}$$

Since (4.285) is, due to the already mentioned exponential increase in model state sequences and thus in the number of mixture components in the a posteriori PDF over time, computationally intractable, multi-model inference algorithms target an approximate but tractable solution to it. A detailed overview of approximate inference algorithms is given in [93], of which the ***G**eneralized **P**seudo* **B**AYSIAN *estimator of order **1*** (GPB1) and the ***I**nteracting **M**ultiple **M**odel* (IMM) estimator will be outlined next.[12]

Both the GPB1 estimator and the IMM estimator first ignore the dependencies among the state vector $\breve{\mathbf{z}}_t^{(\mathsf{l})}$ at the current time instant t and all but the current model state $\breve{m}_t$ in (4.285), i.e., start from

$$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})}\right)=\sum_{i=1}^{M}p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})},m_t=i\right)P_{\breve{m}_t|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}\left(m_t=i\middle|\mathbf{o}_{1:t}^{(\mathsf{l})}\right). \tag{4.286}$$

The state update (4.4) now relates the model-conditioned a posteriori PDF to the model-condition predictive PDF, i.e.,

$$\begin{aligned}&p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})},m_t=i\right)\\&\quad\propto p_{\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{m}_t,\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{z}_t^{(\mathsf{l})},m_t=i,\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right)p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i\right)\end{aligned} \tag{4.287}$$

and the state prediction (4.2) finally the model-condition predictive PDF to the model-

[12] Due to the approximate nature of both multi-model and model-specific inference algorithms, computationally more complex multi-model inference algorithms like the ***G**eneralized **P**seudo* **B**AYSIAN *estimator of order **2*** (GPB2) not necessarily result in improved estimates of the a posteriori PDFs of the LMPSC feature vector of the clean speech signal and are thus excluded from further consideration.

conditioned a posteriori PDF of the previous time instant, i.e.,

$$
\begin{aligned}
&p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_t^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)\\
&=\int\limits_{\mathbb{R}^{(L_C+1)Q}} p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}},\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_t^{\mathsf{(l)}}\middle|\mathbf{z}_{t-1}^{\mathsf{(l)}},\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)p_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)\\
&\qquad\mathrm{d}\mathbf{z}_{t-1}^{\mathsf{(l)}} \qquad (4.288)\\
&\approx\int\limits_{\mathbb{R}^{(L_C+1)Q}} p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_t^{\mathsf{(l)}}\middle|\mathbf{z}_{t-1}^{\mathsf{(l)}},m_t=i\right)p_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)\mathrm{d}\mathbf{z}_{t-1}^{\mathsf{(l)}},\\
&\qquad (4.289)
\end{aligned}
$$

where for the last approximation the dominance of the dependency on the model state over the past observations has been employed (compare (4.11)). Consequently, the predictive PDF $p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}$, required for the application of the decoding rule (3.71), is modeled as the mixture distribution

$$
p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}\left(\mathbf{z}_t^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}}\right)=\sum_{i=1}^{M}p_{\breve{\mathbf{z}}_t^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_t^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)P_{\breve{m}_t|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},}\left(m_t=i\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}}\right). \tag{4.290}
$$

Both the GPB1 estimator and the IMM estimator now approximate the a posteriori PDF of the previous time instant by a single Gaussian distribution. They however differ in the way the moments of this Gaussian are computed.

The GPB1 estimator The GPB1 estimator writes and approximates the model-conditioned a posteriori PDF at time instant $t-1$ as

$$
\begin{aligned}
&p_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right)\\
&=\sum_{j=1}^{M}p_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_{t-1:t}}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_{t-1}=j,m_t=i\right)\\
&\qquad P_{\breve{m}_{t-1}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_t}\left(m_{t-1}=j\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_t=i\right) && (4.291)\\
&\approx\sum_{j=1}^{M}p_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}},\breve{m}_{t-1}}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}}\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}},m_{t-1}=j\right)P_{\breve{m}_{t-1}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}\left(m_{t-1}=j\middle|\mathbf{o}_{1:t-1}^{\mathsf{(l)}}\right) && (4.292)\\
&\approx\mathcal{N}\left(\mathbf{z}_{t-1}^{\mathsf{(l)}};\boldsymbol{\mu}^{\mathsf{(GPB1)}}_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}\middle|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}},\boldsymbol{\Sigma}^{\mathsf{(GPB1)}}_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}\middle|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}\right), && (4.293)
\end{aligned}
$$

i.e., ignores potential statistical dependencies on the current model state in both terms under the sum and further approximates the mixture distribution by a single Gaussian.

The mean vector $\boldsymbol{\mu}^{\mathsf{(GPB1)}}_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}$ and the covariance matrix $\boldsymbol{\Sigma}^{\mathsf{(GPB1)}}_{\breve{\mathbf{z}}_{t-1}^{\mathsf{(l)}}|\breve{\mathbf{o}}_{1:t-1}^{\mathsf{(l)}}}$ are thereby chosen to *match* the first and second moments of the true distribution (4.292). This so-called *moment matching* can also be shown to minimize the KL divergence between (4.292) and the Gaussian approximation (4.293).

Denoting the mean vector and the covariance matrix of the PDF $p_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{t-1}}$ for the model state $m_{t-1} = j$ by $\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j}$, respectively, the moments of a Gaussian distribution minimizing the KL divergence are given by

$$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}\right.} = \sum_{j=1}^{M} P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{t-1} = j \left| \mathbf{o}^{(\mathsf{l})}_{1:t-1}\right.\right) \boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j\right.} \tag{4.294}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}\right.} = \sum_{j=1}^{M} P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{t-1} = j \left| \mathbf{o}^{(\mathsf{l})}_{1:t-1}\right.\right) \Bigg[& \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j\right.} \\ & + \left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j\right.} - \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}\right.}\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},j\right.} - \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}\right.}\right)^{\dagger}\Bigg],\end{aligned} \tag{4.295}$$

respectively. Note that the mean vector and the covariance matrix are independent of the current model state $m_t = i$ and that each model $i \in \{1, M\}$ thus operates on the same *initial* condition. Further, the GPB1 algorithm only specifies a rule to create a common initial condition for the next filter step based on the mean vectors and covariance matrices of the mixture distribution resulting from the most recent filter step. As such, it per se outputs the a posteriori PDF as a mixture distribution at each time instant. However, usually only a single Gaussian with its mean vector and covariance matrix obtained from the GPB1 moment matching specified above is passed to subsequent processing steps.

The same holds for the predictive PDF, whose mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}$ and the covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}$ of the predictive PDF $p_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}$ may be obtained in the same manner, however, employing the model-specific predictive mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2},j}$ and covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2},j}$ and the predictive model probabilities $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}$ in (4.294) and (4.295).

The computation of the a posteriori model probabilities $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ (sometimes also referred to as *merging* probabilities in the context of the GPB1 estimator) will be addressed after introduction of the IMM estimator.

The IMM estimator In contrast to the GPB1 estimator, the IMM estimator writes and approximates the model-conditioned a posteriori PDF at time instant $t-1$ as

$$
\begin{aligned}
&p_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(\mathbf{z}^{(\mathrm{l})}_{t-1}\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right)\\
&\quad=\sum_{j=1}^{M}p_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_{t-1:t}}\left(\mathbf{z}^{(\mathrm{l})}_{t-1}\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_{t-1}=j,m_t=i\right)\\
&\qquad\qquad P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right) &&(4.296)\\
&\quad\approx\sum_{j=1}^{M}p_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_{t-1}}\left(\mathbf{z}^{(\mathrm{l})}_{t-1}\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_{t-1}=j\right)P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right)\\
&&&(4.297)\\
&\quad\approx\mathcal{N}\left(\mathbf{z}^{(\mathrm{l})}_{t-1};\boldsymbol{\mu}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.},\boldsymbol{\Sigma}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}\right), &&(4.298)
\end{aligned}
$$

i.e., ignores potential statistical dependency on the current model state $m_t=i$ only in the first term under the sum. The mean vector $\boldsymbol{\mu}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}$ and the covariance matrix $\boldsymbol{\Sigma}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}$ are again chosen to minimize the KL divergence between the mixture distribution (4.297) and the GAUSSIAN (4.298) and are, in analogy to (4.294) and (4.295), given by

$$
\boldsymbol{\mu}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}=\sum_{j=1}^{M}P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right)\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},j\right.}\tag{4.299}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}&=\sum_{j=1}^{M}P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right)\Bigg[\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},j\right.}\\
&\quad+\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},j\right.}-\boldsymbol{\mu}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},j\right.}-\boldsymbol{\mu}^{(\mathrm{IMM})}_{\breve{\mathbf{z}}^{(\mathrm{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},i\right.}\right)^{\dagger}\Bigg].
\end{aligned}\tag{4.300}
$$

The *mixing* probabilities $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}$ are computed from the a posteriori model state probabilities $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1}}$ and the state transition probabilities $P_{\breve{m}_t|\breve{m}_{t-1}}$ of the switching a priori model by

$$
\begin{aligned}
&P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1},\breve{m}_t}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1},m_t=i\right)\\
&\quad\propto P_{\breve{m}_t|\breve{m}_{t-1}}\left(m_t=i\,|\,m_{t-1}=j\right)P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1}}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1}\right) &&(4.301)\\
&\quad=a_{i|j}P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathrm{l})}_{1:t-1}}\left(m_{t-1}=j\,\middle|\,\mathbf{o}^{(\mathrm{l})}_{1:t-1}\right). &&(4.302)
\end{aligned}
$$

Since the mixing probabilities will in general differ w.r.t. the current model state m_t, in essence, each model $i\in\{1,M\}$ operates on a different initial condition.

Note that the IMM algorithm only specifies a rule to create initial conditions for the next filter step at time instant t based on the mean vectors and covariance matrices of the mixture distribution resulting from the past filter step at time instant $t-1$. As such, it per se outputs the a posteriori PDF and the predictive PDF as a mixture distribution at each time instant. If a single GAUSSIAN is required, the moment matching of the GPB1 algorithm may be used (see (4.294) and (4.295)).

The Calculation of the A Posteriori Model Probabilities Both approximate inference algorithms require the a posteriori model probabilities $P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ to compute the mean vector(s) and the covariance matrix/matrices of the GAUSSIAN approximation(s) to the PDF $p_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t}$.

This quantity may be computed recursively as

$$\begin{aligned} &P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}\left(m_{t-1}=j\left|\mathbf{o}^{(\mathsf{l})}_{1:t-1}\right.\right) \\ &\quad\propto p_{\breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2},\breve{m}_{t-1}}\left(\mathbf{o}^{(\mathsf{l})}_{t-1}\left|\mathbf{o}^{(\mathsf{l})}_{1:t-2},m_{t-1}=j\right.\right)P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}\left(m_{t-1}=j\left|\mathbf{o}^{(\mathsf{l})}_{1:t-2}\right.\right) \end{aligned} \tag{4.303}$$

where

$$\begin{aligned} &P_{\breve{m}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}\left(m_{t-1}=j\left|\mathbf{o}^{(\mathsf{l})}_{1:t-2}\right.\right) \\ &\quad\approx\sum_{i=1}^{M}P_{\breve{m}_{t-1}|\breve{m}_{t-2}}\left(m_{t-1}=j\left|m_{t-2}=i\right.\right)P_{\breve{m}_{t-2}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}\left(m_{t-2}=i\left|\mathbf{o}^{(\mathsf{l})}_{1:t-2}\right.\right) && (4.304) \\ &\quad=\sum_{i=1}^{M}a_{j|i}P_{\breve{m}_{t-2}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}\left(m_{t-2}=i\left|\mathbf{o}^{(\mathsf{l})}_{1:t-2}\right.\right) && (4.305) \end{aligned}$$

is the predicted model probability also required for the determination of the predictive PDF $p_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$ in (4.290). Besides the a posteriori model probabilities $P_{\breve{m}_{t-2}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2}}$ at time instant $t-2$ and the models' state transition probabilities $a_{j|i}$, also the likelihood $p_{\breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2},\breve{m}_{t-1}}$ of the observation at time instant $t-1$ given all past observations and the current model index is required for the evaluation of (4.305).

The calculation of the likelihoods $p_{\breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-2},\breve{m}_{t-1}}$ and the inference of the mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_{t-1}\right.}$ and the covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}\left|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_{t-1}\right.}$ of the a posteriori PDFs $p_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_{t-1}}$ for all model states $m_{t-1}\in\{1,M\}$, as required for the GPB1 merging and the IMM mixing, is carried out in the model-specific inference algorithms discussed next.

4.8.2 Approximate Model-Specific Inference

Given the a posteriori PDF at time instant $t-1$, the model-specific inference now in turn carries out the state prediction step (4.2) and the state updated step (4.4).

The State Prediction With the GAUSSIAN approximations to the a posteriori PDF at time instant $t-1$ given by (4.293) and (4.298) for the GPB1 and IMM estimator, respectively, and the GAUSSIAN a priori model for the state characterized by (4.19) and (4.40), the integral (4.289) exhibits a closed-form solution. The model-conditioned predictive PDF turns into the GAUSSIAN distribution

$$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i\right)=\mathcal{N}\left(\mathbf{z}_t^{(\mathsf{l})};\boldsymbol{\mu}^{(\mathsf{GPB1/IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i},\boldsymbol{\Sigma}^{(\mathsf{GPB1/IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}\right), \tag{4.306}$$

where the mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1/IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}$ and the covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{GPB1/IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}$ are given by

$$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}=\begin{cases}\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}\;\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}+\mathbf{b}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}, & \text{for } t>1\\ \boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i} & \text{for } t=1\end{cases} \tag{4.307}$$

$$\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}=\begin{cases}\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}\;\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\;\mathbf{A}^{\dagger}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}+\mathbf{V}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}, & \text{for } t>1\\ \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i} & \text{for } t=1\end{cases} \tag{4.308}$$

for the GPB1 estimator and by

$$\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}=\begin{cases}\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}\;\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},i}+\mathbf{b}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}, & \text{for } t>1\\ \boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i} & \text{for } t=1\end{cases} \tag{4.309}$$

$$\boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t=i}=\begin{cases}\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}\;\boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}_{t-1}^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},i}\;\mathbf{A}^{\dagger}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}+\mathbf{V}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}, & \text{for } t>1\\ \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i} & \text{for } t=1\end{cases} \tag{4.310}$$

for the IMM estimator. The transition matrix $\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}$, the prediction bias vector $\mathbf{b}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}$, the prediction error covariance matrix $\mathbf{V}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}$ as well as the a mean vector $\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}$ are given in terms of the corresponding parameters of the a priori model for the clean speech LMPSC feature vector and the a priori model for the noise LMPSC feature vector. In particular,

$$\mathbf{A}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}=\begin{bmatrix}\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} & \mathbf{0} & \cdots & \cdots & \mathbf{0}\\ \mathbf{I} & \mathbf{0} & \ddots & & \vdots\\ \mathbf{0} & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \mathbf{0}\\ \vdots & & \ddots & \mathbf{I} & \mathbf{0}\\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0}\end{bmatrix},\quad \mathbf{V}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}=\begin{bmatrix}\mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} & \mathbf{0} & \cdots & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \ddots & & \vdots\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{V}_{\breve{\mathbf{n}}^{(\mathsf{l})}}\end{bmatrix},$$

$$\mathbf{b}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}=\begin{bmatrix}\mathbf{b}^{\dagger}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{b}^{\dagger}_{\breve{\mathbf{n}}^{(\mathsf{l})}}\end{bmatrix}^{\dagger},$$

$$\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}=\begin{bmatrix}\boldsymbol{\mu}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}\\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_0^{(\mathsf{l})}}\\ \vdots\\ \boldsymbol{\mu}_{\breve{\mathbf{x}}_{-L_C+2}^{(\mathsf{l})}}\\ \boldsymbol{\mu}_{\breve{\mathbf{n}}_1^{(\mathsf{l})}}\end{bmatrix},\quad \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}|i}=\begin{bmatrix}\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i} & \mathbf{0} & \cdots & \cdots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_0^{(\mathsf{l})}} & \ddots & & \vdots\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_{-L_C+2}^{(\mathsf{l})}} & \mathbf{0}\\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{\Sigma}_{\breve{\mathbf{n}}^{(\mathsf{l})}}\end{bmatrix}, \tag{4.311}$$

where for matrices the short-hand notations $\mathbf{0} := \mathbf{0}_{Q\times Q}$ and $\mathbf{I} := \mathbf{I}_{Q\times Q}$ and for vectors the short-hand notation $\mathbf{0} := \mathbf{0}_{1\times Q}$ have been employed for ease of readability. Note that an efficient realization of the matrix-matrix and matrix-vector multiplications in (4.307)–(4.310) should take into account the special structure of the involved matrices and in particular of that of the state transition matrices.

The mean vectors $\boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{-L_C+2}}$, ..., $\boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{0}}$ and the covariance matrices $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{-L_C+2}}$, ..., $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{0}}$ should reflect knowledge about the LMPSC feature vectors $\breve{\mathbf{x}}^{(\mathrm{I})}_{-L_C+2:0}$ of the clean speech signal prior to the beginning of the recording (e.g., $\boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{-L_C+2}} = \ldots = \boldsymbol{\mu}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{0}} = -50\cdot\mathbf{1}_{Q\times 1}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{-L_C+2}} = \ldots = \boldsymbol{\Sigma}_{\breve{\mathbf{x}}^{(\mathrm{I})}_{0}} = 10^{-6}\cdot\mathbf{I}_{Q\times Q}$ if absence of speech is assumed for $t < 1$).

The State Update With the model-conditioned predictive PDF $p_{\breve{\mathbf{z}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{m}_t}$ given by (4.306), the state update now aims at finding the model-conditioned a posteriori PDF $p_{\breve{\mathbf{z}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t},\breve{m}_t}$. According to (4.3)/(4.4), this PDF is formally given by

$$\begin{aligned}
&p_{\breve{\mathbf{z}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t},\breve{m}_t}\left(\mathbf{z}^{(\mathrm{I})}_t \middle| \mathbf{o}^{(\mathrm{I})}_{1:t}, m_t\right)\\
&= \frac{p_{\breve{\mathbf{z}}^{(\mathrm{I})}_t,\breve{\mathbf{o}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{m}_t}\left(\mathbf{z}^{(\mathrm{I})}_t, \mathbf{o}^{(\mathrm{I})}_t \middle| \mathbf{o}^{(\mathrm{I})}_{1:t-1}, m_t\right)}{p_{\breve{\mathbf{o}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{m}_t}\left(\mathbf{o}^{(\mathrm{I})}_t \middle| \mathbf{o}^{(\mathrm{I})}_{1:t-1}, m_t\right)} &\quad (4.312)\\
&\propto p_{\breve{\mathbf{o}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{\mathbf{z}}^{(\mathrm{I})}_t,\breve{m}_t}\left(\mathbf{o}^{(\mathrm{I})}_t \middle| \mathbf{z}^{(\mathrm{I})}_t, \mathbf{o}^{(\mathrm{I})}_{1:t-1}, m_t\right) p_{\breve{\mathbf{z}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{m}_t}\left(\mathbf{z}^{(\mathrm{I})}_t \middle| \mathbf{o}^{(\mathrm{I})}_{1:t-1}, m_t\right). &\quad (4.313)
\end{aligned}$$

However, (4.313) cannot be reduced to a known parametric form: neither for the closed-form solution to the observation PDF in the absence of reverberation and the presence of noise given in (4.141) nor for the approximate solutions to the observation PDFs in the presence of reverberation given in (4.96) and (4.130) for the non-recursive observation model and in (4.189) and (4.216) for the recursive observation models, respectively.

Though MONTE CARLO methods could be employed to compute the observation likelihood $p_{\breve{\mathbf{o}}^{(\mathrm{I})}_t|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1},\breve{m}_t}$, the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t\left|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t},m_t\right.}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t\left|\breve{\mathbf{o}}^{(\mathrm{I})}_{1:t},m_t\right.}$ of the RV $\breve{\mathbf{z}}^{(\mathrm{I})}_t$ required to carry out the GPB1 or IMM multi-model inference in the next time step, the *curse of dimensionality* renders its application to the $(L_C+1)\cdot Q$-dimensional state vector $\breve{\mathbf{z}}^{(\mathrm{I})}_t$ quite difficult [94].

The sub-optimal approach to a tractable solution to (4.312) pursued in this work is based on a ***vector* TAYLOR *series*** (VTS) expansion of the functional relation between the observation $\mathbf{o}^{(\mathrm{I})}_t$, the LMPSC feature vectors $\mathbf{x}^{(\mathrm{I})}_{t-L_H:t}$ and $\mathbf{n}^{(\mathrm{I})}_t$ of the clean speech signal and the noise, respectively, and the remaining occurring random variables.[13]

For the non-recursive observation functions (4.91), (4.109) and (4.132) these variables are the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ and, for the former two, the observation error $\breve{\mathbf{v}}^{(\mathrm{I,N})}_{s_t}$ in the presence of reverberation and absence of noise. The corresponding non-recursive

[13] Also see [95], who were the first to study VTS for *environmental-independent* speech recognition.

observation functions are defined as

$$\mathbf{s}_t^{(\mathrm{I})} = g_s^{(\mathrm{I})}\left(\mathbf{z}_t^{(\mathrm{I})}, \mathbf{x}_{t-L_H:t-L_C}^{(\mathrm{I})}, \mathbf{v}_{s_t}^{(\mathrm{I,N})};\ \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_H}^{(\mathrm{I})}}\right) \tag{4.314}$$

$$:= \ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}}\right) + \mathbf{v}_{s_t}^{(\mathrm{I,N})} \tag{4.315}$$

for the absence of noise and the presence of reverberation, as

$$\mathbf{o}_t^{(\mathrm{I})} = g_o^{(\mathrm{I})}\left(\mathbf{z}_t^{(\mathrm{I})}, \mathbf{x}_{t-L_H:t-L_C}^{(\mathrm{I})}, \boldsymbol{\gamma}_t, \mathbf{v}_{s_t}^{(\mathrm{I,N})};\ \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_H}^{(\mathrm{I})}}\right) \tag{4.316}$$

$$:= \ln\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathrm{I,N})}} \circ \sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}} + 2\,\mathrm{erf}\left(\boldsymbol{\gamma}_t\right) \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathrm{I,N})}}{2}} \circ \mathrm{e}^{\frac{\ln\left(\sum_{t'=0}^{L_H} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}}\right)+\mathbf{n}_t^{(\mathrm{I})}}{2}} + \mathrm{e}^{\mathbf{n}_t^{(\mathrm{I})}}\right), \tag{4.317}$$

$$= f_o^{(\mathrm{I})}\left(\mathbf{z}_t^{(\mathrm{I})}, \mathbf{x}_{t-L_H:t-L_C}^{(\mathrm{I})};\ \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_H}^{(\mathrm{I})}}\right) + \mathbf{v}_{o_t}^{(\mathrm{I,N})} \tag{4.318}$$

for the presence of both reverberation and noise and as

$$\mathbf{y}_t^{(\mathrm{I})} = g_y^{(\mathrm{I})}\left(\mathbf{z}_t^{(\mathrm{I})}, \boldsymbol{\gamma}_t\right) \tag{4.319}$$

$$:= \ln\left(\mathrm{e}^{\mathbf{x}_t^{(\mathrm{I})}} + 2\,\mathrm{erf}\left(\boldsymbol{\gamma}_t\right) \circ \mathrm{e}^{\frac{\mathbf{x}_t^{(\mathrm{I})}+\mathbf{n}_t^{(\mathrm{I})}}{2}} + \mathrm{e}^{\mathbf{n}_t^{(\mathrm{I})}}\right) \tag{4.320}$$

for the presence of noise and the absence of reverberation, respectively. Note that the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ has been expressed in terms of the normally distributed RV $\breve{\boldsymbol{\gamma}}_t$.

For the recursive observation functions (4.194) and (4.183) the remaining variables are the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$, the observation error $\breve{\mathbf{v}}_{s_t,L_R}^{(\mathrm{I,R})}$ in the presence of reverberation and absence of noise and, for the latter, also the LMPSC feature vector of the noise at time instant $t-L_R$ and the estimation error $\breve{\mathbf{w}}_{o_t,L_R}^{(\mathrm{I,R})}$ associated with the MMSE estimate of the LMPSC feature vector $\breve{\hat{\mathbf{s}}}_{t-L_R}^{(\mathrm{I,R})}$ of the reverberant speech signal. The corresponding recursive observation functions are defined by

$$\mathbf{s}_t^{(\mathrm{I})} = g_{s,L_R}^{(\mathrm{I,R})}\left(\mathbf{z}_t^{(\mathrm{I})}, \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}, \mathbf{s}_{t-L_R}^{(\mathrm{I})};\ \boldsymbol{\mu}_{\breve{\mathbf{h}}_{0:L_R-1}^{(\mathrm{I})}}\right) \tag{4.321}$$

$$:= \ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{(\mathrm{I})}+\boldsymbol{\mu}_{\breve{\mathbf{h}}_{t'}^{(\mathrm{I})}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}}\,\mathrm{e}^{\mathbf{s}_{t-L_R}^{(\mathrm{I})}}\right) + \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})} \tag{4.322}$$

for the presence of reverberation and the absence of noise and as

$$\mathbf{o}_t^{\text{(I)}} \approx g_{o,L_R}^{\text{(I,R)}}\left(\mathbf{z}_t^{\text{(I)}}, \mathbf{n}_{t-L_R}^{\text{(I)}}, \boldsymbol{\gamma}_t, \mathbf{v}_{s_t,L_R}^{\text{(I,R)}}, \mathbf{o}_{t-L_R}^{\text{(I)}};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{\text{(I)}}}\right), \tag{4.323}$$

$$:= \ln\left(\mathrm{e}^{\mathbf{v}_{s_t,L_R}^{\text{(I,R)}}} \circ \sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{\text{(I)}} + \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{\text{(I)}}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}} \mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{\text{(I,R)}}} + \mathrm{e}^{\mathbf{n}_t^{\text{(I)}}} \right.$$
$$\left. +2\,\mathrm{erf}\left(\boldsymbol{\gamma}_t\right) \circ \mathrm{e}^{\frac{\mathbf{v}_{s_t,L_R}^{\text{(I,R)}}}{2}} \circ \mathrm{e}^{\frac{\ln\left(\sum_{t'=0}^{L_R-1} \mathrm{e}^{\mathbf{x}_{t-t'}^{\text{(I)}} + \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{t'}^{\text{(I)}}}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}} \mathrm{e}^{\hat{\mathbf{s}}_{t-L_R}^{\text{(I,R)}}}\right) + \mathbf{n}_t^{\text{(I)}}}{2}} \right) \tag{4.324}$$

$$= f_{o,L_R}^{\text{(I,R)}}\left(\mathbf{z}_t^{\text{(I)}}, \mathbf{n}_{t-L_R}^{\text{(I)}}, \mathbf{o}_{t-L_R}^{\text{(I)}};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_R-1}^{\text{(I)}}}\right) + \mathbf{v}_{o_t,L_R}^{\text{(I,R)}} \tag{4.325}$$

for the presence of both reverberation of noise, respectively. The MMSE estimate $\hat{\mathbf{s}}_{t-L_R}^{\text{(I,R)}}$ occurring in the latter observation function thereby is a function of $\mathbf{o}_{t-L_R}^{\text{(I)}}$ and $\mathbf{n}_{t-L_R}^{\text{(I)}}$. The approximation (4.323) thereby assumes the recursion length to be large enough to neglect the estimation error $\mathbf{w}_{o_t,L_R}^{\text{(I,R)}}$ associated with $\hat{\mathbf{s}}_{t-L_R}^{\text{(I,R)}}$. Note that the recursion length L_R is assumed to be lower or equal to the number L_C of LMPSC feature vectors of the speech signal in the state vector $\mathbf{z}_t^{\text{(I)}}$ to allow the functional relation to be expressed in terms of the state vector.

Subsuming all variables that are not included in the state vector $\mathbf{z}_t^{\text{(I)}}$ under the *auxiliary* vector $\mathbf{a}_t$, any of the above functional relations may be written as

$$\mathbf{o}_t^{\text{(I)}} = g\left(\mathbf{z}_t^{\text{(I)}}, \mathbf{a}_t, \mathbf{o}_{t-L_R}^{\text{(I)}};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_H}^{\text{(I)}}}\right) \tag{4.326}$$

where $g \in \left\{g_s^{\text{(I)}}, g_o^{\text{(I)}}, g_y^{\text{(I)}}, g_{s,L_R}^{\text{(I,R)}}, g_{o,L_R}^{\text{(I,R)}}\right\}$.

In the ***i**terated **e**xtended* **K**ALMAN ***f**ilter* (IEKF) considered here, the VTS expansion of (4.326) will be truncated to linear terms [93]. While the IEKF assumes the ***h**igher **o**rder **t**erms* (*HOT*) in the VTS to be negligible, higher-order extended KALMAN filters employ the *HOT* up to a specified order to approximate the mean vector and the covariance matrix of the resulting linearization error, which in turn is approximated by a GAUSSIAN distributed RV. However, since the computational demand increases considerably with increasing VTS order and scales unfavorable with the dimension of the state vector [96], all but the ***s**econd-**o**rder **e**xtended* **K**ALMAN ***f**ilter* (SOEKF) are rarely used in practice. Since even the SOEKF suffers from an increased computational load in higher state vector dimensions, only the observation model in the presence of noise and the absence of reverberation with the corresponding observation function (4.319) is a candidate for its application. Its properties will thus be discussed in the respective context, only, and the following considerations focus on the IEKF, only.

With the Gaussian assumption on the predictive model-conditioned PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$ and the approximately Gaussian distributed auxiliary vector $\breve{\mathbf{a}}_t$ with mean vector $\boldsymbol{\mu}_{\breve{\mathbf{a}}_t}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t}$, the IEKF approximates the joint conditional PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$ of the state vector and the observation vector at iteration $\psi \in \{0,\ldots,\Psi-1\}$, where Ψ denotes the total number of employed IEKF iterations, by a Gaussian distribution according to

$$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})},\mathbf{o}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathsf{l})},m_t\right)$$
$$\approx \mathcal{N}\left(\begin{bmatrix}\mathbf{z}_t^{(\mathsf{l})}\\ \mathbf{o}_t^{(\mathsf{l})}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\\ \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} & \boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\\ \boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})},\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} & \boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\end{bmatrix}\right). \tag{4.327}$$

The VTS expansion point component related to the state vector $\mathbf{z}_t^{(\mathsf{l})}$ at iteration ψ will now be denoted by $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}$ and will be initialized to the mean of the predictive model-conditioned PDF, i.e., $\boldsymbol{\mu}^{[0]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}=\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$. Note that his may be either the merged GPB1 estimate (4.294) or the mixed IMM estimate (4.299).

The VTS expansion point component related to the auxiliary vector $\breve{\mathbf{a}}_t$ does not change with the IEKF iterations and will always be set to its expected value, i.e., $\boldsymbol{\mu}_{\breve{\mathbf{a}}_t}$. The new VTS expansion point related to the state vector $\mathbf{z}_t^{(\mathsf{l})}$ at iteration $\psi+1$ and the corresponding covariance matrix are then obtained as

$$\boldsymbol{\mu}^{(\psi+1)}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}=\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}+\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right)^{-1}\left[\mathbf{o}_t^{(\mathsf{l})}-\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right], \tag{4.328}$$

$$\boldsymbol{\Sigma}^{[\psi+1]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}=\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}-\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right)^{-1}\left(\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right)^{\dagger}. \tag{4.329}$$

Eq. (4.328) and (4.329) build the (extended) Kalman filter (measurement) update equations in their generic form[14].

The a posteriori PDF $p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}$ can be obtained from (4.327) and, as a consequence of the inherent Gaussian approximation, is also Gaussian and given by

$$p_{\breve{\mathbf{z}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\left(\mathbf{z}_t^{(\mathsf{l})}\middle|\mathbf{o}_{1:t}^{(\mathsf{l})},m_t\right)=\mathcal{N}\left(\mathbf{z}_t^{(\mathsf{l})};\boldsymbol{\mu}^{[\Psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t},\boldsymbol{\Sigma}^{[\Psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\right). \tag{4.330}$$

The mean vector $\boldsymbol{\mu}^{[\Psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}$ and the associated covariance matrix $\boldsymbol{\Sigma}^{[\Psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}$ are obtained from (4.328) and (4.329) after the last filter iteration, i.e., at $\psi=\Psi-1$.

[14] Note that the update for the covariance matrix needs to be calculated only after the last filter iteration, i.e., $\psi=\Psi-1$, for most of the IEKF variants employed here.

The observation likelihood $p_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}$, required to compute the a posteriori model probabilities $P_{\breve{m}_t|\breve{\mathbf{o}}_{1:t}^{(\mathrm{l})}}$ in (4.305), is finally given by

$$P_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}\left(\mathbf{o}_t^{(\mathrm{l})}\middle|\mathbf{o}_{1:t-1}^{(\mathrm{l})},m_t\right)=\mathcal{N}\left(\mathbf{o}_t^{(\mathrm{l})};\boldsymbol{\mu}^{[0]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}\middle|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t},\boldsymbol{\Sigma}^{[0]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}\middle|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}\right) \tag{4.331}$$

and computed during the initial IEKF iteration. The respective computation of the yet to be determined vectors $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}$ and matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}$ and $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})},\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}$ is subject to the observation function used. In its generic form, these parameters are given by

$$\begin{aligned}
\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t} &:= g\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathrm{l})},\breve{m}_t},\boldsymbol{\mu}_{\breve{\mathbf{a}}_t},\mathbf{o}_{t-L_R}^{(\mathrm{l})};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_H}^{(\mathrm{l})}}\right)\\
&\qquad + J^{[\psi]}_{g,\mathbf{z}_t^{(\mathrm{l})}|m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}-\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathrm{l})},\breve{m}_t}\right), &(4.332)\\
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t} &:= J^{[\psi]}_{g,\mathbf{z}_t^{(\mathrm{l})}|\breve{m}_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}\left(J^{[\psi]}_{g,\mathbf{z}_t^{(\mathrm{l})}|m_t}\right)^{\dagger}+J^{[\psi]}_{g,\mathbf{a}_t|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t}\left(J^{[\psi]}_{g,\mathbf{a}_t|m_t}\right)^{\dagger}, &(4.333)\\
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})},\breve{\mathbf{o}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t} &= \boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{l})},\breve{m}_t}\left(J^{[\psi]}_{g,\mathbf{z}_t^{(\mathrm{l})}|m_t}\right)^{\dagger}, &(4.334)
\end{aligned}$$

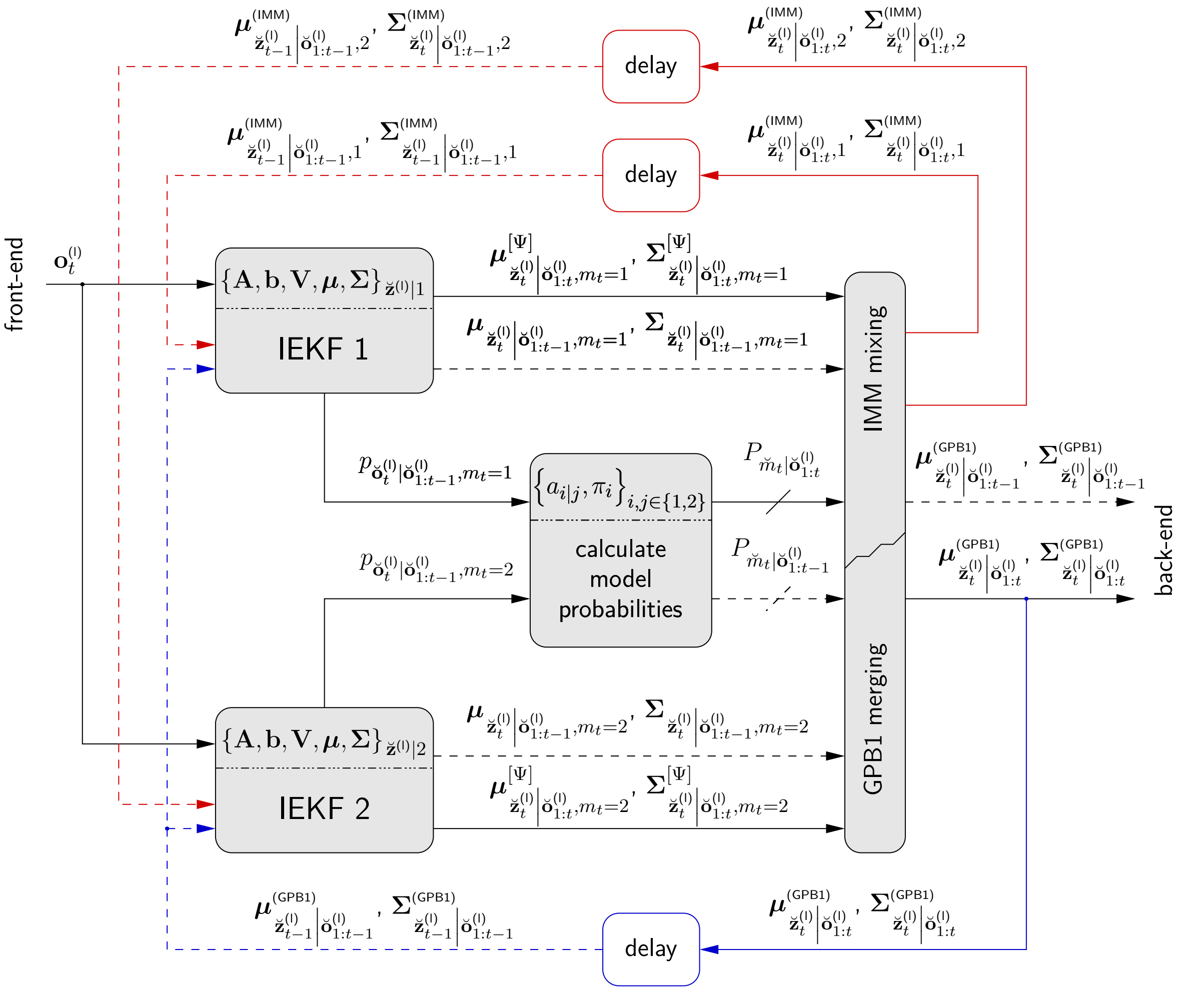
where

$$J^{[\psi]}_{g,\mathbf{z}_t^{(\mathrm{l})}|m_t} := \left.\frac{\partial g\left(\mathbf{z}_t^{(\mathrm{l})},\mathbf{a}_t,\mathbf{o}_{t-L_R}^{(\mathrm{l})};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_H}^{(\mathrm{l})}}\right)}{\partial\mathbf{z}_t^{(\mathrm{l})}}\right|_{\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathrm{l})},\breve{m}_t},\boldsymbol{\mu}_{\breve{\mathbf{a}}_t},\breve{\mathbf{o}}_{t-L_R}^{(\mathrm{l})}} \tag{4.335}$$

$$J^{[\psi]}_{g,\mathbf{a}_t|m_t} := \left.\frac{\partial g\left(\mathbf{z}_t^{(\mathrm{l})},\mathbf{a}_t,\mathbf{o}_{t-L_R}^{(\mathrm{l})};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}_{0:L_H}^{(\mathrm{l})}}\right)}{\partial\mathbf{a}_t}\right|_{\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathrm{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathrm{l})},\breve{m}_t},\boldsymbol{\mu}_{\breve{\mathbf{a}}_t},\breve{\mathbf{o}}_{t-L_R}^{(\mathrm{l})}} \tag{4.336}$$

denote the model-conditioned JACOBIAN matrices of the function g w.r.t. the vector specified in the subscript, both evaluated at the VTS expansion point. A brief overview of the VTS expansion is given in Appendix A.12.

Finally, an overview of the GPB1 and IMM multi-model inference scheme employing $M=2$ IEKFs is given in Fig. 4.30. In the displayed overviews, only the mean vectors and the covariance matrices of the predictive PDF and the a posteriori PDF after the GPB1 moment matching are forwarded to further back-end processing, e.g., calculation of the dynamic features and subsequent speech recognition. However, since both the GPB1 algorithm and the IMM algorithm only specify rules to create common/different initial conditions for all filters in the next filtering step, respectively, also the GMMs at the input



$\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},2}, \boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},2}$
delay
$\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},2}, \boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},2}$
$\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},1}, \boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},1}$
delay
$\boldsymbol{\mu}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},1}, \boldsymbol{\Sigma}^{(\mathsf{IMM})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},1}$
front-end
$\mathbf{o}^{(\mathsf{l})}_t$
$\{\mathbf{A}, \mathbf{b}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}_{\breve{\mathbf{z}}^{(\mathsf{l})}|1}$
IEKF 1
$\boldsymbol{\mu}^{[\Psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},m_t=1}, \boldsymbol{\Sigma}^{[\Psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},m_t=1}$
$\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=1}, \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=1}$
IMM mixing
$p_{\breve{\mathbf{o}}^{(\mathsf{l})}_t|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=1}$
$\left\{a_{i|j}, \pi_i\right\}_{i,j\in\{1,2\}}$
calculate model probabilities
$P_{\breve{m}_t|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}}$
$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}, \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$
$p_{\breve{\mathbf{o}}^{(\mathsf{l})}_t|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=2}$
$P_{\breve{m}_t|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$
$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}}, \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}}$
back-end
GPB1 merging
$\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=2}, \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},m_t=2}$
$\{\mathbf{A}, \mathbf{b}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}_{\breve{\mathbf{z}}^{(\mathsf{l})}|2}$
IEKF 2
$\boldsymbol{\mu}^{[\Psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},m_t=2}, \boldsymbol{\Sigma}^{[\Psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t},m_t=2}$
$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}, \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t-1}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}}$
delay
$\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}}, \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{z}}^{(\mathsf{l})}_{t}|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}}$

Figure 4.30: *Overview of the GPB1 (blue lines) and IMM (red lines) multi-model inference algorithms for $M = 2$ IEKFs. Signal paths employing only knowledge about the past observations are shown in dashed, the ones also incorporating the current observation in solid lines. Note that irrespective of the multi-model inference algorithm, always a single* Gaussian *with its moments calculated by the GPB1 moment matching (merging) is passed to the back-end for further processing.*

of the GPB1/IMM merging/mixing may be employed for further processing. Further note that for the IMM algorithm to produce uni-modal output distributions, the GPB1 moment matching is employed to approximate the respective predictive and a posteriori PDFs by Gaussian distributions.

In the following, the presented generic IEKF state update will be considered for the different observation models presented in Sec. 4.3/Sec. 4.4.

4.8.2.1 The Non-Recursive Observation Model in the Presence of Reverberation

For the non-recursive observation model in the presence of reverberation, the state vector is given by

$$\mathbf{z}_t^{\text{(l)}} = \left[\left(\mathbf{x}_t^{\text{(l)}}\right)^\dagger \; \cdots \; \left(\mathbf{x}_{t-L_C+1}^{\text{(l)}}\right)^\dagger \right]^\dagger \tag{4.337}$$

and the auxiliary vector $\mathbf{a}_t$ is composed as

$$\mathbf{a}_t := \left[\left(\mathbf{x}_{t-L_C}^{\text{(l)}}\right)^\dagger \; \cdots \; \left(\mathbf{x}_{t-L_H}^{\text{(l)}}\right)^\dagger \; \left(\mathbf{v}_{s_t}^{\text{(l,N)}}\right)^\dagger \right]^\dagger . \tag{4.338}$$

While the mean vector and the (diagonal) covariance matrix of the observation error $\breve{\mathbf{v}}_{s_t}^{\text{(l,N)}}$ can be obtained from training data, the mean vector and the covariance matrix of the LMPSC feature vector made up of the LMPSC feature vectors of the clean speech signal in the sequence $\breve{\mathbf{x}}_{t-L_H:t-L_C}^{\text{(l)}}$ can only be inferred during the filtering process. However, since the required LMPSC feature vectors are not part of the state vector, only lagged estimate of the respective mean vectors and covariance matrices are available. Hence, the VTS expansion point w.r.t. the auxiliary vector $\breve{\mathbf{a}}_t$ and the associated (block-diagonal[15]) covariance matrix, denoted by $\boldsymbol{\mu}_{\breve{\mathbf{a}}_t} = E\left[\breve{\mathbf{a}}_t\right]$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t} = E\left[\left(\breve{\mathbf{a}}_t - E\left[\breve{\mathbf{a}}_t\right]\right)\left(\breve{\mathbf{a}}_t - E\left[\breve{\mathbf{a}}_t\right]\right)^\dagger\right]$ are given by

$$\boldsymbol{\mu}_{\breve{\mathbf{a}}_t} := \begin{bmatrix} \boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_C} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-1}} \\ \boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_C-1} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-2}} \\ \vdots \\ \boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_H} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-(L_H-L_C+1)}} \\ \boldsymbol{\mu}_{\breve{\mathbf{v}}^{\text{(l,N)}}_{s_t}} \end{bmatrix}, \; \boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t} := \mathrm{blockdiag}\left(\begin{bmatrix} \boldsymbol{\Sigma}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_C} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-1}} \\ \boldsymbol{\Sigma}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_C-1} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-2}} \\ \vdots \\ \boldsymbol{\Sigma}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_H} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-(L_H-L_C+1)}} \\ \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{\text{(l,N)}}_{s_t}} \end{bmatrix} \right). \tag{4.339}$$

Note that for both GPB1 and IMM multi-model inference algorithms the moments computed with the moment matching of the GPB1 algorithm are employed here. Further, any correlation between the sub-vectors in the auxiliary state are neglected by considering the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t}$ to have block-diagonal structure. Any correlations between the auxiliary state and the state vector are neglected, too. For readability purposes, the short-hand notation

$$\boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_H:t-L_C}} := \boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_C} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-1}}, \ldots, \boldsymbol{\mu}^{\text{(GPB1)}}_{\breve{\mathbf{x}}^{\text{(l)}}_{t-L_H} \middle| \breve{\mathbf{o}}^{\text{(l)}}_{t-(L_H-L_C+1)}} \tag{4.340}$$

will be employed in the following.

[15] For readability purposes, the $\mathrm{blockdiag}(\cdot)$ operator is introduced here. It takes a *vector of matrices* as argument and builds up a matrix with the vector components on its diagonal and zeros on all other positions.

With (4.339), the vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{I})},\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}$ may now be computed as

$$\begin{aligned}
\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} :=& f_s^{(\mathsf{I})}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{I})},\breve{m}_t}, \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_H:t-L_C}};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{I})}_{0:L_H}}\right) \\
&+ J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{I})},\breve{m}_t}\right) + \boldsymbol{\mu}_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{I,N})}}, \qquad (4.341)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} :=& J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}\left(J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t}\right)^\dagger \\
&+ \sum_{t'=L_C}^{L_H} J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{x}_{t-t'}^{(\mathsf{I})}|m_t}\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_{t-t'}^{(\mathsf{I})}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{t-(t'-L_C+1)}}\left(J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{x}_t^{(\mathsf{I})}|m_t}\right)^\dagger + \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{I,N})}},
\end{aligned}$$

$$(4.342)$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{I})},\breve{\mathbf{o}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} := \boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{I})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}\left(J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t}\right)^\dagger. \qquad (4.343)$$

Thereby, use has been made of the block-diagonal structure of the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t}$ and the fact that the JACOBIAN matrix w.r.t. the observation error $\breve{\mathbf{v}}_{s_t}^{(\mathsf{I,N})}$ is just the identity matrix and that the identities

$$J^{[\psi]}_{g_s^{(\mathsf{I})},\mathbf{x}_{t-t'}^{(\mathsf{I})}|m_t} = J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{x}_{t-t'}^{(\mathsf{I})}|m_t}, \qquad J^{[\psi]}_{g_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t} = J^{[\psi]}_{f_s^{(\mathsf{I})},\mathbf{z}_t^{(\mathsf{I})}|m_t} \qquad (4.344)$$

hold for all $t' \in \{L_C, \ldots, L_H\}$ Further, the observation function $g_s^{(\mathsf{I})}$ has been expressed by means of the observation mapping $f_s^{(\mathsf{I})}$

4.8.2.2 The Non-Recursive Observation Model in the Presence of Reverberation and Background Noise

For the non-recursive observation model in the presence of both reverberation and background noise, the state vector is given by

$$\mathbf{z}_t^{(\mathsf{I})} = \left[\left(\mathbf{x}_t^{(\mathsf{I})}\right)^\dagger \ \ \ldots \ \ \left(\mathbf{x}_{t-L_C+1}^{(\mathsf{I})}\right)^\dagger \ \ \left(\mathbf{n}_t^{(\mathsf{I})}\right)^\dagger\right]^\dagger \qquad (4.345)$$

and the auxiliary vector $\mathbf{a}_t$ is composed as

$$\mathbf{a}_t := \left[\left(\mathbf{x}_{t-L_C}^{(\mathsf{I})}\right)^\dagger \ \ \ldots \ \ \left(\mathbf{x}_{t-L_H}^{(\mathsf{I})}\right)^\dagger \ \ (\boldsymbol{\gamma}_t)^\dagger \ \ \left(\mathbf{v}_{s_t}^{(\mathsf{I,N})}\right)^\dagger\right]^\dagger. \qquad (4.346)$$

Its mean vector and its covariance matrix are given by

$$\boldsymbol{\mu}_{\breve{\mathbf{a}}_t} := \begin{bmatrix} \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}} \\ \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C-1} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-2}} \\ \vdots \\ \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-(L_H-L_C+1)}} \\ \boldsymbol{\mu}_{\breve{\gamma}_t} = \mathbf{0} \\ \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{s_t}} \end{bmatrix}, \; \boldsymbol{\Sigma}_{\breve{\mathbf{a}}_t} := \operatorname{blockdiag}\left(\begin{bmatrix} \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-1}} \\ \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_C-1} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-2}} \\ \vdots \\ \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-(L_H-L_C+1)}} \\ \boldsymbol{\Sigma}_{\breve{\gamma}_t} \\ \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{s_t}} \end{bmatrix} \right). \quad (4.347)$$

With (4.347), the vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t, \breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t}$ may now be computed as

$$\begin{aligned} \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} &:= g^{(\mathsf{l})}_o \left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}, \breve{m}_t}, \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-L_H:t-L_C}}, \boldsymbol{\mu}_{\breve{\gamma}_t}, \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{s_t}};\; \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{l})}_{0:L_H}} \right) \\ &\quad + J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{z}^{(\mathsf{l})}_t | m_t} \left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t}, \breve{m}_t} \right), & (4.348) \\ \boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} &:= J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{z}^{(\mathsf{l})}_t | m_t} \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} \left(J^{[\psi]}_{g^{(\mathsf{l})}_s, \mathbf{z}^{(\mathsf{l})}_t | m_t} \right)^{\dagger} \\ &\quad + \sum_{t'=L_C}^{L_H} J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{x}^{(\mathsf{l})}_{t-t'} | m_t} \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{l})}_{t-t'} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{t-(t'-L_C+1)}} \left(J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{x}^{(\mathsf{l})}_t | m_t} \right)^{\dagger} \\ &\quad + J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{v}^{(\mathsf{l,N})}_{s_t} | m_t} \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{s_t}} \left(J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{v}^{(\mathsf{l,N})}_{s_t} | m_t} \right)^{\dagger} + J^{[\psi]}_{g^{(\mathsf{l})}_o, \gamma_t | m_t} \boldsymbol{\Sigma}_{\breve{\gamma}} \left(J^{[\psi]}_{g^{(\mathsf{l})}_o, \gamma_t | m_t} \right)^{\dagger}, \\ & & (4.349) \\ \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t, \breve{\mathbf{o}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} &:= \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{l})}_t \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t} \left(J^{[\psi]}_{g^{(\mathsf{l})}_o, \mathbf{z}^{(\mathsf{l})}_t | m_t} \right)^{\dagger}. & (4.350) \end{aligned}$$

Since the observation error $\mathbf{v}^{(\mathsf{l,N})}_{s_t}$ exponentially weights the contribution of the clean speech feature vectors in the observation function $g^{(\mathsf{l})}_o$ (compare (4.317)), the VTS approximation of the observation function $g^{(\mathsf{l})}_o$ may be even more susceptible to the chosen expansion point than it already is for the observation model in the presence of reverberation and the absence of noise.

To circumvent this issue, a VTS expansion of the observation mapping $f^{(\mathsf{l})}_o$ in (4.318) may be considered, instead. The error of truncating the VTS expansion after the linear term is than modeled by the observation error $\breve{\mathbf{v}}^{(\mathsf{l,N})}_{o_t}$. Application of the extended KALMAN filter thus calls for the computation of its mean vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{o_t} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t}$ and its covariance matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{l,N})}_{o_t} \middle| \breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1}, \breve{m}_t}$, i.e., conditioned on the past observations $\mathbf{o}^{(\mathsf{l})}_{1:t-1}$ and the current model state m_t in iteration ψ of the IEKF.

With the analysis of the distribution of the observation error components $\breve{v}_{ot}^{(\mathsf{I,N})}(q)$ carried out in Sec. 4.7.2, a naive approximation to the mean vector and the covariance matrix is given by

$$\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} \approx \boldsymbol{\mu}_{\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}} \tag{4.351}$$

$$\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t} \approx \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}}, \tag{4.352}$$

i.e., the mean vector and the covariance matrix of the observation error $\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}$ in the additional presence of noise are approximated by the mean vector and the covariance matrix of the observation error $\breve{\mathbf{v}}_{st}^{(\mathsf{I,N})}$ in the absence of noise. While the approximation of the mean vector by (4.351) may be considered quite reasonable, the approximation of the covariance matrix by (4.352) is clearly sub-optimal, especially at very low values of the IRNR. However, approximation (4.352) may be considered optimal in terms of practical considerations, as it i) does provide the maximum degree of uncertainty about the observation and ii) is insensitive to errors in the estimation of the state vector. The latter is of practical importance, since the quality of the state estimate (and its prediction) controls the quality of the linear truncation of the VTS expansion[16].

Since the observation error $\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}$ for a given IRNR may be approximated by a Gaussian distribution, whose mean vector and covariance matrix are computed by (4.122) and (4.123), which are related to the auxiliary functions $\xi\left(\mathbf{r}_t^{(\mathsf{I,N})}\right)$ and $\zeta\left(\mathbf{r}_t^{(\mathsf{I,N})}\right)$ via (4.124) and (4.125), an alternative approximation of the mean vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}$ and the covariance matrix $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}_{ot}^{(\mathsf{I,N})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t}$ may be obtained by calculating the conditional expectation of (4.124) and (4.125). In particular, this amounts to determining the expectation values

$$E\left[\xi\left(\breve{r}_t^{(\mathsf{I})}(q)\right)\Big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t;\psi\right], \tag{4.353}$$

$$E\left[\xi\left(\breve{r}_t^{(\mathsf{I})}(q)\right)\xi\left(\breve{r}_t^{(\mathsf{I})}(q')\right)\Big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t;\psi\right], \tag{4.354}$$

$$E\left[\zeta\left(\breve{r}_t^{(\mathsf{I})}(q)\right)\zeta\left(\breve{r}_t^{(\mathsf{I})}(q')\right)\Big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{I})},\breve{m}_t;\psi\right], \tag{4.355}$$

for which no closed-form solution exists. Both $\xi\left(r_t^{(\mathsf{I})}(q)\right)$ and $\zeta\left(r_t^{(\mathsf{I})}(q)\right)$ may, however, also be written as functions of $\mathrm{e}^{\frac{\ln(10)}{10}r_t^{(\mathsf{I})}(q)}$ (compare (4.106) and (4.107)). Denoting those functions by $\tilde{\xi}\left(\mathrm{e}^{\frac{\ln(10)}{10}r_t^{(\mathsf{I})}(q)}\right)$ and $\tilde{\zeta}\left(\mathrm{e}^{\frac{\ln(10)}{10}r_t^{(\mathsf{I})}(q)}\right)$, the sub-optimal solution considered here approximates the expectation of these functions by evaluating the functions at the expec-

[16] The quality of the state prediction may, e.g., be measured in terms of the *spectral radius* of the model-conditioned predictive PDF

tation $E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]$, i.e.,

$$\begin{aligned}
&E\left[\xi\left(\breve{r}_t^{(\mathsf{l})}(q)\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&=E\left[\tilde{\xi}\left(\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&\approx\tilde{\xi}\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\right) \qquad (4.356)\\
&E\left[\xi\left(\breve{r}_t^{(\mathsf{l})}(q)\right)\xi\left(\breve{r}_t^{(\mathsf{l})}\left(q'\right)\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&=E\left[\tilde{\xi}\left(\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\right)\tilde{\xi}\left(\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q')}\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&\approx\tilde{\xi}\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\right)\tilde{\xi}\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q')}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\right), \qquad (4.357)\\
&E\left[\zeta\left(\breve{r}_t^{(\mathsf{l})}(q)\right)\zeta\left(\breve{r}_t^{(\mathsf{l})}\left(q'\right)\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&=E\left[\tilde{\zeta}\left(\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\right)\tilde{\zeta}\left(\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q')}\right)\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&\approx\tilde{\zeta}\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\right)\tilde{\zeta}\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q')}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\right). \qquad (4.358)
\end{aligned}$$

With the Gaussian approximation to the joint distribution of all involved RVs, the required expectation value may be computed in closed form as

$$\begin{aligned}
&E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{l})}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right]\\
&=E\left[\sum_{t'=0}^{L_H}\mathrm{e}^{\breve{x}^{(\mathsf{l})}_{t-t'}(q)+\mu_{\breve{h}^{(\mathsf{l})}_{t'}(q)}-\breve{n}^{(\mathsf{l})}_t(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t;\psi\right] \qquad (4.359)\\
&=\sum_{t'=0}^{L_C-1}\mathrm{e}^{\mu^{[\psi]}_{\breve{x}^{(\mathsf{l})}_{t-t'}(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}+\mu_{\breve{h}^{(\mathsf{l})}_{t'}(q)}-\mu^{[\psi]}_{\breve{n}^{(\mathsf{l})}_t(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}}\\
&\qquad\mathrm{e}^{\frac{1}{2}\left(\sigma^{2,[\psi]}_{\breve{x}^{(\mathsf{l})}_{t-t'}(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}+\sigma^{2,[\psi]}_{\breve{n}^{(\mathsf{l})}_t(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}-2\sigma^{[\psi]}_{\breve{x}^{(\mathsf{l})}_t(q),\breve{n}^{(\mathsf{l})}_t(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}\right)}\\
&\quad+\sum_{t'=L_C}^{L_H}\mathrm{e}^{\mu^{(\mathsf{GPB1})}_{\breve{x}^{(\mathsf{l})}_{t-t'}(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-(t'-L_C+1)}\right.}+\mu_{\breve{h}^{(\mathsf{l})}_{t'}(q)}-\mu^{[\psi]}_{\breve{n}^{(\mathsf{l})}_t(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}}\\
&\qquad\mathrm{e}^{\frac{1}{2}\left(\sigma^{2,(\mathsf{GPB1})}_{\breve{x}^{(\mathsf{l})}_{t-t'}(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-(t'-L_C+1)}\right.}+\sigma^{2,[\psi]}_{\breve{n}^{(\mathsf{l})}_t(q)\left|\breve{\mathrm{o}}^{(\mathsf{l})}_{1:t-1},\breve{m}_t\right.}\right)}, \qquad (4.360)
\end{aligned}$$

where again the relation of the mean of a log-normally distributed RV to the mean and the variance of the corresponding normally distributed variable has been employed (see Appendix A.5 for details). Further, any correlation between the state vector components and those of the auxiliary vector are neglected. Although the quality of approximating (4.353)-(4.355) by (4.356)-(4.358) may be debatable, (4.359) exhibits some very desirable properties: If the uncertainty about the VTS expansion points is *large* (large spectral radius of the associated covariance matrices), (4.356) and (4.357) will approximately be one and

(4.358) will approximately be zero and the mean vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ and the covariance matrix $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ of the error due to the truncation of the VTS expansion to linear terms will be approximated by (4.351) and (4.352), respectively. On the other hand, as the uncertainty about the VTS expansion point becomes sufficiently low (small spectral radius of the associated covariance matrices), the mean vector and the covariance matrix of the linearization error are approximately equal to the true observation error's moments given in (4.122) and (4.123), respectively. In general, the variances of the linearization error are thus rather overestimated than underestimated by employing (4.359) and it may be assumed that the filtering process benefits from this rather *conservative* approach.

Since employing the approximation (4.359) in the computation of (4.356)-(4.358) may also be considered as employing the estimate $\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}$ of the IRNR $\mathbf{r}_t^{(\mathsf{I,N})}$, composed of the components

$$\hat{r}_t^{(\mathsf{I,N}),[\psi]}(q) := \frac{10}{\ln(10)} \ln\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}_t^{(\mathsf{I})}(q)}\bigg|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t;\psi\right]\right), \tag{4.361}$$

in the computation of the mean vector and covariance matrix (4.122) and (4.123), respectively, the final estimate of the mean vector and the covariance matrix of the linearization error are given by

$$\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} \approx \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}}\left(\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}\right) \tag{4.362}$$

$$\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} \approx \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}}\left(\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}\right). \tag{4.363}$$

In either case, the vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t,\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ may now be computed as

$$\begin{aligned}\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} &:= f_o^{(\mathsf{I})}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t},\breve{m}_t}, \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{I})}_{t-L_H:t-L_C}};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{I})}_{0:L_H}}\right)\\ &\quad + J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{z}^{(\mathsf{I})}_t|m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t},\breve{m}_t}\right) + \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t},\end{aligned} \tag{4.364}$$

$$\begin{aligned}\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} &:= J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{z}^{(\mathsf{I})}_t|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}\left(J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{z}^{(\mathsf{I})}_t|m_t}\right)^{\dagger}\\ &\quad + \sum_{t'=L_C}^{L_H} J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{x}^{(\mathsf{I})}_{t-t'}|m_t}\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}^{(\mathsf{I})}_{t-t'}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{t-(t'-L_C+1)}}\left(J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{x}^{(\mathsf{I})}_t|m_t}\right)^{\dagger} + \boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t},\end{aligned} \tag{4.365}$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t,\breve{\mathbf{o}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t} := \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\big|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}\left(J^{[\psi]}_{f_o^{(\mathsf{I})},\mathbf{z}^{(\mathsf{I})}_t|m_t}\right)^{\dagger}. \tag{4.366}$$

Since the mean of the observation error in the presence of reverberation and the absence of noise is approximately zero, the linearization of the observation function $f_o^{(\mathsf{I})}$ (see (4.316)) and the observation mapping $g_o^{(\mathsf{I})}$ (see (4.111) majorly differs in the way the covariance matrix $\mathbf{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t}$ is computed. This may best be seen by comparing (4.349) and (4.365) for the first IEKF iteration while assuming equal VTS expansion points.

With the JACOBIAN matrices of the observation function $f_o^{(\mathsf{I})}$ and the observation mapping $g_o^{(\mathsf{I})}$ being approximately equal if evaluated at the respective linearization vectors, the first two summands in (4.349) and (4.365) are approximately equal. The three proposed IEKF schemes thus compute the additional contribution to the covariance matrix $\mathbf{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t}$, further denoted by $\Delta\mathbf{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t}$, differently. In (4.349) it is computed as

$$\Delta\mathbf{\Sigma}^{\mathsf{AUG},[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} = J^{[\psi]}_{g_o^{(\mathsf{I})}, \mathbf{v}^{(\mathsf{I,N})}_{s_t} | m_t} \mathbf{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{s_t}} \left(J^{[\psi]}_{g_o^{(\mathsf{I})}, \mathbf{v}^{(\mathsf{I,N})}_{s_t} | m_t} \right)^{\dagger} + J^{[\psi]}_{g_o^{(\mathsf{I})}, \boldsymbol{\gamma}_t | m_t} \mathbf{\Sigma}_{\breve{\boldsymbol{\gamma}}} \left(J^{[\psi]}_{g_o^{(\mathsf{I})}, \boldsymbol{\gamma}_t | m_t} \right)^{\dagger} \tag{4.367}$$

$$= \mathrm{diag}\left(\xi \left(\mathbf{r}_t^{(\mathsf{I,N}),[\psi]} \right) \right) \mathbf{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{s_t}} \mathrm{diag}\left(\xi \left(\mathbf{r}_t^{(\mathsf{I,N}),[\psi]} \right) \right)$$
$$+ \frac{16}{\pi} \mathrm{diag}\left(\zeta \left(\mathbf{r}_t^{(\mathsf{I,N}),[\psi]} \right) \right) \mathbf{\Sigma}_{\breve{\boldsymbol{\gamma}}} \mathrm{diag}\left(\xi \left(\mathbf{r}_t^{(\mathsf{I,N}),[\psi]} \right) \right), \tag{4.368}$$

where $\mathbf{r}_t^{(\mathsf{I,N}),[\psi]}$ denotes the IRNR that is obtained by plugging the components of the linerization vector into the definition of the IRNR given in (4.108). Further, the diag operator builds a diagonal matrix from the vector-valued argument. Note that $\mathbf{r}_t^{(\mathsf{I,N}),[\psi]}$ and $\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}$ differ in the fact that the latter also takes into account the covariance matrix associated with the VTS expansion point, which is completely ignored by the former one.

In (4.365), it is given by

$$\Delta\mathbf{\Sigma}^{\mathsf{TV},[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} = \mathbf{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t} \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} \approx \mathbf{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}} \left(\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]} \right) \tag{4.369}$$

and, as the most conservative setting, may also be approximated by

$$\Delta\mathbf{\Sigma}^{\mathsf{TI},[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} = \mathbf{\Sigma}^{[\psi]}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t} \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} \tag{4.370}$$

$$\approx \mathbf{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{o_t}} \left(\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]} = \infty \right) = \mathbf{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,N})}_{s_t}}. \tag{4.371}$$

Thereby, to distinguish between the different IEKF schemes, the first is denoted by AUG, since the linearization considers the observation error and the vector of phase factors in the (AUG)gmented state vector. The second is denoted by TV, since it in essence directly takes into account the (T)ime-(V)ariant characteristic of the statistics of the observation error in the presence of reverberation and noise. Consequently, since the last (conservative) approximation considers the observation error statistics in the presence of both reverberation and noise to be (T)ime-(I)nveriant, it is denoted by TI.

To illustrate the difference between the three IEKF schemes, the *additional* variances are given in Fig. 4.31 as a function of the IRNRs $\mathbf{r}_t^{(\mathsf{I,N}),[\psi]}$ and $\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}$, respectively, and different mel indices q. While the additional variance of both AUG and TV converges to

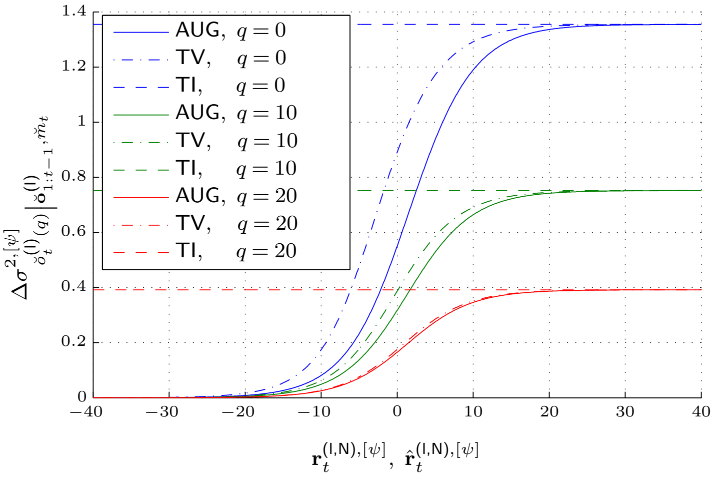


Figure 4.31: *The* additional *variances computed by the three IEKF schemes AUG (solid lines), TV (dash-dotted lines) and TI (dashed lines) as a function of the IRNRs $\mathbf{r}_t^{(I,N),[\psi]}$ and $\hat{\mathbf{r}}_t^{(I,N),[\psi]}$, respectively, and for $q \in \{0,10,20\}$.*

the variance of the TI scheme for very high and very low IRNR values, they considerably differ at mid-level IRNR values. However, since the computation of $\hat{\mathbf{r}}_t^{(\mathsf{I,N}),[\psi]}$ takes in account the covariance matrix associated with the VTS expansion point it will, in tendency, always be larger than the IRNR $\mathbf{r}_t^{(\mathsf{I,N}),[\psi]}$, which is just employing the VTS expansion point itself.

4.8.2.3 The (Non-Recursive) Observation Model in the Absence of Reverberation and the Presence of Background Noise

For the (non-recursive) observation model in the absence of reverberation and the presence of background noise, the state vector is given by

$$\mathbf{z}_t^{(\mathsf{l})} = \left[\left(\mathbf{x}_t^{(\mathsf{l})}\right)^{\dagger} \quad \left(\mathbf{n}_t^{(\mathsf{l})}\right)^{\dagger} \right]^{\dagger}, \tag{4.372}$$

i.e., $L_C = 1$, and the auxiliary vector $\mathbf{a}_t$ is just

$$\mathbf{a}_t := \boldsymbol{\gamma}_t \tag{4.373}$$

with mean vector $\boldsymbol{\mu}_{\breve{\boldsymbol{\gamma}}_t} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\breve{\boldsymbol{\gamma}}_t}$. The vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}$ may now be computed as

$$\begin{aligned}
\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= f_y^{(\mathsf{l})}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}, \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{n}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\right) \\
&\quad + J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})},\breve{m}_t}\right), && (4.374)\\
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\right)^\dagger + J^{[\psi]}_{g_y^{(\mathsf{l})},\boldsymbol{\gamma}_t|m_t}\boldsymbol{\Sigma}_{\breve{\boldsymbol{\gamma}}}\left(J^{[\psi]}_{g_y^{(\mathsf{l})},\boldsymbol{\gamma}_t|m_t}\right)^\dagger, && (4.375)\\
\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= \boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\right)^\dagger. && (4.376)
\end{aligned}$$

Thereby, use has been made of the identity

$$J^{[\psi]}_{g_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t} = J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}, \tag{4.377}$$

which arises from the zero-mean of the (transformed) vector of phase factors RV. Further, the observation function $g_y^{(\mathsf{l})}$ has been expressed by means of the observation mapping $f_y^{(\mathsf{l})}$

As the total dimension of the *augmented* state vector $\boldsymbol{\chi}_t^{(\mathsf{l})} \in \mathbb{R}^{3Q}$, defined by

$$\boldsymbol{\chi}_t^{(\mathsf{l})} := \left[\left(\breve{\mathbf{z}}_t^{(\mathsf{l})}\right)^\dagger \quad (\breve{\mathbf{a}}_t)^\dagger\right]^\dagger \tag{4.378}$$

is considerably lower than for the observation models in the presence of noise, the IEKF update equations (4.374)-(4.376) may also be replaced by the update equations of the SOEKF, which are given by [93]

$$\begin{aligned}
\boldsymbol{\mu}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= f_y^{(\mathsf{l})}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}, \boldsymbol{\mu}_{\breve{\mathbf{n}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{Q}\mathbf{e}_i \operatorname{tr}\left(H^i_{g_y^{(\mathsf{l})},\boldsymbol{\chi}_t^{(\mathsf{l})}}\boldsymbol{\Sigma}_{\boldsymbol{\chi}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right) && (4.379)\\
\boldsymbol{\Sigma}_{\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= J_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(J_{g_s^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\right)^\dagger + J_{g_y^{(\mathsf{l})},\boldsymbol{\gamma}_t|m_t}\boldsymbol{\Sigma}_{\breve{\boldsymbol{\gamma}}}\left(J_{g_y^{(\mathsf{l})},\boldsymbol{\gamma}_t|m_t}\right)^\dagger && (4.380)\\
&\quad + \frac{1}{2}\sum_{i,j=1}^{Q}\mathbf{e}_i(\mathbf{e}_j)^\dagger \operatorname{tr}\left(H^i_{g_y^{(\mathsf{l})},\boldsymbol{\chi}_t^{(\mathsf{l})}}\boldsymbol{\Sigma}_{\boldsymbol{\chi}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}H^j_{g_y^{(\mathsf{l})},\boldsymbol{\chi}_t^{(\mathsf{l})}}\boldsymbol{\Sigma}_{\boldsymbol{\chi}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\right) && (4.381)\\
\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t} &:= \boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathsf{l})}\big|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{m}_t}\left(J^{[\psi]}_{f_y^{(\mathsf{l})},\mathbf{z}_t^{(\mathsf{l})}|m_t}\right)^\dagger. && (4.382)
\end{aligned}$$

Thereby $H^{i}_{g_y^{(\mathrm{I})},\boldsymbol{\chi}_t^{(\mathrm{I})}}$ denotes the HESSIAN matrix of the i-th component of the function $g_y^{(\mathrm{I})}\left(\boldsymbol{\chi}_t^{(\mathrm{I})}\right)$, evaluated at the VTS expansion vector, and $\mathbf{e}_i$ denotes the i-th CARTESIAN basis vector in $\mathbb{R}^{3Q}$. Further, the operator $\mathrm{tr}\left(\cdot\right)$ computes the trace of a given matrix. Note that the SOEKF may also be iterated, however, the computational effort increases considerably, and, as such, is not considered in this work.

Comparing (4.379) and (4.380) with the corresponding updated equations of the IEKF (4.374) and (4.375) for the first iteration, it can be seen, that the second central moments of the augmented state vector $\breve{\boldsymbol{\chi}}_t^{(\mathrm{I})}$ are used to correct the prediction of the observation $\boldsymbol{\mu}_{\breve{\mathbf{o}}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$ and that the fourth central moments of the augmented state vector $\breve{\boldsymbol{\chi}}_t^{(\mathrm{I})}$ are used to correct to prediction covariance $\boldsymbol{\Sigma}_{\breve{\mathbf{o}}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$.

At this point, it is worth noting that the compact form of the correction term in (4.380) is only valid for a GAUSSIAN distributed augmented state vector, which is the reason why the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ has been replaced by the transformed vector of phase factors $\breve{\boldsymbol{\gamma}}_t$ in the first place. For a practical realization of the corrections in (4.379) and (4.380) the special structure of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\chi}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$, given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\chi}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.} := \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{z}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\breve{\boldsymbol{\gamma}}} \end{bmatrix}, \tag{4.383}$$

and the fact that the i-component of the observation function $g_y^{(\mathrm{I})}$ only depends on the i-th components of the LMPSC feature vector of the speech signal, the noise signal and the transformed vector of phase factors, i.e., that the HESSIAN matrices as such are rather sparse, should be employed.

4.8.2.4 The Recursive Observation Model in the Presence of Reverberation and the Absence of Background Noise

For the recursive observation model in the presence of reverberation, the state vector is given by

$$\mathbf{z}_t^{(\mathrm{I})} = \left[\left(\mathbf{x}_t^{(\mathrm{I})}\right)^{\dagger} \quad \ldots \quad \left(\mathbf{x}_{t-L_R+1}^{(\mathrm{I})}\right)^{\dagger}\right]^{\dagger} \tag{4.384}$$

and the auxiliary vector $\mathbf{a}_t$ is just

$$\mathbf{a}_t := \mathbf{v}_{s_t,L_R}^{(\mathrm{I,R})}. \tag{4.385}$$

Its mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathrm{I,R})}}$ and the (diagonal) covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_{s,L_R}^{(\mathrm{I,R})}}$ may be obtained from (artificially reverberated) training data.

The vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}_t^{(\mathrm{I})},\breve{\mathbf{o}}_t^{(\mathrm{I})}\left|\breve{\mathbf{o}}_{1:t-1}^{(\mathrm{I})},\breve{m}_t\right.}$ may now be computed as

$$\begin{aligned}\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} &:= f^{(\mathsf{I,R})}_{s,L_R}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t}, \breve{m}_t}, \mathbf{o}^{(\mathsf{I})}_{t-L_R};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{I})}_{0:L_R}}\right) \\ &\quad + J^{[\psi]}_{f^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t}, \breve{m}_t}\right) + \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}, \end{aligned} \tag{4.386}$$

$$\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} := J^{[\psi]}_{f^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t} \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} \left(J^{[\psi]}_{f^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t}\right)^{\dagger} + \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}, \tag{4.387}$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t, \breve{\mathbf{o}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} := \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1}, \breve{m}_t} \left(J^{[\psi]}_{f^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t}\right)^{\dagger}. \tag{4.388}$$

Thereby, use has been made of the identities

$$J^{[\psi]}_{g^{(\mathsf{I,R})}_{s,L_R}, \mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R} | m_t} = \mathbf{I}, \qquad J^{[\psi]}_{g^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t} = J^{[\psi]}_{f^{(\mathsf{I,R})}_{s,L_R}, \mathbf{z}^{(\mathsf{I})}_t | m_t}. \tag{4.389}$$

Further, the observation function $g^{(\mathsf{I,R})}_{s,L_R}$ has been expressed by means of the observation mapping $f^{(\mathsf{I,R})}_{s,L_R}$. Note that in comparison to (4.341) and (4.342), even for $L_C = L_R$, the evaluation of (4.386) and (4.387) is considerably more efficient in terms of computational complexity and also memory requirement.

4.8.2.5 The Recursive Observation Model in the Presence of Reverberation and Background Noise

For the recursive observation model in the presence of both reverberation and background noise, the state vector is given by

$$\mathbf{z}^{(\mathsf{I})}_t = \left[\left(\mathbf{x}^{(\mathsf{I})}_t\right)^{\dagger} \ \cdots \ \left(\mathbf{x}^{(\mathsf{I})}_{t-L_R+1}\right)^{\dagger} \ \left(\mathbf{n}^{(\mathsf{I})}_t\right)^{\dagger}\right]^{\dagger} \tag{4.390}$$

and the auxiliary vector $\mathbf{a}_t$ is composed as

$$\mathbf{a}_t := \left[\left(\mathbf{n}^{(\mathsf{I})}_{t-L_R}\right)^{\dagger} \ \left(\boldsymbol{\gamma}_t\right)^{\dagger} \ \left(\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}\right)^{\dagger}\right]^{\dagger}. \tag{4.391}$$

Its mean vector and its covariance matrix are given by

$$\boldsymbol{\mu}_{\breve{\mathbf{a}}_t} := \begin{bmatrix} \boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R} \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{t-L_R}} \\ \boldsymbol{\mu}_{\breve{\boldsymbol{\gamma}}_t} = \mathbf{0} \\ \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}} \end{bmatrix}, \ \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}} := \mathrm{blockdiag}\left(\begin{bmatrix} \boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R} \big| \breve{\mathbf{o}}^{(\mathsf{I})}_{t-L_R}} \\ \boldsymbol{\Sigma}_{\breve{\boldsymbol{\gamma}}_t} \\ \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}} \end{bmatrix}\right). \tag{4.392}$$

Note that, since the LMPSC feature vector of the noise at time instant $t - L_R$ is not part of the state vector, only the lag-L_R estimate and the associated covariance matrix are available for the VTS expansion.

With (4.392), the vector $\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ and the matrices $\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t,\breve{\mathbf{o}}^{(\mathsf{I})}_t|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t}$ may now be computed as

$$\begin{aligned}
\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.} &:= g^{(\mathsf{I,R})}_{o,L_R}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t},\breve{m}_t\right.},\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{n}}^{(\mathsf{I})}_{t-L_R}\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{t-L_R}\right.},\boldsymbol{\mu}_{\breve{\gamma}_t},\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}},\mathbf{o}^{(\mathsf{I})}_{t-L_R};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathsf{I})}_{0:L_R}}\right)\\
&\quad + J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{z}^{(\mathsf{I})}_t|m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.}-\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t},\breve{m}_t\right.}\right), &(4.393)\\
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.} &:= J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{z}^{(\mathsf{I})}_t|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.}\left(J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{z}^{(\mathsf{I})}_t|m_t}\right)^{\dagger}\\
&\quad + J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}|m_t}\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I,R})}_{s_t,L_R}}\left(J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}|m_t}\right)^{\dagger}\\
&\quad + J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\gamma_t|m_t}\boldsymbol{\Sigma}_{\breve{\gamma}}\left(J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\gamma_t|m_t}\right)^{\dagger}, &(4.394)\\
\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t,\breve{\mathbf{o}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.} &:= \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathsf{I})}_t\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t\right.}\left(J^{[\psi]}_{g^{(\mathsf{I,R})}_{o,L_R},\mathbf{z}^{(\mathsf{I})}_t|m_t}\right)^{\dagger}. &(4.395)
\end{aligned}$$

As in the non-recursive variant, the observation error $\mathbf{v}^{(\mathsf{I,R})}_{s_t,L_R}$ exponentially weights the contribution of the clean speech feature vectors in the observation function $g^{(\mathsf{I,R})}_{o,L_R}$ (compare (4.323)) and the VTS approximation of the observation function $g^{(\mathsf{I,R})}_{o,L_R}$ may be even more susceptible to the chosen expansion point than it already is for the observation model in the presence of reverberation and the absence of noise.

The alternative variants thus (again) consider the VTS expansion of the observation mapping $f^{(\mathsf{I,R})}_{o,L_R}$ instead of the observation function $g^{(\mathsf{I,R})}_{o,L_R}$ and approximate the mean vector and the covariance matrix of the remaining terms by employing a point estimate of the IRNR vector $\breve{\mathbf{r}}^{(\mathsf{I,R})}_{t,L_R}$ in the computation of the observation error's mean vector and covariance matrix according to (4.209) and (4.210), respectively.

The update equations for the prediction of the observation, the associated covariance matrix and the cross-covariance matrix between the state vector and the observation are thus given by

$$
\begin{aligned}
\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} &:= f^{(\mathrm{I,R})}_{o,L_R}\left(\boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t}, \breve{m}_t}, \boldsymbol{\mu}^{(\mathrm{GPB1})}_{\breve{\mathbf{n}}^{(\mathrm{I})}_{t-L_R} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{t-L_R}}, \mathbf{o}^{(\mathrm{I})}_{t-L_R};\ \boldsymbol{\mu}_{\breve{\bar{\mathbf{h}}}^{(\mathrm{I})}_{0:L_R}}\right) \\
&\quad + J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{z}^{(\mathrm{I})}_t | m_t}\left(\boldsymbol{\mu}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} - \boldsymbol{\mu}^{[\psi]}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t}, \breve{m}_t}\right) + \boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathrm{I,R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathrm{I,R}),[\psi]}_{t,L_R}\right),
\end{aligned} \tag{4.396}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}^{[\psi]}_{\breve{\mathbf{o}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} &:= J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{z}^{(\mathrm{I})}_t | m_t} \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t}\left(J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{z}^{(\mathrm{I})}_t | m_t}\right)^{\dagger} \\
&\quad + J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{n}^{(\mathrm{I})}_{t-L_R} | m_t} \boldsymbol{\Sigma}^{(\mathrm{GPB1})}_{\breve{\mathbf{n}}^{(\mathrm{I})}_{t-L_R} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{t-L_R}}\left(J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{n}^{(\mathrm{I})}_{t-L_R} | m_t}\right)^{\dagger} \\
&\quad + \boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathrm{I,R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathrm{I,R}),[\psi]}_{t,L_R}\right),
\end{aligned} \tag{4.397}
$$

$$
\boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t, \breve{\mathbf{o}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} := \boldsymbol{\Sigma}_{\breve{\mathbf{z}}^{(\mathrm{I})}_t \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t}\left(J^{[\psi]}_{f^{(\mathrm{I,R})}_{o,L_R}, \mathbf{z}^{(\mathrm{I})}_t | m_t}\right)^{\dagger}. \tag{4.398}
$$

The components of the estimate of the IRNR vector $\hat{\mathbf{r}}^{(\mathrm{I,R}),[\psi]}_{t,L_R}$ are given by

$$
\hat{r}^{(\mathrm{I,R}),[\psi]}_{t,L_R}(q) := \frac{10}{\ln(10)} \ln\left(E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}^{(\mathrm{I,R})}_{t,L_R}(q)} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t; \psi\right]\right), \tag{4.399}
$$

where the required expectation is given by

$$
\begin{aligned}
&E\left[\mathrm{e}^{\frac{\ln(10)}{10}\breve{r}^{(\mathrm{I,R})}_{t,L_R}(q)} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t; \psi\right] \\
&= E\left[\sum_{t'=0}^{L_R-1} \mathrm{e}^{x^{(\mathrm{I})}_{t-t'}(q) - n^{(\mathrm{I})}_t(q) + \mu_{\breve{\bar{h}}^{(\mathrm{I})}_{t'}(q)}} + \mathrm{e}^{-\frac{2L_R B}{\tau_h}} \mathrm{e}^{\hat{s}^{(\mathrm{I,R})}_{t-L_R}(q) - n^{(\mathrm{I})}_t(q)} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t; \psi\right] \qquad (4.400) \\
&= \sum_{t'=0}^{L_R-1} \mathrm{e}^{\mu^{[\psi]}_{\breve{x}^{(\mathrm{I})}_{t-t'}(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} - \mu^{[\psi]}_{\breve{n}^{(\mathrm{I})}_{t-t'}(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} + \mu_{\breve{\bar{h}}^{(\mathrm{I})}_{t'}(q)}} \\
&\qquad \cdot \mathrm{e}^{\frac{1}{2}\left(\sigma^{2,[\psi]}_{\breve{x}^{(\mathrm{I})}_{t-t'}(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} + \sigma^{2,[\psi]}_{\breve{n}^{(\mathrm{I})}_{t-t'}(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} - 2\sigma^{2,[\psi]}_{\breve{x}^{(\mathrm{I})}_{t-t'}(q), \breve{n}^{(\mathrm{I})}_{t-t'}(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t}\right)} \\
&\quad + \mathrm{e}^{-\frac{2L_R B}{\tau_h}} E\left[\mathrm{e}^{\hat{s}^{(\mathrm{I,R})}_{t-L_R}(q)} \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t; \psi\right] \mathrm{e}^{-\mu^{[\psi]}_{n^{(\mathrm{I})}_t(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t} + \frac{1}{2}\sigma^{2,[\psi]}_{n^{(\mathrm{I})}_t(q) \middle| \breve{\mathbf{o}}^{(\mathrm{I})}_{1:t-1}, \breve{m}_t}}. \qquad (4.401)
\end{aligned}
$$

For the last equality, $\breve{n}^{(\mathrm{I})}_t(q)$ and $\breve{n}^{(\mathrm{I})}_{t-L_R}(q)$ (as the only RV in the MMSE estimate (4.193)) are assumed to be uncorrelated. The expectation of the MMSE estimate $\mathrm{e}^{\hat{s}^{(\mathrm{I,R})}_{t-L_R}(q)}$ will be approximated by

$$\begin{aligned}
&E\left[\mathrm{e}^{\hat{s}^{(\mathsf{I},\mathsf{R})}_{t-L_R}(q)}\middle|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-1},\breve{m}_t;\psi\right]\\
&\approx\left(2\sigma^2_{\breve{\alpha}_q}-1\right)\min\left\{\mathrm{e}^{\mu^{(\mathsf{GPB1})}_{\breve{n}^{(\mathsf{I})}_{t-L_R}(q)\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-L_R}\right.}+\frac{1}{2}\sigma^{2,(\mathsf{GPB1})}_{\breve{n}^{(\mathsf{I})}_{t-L_R}(q)\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-L_R}\right.}},\mathrm{e}^{o^{(\mathsf{I})}_{t-L_R}(q)}\right\}\\
&\quad+\max\left\{\mathrm{e}^{\mu^{(\mathsf{GPB1})}_{\breve{n}^{(\mathsf{I})}_{t-L_R}(q)\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-L_R}\right.}+\frac{1}{2}\sigma^{2,(\mathsf{GPB1})}_{\breve{n}^{(\mathsf{I})}_{t-L_R}(q)\left|\breve{\mathbf{o}}^{(\mathsf{I})}_{1:t-L_R}\right.}},\mathrm{e}^{o^{(\mathsf{I})}_{t-L_R}(q)}\right\}
\end{aligned}\tag{4.402}$$

instead of computing the exact expectation values of the maximum and the minimum occurring in the definition of the estimate in (4.193), respectively.[17] While the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}\right)$ is approximately zero, anyway, the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}\right)$ computed from the estimate of the IRNR vector $\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}$ again exhibits the desirable property of being rather *conservative* in the presence of uncertainty about the VTS expansion point as it, e.g., in general rather overestimates the true variances.

The most conservative setup may again be achieved by setting the IRNR vector $\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}$ to $+\infty$, resulting in the approximation

$$\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}\right)\approx\boldsymbol{\mu}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{s_t,L_R}}\,,\tag{4.403}$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{o_t,L_R}}\left(\hat{\mathbf{r}}^{(\mathsf{I},\mathsf{R}),[\psi]}_{t,L_R}\right)\approx\boldsymbol{\Sigma}_{\breve{\mathbf{v}}^{(\mathsf{I},\mathsf{R})}_{s_t,L_R}}\,.\tag{4.404}$$

[17] For the considered GAUSSIAN approximation, closed-form solutions may be given in terms of the means of lower- and upper-tail truncated Normal distributions (see [80, ch. 1, p. 81, (79)]).

5 Evaluation

The assessment of the different observation models to be employed in the Bayesian inference of the clean speech feature posterior in the presence of either reverberation or noise and the presence of both reverberation and noise calls for suitable databases.

The following sections therefore first summarize the databases and corresponding system setups used for the evaluation of the algorithms in

- the absence of reverberation and presence of noise: the AURORA 2 and AURORA 4 corpus (see Sec. 5.1 and Sec. 5.2),
- the presence of reverberation and the optional presence of noise: the AURORA 5 corpus (see Sec. 5.3) and
- the presence of both reverberation and noise: the WSJCAM 0 and the MC-WSJ-AV corpus (see Sec. 5.4).

While the three AURORA databases (AURORA 2, AURORA 4 and AURORA 5) provide evaluation data that have artificially been distorted, the MC-WSJ-AV database provides real recordings in an adverse environment.

To set the results obtained with the **B**ayesian ***F**eature* ***E**nhancement* (BFE) into proper context, each section also includes *baseline results* obtained with either the ETSI standard front-end (denoted by SFE) or the ETSI advanced front-end (denoted by AFE). Though the ETSI advanced front-end has primarily been designed for feature extraction of noisy speech, for which it has already been shown to be quite effective, its performance will also be evaluated on the databases comprising noisy reverberant speech for comparison purposes.

Further the ETSI standard front-end will be applied with subsequent CMN [97] on the complete feature vectors (i.e., including the static and dynamic components) of each utterance (denoted by SFE+CMN).

CMN is based on the MTFA [30] and as such is capable of compensating for AIRs of comparatively (w.r.t. the length of the analysis window) short duration, only [61, ch. 33, p. 658]. In particular, CMN may not be able to compensate for AIRs typically encountered in a reverberant environment. However, it may, at least to some extent, compensate for constant biases possibly introduced by the sub-optimal inference algorithms employed for Bayesian feature enhancement.

Note that for all baseline experiments reported here, the acoustic model has been matched to the front-end feature extraction scheme, i.e., if, e.g., the SFE+CMN is used, the acoustic model is trained on features extracted with the SFE+CMN, too. However, for the BFE schemes, always the acoustic models trained on the features of the clean training data extracted with SFE+CMN are employed.

Since an in-depth analysis of all the possible combinations of BFE parameters is infeasible, the following analyses majorly aim at

- highlighting the sensitivity of the BFE w.r.t. the chosen a priori model for the speech feature vectors,
- elaborating on the importance of the vector of phase factors in the model-specific inference and
- examining and evaluating the different observation models and the employed model-specific inference schemes on the appropriate recognition tasks.

5.1 AURORA 2 task

In this section, the performance of the proposed inference schemes for the absence of reverberation and the presence of noise is investigated on a small vocabulary recognition task. The employed AURORA 2 database is described in Sec. 5.1.1 in full detail. The recognizer setup used throughout the experiments is briefly summarized in Sec. 5.1.2 and is followed by baseline recognition results presented in Sec. 5.1.3. The BFE setup is briefly summarized in Sec. 5.1.4 and the considered inference schemes are finally examined in Sec. 5.1.5.

5.1.1 AURORA 2 Database Description

The AURORA 2 database [98] is a subset of the TIDigits database [99], comprising connected digits spoken in American English recorded at $T_S^{-1} = 20$ kHz, which has been decimated to $T_S^{-1} = 8$ kHz by applying an "ideal" low-pass filter. The database defines two training sets (*clean* and *multi-condition*) and three test sets (A, B and C) and has been designed for the evaluation of ASR systems in noisy conditions.

The data for the *clean* condition training comprise $8,440$ utterances from the training part of the TIDigits database and are uttered by 55 male and 55 female speakers.

The data for the *multi-condition* training are based on the $8,440$ utterances of the data for the *clean* condition training. They are split into 20 subsets of 422 utterances to which real-world noises recorded in 4 different environments, denoted by *car*, *suburban train*, *crowd of people* (*babble*) and *exhibition hall*, have artificially been added at 5 different global broadband SNRs, i.e., ∞ dB (no noise added), 20 dB, 15 dB, 10 dB and 5 dB. Each subset thereby covers all speakers.

The data for the test sets consist of a subset of $4,004$ utterances taken from the test part of the TIDigits database and are uttered by 52 male and 52 female speakers. After splitting the data into 4 subsets of $1,001$ utterances, the test set A and B are obtained as follows:

Test set A is obtained by artificially adding four different noise types, namely *suburban train*, *babble*, *car* and *exhibition hall* noise, to the clean data. The noise signals are added to each subset at a global broadband SNR of either ∞ dB (no noise added), 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB, giving a total of $7,007$ utterances per noise type to be processed.

Test set B is created in the same way, however, with four different noise types, namely *restaurant*, *street*, *airport* and *train station* noise.

While some of the noises are quite stationary, e.g., *car* and *exhibition hall* noise, others contain non-stationary segments, e.g., *street* and *airport* noise.

Test set C, which has been filtered with a different frequency characteristic, is excluded from all recognition experiments.

5.1.2 Recognizer Setup

The acoustic model for the AURORA 2 task comprises whole-word HMMs for the digits 0-9, of which the digit 0 is represented by two models, namely *zero* and *oh*, and two models to represent the silence at the beginning and the end of each utterance (denoted by *sil*) and the short-pause between words (denoted by *sp*), respectively.

The digit models employ $16/18$ HMM states (16 emitting states, 2 non-emitting states) with the respective emission density modeled by diagonal-covariance GMMs with 20 mixture components, each. The HMM topology is strictly left-to-right (linear) and no skips are allowed.

The *sil* model employs $3/5$ HMM states with diagonal-covariance GMMs of 36 mixture components. A deviation from the left-to-right HMM topology employed for the digit models is introduced by allowing state *skips* and an additional transition from the last emitting state to the first emitting state to also model periods of short and perseverative speech absence, respectively.

The *sp* model employs $1/3$ HMM states and the only emitting state is tied to the second emitting state of the *sil* model. The HMM topology is again left-to-right and allows continuously spoken digits, i.e., no short-pause between them, by introducing a transition from the starting non-emitting state to the ending non-emitting state.

Training of the HMMs is carried out using the ***h****idden* MARKOV *model* ***t****ool****k****it* (*HTK*) [100] employing the ML criterion on either the *clean* or *multi-condition* training data.

The recognition employs a zero-gram language model with language model scale factor $\alpha_{\mathsf{LMS}} = 1$ and word insertion penalty $\alpha_{\mathsf{WIP}} = 0$. The latter just means that explicit modeling of the sentence length is neither considered in (3.53) nor in (3.71).

Recognition results on the AURORA 2 database are reported in terms of word accuracies (see (3.55)) without the results on the noises added at a global broadband SNR of -5 dB.

5.1.3 Baseline Results

The baseline results obtained on the AURORA 2 database with a *clean* acoustic model are listed in Tab. 5.1 for the aforementioned front-end feature extraction schemes, namely the SFE, the SFE+CMN and AFE. Clearly the performance with the SFE suffers most from the additional presence of noise and considerably decreases with decreasing SNR. The application of the SFE+CMN is only partly capable of increasing the recognition accuracy and the performance is, in particular, highly sensitive to the noise type and SNR condition. The AFE, due to its two-staged WIENER filter driven noise-reduction, is eventually capable of reducing the mismatch between the acoustic model trained on the clean speech training data and the features extracted from the noisy test data to an impressive extent.

Table 5.1: Baseline recognition accuracies λ_{ACC} [%] on test set A and test set B of the AURORA 2 database obtained with the SFE (a), the SFE+CMN (b) and the AFE (c) with the clean *acoustic model.*

(a) SFE

SNR			test set A					test set B		
[dB]	subway	babble	car	exhibition	$\mathrm{AVG}^{\mathrm{exh.}}_{\mathrm{sub.}}$	restaurant	street	airway	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{rest.}}$
∞	99.79	99.49	99.70	99.78	99.69	99.79	99.49	99.70	99.78	99.69
20	98.53	90.54	98.99	98.15	96.55	92.75	97.31	93.23	95.96	94.81
15	95.67	75.21	93.26	95.16	89.82	79.61	92.81	80.11	87.16	84.92
10	85.20	49.09	73.78	84.39	73.12	58.74	74.82	56.81	66.28	64.16
5	56.95	21.49	37.73	54.83	42.75	30.40	46.49	29.08	33.11	34.77
0	25.36	4.11	15.93	24.13	17.38	6.79	21.04	12.38	12.74	13.24
$\mathrm{AVG}^{20\,\mathrm{dB}}_{0\,\mathrm{dB}}$	72.34	48.09	63.94	71.33	63.92	53.66	66.49	54.32	59.05	58.38

(b) SFE+CMN

SNR			test set A					test set B		
[dB]	subway	babble	car	exhibition	$\mathrm{AVG}^{\mathrm{exh.}}_{\mathrm{sub.}}$	restaurant	street	airway	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{rest.}}$
∞	99.72	99.64	99.55	99.72	99.66	99.72	99.64	99.55	99.72	99.66
20	97.45	97.97	97.76	97.04	97.56	98.50	97.88	98.39	98.09	98.22
15	92.42	94.74	93.59	91.48	93.06	95.89	94.47	96.09	95.43	95.47
10	77.80	84.16	77.57	75.50	78.76	86.49	81.41	88.16	84.26	85.08
5	49.52	57.35	42.71	44.21	48.45	62.94	54.84	63.41	55.41	59.15
0	18.42	20.62	8.29	11.08	14.60	28.00	18.14	26.36	17.00	22.38
$\mathrm{AVG}^{20\,\mathrm{dB}}_{0\,\mathrm{dB}}$	67.12	70.97	63.98	63.86	66.48	74.36	69.35	74.48	70.04	72.06

(c) AFE

SNR			test set A					test set B		
[dB]	subway	babble	car	exhibition	$\mathrm{AVG}^{\mathrm{exh.}}_{\mathrm{sub.}}$	restaurant	street	airway	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{rest.}}$
∞	99.66	99.67	99.67	99.78	99.69	99.66	99.67	99.67	99.78	99.69
20	98.83	98.88	99.19	98.95	98.96	98.74	98.76	99.11	99.11	98.93
15	97.24	97.13	98.36	97.81	97.64	96.68	97.55	98.15	97.84	97.56
10	94.14	92.59	96.66	94.88	94.57	92.51	93.92	95.20	95.46	94.27
5	86.25	80.74	90.81	86.36	86.04	79.18	84.92	86.25	87.26	84.40
0	66.78	50.82	71.25	65.72	63.64	52.32	63.66	63.64	66.95	61.64
$\mathrm{AVG}^{20\,\mathrm{dB}}_{0\,\mathrm{dB}}$	88.65	84.03	91.25	88.74	88.17	83.89	87.76	88.47	89.32	87.36

These differences between the listed front-end feature extraction schemes can also be observed when a *multi-condition* acoustic model is employed. The results obtained on the AURORA 2 database with a *multi-condition* acoustic model are listed in Tab. 5.3 for the aforementioned front-end feature extraction schemes, namely the SFE, the SFE+CMN and AFE. Since the AFE is also applied to the multi-condition training data, the recognition accuracies can be improved to 94.18 % and 93.35 % on the test set A and B, respectively. However, it can also be observed that the performance difference between the SFE+CMN and the AFE is not as big as with the clean acoustic model and in particular also depends on the present noise type.

Table 5.3: *Baseline recognition accuracies λ_{ACC} [%] on test set A and test set B of the AURORA 2 database obtained with the SFE (a), the SFE+CMN (b) and the AFE (c) with the* multi-condition *acoustic model.*

(a) SFE

SNR			test set A					test set B		
[dB]	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\text{train}}_{\text{rest.}}$
∞	99.54	99.37	99.46	99.51	99.47	99.54	99.37	99.46	99.51	99.47
20	99.08	98.91	99.19	98.86	99.01	98.93	98.88	98.66	98.70	98.79
15	98.43	98.61	99.05	98.55	98.66	98.28	98.22	97.73	97.72	97.99
10	97.64	97.46	97.67	96.58	97.34	96.13	95.89	95.35	94.60	95.49
5	93.92	91.93	92.90	91.36	92.53	90.08	88.88	90.52	88.03	89.38
0	76.54	68.05	69.52	75.01	72.28	71.11	70.10	73.78	68.71	70.92
$\mathbf{AVG}^{20\text{dB}}_{0\text{dB}}$	93.12	90.99	91.67	92.07	91.96	90.91	90.39	91.21	89.55	90.52

(b) SFE+CMN

SNR			test set A					test set B		
[dB]	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\text{train}}_{\text{rest.}}$
∞	99.42	99.33	99.40	99.54	99.42	99.42	99.33	99.40	99.54	99.42
20	99.23	99.06	99.22	98.98	99.12	98.99	98.88	98.99	99.20	99.02
15	98.80	98.67	98.72	98.58	98.69	98.83	98.61	98.42	98.43	98.57
10	97.42	97.82	97.52	96.98	97.44	97.33	96.77	97.17	97.04	97.08
5	94.47	93.05	92.78	91.39	92.92	92.26	92.35	93.41	91.92	92.48
0	82.01	72.76	71.88	77.91	76.14	75.62	75.63	78.35	73.25	75.71
$\mathbf{AVG}^{20\text{dB}}_{0\text{dB}}$	94.39	92.27	92.02	92.77	92.86	92.61	92.45	93.27	91.97	92.57

(c) AFE

SNR			test set A					test set B		
[dB]	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\text{train}}_{\text{rest.}}$
∞	99.42	99.43	99.64	99.57	99.52	99.42	99.43	99.64	99.57	99.52
20	99.29	99.27	99.52	99.32	99.35	99.45	99.24	99.40	99.63	99.43
15	98.83	98.76	99.14	98.80	98.88	98.77	98.64	98.87	99.01	98.82
10	97.42	97.55	98.18	97.04	97.55	96.93	96.77	97.29	97.59	97.15
5	94.69	93.32	95.14	93.34	94.12	92.08	92.41	93.56	93.06	92.78
0	82.13	74.82	85.24	81.73	80.98	74.46	78.87	80.47	80.53	78.58
$\mathbf{AVG}^{20\text{dB}}_{0\text{dB}}$	94.47	92.74	95.44	94.05	94.18	92.34	93.19	93.92	93.96	93.35

5.1.4 BFE Setup

For the BFE on the AURORA 2 task both GMMs and MSLDMs as a priori models for the clean speech feature vectors are employed. Both kinds of models are trained on the clean training data under the EM framework presented in Sec. 4.2.1.

Starting with the ML estimate of the model parameters for $M = 1$, the *model-splitting* approach pursued in this work iteratively increases the number of dynamic states from $M = 1$ to $M = 2$, from $M = 2$ to $M = 4$, etc.. At each splitting step, the EM-algorithm is iterated until either the maximum of 20 EM iterations is reached or the relative improvement of the average likelihood, computed over all training utterances, falls below a predefined threshold of 10%.

Note that the number of parameters in an MSLDM with M dynamic states in Q dimension is $M(1+M+Q^2+Q(Q+1)+2Q)$ and as such considerably larger than the number of parameters in a GMM with an equivalent number of dynamic states, i.e., $M(1+Q(Q+$

$1)/2+Q)$.

The single GAUSSIAN a priori model for the noise feature vector trajectory is trained on a per-utterance basis. This procedure allows the model to capture noise properties that are specific to the current utterance. Since the noise model is trained on just a single utterance, equal mean vectors and covariances matrices are chosen for the first and all other time instances, i.e., $\boldsymbol{\mu}_{\breve{\mathbf{n}}_1^{(\mathsf{l})}} = \mathbf{b}_{\breve{\mathbf{n}}^{(\mathsf{l})}}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{n}}_1^{(\mathsf{l})}} = \mathbf{V}_{\breve{\mathbf{n}}^{(\mathsf{l})}}$. The respective parameters are trained on the first and last 20 feature vectors of each test utterance in an ML fashion.

The following model-specific inference algorithms are examined under the GPB1 multi-model inference framework:

- IEKF+CMN: The model-specific inference is carried out by employing the IEKF presented in Sec. 4.8.2.3, however, with the zero-mean/zero-variance assumption utilized in [101] posed on the phase factors. Further, only the causal estimate of the mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ of the GPB1 matched single a posteriori PDF is forwarded to the recognizer. Since CMN has been found to substantially improve the baseline results over the ETSI standard front-end alone (compare Sec. 5.1.3), it is applied to the enhanced feature vector after computation of the dynamic features, too.

- IEKF-α+CMN: As the IEKF+CMN, however, this time the model-specific IEKF inference employs the analytically determined moments of the phase factors derived in Sec. 4.6.4.

- SOEKF-α+CMN: As with the IEKF-α+CMN, the analytically determined moments of the phase factors are employed. However, the model-specific inference is carried out by the SOEKF instead of the IEKF.

Further, for the model-specific inference scheme with the best performance, also the causal estimate of the (diagonal) covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ of the GPB1 matched single a posteriori PDF is forwarded to the recognizer to also exploit the uncertainty about the estimate of the mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$. The practically realizable UD rules summarized in Sec. 3.6.1 will thereby be denoted by

- causal UD-p: The causal UD rule (3.71) employing the mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ and the (diagonal) covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ of the predictive PDF.

- causal UD-m: The causal variant of the UD rule (3.77) employing the global mean vector and the (diagonal) covariance matrix of the clean training data as the moments of the GAUSSIAN marginal a priori PDF.

- causal UD-n: The causal variant of the UD rule (3.77), however, this time the a priori distribution is (n)eglected, which is equivalent to approximating the equivalent mean vectors and the equivalent covariance matrices by the corresponding mean vectors and covariance matrices of the a posteriori PDF.

The respective inference scheme will, e.g., be denoted by SOEKF-α+CMN+UD-p, SOEKF-α+CMN+UD-m and SOEKF-α+CMN+UD-n.

Since, in general, severe approximations are required at all stages of the inference of the a posteriori PDF of the clean speech feature vector, i.e., at the a priori models, the multi-model and the model-specific inference, the quality of the estimates, i.e., the mean vectors and covariance matrices of the a posteriori PDF, is quite questionable. However, with the estimate of the respective mean vectors being considerably more reliable than the estimate of the corresponding covariance matrices, this issue mainly affects the application of the uncertainty decoding rules. Thus, for instance, it has been reported in [78] and [102] that posing an upper bound on the variances of the a posteriori PDF of the clean speech feature w.r.t. its a priori PDF is beneficial for the performance of a recognizer in the uncertainty decoding framework. Following these heuristics, the respective variances of the a posteriori PDF will be upper bounded by 5% of the corresponding variances of the a priori PDF in all experiments employing the UD rules that employ an a priori PDF, i.e., UD-p (predictive prior) and UD-m (marginal prior).

5.1.5 Results with Bayesian Feature Enhancement

The Bayesian feature enhancement will first be investigated with a GMM a priori model and then with an MSLDM a priori model.

5.1.5.1 GMM A Priori Speech Model

The recognition results obtained with the BFE employing GMM a priori models for the speech feature vectors and a *clean* acoustic model are listed in Tab. 5.5 for the different model-specific inference schemes and a varying number of dynamic states in the a priori model. While all model-specific inference schemes, e.g., the IEKF+CMN, the IEKF-α+CMN and the SOEKF-α+CMN, with a GMM consisting of a single dynamic state, i.e., $M = 1$, result in a recognition accuracy that is lower than that with the SFE+CMN, it is already with $M = 2$ that the baseline recognition results are exceeded. In particular, a steady increase of the recognition performance can be observed with an increasing number of dynamic states M.

From a comparison of the results obtained with the IEKF+CMN and the IEKF-α+CMN the importance of considering the vector of phase factor becomes apparent. This can best be seen by looking at Tab. 5.7 which lists the recognition results for $M = 128$ dynamic states for the two model-specific inference schemes in more detail. While the IEKF+CMN and the IEKF-α+CMN perform approximately equal at a high global broadband SNR, an increasing benefit of considering the vector of phase factors in the IEKF-α+CMN can be observed with decreasing global broadband SNR. This benefit becomes most pronounced at a global broadband SNR of 0 dB to 5 dB and consistently match the finding in Sec. 4.7.2 where the analysis of the distribution of the observation error is carried out.

Application of the SOEKF in the SOEKF-α+CMN scheme further increases the recognition accuracies over those obtained with the IEKF-α+CMN scheme and provides a performance that is equivalent or superior to the AFE starting with $M = 16$. The best recognition accuracies of 88.36 % and 89.45 % are achieved with the SOEKF-α+CMN at $M = 128$.

Table 5.5: *Averaged recognition accuracies λ_{ACC} [%] (averaged over the global broadband SNRs* 20 dB-0 dB*) on test set A and test set B of the AURORA 2 database obtained with the IEKF+CMN (a), the IEKF-α+CMN (b) and the SOEKF-α+CMN (c) with the* clean *acoustic model and GMM a priori models for the speech feature vectors with* $M \in \{1,2,4,8,16,32,64,128\}$ *dynamic states.*

(a) IEKF+CMN

	test set A					test set B				
M	subway	babble	car	exhibition	$\mathsf{AVG}^{\mathsf{exh.}}_{\mathsf{sub.}}$	restaurant	street	airway	train	$\mathsf{AVG}^{\mathsf{train}}_{\mathsf{rest.}}$
1	64.29	67.93	68.45	69.02	67.42	71.57	65.18	74.75	71.84	70.84
2	79.34	76.22	81.52	76.06	78.28	78.25	78.08	81.59	81.96	79.97
4	83.10	79.29	84.60	78.81	81.45	80.83	81.47	84.18	84.34	82.70
8	84.98	80.94	86.50	80.94	83.34	82.72	82.94	85.74	86.45	84.46
16	86.44	83.20	88.27	82.69	85.15	84.12	85.14	87.27	88.24	86.19
32	87.20	84.02	88.95	83.73	85.98	84.77	86.11	87.92	88.68	86.87
64	87.80	84.55	89.48	84.37	86.55	85.07	86.27	88.44	89.32	87.28
128	88.13	84.87	89.93	84.97	86.97	85.79	86.75	88.92	89.81	87.82

(b) IEKF-α+CMN

	test set A					test set B				
M	subway	babble	car	exhibition	$\mathsf{AVG}^{\mathsf{exh.}}_{\mathsf{sub.}}$	restaurant	street	airway	train	$\mathsf{AVG}^{\mathsf{train}}_{\mathsf{rest.}}$
1	64.97	68.43	68.50	69.37	67.82	72.26	65.58	75.20	72.18	71.30
2	80.73	78.41	83.01	77.31	79.86	80.46	80.00	83.91	83.37	81.94
4	83.98	81.14	85.72	79.90	82.69	82.41	82.68	85.79	85.60	84.12
8	86.11	82.73	87.40	82.38	84.66	84.25	84.34	87.04	87.41	85.76
16	87.16	84.55	89.01	83.78	86.13	85.38	86.14	88.30	88.89	87.18
32	88.05	85.29	89.70	84.53	86.89	86.17	86.97	88.94	89.24	87.83
64	88.60	85.76	90.11	85.26	87.43	86.29	87.42	89.48	89.91	88.28
128	88.80	85.98	90.63	85.67	87.77	86.94	87.79	89.84	90.27	88.71

(c) SOEKF-α+CMN

	test set A					test set B				
M	subway	babble	car	exhibition	$\mathsf{AVG}^{\mathsf{exh.}}_{\mathsf{sub.}}$	restaurant	street	airway	train	$\mathsf{AVG}^{\mathsf{train}}_{\mathsf{rest.}}$
1	46.22	46.87	48.06	52.98	48.53	49.43	43.42	54.16	51.31	49.58
2	80.51	78.77	81.22	77.76	79.57	80.90	78.65	83.89	82.06	81.37
4	83.84	82.42	85.40	80.21	82.97	83.83	82.82	86.99	85.26	84.72
8	86.44	84.32	87.43	83.39	85.39	85.99	84.81	88.76	87.32	86.72
16	87.76	85.89	89.36	84.67	86.92	87.42	86.87	89.72	89.34	88.34
32	88.43	86.64	90.00	85.29	87.59	87.84	87.59	90.24	89.61	88.82
64	88.81	86.90	90.52	85.77	88.00	88.11	88.02	90.73	90.09	89.24
128	89.34	87.11	90.90	86.09	88.36	88.28	88.38	90.89	90.26	89.45

5.1.5.2 MSLDM A Priori Speech Model

The recognition results obtained with the BFE employing MSLDM a priori models for the speech feature vectors and a *clean* acoustic model are listed in Tab. 5.9 for the different model-specific inference schemes and a varying number of dynamic states in the a priori model. The findings with an MSLDM a priori model for the clean speech feature vectors are quite distinct from those employing a GMM. Although the IEKF-α+CMN again performs better than the IEKF+CMN and the SOEKF-α+CMN in general again better than the IEKF-α+CMN, the respective recognition results with a varying number of dynamic states are rather inconsistent and in particular lower than with GMM a priori models.

While the recognition results already exceed those obtained with the SFE+CMN with

Table 5.7: *Recognition accuracies λ_{ACC} on test set A and B of the AURORA 2 database obtained with the IEKF+CMN (a) and the IEKF-α+CMN (b) with the* clean *acoustic model and a GMM a priori model for the speech feature vectors with $M = 128$ dynamic states.*

(a) IEKF+CMN

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	**AVG$^{\text{exh.}}_{\text{sub.}}$**	**restaurant**	**street**	**airway**	**train**	**AVG$^{\text{train.}}_{\text{rest.}}$**
∞	99.63	99.52	99.49	99.66	99.57	99.63	99.52	99.49	99.66	99.57
20	99.02	98.49	99.43	98.33	98.82	98.34	98.43	98.81	98.83	98.60
15	97.64	96.83	98.99	96.48	97.48	97.14	97.43	97.70	98.09	97.59
10	93.43	93.56	96.45	91.42	93.72	93.18	92.99	95.17	95.80	94.29
5	84.99	81.44	88.49	79.85	83.69	82.19	83.62	86.16	88.31	85.07
0	65.58	54.02	66.30	58.75	61.16	58.09	61.28	66.75	68.03	63.54
AVG$^{20\,\text{dB}}_{0\,\text{dB}}$	88.13	84.87	89.93	84.97	86.97	85.79	86.75	88.92	89.81	87.82

(b) IEKF-α+CMN

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	**AVG$^{\text{exh.}}_{\text{sub.}}$**	**restaurant**	**street**	**airway**	**train**	**AVG$^{\text{train.}}_{\text{rest.}}$**
∞	99.60	99.52	99.52	99.63	99.57	99.60	99.52	99.52	99.63	99.57
20	99.14	98.82	99.49	98.40	98.96	98.71	98.49	98.90	99.04	98.79
15	98.00	97.25	99.11	96.98	97.84	97.45	97.64	97.97	98.27	97.83
10	94.04	93.92	96.75	91.61	94.08	94.17	93.86	95.82	95.96	94.95
5	86.15	82.98	89.59	80.72	84.86	83.85	85.13	87.47	88.74	86.30
0	66.69	56.92	68.21	60.66	63.12	60.52	63.85	69.04	69.36	65.69
AVG$^{20\,\text{dB}}_{0\,\text{dB}}$	88.80	85.98	90.63	85.67	87.77	86.94	87.79	89.84	90.27	88.71

just a single dynamic state, i.e., $M = 1$ (note: a similar performance with a GMM a prior model already requires $M = 4$ dynamic states, as can be inferred from Tab. 5.5), they first drop when increasing the number of dynamic states to $M = 2$. With a further increasing M, the recognition results improve and reach a maximum at $M = 16$ dynamic states before turning worse again at $M = 32$. This inconsistency may be considered an indicator that an MSLDM may be more susceptible to the approximations underlying the employed multi-model inference scheme than a GMM.

The best recognition accuracies of 85.80 % and 86.73 % are achieved with the SOEKF-α+CMN at $M = 16$ dynamic states.

5.1.5.3 UD Variants

Finally, the best performing combinations of *a priori model* and *model-specific inference scheme* are employed in the UD framework. Table 5.11 summarizes the results without and with application of the different UD rules for the SOEKF-α+CMN scheme employing a GMM a priori model with $M = 128$ dynamic states. It can be observed that the application of any of the considered UD rules results in a slightly improved average recognition accuracy (about 0.5 % absolute). These gains are majorly achieved at low global broadband SNR values. Though all UD rules perform equally well, the SOEKF-α+CMN+UD-p configuration employing the GPB1-matched predictive distribution as the a priori PDF can be found to give marginally better results than the other two.

The corresponding results for the best model-specific inference scheme employing an

Table 5.9: *Averaged recognition accuracies λ_{ACC} [%] on test set A and test set B of the AURORA2 database obtained with the IEKF+CMN (a), the IEKF-α+CMN (b) and the SOEKF-α+CMN (c) with the* clean *acoustic model and MSLDM a priori models for the speech feature vectors with $M \in \{1,2,4,8,16,32\}$ dynamic states.*

(a) IEKF+CMN

	test set A					test set B				
M	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\mathbf{exh.}}_{\mathbf{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\mathbf{train}}_{\mathbf{rest.}}$
1	81.53	79.78	85.82	81.28	82.10	81.62	81.55	84.77	85.32	83.32
2	78.81	77.43	82.24	80.66	79.78	79.88	77.54	83.31	82.79	80.88
4	81.04	79.47	83.36	81.77	81.41	81.57	79.29	84.78	84.17	82.45
8	82.76	79.75	85.29	81.96	82.44	81.24	80.80	84.90	85.46	83.10
16	82.74	80.14	85.54	82.06	82.62	81.58	80.71	85.08	85.84	83.30
32	80.74	77.63	82.11	81.00	80.37	79.64	77.92	83.01	83.47	81.01

(b) IEKF-α+CMN

	test set A					test set B				
M	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\mathbf{exh.}}_{\mathbf{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\mathbf{train}}_{\mathbf{rest.}}$
1	81.68	79.93	85.90	81.39	82.22	81.83	81.64	84.99	85.44	83.48
2	79.26	77.47	82.28	80.91	79.98	80.09	77.65	83.32	82.90	80.99
4	82.10	80.05	84.10	82.25	82.13	81.98	80.02	85.30	84.75	83.01
8	84.54	81.41	86.88	82.98	83.95	82.43	82.74	86.01	86.83	84.50
16	84.79	82.01	87.40	82.95	84.29	82.86	83.28	86.38	87.45	84.99
32	83.51	80.54	85.14	82.17	82.84	82.06	81.56	85.06	85.92	83.65

(c) SOEKF-α+CMN

	test set A					test set B				
M	**subway**	**babble**	**car**	**exhibition**	$\mathbf{AVG}^{\mathbf{exh.}}_{\mathbf{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\mathbf{AVG}^{\mathbf{train}}_{\mathbf{rest.}}$
1	80.61	79.61	84.88	81.86	81.74	81.76	80.41	84.74	84.30	82.80
2	77.04	76.06	80.56	80.91	78.64	78.91	75.77	82.25	81.21	79.54
4	82.40	81.05	84.75	83.25	82.86	82.84	80.72	85.84	85.05	83.61
8	85.94	83.40	88.63	84.02	85.50	83.97	84.67	87.44	88.14	86.05
16	86.14	83.97	89.35	83.74	85.80	84.26	85.63	87.90	89.15	86.73
32	86.23	83.92	88.92	83.50	85.64	84.19	85.25	87.72	88.75	86.48

MSLDM, here the IEKF-α+CMN with an a priori model with $M = 16$ dynamic states, are given in Tab. 5.13. As with the GMM a priori model, the recognition performance improves with the additional application of the UD rules. This time, the IEKF-α+CMN+UD-n configuration performs remarkably better than the UD rules employing an a priori distribution.

In summary, the consideration of the uncertainty in the estimate of the clean speech feature vector provides recognition results that are equivalent (MSLDM a priori model) or superior (GMM a priori model) to the AFE. The overall best recognition results with 89.42 % and 90.15 % on test set A and B, respectively, are thereby obtained with the SOEKF-α+CMN+UD-p configuration.

Table 5.11: *Recognition accuracies λ_{ACC} [%] on test set A and test set B of the AURORA 2 database obtained with the SOEKF-α+CMN (a) and the UD variants SOEKF-α+CMN-p (b), SOEKF-α+CMN-m (c) and SOEKF-α+CMN-n (d) with the* clean *acoustic model and GMM a priori models for the speech feature vectors with $M = 128$ dynamic states.*

(a) SOEKF-α+CMN

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.66	99.55	99.52	99.57	99.57	99.66	99.55	99.52	99.57	99.57
20	99.14	99.03	99.40	98.55	99.03	99.08	98.67	99.34	99.11	99.05
15	98.28	97.79	99.14	96.91	98.03	98.37	97.82	98.99	98.33	98.38
10	94.32	94.89	97.08	92.72	94.75	95.43	94.41	96.96	95.83	95.66
5	87.10	85.10	89.92	81.67	85.95	85.75	86.00	89.29	88.77	87.45
0	67.85	58.74	68.98	60.60	64.04	62.76	65.02	69.88	69.27	66.73
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	89.34	87.11	90.90	86.09	88.36	88.28	88.38	90.89	90.26	89.45

(b) SOEKF-α+CMN+UD-p (predictive prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.66	99.55	99.55	99.57	99.58	99.66	99.55	99.55	99.57	99.58
20	99.32	98.97	99.37	98.61	99.07	99.08	98.79	99.46	99.20	99.13
15	98.71	97.82	99.22	97.66	98.35	98.34	97.97	98.84	98.49	98.41
10	95.39	94.98	97.55	93.15	95.27	95.46	94.92	96.87	96.08	95.83
5	88.58	85.64	91.47	83.12	87.20	86.12	86.97	90.25	89.82	88.29
0	71.11	61.25	73.01	63.50	67.22	64.20	67.93	71.79	72.32	69.06
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	90.62	87.73	92.12	87.21	89.42	88.64	89.32	91.44	91.18	90.15

(c) SOEKF-α+CMN+UD-m (marginal prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.66	99.52	99.58	99.57	99.58	99.66	99.52	99.58	99.57	99.58
20	99.29	98.97	99.37	98.58	99.05	99.08	98.82	99.43	99.17	99.12
15	98.71	97.85	99.22	97.53	98.33	98.40	97.82	98.84	98.36	98.36
10	95.21	94.95	97.38	92.97	95.13	95.49	94.80	96.87	95.99	95.79
5	88.39	85.52	91.23	83.00	87.03	86.03	87.06	90.10	89.66	88.21
0	70.71	60.49	72.23	62.94	66.59	63.74	67.32	71.43	71.43	68.48
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	90.46	87.56	91.89	87.00	89.23	88.55	89.16	91.33	90.92	89.99

(d) SOEKF-α+CMN+UD-n (neglected prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.69	99.58	99.55	99.66	99.62	99.69	99.58	99.55	99.66	99.62
20	99.17	98.94	99.40	98.73	99.06	98.99	98.88	99.37	99.32	99.14
15	98.34	97.85	99.08	97.78	98.26	98.10	97.82	98.81	98.40	98.28
10	95.49	94.23	97.23	93.95	95.22	95.24	95.25	96.84	96.11	95.86
5	87.96	84.43	93.05	83.68	87.28	84.99	86.79	90.07	89.97	87.96
0	70.77	58.74	75.31	61.83	66.66	61.56	66.69	70.47	72.66	67.84
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	90.35	86.84	92.81	87.19	89.30	87.78	89.09	91.11	91.29	89.82

Table 5.13: *Recognition accuracies λ_{ACC} [%] on test set A and test set B of the AURORA 2 database obtained with the IEKF-α+CMN (a) and the UD variants IEKF-α+CMN+UD-p (b), IEKF-α+CMN+UD-m (c) and IEKF-α+CMN+UD-n (d) with the* clean *acoustic model and MSLDM a priori models for the speech feature vectors with $M = 16$ dynamic states.*

(a) IEKF-α+CMN

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.26	99.09	99.34	99.54	99.31	99.26	99.09	99.34	99.54	99.31
20	98.86	98.22	99.22	98.18	98.62	98.25	98.70	98.66	98.92	98.63
15	97.39	96.22	98.69	96.36	97.17	96.16	97.10	97.52	97.96	97.18
10	93.37	91.93	95.94	91.36	93.15	91.53	92.59	93.92	95.22	93.31
5	83.24	80.65	88.34	78.31	82.63	79.61	81.86	85.06	87.69	83.56
0	57.84	52.81	64.57	54.49	57.43	55.73	57.89	64.33	65.94	60.97
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	86.14	83.97	89.35	83.74	85.80	84.26	85.63	87.90	89.15	86.73

(b) IEKF-α+CMN+UD-p (predictive prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.29	99.09	99.34	99.54	99.32	99.29	99.09	99.34	99.54	99.32
20	98.89	98.10	99.19	98.27	98.61	98.25	98.70	98.69	98.95	98.65
15	97.67	96.31	98.78	96.45	97.30	96.28	97.37	97.55	98.12	97.33
10	93.74	92.17	96.06	91.82	93.45	91.71	92.87	94.01	95.22	93.45
5	84.19	81.35	89.26	79.57	83.59	80.53	82.95	86.01	88.00	84.37
0	59.56	53.75	65.82	56.06	58.80	56.83	59.04	65.25	67.36	62.12
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	86.81	84.34	89.82	84.43	86.35	84.72	86.19	88.30	89.53	87.18

(c) IEKF-α+CMN+UD-m (marginal prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.29	99.06	99.34	99.54	99.31	99.29	99.06	99.34	99.54	99.31
20	98.93	97.97	99.22	98.21	98.58	98.07	98.61	98.48	98.83	98.50
15	97.64	96.04	98.78	96.48	97.24	95.73	97.25	97.35	98.21	97.14
10	94.32	91.96	96.48	91.92	93.67	91.53	93.14	93.68	95.37	93.43
5	84.89	81.74	90.01	80.22	84.22	80.69	83.31	86.34	87.97	84.58
0	60.36	54.35	66.98	56.62	59.58	57.41	59.76	65.82	67.76	62.69
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	87.23	84.41	90.29	84.69	86.66	84.69	86.41	88.33	89.63	87.27

(d) IEKF-α+CMN+UD-n (neglected prior)

	test set A					test set B				
SNR	**subway**	**babble**	**car**	**exhib.**	$\textbf{AVG}^{\text{exh.}}_{\text{sub.}}$	**restaurant**	**street**	**airway**	**train**	$\textbf{AVG}^{\text{train.}}_{\text{rest.}}$
∞	99.48	99.15	99.40	99.57	99.40	99.48	99.15	99.40	99.57	99.40
20	98.93	98.43	99.25	98.83	98.86	98.43	98.82	98.99	99.01	98.81
15	97.94	96.77	98.93	97.32	97.74	96.87	97.70	97.91	98.46	97.73
10	94.66	93.08	97.26	93.40	94.60	93.00	93.80	95.17	95.62	94.40
5	85.66	83.49	91.05	82.44	85.66	83.48	83.65	88.25	89.32	86.17
0	60.82	56.98	69.67	57.11	61.14	59.78	59.58	68.51	69.36	64.31
$\textbf{AVG}^{20\,\text{dB}}_{0\,\text{dB}}$	87.60	85.75	91.23	85.82	87.60	86.31	86.71	89.77	90.35	88.29

5.2 AURORA 4 task

In this section, the performance of the proposed inference schemes for the absence of reverberation and the presence of noise is investigated on a large vocabulary recognition task. The employed AURORA 4 database is described in Sec. 5.2.1 in full detail. The recognizer setup used throughout the experiments is briefly summarized in Sec. 5.2.2 and is followed by baseline recognition results presented in Sec. 5.2.3. The BFE setup is briefly described in Sec. 5.2.4 and the considered inference schemes are finally examined in Sec. 5.2.5.

5.2.1 AURORA 4 Database Description

The AURORA 4 database [103] is based on the ***D****efense* ***A****dvanced* ***R****esearch* ***P****rojects* ***A****gency* (DARPA) ***W****all* ***S****treet* ***J****ournal* (WSJ) (WSJ 0) corpus [104] and has been designed for evaluation of ASR systems for American English on the specified $5,000$-word closed-loop vocabulary task. Two training conditions (*clean* and *multi-condition*) and 7 test conditions are defined.

The data for the *clean* training are taken from the clean data of the SI-84 WSJ 0 training set recorded with a Sennheiser microphone at a sample rate of $T_S^{-1} = 16$ kHz and are decimated to $T_S^{-1} = 8$ kHz. The $7,138$ utterances uttered by 83 speakers comprise a total of 14 h of speech recordings.

The data for the *multi-condition* training are based on the same utterances as the data employed for the *clean* training, however, recorded by 18 different microphone types and distorted by 6 different noise type, namely, *car*, *babble*, *restaurant*, *street*, *airport* and *train* noise, added at global broadband SNRs of ∞ dB (no noise added) and 20 dB to 10 dB.

The test data consist of the WSJ 0 November'92 ***N****ational* ***I****nstitute of* ***S****tandards and* ***T****echnology* (*NIST*) evaluation test set consisting of 330 utterances from 8 different speakers, totaling 40 min of recorded speech. The data have been recorded by the same Sennheiser microphone also employed for the recording of the *clean* training data and by a secondary microphone not specified further. The recordings of the Sennheiser microphone have artificially been distorted by adding noise of 6 different types, namely, *car*, *babble*, *restaurant*, *street*, *airport* and *train* noise, at varying global broadband SNRs between 15 dB-5 dB, resulting in 6 subsets. The clean data finally build the 7th subset.

The recordings of the second microphone have been distorted in the same way, however, will not be considered in this work.

Also, a subset of 166 utterances representative of the complete test set of 330 utterances has been specified to speed up evaluation.

5.2.2 Recognizer Setup

The acoustic model for the AURORA 4 task comprises word-internal triphone HMMs with $3/5$ states in a linear topology. Each state employs a diagonal-covariance GMM with 10 mixture components.

An additional silence model also employs $3/5$ HMM states, however, with diagonal-covariance GMMs of 20 mixture components. Further, its topology allows an emitting state to be skipped and an additional transition from the last emitting state to the first emitting state to also model periods of short and perseverative speech absence, respectively.

The short-pause model is tied to the three states of the silence model, however, its HMM topology is extended by a skip transitions connecting the starting non-emitting state to the ending non-emitting state.

The pronunciations are taken from the ***C**arnegie **M**ellon **U**niversity* (CMU) dictionary (version 0.6) available at http://www.speech.cs.cmu.edu.

Again, training of the HMMs is carried out using the *HTK* employing the ML criterion on either the *clean* or *multi-condition* training data.

The recognition further employs the standard ***M**assachusetts **I**nstitute of **T**echnology* (MIT) Lincoln Laboratories 5 k compact back-off bigram language model initially provided with the WSJ 0 database. The language model scale factor is set to $\alpha_{\text{LMS}} = 16$ and the word insertion penalty to $\alpha_{\text{WIP}} = 0$.

Recognition results are usually given in terms of insertion, substitution, deletion and word errors rates for each of the 7 test subsets (clean + 6 noise types).

5.2.3 Baseline Results

The baseline results obtained on the AURORA 4 database with a *clean* acoustic model are listed in Tab. 5.15 for the SFE, the SFE+CMN and the AFE, respectively. The overall performance obtained with the three front-end feature extraction schemes follows the trend observed on the AURORA 2 database. The performance with the SFE again suffers most

Table 5.15: *Baseline recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the AURORA 4 database obtained with the SFE (a), the SFE+CMN (b) and the AFE (c) with the* clean *acoustic model.*

(a) SFE

error statistic	clean	airport	babble	car	restaurant	street	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{clean}}$
λ_{SUB}	8.77	37.83	39.30	23.50	36.35	36.69	36.54	31.28
λ_{DEL}	1.36	11.16	10.42	3.06	13.85	14.25	16.91	10.14
λ_{INS}	2.47	12.23	10.68	12.52	8.51	7.22	6.67	8.61
λ_{WER}	12.60	61.22	60.41	39.08	58.71	58.16	60.11	50.04

(b) SFE+CMN

error statistic	clean	airport	babble	car	restaurant	street	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{clean}}$
λ_{SUB}	8.25	28.36	27.29	14.95	29.02	31.23	30.24	24.19
λ_{DEL}	1.22	13.92	14.03	4.60	14.88	15.47	19.37	11.93
λ_{INS}	2.50	4.24	2.36	2.50	3.39	2.39	2.36	2.82
λ_{WER}	11.97	46.52	43.68	22.06	47.29	49.10	51.97	38.94

(c) AFE

error statistic	clean	airport	babble	car	restaurant	street	train	$\mathrm{AVG}^{\mathrm{train}}_{\mathrm{clean}}$
λ_{SUB}	8.55	23.54	21.10	12.38	23.13	21.22	20.66	18.65
λ_{DEL}	1.33	4.01	4.01	2.03	4.86	5.19	4.94	3.77
λ_{INS}	2.21	8.03	5.93	2.87	7.73	5.38	4.13	5.18
λ_{WER}	12.08	35.58	31.05	17.27	35.73	31.79	29.72	27.60

from the presence of noise, followed by the SFE+CMN. The best recognition result, i.e., the lowest average word error rate λ_{WER}, is achieved with the AFE.

These differences between the listed front-end feature extraction schemes can again be observed when a *multi-condition* acoustic model is employed. The results obtained on the AURORA 4 database with a *multi-condition* acoustic model are listed in Tab. 5.17 for the SFE, the SFE+CMN and the AFE. The application of the AFE to the training and the testing data again results in a greatly reduced mismatch between the training and the testing data. The obtained error rate drops to 23.15 %.

Table 5.17: *Baseline recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the AURORA 4 database obtained with the SFE (a), the SFE+CMN (b) and the AFE (c) with the* multi-condition *acoustic model.*

(a) SFE

error statistic	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	9.43	22.28	22.39	13.33	24.86	23.57	24.60	20.07
λ_{DEL}	1.18	8.25	7.55	2.14	9.76	11.71	13.48	7.72
λ_{INS}	3.02	3.24	2.76	2.80	3.35	2.39	2.28	2.83
λ_{WER}	13.63	33.78	32.71	18.27	37.97	37.68	40.37	30.63

(b) SFE+CMN

error statistic	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	9.39	21.51	21.14	12.85	23.57	23.87	25.41	19.68
λ_{DEL}	1.18	7.96	9.06	2.73	9.80	10.53	12.23	7.64
λ_{INS}	2.80	2.80	2.84	2.36	3.35	2.43	2.25	2.69
λ_{WER}	13.37	32.27	33.04	17.94	36.72	36.83	39.89	30.01

(c) AFE

error statistic	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.73	18.49	17.27	10.39	19.56	19.08	18.82	16.05
λ_{DEL}	1.22	4.35	4.60	1.92	5.05	4.53	5.19	3.84
λ_{INS}	2.54	4.42	3.35	2.47	4.20	2.69	3.17	3.26
λ_{WER}	12.49	27.26	25.23	14.77	28.80	26.30	27.18	23.15

5.2.4 BFE Setup

The BFE setup for the AURORA 4 recognition task is equivalent to the one for the AURORA 2 task described in Sec. 5.1.5. Again, GMMs and MSLDMs as a priori models for the clean speech feature vectors are employed. Both kinds of models are trained on the clean training data under the EM framework presented in Sec. 4.2.1.

The parameters of the single GAUSSIAN a priori model for the noise are obtained by means of ML training from the first and last 20 frames of each utterance.

As model-specific inference algorithms, the previously described IEKF+CMN, IEKF-α+CMN and SOEKF-α+CMN are again employed under the GPB1 multi-model inference scheme. The best performing model-specific inference algorithm are again combined with additional uncertainty decoding.

5.2.5 Results with Bayesian Feature Enhancement

5.2.5.1 GMM A Priori Speech Model

The recognition results obtained with the BFE employing GMM a priori models for the speech feature vectors and a *clean* acoustic model are listed in Tab. 5.19 for the different model-specific inference schemes and a varying number of dynamic states in the a priori model. As with the AURORA 2 task, it can again be observed that an increase in the number of dynamic states composing the a priori model for the speech feature vector trajectory leads to a decrease in the word error rates. Thereby the IEKF-α+CMN again turns out to be superior to the IEKF+CMN. Although the AURORA 4 database does not provide access to the global broadband SNR the speech signal and the noise mix at, this observation again indicates that a statistically sound description of the observation model in terms of considering the contribution of the vector of phase factors is beneficial for the recognition task.

Table 5.19: *Word error rates λ_{WER} [%] on the AURORA 4 database obtained with the IEKF+CMN (a), the IEKF-α+CMN (b) and the SOEKF-α+CMN (c) with the* clean *acoustic model and GMM a priori models for the speech feature vectors with $M \in \{1,2,4,8,16,32,64,128\}$ dynamic states.*

(a) IEKF+CMN

M	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
1	12.93	44.42	42.03	21.07	50.83	53.59	55.99	40.12
2	12.52	44.94	42.87	21.62	49.32	45.89	49.98	38.16
4	12.45	42.25	38.97	17.50	45.67	41.14	44.16	34.59
8	12.45	40.77	38.97	18.16	44.64	40.07	43.35	34.06
16	12.41	39.96	38.23	16.54	42.65	39.19	40.07	32.72
32	12.41	39.56	37.35	15.80	42.10	38.05	40.15	32.20
64	12.19	37.72	35.80	16.32	42.39	37.35	38.71	31.50
128	12.15	37.79	35.76	16.76	41.62	36.32	38.20	31.23

(b) IEKF-α+CMN

M	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
1	12.82	43.54	42.10	19.63	50.64	53.44	55.76	39.70
2	12.60	42.65	40.00	20.63	46.52	42.87	44.38	35.66
4	12.45	40.59	37.05	18.01	43.02	40.22	42.03	33.34
8	12.49	39.01	37.20	18.16	43.20	38.78	40.96	32.83
16	12.45	39.12	37.16	16.24	42.39	37.16	38.60	31.87
32	12.41	37.79	36.17	15.73	40.66	36.21	38.16	31.02
64	12.45	37.46	35.58	16.06	41.14	35.91	38.05	30.95
128	11.82	36.69	35.25	16.21	41.10	35.54	37.27	30.55

(c) SOEKF-α+CMN

M	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
1	12.74	55.51	55.80	22.17	64.46	66.52	68.55	49.39
2	12.49	48.80	50.87	21.95	53.55	54.22	54.70	42.37
4	12.15	44.53	44.71	19.45	49.21	49.32	51.09	38.64
8	12.34	41.40	41.66	20.37	47.11	45.52	47.66	36.58
16	12.15	38.53	37.94	18.53	43.54	41.40	42.73	33.55
32	11.97	36.80	36.65	17.24	41.92	38.86	39.82	31.89
64	12.34	36.35	35.62	17.09	41.66	38.34	40.00	31.63
128	11.93	36.24	34.92	17.20	41.25	37.64	40.18	31.34

Opposed to the AURORA2 task, the SOEKF-α+CMN this time does, however, not contribute to a decrease in the recognition errors. In fact, it causes the error rates to be slightly higher than with the IEKF+CMN. This observation again brings into mind that a putative better model-specific inference scheme does not necessarily result in an improved recognition performance in a multi-model inference framework that is subject to many approximations.

The lowest average word error rate of 30.55 % is achieved with the IEKF-α+CMN at $M = 128$.

5.2.5.2 MSLDM A Priori Speech Model

The recognition results obtained with the BFE employing MSLDM a priori models for the speech feature vectors and a *clean* acoustic model are listed in Tab. 5.21 for the different model-specific inference schemes and a varying number of dynamic states in the a priori model. With an increase in the number of dynamic states in the a priori model a decrease in the word error rate can be observed for the three considered model-specific inference schemes. While the application of the IEKF-α+CMN scheme again improves the recognition performance obtained with the IEKF+CMN scheme, the SOEKF-α+CMN now

Table 5.21: *Word error rates λ_{WER} [%] on the AURORA4 database obtained with the IEKF+CMN (a), the IEKF-α+CMN (b) and the SOEKF-α+CMN (c) with the* clean *acoustic model and MSLDM a priori models for the speech feature vectors with $M \in \{1,2,4,8,16,32\}$ dynamic states.*

(a) IEKF+CMN

M	clean	airport	babble	car	restaurant	street	train	$\text{AVG}_{\text{clean}}^{\text{train}}$
1	13.59	44.68	38.31	20.29	46.48	42.17	40.00	35.07
2	13.89	44.42	37.68	20.18	45.41	42.80	39.34	34.82
4	13.04	43.72	35.58	18.27	43.83	40.26	38.67	33.34
8	12.52	43.17	36.50	16.80	43.28	38.45	37.31	32.58
16	12.56	42.80	34.84	16.72	43.87	37.53	36.50	32.12
32	12.52	42.32	34.07	15.99	43.50	36.54	35.17	31.44

(b) IEKF-α+CMN

M	clean	airport	babble	car	restaurant	street	train	$\text{AVG}_{\text{clean}}^{\text{train}}$
1	13.55	44.46	38.16	20.52	46.34	41.77	39.56	34.91
2	13.70	45.08	37.64	20.59	46.11	42.43	39.48	35.00
4	13.04	43.83	35.32	17.42	44.24	39.96	38.34	33.16
8	12.49	42.17	35.14	16.94	43.79	37.20	36.02	31.96
16	12.41	41.66	33.81	16.72	42.58	36.10	35.17	31.21
32	12.45	41.58	33.89	16.32	43.17	35.54	34.48	31.06

(c) SOEKF-α+CMN

M	clean	airport	babble	car	restaurant	street	train	$\text{AVG}_{\text{clean}}^{\text{train}}$
1	14.48	46.59	39.45	21.47	49.91	44.31	42.17	36.91
2	14.59	46.70	39.04	21.33	49.39	45.30	44.05	37.20
4	13.15	42.91	35.73	17.94	46.70	39.93	39.04	33.63
8	12.74	42.28	34.81	16.57	44.31	37.31	35.58	31.94
16	12.56	42.17	33.66	16.35	43.35	36.69	34.62	31.34
32	12.63	42.25	33.26	16.28	42.73	35.95	34.59	31.10

results in a performs that is approximately as good as with the IEKF-α+CMN scheme if the number of dynamic states M is chosen appropriately.

Opposed to the AURORA 2 task, where a GMM a priori model for the speech feature vectors performed significantly better than an MSLDM, an MSLDM a priori model for the speech feature vector trajectory this time performs equally well. Considering the same number of dynamic states M, it can even be found to be slightly superior to a GMM. Though an MSLDM with M dynamic states exhibits a (possibly much) larger number of parameters to be trained than a GMM, the computational burden in the final multi-model inference scheme is approximately equivalent.

5.2.5.3 UD Variants

Finally, the best performing combinations of *a priori model* and *model-specific inference scheme* are employed in the UD framework. Table 5.23 summarizes the results without and with application of the different UD rules for the IEKF-α+CMN scheme employing a GMM a priori model with $M = 128$ dynamic states. While the uncertainty decoding schemes UD-p and UD-m lead to only slightly improved recognition results over the reference BFE scheme, application of the UD-n scheme drastically reduces the error rate from 30.55 % to 26.94 %.

The corresponding results for the best model-specific inference scheme employing an MSLDM with an a priori model consisting of $M = 32$ dynamic states are given in Tab. 5.25. Again, the UD-p and UD-m scheme perform equally well but only lead to a minor error reduction over the reference BFE scheme. The best result is once more obtained with the UD-n scheme, eventually improving the error rate from 31.06 % to 27.74 %.

Table 5.23: *Recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the AURORA 4 database obtained with the IEKF-α+CMN (a) and the UD variants IEKF-α+CMN+UD-p (b), IEKF-α+CMN+UD-m (c) and IEKF-α+CMN+UD-n (d) with the* clean *acoustic model and a GMM a priori model for the speech feature vectors with $M = 128$ dynamic states.*

(a) *IEKF-α+CMN*

error stat.	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.14	25.38	24.97	11.68	27.11	24.79	26.74	21.26
λ_{DEL}	1.18	3.90	2.91	1.47	4.01	4.09	3.79	3.05
λ_{INS}	2.50	7.40	7.37	3.06	9.98	6.67	6.74	6.25
λ_{WER}	11.82	36.69	35.25	16.21	41.10	35.54	37.27	30.55

(b) *IEKF-α+CMN+UD-p (predictive prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.32	24.38	24.05	11.49	25.86	23.61	26.11	20.55
λ_{DEL}	1.18	3.68	2.80	1.51	3.87	3.94	3.46	2.92
λ_{INS}	2.50	7.70	7.62	3.06	10.42	6.37	6.52	6.31
λ_{WER}	12.01	35.76	34.48	16.06	40.15	33.92	36.10	29.78

(c) *IEKF-α+CMN+UD-m (marginal prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.25	24.71	24.24	11.49	26.34	23.79	26.41	20.75
λ_{DEL}	1.18	3.57	2.84	1.55	3.50	3.90	3.43	2.85
λ_{INS}	2.58	7.73	7.62	3.20	10.17	6.48	6.45	6.32
λ_{WER}	12.01	36.02	34.70	16.24	40.00	34.18	36.28	29.92

(d) *IEKF-α+CMN+UD-n (neglected prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	AVG$^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.36	21.73	20.92	11.23	23.43	22.21	24.75	18.95
λ_{DEL}	1.18	4.01	3.90	1.66	5.16	4.53	4.86	3.61
λ_{INS}	2.62	5.52	5.05	2.84	6.08	4.27	4.27	4.38
λ_{WER}	12.15	31.27	29.87	15.73	34.66	31.01	33.89	26.94

Table 5.25: *Recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the AURORA 4 database obtained with the IEKF-α+CMN (a) and the UD variants IEKF-α+CMN+UD-p (b), IEKF-α+CMN+UD-m (c) and IEKF-α+CMN+UD-n (d) with the* clean *acoustic model and a MSLDM a priori model for the speech feature vectors with $M = 32$ dynamic states.*

(a) *IEKF-α+CMN*

error stat.	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.58	27.73	23.35	10.94	29.72	24.90	24.86	21.44
λ_{DEL}	1.22	5.52	4.38	1.80	5.23	4.71	4.97	3.98
λ_{INS}	2.65	8.32	6.15	3.57	8.21	5.93	4.64	5.64
λ_{WER}	12.45	41.58	33.89	16.32	43.17	35.54	34.48	31.06

(b) *IEKF-α+CMN+UD-p (predictive prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.58	27.00	23.02	10.94	28.73	24.31	24.35	20.99
λ_{DEL}	1.22	5.41	4.01	1.73	4.83	4.60	5.05	3.84
λ_{INS}	2.69	8.40	5.75	3.57	7.73	6.04	4.60	5.54
λ_{WER}	12.49	40.81	32.78	16.24	41.29	34.95	34.00	30.37

(c) *IEKF-α+CMN+UD-m (marginal prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.66	27.03	22.98	10.87	28.47	23.65	23.24	20.70
λ_{DEL}	1.22	5.08	3.90	1.80	4.57	4.68	4.94	3.74
λ_{INS}	2.76	8.91	6.41	3.87	8.29	6.30	4.38	5.85
λ_{WER}	12.63	41.03	33.30	16.54	41.33	34.62	32.56	30.29

(d) *IEKF-α+CMN+UD-n (neglected prior)*

error stat.	clean	airport	babble	car	restaurant	street	train	$\text{AVG}^{\text{train}}_{\text{clean}}$
λ_{SUB}	8.43	23.50	20.66	10.72	24.68	22.62	23.17	19.11
λ_{DEL}	1.22	5.12	4.60	1.80	5.45	4.79	5.75	4.10
λ_{INS}	2.62	6.56	4.27	3.20	6.59	4.68	3.76	4.53
λ_{WER}	12.27	35.17	29.54	15.73	36.72	32.08	32.67	27.74

5.3 AURORA 5 task

In this section, the performance of the proposed inference schemes for the presence of reverberation and the optional presence of noise is investigated on a small vocabulary recognition task. The employed AURORA 5 database is described in Sec. 5.3.1 in full detail. The recognizer setup used throughout the experiments is summarized in Sec. 5.3.2 and is followed by baseline recognition results presented in Sec. 5.3.3. The BFE setup is briefly summarized in Sec. 5.3.4 and the considered inference schemes are finally examined in Sec. 5.3.5.

5.3.1 AURORA 5 Database Description

In contrast to the previously described databases, the AURORA 5 database [105] is designed for the evaluation of ASR systems in the presence of both reverberation and noise and thus focuses on realistic application scenarios for hands-free speech input. Note, however, that the database is artificially composed. As the AURORA 2 database, it is based on the TIDigits corpus whose data are decimated to $T_S^{-1} = 8$ kHz. Two main scenarios are considered; one for the application of a hands-free system in a car and one for the respective application in interiors. This work focuses only on the latter, since the extent and impact of reverberation faced in interiors may be considered significantly more pronounced than for in-car scenarios.

For the *interior* scenario, two training conditions (*clean* and *multi-condition*) and two test conditions (*office* and *living room*) are defined.

The data employed for the *clean* training comprise $8,623$ utterances from the clean data of the TIDigits database. The data used for the *multi-condition* training are based on the ones also used for the *clean* training. However, the clean speech signals are reverberated with artificially generated AIRs exhibiting a reverberation time of 300 ms to 500 ms and noise taken from recordings of 5 different interiors, namely a *shopping mall*, a *restaurant*, an *exhibition hall*, an *office* and a *hotel lobby*, is added at a global broadband RNR of ∞ dB, 15 dB, 10 dB, 5 dB, and 0 dB.

The data for the test sets consist of all $8,700$ utterances taken from the test part of the TIDigits database. The data are, in the same ways as the multi-condition training data, artificially distorted by reverberation and noise. For the *office* scenario, artificially generated AIRs with reverberation times in the range of 300 ms-400 ms have been employed. For the *living room* scenario, artificially generated AIRs with reverberation times in the range of 400 ms-500 ms have been employed. The ***d**irect-to-**r**everberant energy **r**atio* (*DRR*) of the AIRs is about -6 dB in all cases.

More details about the employed simulation of realistic acoustic input scenarios can be found in [106].

5.3.2 Recognizer Setup

The acoustic model for the AURORA 5 task comprises, as the ones trained on the AURORA 2 database, whole-word HMMs for the digits 0-9, of which the digit 0 is represented by two models, namely *zero* and *oh*. However, it only contains a single model to represent the silence at the beginning and the end of each utterance (denoted by *sil*) and in particular no short-pause model.

The digit models employ $16/18$ HMM states (16 emitting states, 2 non-emitting states) with the respective emission density modeled by diagonal-covariance GMMs with 4 mixture components, each. The HMM topology is strictly left-to-right (linear) and no skips are allowed.

The *sil* model employs $3/5$ HMM states with diagonal-covariance GMMs of 36 mixture components. A deviation from the left-to-right HMM topology employed for the digit models is introduced by allowing state *skips* and an additional transition from the last emitting state to the first emitting state to also model periods of short and perseverative speech absence, respectively.

Training of the HMMs is carried out using the *HTK* employing the ML criterion on either the *clean* or *multi-condition* training data.

The recognition again employs a zero-gram language model with language model scale factor $\alpha_{\mathsf{LMS}} = 1$ and word insertion penalty $\alpha_{\mathsf{WIP}} = 0$.

Recognition results are given in terms of word error rates (see (3.54)).

5.3.3 Baseline Results

The baseline results obtained on the AURORA 5 database with a *clean* acoustic model are listed in Tab. 5.27 for the SFE, the SFE+CMN and the AFE, respectively.

Table 5.27: *Baseline recognition error rates λ_{WER} [%] on the* office *and the* living room *test set of the AURORA 5 database obtained with the SFE, the SFE+CMN and the AFE with the respective* clean *acoustic models.*

	SFE		**SFE+CMN**		**AFE**	
RNR [dB]	**office**	**living room**	**office**	**living room**	**office**	**living room**
∞	6.18	14.85	7.00	16.25	6.24	13.95
15	22.34	35.56	23.86	40.82	10.74	19.50
10	44.75	57.40	41.10	58.84	17.25	26.72
5	71.70	78.99	65.78	79.66	29.47	40.60
0	88.10	89.71	88.45	94.55	49.89	60.09

The corresponding error rates λ_{WER} on the noise-free, non-reverberant test data with the *clean* acoustic model are given by 0.64 %, 0.62 %, and 0.61 % for the SFE, the SFE+CMN and the AFE, respectively. The overall error rates can be found to considerably rise in the presence of reverberation and in the additional presence of background noise. Thereby, the AFE again seems to be better suited to compensate for the acoustic model mismatch than the SFE and the SFE+CMN under low global broadband RNR conditions. However, in the presence of reverberation and the absence of noise it does not provide reduced error rates. While it has been designed to counteract the influence of additive background noise on the extracted feature vectors, its deficiency in counteracting the influence of reverberation comes with no surprise.

It can further be observed that the additional application of CMN in general is incapable of improving the recognition performance over the SFE. In particular, it does not reduce the error rates in the presence of reverberation and the absence of noise, which, as discussed in the introductory section of this chapter, may be attributed to the invalidity of the MTFA for the two scenarios with a reverberation time of approximately $T_{60} \approx 350\,\text{ms}$ and $T_{60} \approx 450\,\text{ms}$ for the *office* and the *living room*, respectively.

The corresponding recognition results obtained with the *multi-condition* acoustic model are listed in Tab. 5.28. With the acoustic model matched to the noisy reverberant test data to a certain extent, the recognition errors are greatly reduced, even in the presence of noise at a low global broadband RNR. Under the matched condition, all front-end feature extraction schemes perform equally well along the different global broadband RNR conditions.

Table 5.28: *Baseline recognition error rates λ_{WER} [%] on the* office *and the* living room *test set of the AURORA 5 database obtained with the SFE, the SFE+CMN and the AFE with the respective* multi-condition *acoustic models.*

	SFE		SFE+CMN		AFE	
RNR [dB]	**office**	**living room**	**office**	**living room**	**office**	**living room**
∞	2.23	3.53	2.31	3.80	2.42	3.94
15	5.63	7.52	5.16	7.01	4.38	6.20
10	9.58	12.62	8.63	11.93	7.91	10.57
5	18.80	23.44	16.92	21.89	15.73	20.45
0	37.46	45.01	34.92	42.39	33.27	40.51

5.3.4 BFE Setup

The BFE setup for the AURORA 5 task is almost equivalent to the AURORA 2 task. However, only the MSLDM a priori models with $M \in \{1,2,4,8,16,32\}$ are considered. Though preliminary experiments have also been carried out with GMMs as a priori models for the speech feature vector trajectory, the obtained recognition results turned out to be even worse than the ones with the corresponding baseline system and thus are not reported here. The performance loss due to a GMM a priori speech model may be reasoned as follows: Unlike an MSLDM, a GMM is not capable of capturing the temporal correlations among adjacent LMPSC feature vectors of the speech signal. This, however, is a prerequisite of being able to discriminate between the intrinsic correlation of the speech feature vectors from that introduced by reverberation.

The single Gaussian a priori model for the noise feature vector trajectory is again trained on a per-utterance basis. However, this time only the first and last 15 frames of each utterance are employed to train its parameters in an ML fashion.

Since the state vector incorporates a total of L_C feature vectors of the clean speech signal for the non-recursive and the recursive observation model[1], only the IEKF is applied in the model-specific inference. While this decision has primarily been motivated by the increased computational burden coming along with the application of the SOEKF, the results on the AURORA 4 recognition task further revealed that it does not necessarily outperform the IEKF.

This definition of the state vector further allows a non-causal estimation of the a posteriori PDF of the clean speech feature vector by introducing a lag of $L_C - 1$ frames into the

[1] For the recursive observation model, the recursion length is formally set to and $L_R = L_C$

inference. Hence, not only the GPB1-matched mean vector $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ may be forwarded to the recognizer, but also the fixed-lag smoothed MMSE estimates $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t+L_C-1}^{(\mathsf{l})}}$.

Equivalently, in case uncertainty decoding is considered, also the corresponding covariance matrices $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ and $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ for the causal UD rule and $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}|\breve{\mathbf{o}}_{1:t+L_C-1}^{(\mathsf{l})}}$ for the non-causal UD rule may be employed.

To distinguish the different observation models employed for the AURORA 5 task, the following inference schemes are defined:

- IEKF+CMS: The IEKF+CMS employs the IEKF in the BFE and will be applied in the presence of reverberation and the absence of noise. The mean vector and (diagonal) covariance matrix of the observation errors are trained on artificially reverberated training data (here: 800 utterances) in the ML fashion.

- IEKF-TI+CMN: The IEKF-TI employs the IEKF in the BFE and will be applied in the presence of reverberation and additional background noise. The mean vector and (diagonal) covariance matrix of the observation errors are the fixed, (T)ime-(I)nvariant mean vector and (diagonal) covariance matrix already used for the IEKF+CMS.

- IEKF-α-AUG+CMN: The IEKF-α-AUG employs the (AUG)mented state vector in the IEKF and will be applied in the presence of reverberation and additional background noise. The linearization of the observation functions is carried out w.r.t. the vector of phase factors and the observation error in the presence of reverberation and the absence of noise. For the TAYLOR series expansion w.r.t. the latter, the fixed, time-invariant mean vector and (diagonal) covariance matrix trained for the IEKF are employed.

- IEKF-α-TV+CMN: The IEKF-α-TV employs the IEKF in the BFE and will be applied in the presence of reverberation and additional background noise. The mean vector and (diagonal) covariance matrix of the observation error are approximated by plugging the IRNR estimates (4.361) and (4.399) for the non-recursive and recursive observation model, respectively, into the respective definition of the mean vector and covariance matrix.

  For the non-recursive observation model these definitions are given in (4.122) and (4.123), for the recursive variants in (4.209) and (4.210), respectively. Note that the observation mapping is the same as for the IEKF-TI, however the mean and the (diagonal) covariance are now made (T)ime-(V)ariant.

CMN is thereby applied to the enhanced feature vector after computation of the dynamic features to all of the above schemes.

The best BFE scheme will again be employed with subsequent UD decoding, e.g., denoted by IEKF-TI+CMN-UD-m for the non-causal uncertainty decoding scheme employing the marginal a priori PDF.

5.3.5 Results with Bayesian Feature Enhancement

Since the AURORA 5 database is artificially composed, it allows to investigate the influence of reverberation on the recognition performance independent of that of additional background noise. The first series of experiments in Sec. 5.3.5.1 is thus dedicated to the reverberant-only case before focusing on the noisy reverberant case in Sec. 5.3.5.2.

5.3.5.1 Presence of Reverberation and Absence of Background Noise

In a first experiment, the observation models for reverberant-only speech are employed on the reverberant but noise-free test data of the AURORA 5 database. The IEKF+CMS scheme is investigated for $M \in \{1,2,4,8,16,32\}$ of dynamic states in the MSLDM a priori model and $L_C \in \{1,2,3,4,5,6\}$ LMPSC vectors of the clean speech signal in the state vector. Further, both the causal estimate $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})} \big| \breve{\mathbf{o}}_{1:t}^{(\mathsf{l})}}$ and the non-causal estimate $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})} \big| \breve{\mathbf{o}}_{1:t+L_C-1}^{(\mathsf{l})}}$ are employed in the non-recursive and recursive observation model, respectively. The obtained recognition results are listed in Tab. 5.29.

Comparing the use of the causal and the non-causal estimate, first, it can be observed that employing the non-causal estimate with $L_C = L_R > 1$ always provides results that are superior to those obtained with the causal estimate[2]. This benefit may be attributed to the observation model, which ties the current observation to the most recent L_H LMPSC vectors of the clean speech signal, of which the most recent L_C are part of the state vector. Hence, by introducing a lag of $L_C - 1$ into the estimation, the uncertainty about the actual value of the LMPSC vector of the clean speech signal may be reduced, eventually resulting in an improved MMSE estimate that is forwarded to the recognizer.

In tendency, for all considered number of dynamic states M in the a priori model, the increase of the number L_C of LMPSC feature vectors of the clean speech signal in the state vector leads to a reduction of the error rate. The most dominant improvement is thereby observable when increasing L_C from one to two.

Interestingly, an increase in the number of dynamic states M only leads to reduced error rates if the non-causal estimation is considered. With the causal estimate, the recognition results become slightly worse with increasing M. This observation again highlights the sensitivity of the inference w.r.t. the employed a priori model. With $M = 1$ dynamic state, the only approximation lies in the model-specific inference, i.e., the IEKF. By increasing M, the approximations required for a feasible multi-model inference compromise a causal estimation. This effect may, however, be overcompensated by targeting the non-causal estimation, instead.

Table 5.29 further allows to access the sensitivity of the inference w.r.t. the employed observation model. While the recursive observation model results in superior recognition results for lower L_C, the non-recursive observation model excels at higher values of L_C. Thereby, the recursive observation model is slightly more efficient in terms of both computational complexity and memory requirements (see also [82]).

The on average best results are obtained with the non-causal estimate of the LMPSC vector of the clean speech signal of the BFE scheme with the non-recursive observation model employing an MSLDM a priori model with $M = 32$ dynamic states estimate and

[2] Note that for $L_C = L_R = 1$ the non-causal estimate reduces to the causal estimate.

Table 5.29: *Word error rates λ_{WER} [%] obtained with the IEKF+CMS employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ dynamic states and $L_C \in \{1,2,3,4,5,6\}$ LMPSC feature vectors of the clean speech signal in the state vector applied to the reverberant but noise-free test utterances of the AURORA 5 database. Both the non-recursive and the recursive observation model is employed with either the causal or the non-causal MMSE estimate, i.e., the GPB1-matched mean vector of the* Gaussian *a posteriori PDF.*

		causal estimate				non-causal estimate			
		non-recursive observation model		recursive observation model		non-recursive observation model		recursive observation model	
M	**$L_C = L_R$**	**office**	**living room**	**office**	**living room**	**office**	**living room**	**office**	**living room**
1	**1**	5.81	16.11	3.49	13.26	5.81	16.11	3.49	13.26
	2	3.31	10.28	2.69	9.57	3.05	8.76	2.52	7.79
	3	2.95	9.48	2.68	8.76	2.49	7.11	2.47	6.75
	4	2.79	9.16	2.71	8.60	2.20	6.06	2.55	6.46
	5	2.72	9.01	2.75	8.53	2.04	5.52	2.41	5.99
	6	2.69	8.95	2.67	8.67	1.97	5.11	2.28	5.63
2	**1**	5.34	15.70	3.67	14.83	5.34	15.70	3.67	14.83
	2	3.12	9.93	2.57	10.07	2.57	8.24	2.21	7.48
	3	2.80	9.31	2.59	9.04	2.07	6.36	2.14	6.01
	4	2.77	9.22	2.57	8.75	1.82	5.33	2.09	5.64
	5	2.63	9.04	2.66	8.93	1.75	4.86	1.98	5.29
	6	2.63	9.08	2.69	8.81	1.73	4.56	1.95	5.12
4	**1**	4.54	12.92	3.74	12.19	4.54	12.92	3.74	12.19
	2	3.02	9.09	2.49	8.46	2.68	7.69	2.13	6.62
	3	2.90	8.74	2.66	8.23	2.21	6.14	2.13	5.70
	4	2.74	8.70	2.74	8.46	1.96	5.16	2.15	5.43
	5	2.79	8.73	2.84	8.61	1.83	4.62	2.06	5.02
	6	2.79	8.74	2.89	8.72	1.82	4.19	2.00	4.77
8	**1**	4.63	12.18	3.33	10.67	4.63	12.18	3.33	10.67
	2	3.10	8.57	2.34	7.78	2.65	7.23	2.03	6.04
	3	2.92	8.18	2.49	7.81	2.20	5.62	1.96	5.21
	4	2.80	8.13	2.73	8.00	2.00	4.73	2.00	4.90
	5	2.84	8.24	2.87	8.33	1.79	4.13	1.96	4.64
	6	2.83	8.20	2.93	8.33	1.77	3.84	1.91	4.50
16	**1**	4.08	11.17	3.33	10.22	4.08	11.17	3.33	10.22
	2	3.16	8.34	2.47	7.48	2.68	7.11	2.03	5.79
	3	2.91	8.04	2.64	7.52	2.20	5.52	1.93	5.01
	4	2.88	8.07	2.78	7.98	1.92	4.58	1.96	4.76
	5	2.94	8.13	3.01	8.34	1.75	3.95	1.86	4.61
	6	2.95	8.24	3.11	8.34	1.71	3.84	1.80	4.28
32	**1**	5.14	12.00	3.60	10.10	5.14	12.00	3.60	10.10
	2	4.07	8.87	2.62	7.14	3.37	7.48	2.18	5.68
	3	3.69	8.30	3.03	7.56	2.44	5.76	2.20	4.94
	4	3.45	8.30	3.29	8.19	2.03	4.71	2.16	4.68
	5	3.41	8.43	3.48	8.58	1.81	3.90	1.94	4.33
	6	3.48	8.61	3.66	8.95	1.74	3.75	1.90	4.00

$L_C = 6$. The respective error rates for the office and the living room environment are 1.74 % and 3.75 % – a relative improvement of 75 % and 80 % over the baseline! With the same setup, use of the recursive observation model gives approximately equivalent recognition results of 1.90 % and 4.00 % for the office and the living room environment, respectively.

Special note has to be taken of the fact that these relatively small improvements over the error rate at $M = 1$ and $L_C = 6$ come with a large (approximately 32-fold) increase of the computational complexity and the memory requirements.

By additionally exploiting the remaining uncertainty of the estimate in the non-causal UD-m and UD-n scheme[3], the error rates for the office environment can further be reduced (see Tab. 5.30). However, the very small improvement of the error rates comes with a large increase of the computational complexity – this time in the back-end of the ASR system.

Table 5.30: *Word error rates λ_{WER} [%] obtained with the IEKF+CMS and the uncertainty decoding variants employing an MSLDM with $M = 32$ dynamic states and $L_C = 6$ LMPSC feature vectors of the clean speech signal in the state vector applied to the reverberant but noise-free test utterances of the AURORA 5 database. Both the non-recursive and the recursive observation model are employed with the non-causal MMSE estimate in the UD-m and UD-n scheme.*

	non-recursive observation model		recursive observation model	
processing scheme	**office**	**living room**	**office**	**living room**
IEKF+CMS	1.74	3.75	1.90	4.00
IEKF+CMS+UD-m	1.70	3.62	1.75	3.86
IEKF+CMS+UD-n	1.61	3.90	1.57	4.15

5.3.5.2 Presence of Reverberation and Background Noise

In this second experiment, the observation models for noisy reverberant speech are employed on the complete, noisy reverberant test data of the AURORA 5 database. The IEKF-TI+CMN scheme, the IEKF-α-AUG+CMN scheme and the IEKF-α-TV+CMN scheme is investigated for $M \in \{1, 2, 4, 8, 16, 32\}$ dynamic states in the MSLDM a priori model and $L_C = 6$ LMPSC vectors of the clean speech signal in the state vector.

Since both IEKF-α-AUG+CMN and IEKF-α-TV+CMN are based on the theoretically exact relation of the respective variables in the state vector and the observation, each represents a putative better way of modeling the observation since they both, as illustrated in Fig. 4.31, are capable of modeling the covariance matrix of the observation error as a function of the IRNR estimates[4]. While the IEKF-α-AUG+CMN thereby only relies on the VTS expansion points, the IEKF-α-TV+CMN also takes into account the associated covariance matrices.

[3] Since the use of the non-causal estimate is considerably superior to the use of the causal estimate, the causal UD schemes, which are targeting a causal estimation, are not taken into account here.

[4] Strictly speaking, the IEKF-α-AUG+CMN does not incorporate an observation error. However, as illustrated in Fig. 4.31 and outlined in the corresponding discussion, the *additional* covariance matrix employed in the IEKF update formulas of the IEKF-α-AUG+CMN may well be compared to the observation error covariance matrix computed by the IEKF-TI+CMN.

In contrast, the IEKF-TI+CMN models the covariance matrix of the observation error a fixed, time-invariant quantity independent of the IRNR (and also independent of the current dynamic state m_t of the a priori model). While this is certainly sub-optimal when looking at the observation model alone (recall the discussion on the observation error in Sec. 4.7.2), the covariance matrix of the observation error will always be *larger* than or equal to the covariance matrices estimated by the IEKF-α-AUG+CMN and IEKF-α-TV+CMN (compare Fig. 4.31). This additional uncertainty may, however, be useful to model the additional error arising due to the linearization of the observation function/mapping.

The obtained recognition results are listed in Tab. 5.31 and Tab. 5.32 for the non-recursive and recursive observation model, respectively. Since the non-causal estimate has been found to provide substantially better results on the reverberant-only data, only the non-causal estimate $\boldsymbol{\mu}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t+L_C-1}}$ and, later on, the corresponding covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{GPB1})}_{\breve{\mathbf{x}}_t^{(\mathsf{l})}\middle|\breve{\mathbf{o}}^{(\mathsf{l})}_{1:t+L_C-1}}$ are considered here.

Looking at the non-recursive observation models, first, it can be observed that the recognition results listed in Tab. 5.31 are highly depending on the employed observation model, the number M of dynamic states in the a priori model and the global broadband RNR the reverberant speech signal and the noise mix at.

For $M = 1$ and $M = 32$, the IEKF-TI+CMN and the IEKF-α-TV+CMN can be found to perform equally well over the considered range of global broadband RNRs. Since the two schemes only differ in the way the covariance matrix of the observation error is modeled, this observation may be reasoned as follows.

With $M = 1$ the prediction of the state vector is associated with a large uncertainty majorly because a single a priori model is not capable of predicting the state vector very accurately. As a consequence, the spectral radius of the covariance matrix of the predictive PDF is large and causes the estimate of the IRNR employed in the IEKF-α-TV+CMN to approach infinity. Thus, the corresponding covariance matrix of the observation error approaches the time-invariant covariance matrix employed in the IEKF-TI+CMN (compare Fig. 4.31) and the two BFE schemes turn out to be essentially equivalent.

On the other hand, with increasing M, the prediction of the state vector is associated with a larger uncertainty because of the sub-optimality of the multi-model inference, i.e., the employed IMM mixing. Even though the spectral radii of the model-specific prediction error covariance matrices are getting smaller with increasing M, the diversity of the model-specific state estimates employed in the calculation of the *mixing* covariance matrices (see (4.300)) renders the spectral radius of the covariance matrix of the model-specific predictive PDF large, again. The corresponding covariance matrix of the observation error then again approaches the time-invariant covariance matrix employed in the IEKF-TI+CMN (compare Fig. 4.31) and the two BFE schemes turn out to be essentially equivalent.

Interestingly, for $M \in \{4, 8, 16\}$, these two effects are not that dominant and the benefit of modeling the observation error as a function of the (estimate of the) IRNR in the IEKF-α-TV+CMN becomes visible. While the two inference schemes do not differ much for high values of the global broadband RNR, a superiority of the IEKF-α-TV+CMN can be observed for low and mid-level global broadband RNRs.

Similar observations can be made when looking at the performance of the IEKF-α-AUG+CMN. Since the modeling of the observation error covariance matrix thereby does not take into account the uncertainty associated with the VTS expansion points, it is not affected by possible errors in the estimation/inference of the associated covariance matrix. Together with the simple, single Gaussian a priori model for the noise, this causes the IEKF-α-AUG+CMN to outperform the IEKF-α-TV+CMN and the IEKF-TI+CMN for $M \in \{4, 8\}$ at low global broadband RNR values. It, however, is always inferior to the two other schemes at high values of the global broadband RNR. Further, at $M = 32$ it is much worse than the IEKF-TI+CMN and the IEKF-α-TV+CMN.

This and the irregularity of the results at $M = 2$ again highlight the sensitivity of the inference of the a posteriori PDF of the clean speech LMPSC feature vector not only w.r.t. the employed observation model but also w.r.t. the chosen a priori model. While the above analysis in this regard only targets the number of dynamic states M building the MSLDM, this sensitivity also extends to different training approaches, as can be inferred from [107].

Table 5.31: *Word error rates λ_{WER} [%] on the* office *and the* living room *test set of the AURORA 5 database obtained with the non-recursive observation model with $L_C = 6$ in the IEKF-TI+CMN, the IEKF-α-AUG+CMN and the IEKF-α-TV+CMN model-specific inference. The a priori model for the speech feature vectors is an MSLDM with $M \in \{1, 2, 4, 8, 16, 32\}$ dynamic states. The non-causal (lag $L_C = 6$) MMSE estimate is passed to the recognizer, which employs the* clean *acoustic model.*

		IEKF-TI+CMN		IEKF-α-AUG+CMN		IEKF-α-TV+CMN	
M	RNR [dB]	office	living room	office	living room	office	living room
1	∞	2.37	6.00	3.55	7.17	2.50	6.16
	15	11.29	19.44	18.75	27.18	11.47	19.67
	10	21.41	31.52	32.25	41.42	21.66	31.83
	5	39.99	50.35	51.93	60.16	40.16	50.57
	0	66.18	72.71	73.17	78.51	65.87	72.55
2	∞	6.84	11.25	4.47	8.51	12.50	20.81
	15	21.32	30.64	21.00	29.23	33.08	41.68
	10	41.21	51.35	36.27	45.35	51.40	59.79
	5	69.34	77.12	58.15	65.85	74.54	80.95
	0	91.28	94.24	79.40	83.75	92.21	94.71
4	∞	2.37	5.13	2.35	5.27	2.60	6.02
	15	8.78	16.29	10.85	17.14	9.25	16.69
	10	20.46	30.89	21.19	29.37	19.19	29.01
	5	43.62	55.35	39.80	48.76	39.14	50.26
	0	74.69	81.33	64.24	70.21	68.25	75.71
8	∞	2.09	4.56	2.50	4.95	2.40	5.64
	15	9.18	16.52	11.06	16.77	8.39	15.11
	10	21.39	31.25	20.94	28.48	17.65	27.17
	5	45.73	56.50	39.29	48.27	37.48	48.58
	0	76.52	82.37	63.14	69.47	66.00	74.21
16	∞	1.77	3.84	1.91	4.08	1.90	4.33
	15	6.35	12.22	11.23	16.30	7.50	13.29
	10	13.75	22.97	21.65	28.81	15.10	23.62
	5	31.83	43.41	40.32	49.01	31.29	41.86
	0	62.11	70.08	64.55	70.92	57.70	66.37
32	∞	1.68	3.75	1.96	3.98	2.00	4.33
	15	6.59	11.80	12.87	17.26	7.92	12.71
	10	13.40	21.66	24.04	30.27	14.93	22.77
	5	28.17	38.58	42.90	50.58	30.00	39.70
	0	52.48	62.03	65.81	71.97	53.92	62.50

There, a *kmeans++*-inspired initialization has been employed to train an MSLDM with $M=4$ dynamic states. The application of the IEKF-α-TV+CMN inference scheme results in a recognition performance that can only be achieved with an MSLDM with $M=16$ dynamic states trained with the splitting scheme used in this work – notably, the former model achieves this results with only a quarter of the computational complexity required by the latter.

Without any further discussion, these findings also apply to the recursive counterparts of the three observation models, whose results are listed in Tab. 5.32. In a last experiment on the AURORA 5 database, the best performing combination of a priori model and observation model – the IEKF-TI+CMN with $M=32$ – is combined with additional uncertainty decoding in the non-causal UD-m and UD-n scheme. The results are listed in Tab. 5.33 for the respective non-recursive and recursive variant. Again, only the UD-m scheme results

Table 5.32: *Word error rates λ_{WER} [%] on the* office *and the* living room *test set of the AURORA 5 database obtained with the recursive observation model with $L_C = L_R = 6$ in the IEKF-TI+CMN, the IEKF-α-AUG+CMN and the IEKF-α-TV+CMN model-specific inference. The a priori model for the speech feature vectors is an MSLDM with $M \in \{1,2,4,8,16,32\}$ dynamic states. The non-causal (lag $L_C = L_R = 6$) MMSE estimate is passed to the recognizer, which employs the* clean *acoustic model.*

		IEKF-TI+CMN		**IEKF-α-AUG+CMN**		**IEKF-α-TV+CMN**	
M	**RNR** [dB]	**office**	**living room**	**office**	**living room**	**office**	**living room**
1	∞	2.81	6.69	4.50	8.56	2.96	6.99
	15	11.03	18.89	17.26	25.50	11.44	19.32
	10	20.78	31.21	30.35	39.41	21.48	31.57
	5	39.01	49.73	48.89	58.17	39.49	50.01
	0	64.58	72.67	71.07	77.26	64.39	72.47
2	∞	4.55	8.59	5.54	10.08	16.22	23.72
	15	21.92	31.76	24.90	35.84	34.47	43.59
	10	42.06	53.68	42.42	54.41	51.63	61.04
	5	69.58	79.36	65.73	76.56	73.82	81.21
	0	91.15	94.87	86.81	92.28	91.15	94.32
4	∞	2.26	5.48	2.50	5.78	3.04	7.10
	15	8.61	15.88	10.09	16.24	10.29	17.96
	10	19.94	30.55	20.30	29.39	20.31	31.12
	5	42.41	55.27	39.09	51.18	40.40	53.44
	0	73.07	81.26	66.20	76.15	69.37	78.75
8	∞	2.05	5.11	2.35	5.50	2.74	6.84
	15	8.88	15.97	9.63	15.31	9.44	16.61
	10	20.75	31.28	19.27	28.21	19.08	29.43
	5	44.59	56.65	38.95	51.35	39.12	52.10
	0	75.07	82.69	66.48	76.76	67.78	77.68
16	∞	1.85	4.56	2.08	5.00	2.10	5.31
	15	6.59	11.81	10.63	15.11	8.24	13.63
	10	13.91	22.48	20.11	26.44	16.07	24.39
	5	31.27	42.76	38.08	44.94	32.69	43.69
	0	60.69	70.10	61.65	67.99	59.26	68.33
32	∞	1.98	4.44	2.27	4.74	2.39	5.29
	15	6.85	11.51	12.55	16.27	8.56	12.88
	10	13.81	21.16	23.05	28.21	15.78	22.54
	5	28.28	37.80	40.99	47.12	30.65	39.85
	0	52.50	61.27	63.31	68.02	54.40	62.82

Table 5.33: Word error rates λ_{WER} [%] obtained with the IEKF-TI+CMN and its UD variants employing an MSLDM with $M = 32$ dynamic states and $L_C = 6$ LMPSC feature vectors of the clean speech signal in the state vector applied to the noisy reverberant test utterances of the AURORA 5 database. Both the non-recursive and the recursive observation model are employed with the non-causal MMSE estimate in the UD-m and UD-n scheme.

		non-recursive observation model		recursive observation model	
processing scheme	**RNR [dB]**	**office**	**living room**	**office**	**living room**
IEKF-TI+CMN	∞	1.68	3.75	1.98	4.44
	15	6.59	11.80	6.85	11.51
	10	13.40	21.66	13.81	21.16
	5	28.17	38.58	28.28	37.80
	0	52.48	62.03	52.50	61.27
IEKF-TI+CMN+UD-m	∞	1.70	3.64	1.89	4.33
	15	6.36	11.00	6.75	10.75
	10	12.75	20.40	13.23	19.97
	5	27.17	37.50	27.13	36.47
	0	51.75	61.17	51.70	60.64
IEKF-TI+CMN+UD-n	∞	1.66	4.05	1.77	4.89
	15	6.57	11.88	6.56	11.37
	10	13.97	22.51	13.91	21.46
	5	29.68	40.51	29.31	39.44
	0	54.80	64.20	54.71	63.58

in an improved recognition performance compared to the IEKF-TI+CMN without UD. The application of the UD-n scheme even slightly worsens the results.

5.4 MC-WSJ-AV task

In this section, the performance of the proposed inference schemes for the presence of reverberation and noise is investigated on a large vocabulary recognition task. The employed WSJCAM 0 database (training) and MC-WSJ-AV database (testing) are described in Sec. 5.4.1 and Sec. 5.4.2 in full detail. The recognizer setup used throughout the experiments is summarized in Sec. 5.4.3 and is followed by baseline recognition results on the WSJCAM 0 database presented in Sec. 5.4.4 and baseline recognition results on the MC-WSJ-AV database presented in Sec. 5.4.5 and The BFE setup is briefly summarized in Sec. 5.4.6 and the considered inference schemes are finally examined in Sec. 5.4.7.

5.4.1 WSJCAM 0 Database Description

The WSJCAM 0 database [108] is designed for the evaluation of ASR systems for British English. It is based on the WSJ 0 corpus, which has partly been re-recorded with 140 British English speaking speakers.

Opposed to the aforementioned databases, the WSJCAM 0 corpus only provides the *raw* recordings obtained with both a headset and a desk microphone. The data are recorded at a sampling rate of $T_S^{-1} = 16$ kHz and exhibit a global broadband SNR of 35 dB-45 dB and 20 dB-25 dB for the headset and the desk microphone, respectively. However, only the data from the headset microphone are considered further.

The data for a *clean* training condition comprise $7,861$ utterances from 92 different speakers (53 male speakers and 39 female speakers), all with British English accent.

For test purposes, a $5,000$ words closed vocabulary task and a $20,000$ words open vocabulary task are defined. The test utterances are spoken by 48 different speakers and the recordings are further divided into two development and two evaluation subsets. For the $5,000$ word task, the development subsets comprise 368 and 374 utterances, respectively, and the evaluation subsets 538 and 550 utterances, respectively.

Since the WSJCAM 0 corpus only serves the purpose of providing training data for the acoustic model of the speech recognizer and the a priori model for the BAYESIAN inference of the clean speech feature posterior, which eventually will both be applied to the MC-WSJ-AV task described further below, only the training data are employed in this work. Recognition results reported on the development and evaluation test data aim only at characterizing the trained acoustic model.

To also obtain *multi-condition* training data, the clean WSJCAM 0 training data are artificially distorted by reverberation and additive noise. The AIRs to convolve the clean training data with and the noises to add are essentially the ones provided with the ***RE**verberant **V**oice **E**nhancement and **R**ecognition **B**enchmark* (REVERB) challenge [109]. The (measured) AIRs thereby exhibit a reverberation time of roughly 200 ms-800 ms. Noise is only added a moderate global broadband RNR of 20 dB.

5.4.2 MC-WSJ-AV Database Description

The MC-WSJ-AV corpus is a collection of read WSJ sentences taken from the development and evaluation test sets of the WSJCAM 0 database (see Sec. 5.4.1 above). The data were recorded in a number of instrumented meeting rooms constructed within the framework of the *European **A**ugmented **M**ulti-party **I**nteraction* (AMI) project [110]. Three different scenarios were considered during the recording of the database, namely, *single stationary speaker*, *single moving speaker* and *two stationary, overlapping speakers*, of which the *single stationary speaker* subset is chosen for the experiments reported here.

For this condition, the speakers read sentences from six different positions within the meeting room — four seated around a table, one standing at a whiteboard, and one standing at a presentation screen. Data were recorded simultaneously by a headset microphone, a lapel microphone and two 8-element circular microphone arrays positioned on the table, all at a sampling rate of $T_S^{-1} = 16$ kHz. The test set used in the experiments reported here consists of the two evaluation sets (EVAL 1 and EVAL 2) recorded at the University of Edinburgh. The EVAL 1 (here denoted by evaluation 1) set features 189 sentences, approximately 3100 words and a total length of 21 min. The EVAL 2 (here denoted by evaluation 2) set features 183 sentences, approximately 3200 words and a total length of 19 min.

The global broadband RNR at the input of the microphone (the microphone denoted by "array1-1" in [110] is employed here) of the circular array was about $15-20$ dB on average. Estimates of the room reverberation time range from 380 ms [111] to 700 ms [110].

5.4.3 Recognizer Setup

The acoustic model trained on the WSJCAM 0 database exhibit the same topology and structure as the ones trained on the AURORA 4 database. Training of the HMMs is carried out using the *HTK* employing the ML criterion on either the *clean* or *multi-condition* training data. However, the pronunciations are taken from the *__B__ritish __E__nglish __E__xample __P__ronunciation* (BEEP) dictionary (version 1.0), available at http://www.speech.cs.cmu.edu.

The recognition again employs the standard MIT Lincoln Laboratories 5 k compact back-off bigram language model initially provided with the WSJ 0 database. The language model scale factor is set to $\alpha_{\mathsf{LMS}} = 16$ and the word insertion penalty to $\alpha_{\mathsf{WIP}} = 0$.

Recognition results are given in terms of word error rates.

5.4.4 WSJCAM 0 Baseline Results

The baseline recognition results obtained on the development and evaluation test sets of the WSJCAM 0 database with a *clean* acoustic model are listed in Tab. 5.34 for the SFE, the SFE+CMN and AFE, respectively. Though no further experiments will be carried out on the WSJCAM 0 database, the presented results characterize the quality of the acoustic model to be used for the recognition on the MC-WSJ-AV database. Since the data of the WSJCAM 0 database are neither affected by reverberation nor by additive background noise, the recognition error rates are almost identical among the three considered front-ends.

Table 5.34: *Baseline recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the WSJCAM 0 database obtained with the SFE, the SFE+CMN and the AFE with the* clean *acoustic model.*

	SFE				**SFE+CMN**				**AFE**			
error	**development**		**evaluation**		**development**		**evaluation**		**development**		**evaluation**	
statistic	**set 1**	**set 2**	**set 1**	**set 2**	**set 1**	**set 2**	**set 1**	**set 2**	**set 1**	**set 2**	**set 1**	**set 2**
λ_{SUB}	8.28	8.94	10.66	8.51	7.64	8.22	10.42	8.22	8.33	8.98	10.52	8.55
λ_{DEL}	2.57	2.30	2.59	2.29	2.41	2.23	2.73	2.30	2.96	3.04	2.82	2.32
λ_{INS}	1.34	1.34	1.99	1.54	1.44	1.27	2.09	1.43	1.28	1.40	2.07	1.54
λ_{WER}	12.20	12.58	15.23	12.34	11.50	11.72	15.24	11.96	12.57	13.42	15.41	12.42

5.4.5 MC-WSJ-AV Baseline Results

The baseline results obtained on the MC-WSJ-AV database with a *clean* and *multi-condition* acoustic model trained on the WSJCAM 0 database are listed in Tab. 5.35 and Tab. 5.37 for the SFE, the SFE+CMN and AFE, respectively.[5] While the recognition performance employing the *headset* microphone is only a little worse than on the WSJCAM 0 database (indicating a sufficient match of the acoustic model), it considerably decreases when the *lapel* microphone or the *array1-1* microphone is used. In fact, with the single distant

[5] For the SFE and the AFE, the audio data of each utterance have been scaled to ensure the power of the underlying reverberant test data and the (clean) training data to be equal. Without this scaling, which is also employed to compute the energy parameter $\sigma_{\check{h}}$ in (4.157), the recognition performance is, due to a severe model mismatch, much worse.

Table 5.35: *Baseline recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the MC-WSJ-AV database obtained with the SFE, the SFE+CMN and the AFE with the* clean *acoustic model.*

(a) SFE

error statistic	development			evaluation 1			evaluation 2		
	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB}	56.07	18.30	41.36	61.43	11.93	38.47	66.30	22.93	53.23
λ_{DEL}	36.53	3.75	12.26	29.07	2.62	7.95	27.33	3.43	8.90
λ_{INS}	2.22	3.62	5.48	3.85	2.85	6.08	3.03	5.27	8.97
λ_{WER}	94.81	25.67	59.10	94.34	17.39	52.51	96.67	31.63	71.10

(b) SFE+CMN

error statistic	development			evaluation 1			evaluation 2		
	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB}	46.97	14.19	21.62	50.79	9.93	15.65	57.63	19.13	30.57
λ_{DEL}	36.56	3.33	7.47	28.55	3.14	5.30	31.57	3.37	7.37
λ_{INS}	1.70	2.94	2.38	3.01	1.94	2.17	1.87	4.07	4.17
λ_{WER}	85.23	20.45	31.47	82.35	15.00	23.12	91.07	26.57	42.10

(c) AFE

error statistic	development			evaluation 1			evaluation 2		
	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB}	61.12	23.65	28.25	58.75	12.22	21.08	70.90	23.27	35.70
λ_{DEL}	17.16	5.48	7.18	13.51	3.17	4.40	14.10	3.43	5.37
λ_{INS}	3.69	4.04	2.87	6.27	2.78	3.56	5.53	4.43	6.30
λ_{WER}	81.96	33.17	38.29	78.53	18.17	29.03	90.53	31.13	47.37

microphone *array1-1*, the word error rate increases to about 80% for the (on average) best front-end, namely SFE+CMN. This decrease in recognition performance ultimately indicates the need to counteract the detrimental effect of reverberation and noise.

With the multi-condition model targeting noisy reverberant speech, the recognition performance on the headset can first be observed to considerably drop compared to the use of a clean acoustic model. However, the multi-condition model shows a better match with the noisy reverberant data of the array1-1 and the lapel microphone than the clean acoustic model. Nevertheless, the word error rate is still about or above 50 % for the three considered baseline front-end schemes, of which the SFE+CMN again yields the best performance.

5.4.6 BFE Setup

The BFE setup for the MC-WSJ-AV task is almost equivalent to that employed for the AURORA 5 task.

The single Gaussian a priori model for the LMPSC vectors of the noise is again trained on a per-utterance basis. However, this time only the first 12 frames of each utterance are employed for training of its parameters.

Opposed to the AURORA 5 database, where the (average) reverberation time has been assumed to be known in beforehand (office: $T_{60} = 350\,$ms, living room: $T_{60} = 450\,$ms), the reverberation time is not known for the room the recordings for the considered test sets

Table 5.37: *Baseline recognition statistics λ_{SUB} [%], λ_{DEL} [%], λ_{INS} [%] and λ_{WER} [%] on the MC-WSJ-AV database obtained with the SFE (a), the SFE+CMN (b) and the AFE (c) with the* multi-condition *acoustic model.*

(a) SFE

error	development			evaluation 1			evaluation 2		
stat.	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB} [%]	49.25	31.51	40.77	46.46	23.12	34.72	56.67	38.33	50.17
λ_{DEL} [%]	20.12	3.69	7.53	17.65	2.85	5.98	18.10	3.17	6.33
λ_{INS} [%]	4.01	13.76	9.26	5.27	7.40	8.37	4.17	15.00	11.17
λ_{WER} [%]	73.39	48.96	57.57	69.38	33.37	49.08	78.93	56.50	67.67

(b) SFE+CMN

error	development			evaluation 1			evaluation 2		
stat.	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB} [%]	32.71	23.22	18.98	30.46	16.62	14.65	42.67	32.17	27.07
λ_{DEL} [%]	19.41	3.39	5.54	15.26	2.62	3.94	18.40	3.20	5.47
λ_{INS} [%]	1.30	7.08	2.25	1.75	4.49	2.23	2.10	10.30	4.53
λ_{WER} [%]	53.42	33.69	26.78	47.46	23.73	20.82	63.17	45.67	37.07

(c) AFE

error	development			evaluation 1			evaluation 2		
stat.	array1-1	headset	lapel	array1-1	headset	lapel	array1-1	headset	lapel
λ_{SUB} [%]	39.20	31.18	29.94	36.40	21.56	22.21	48.00	33.60	35.33
λ_{DEL} [%]	13.67	5.90	6.78	11.87	3.52	4.62	13.20	3.83	5.80
λ_{INS} [%]	5.09	7.80	4.76	4.95	6.43	5.04	4.87	9.37	7.63
λ_{WER} [%]	57.96	44.88	41.49	53.22	31.52	31.88	66.07	46.80	48.77

of the MC-WSJ-AV database have been made in. Since a sensitive blind estimation of the reverberation time is out of scope of this work, the sensitivity of the proposed inference schemes w.r.t. the estimated reverberation time $\hat{T}_{60}$ is investigated here. Therefore, the estimated reverberation time $\hat{T}_{60}$ is varied between 400 ms and 800 ms in steps of 50 ms.

Further, though the experiments carried out on the AURORA 5 database indicate $L_C = L_R = 6$ to give reasonable results if the reverberation time is known, the effect of different $L_C = L_R$ on the recognition performance is investigated if the reverberation time is estimated rather than known.

For the same reason the considered multi-model inference schemes again comprise the IEKF-TI+CMN, the IEKF-α-AUG+CMN and the IEKF-α-TV+CMN, although at moderate noise conditions as present in the the MC-WSJ-AV database (the global broadband RNR lies in the range of 20-15 dB), e.g., the IEKF-TI+CMN and IEKF-α-TV+CMN have been found to differ only marginally on the AURORA 5 task. All considered BFE schemes are accessed on the two evaluation sets for $M \in \{1,4,8,16,32\}$ dynamic states in the MSLDM a priori and $L_C = L_R \in \{1,2,3,4,5,6\}$ clean speech LMPSC feature vectors in the state vector. Since only the non-causal estimation is considered, the latter also specifies the employed lag in the estimation.

5.4.7 Results on MC-WSJ-AV with Bayesian Feature Enhancement

Detailed recognition results obtained with the non-recursive and recursive observation model are given in Tab. 5.39 and Tab. 5.40 for the IEKF-TI+CMN scheme, respectively, in Tab. 5.41 and Tab. 5.42 for the IEKF-α-AUG+CMN scheme and in Tab. 5.43 and Tab. 5.44 for the IEKF-α-TV+CMN scheme.

A strong dependency of the recognition performance on the chosen number M of dynamic states in the a priori model, the number L_C of clean speech LMPSC feature vectors in the state vector and the estimate $\hat{T}_{60}$ of the reverberation time can be observed for all employed BFE schemes.

The lowest word error rates (marked in bold) are thereby achieved with the non-recursive IEKF-α-TV+CMN scheme, closely followed by it recursive counterpart and the non-recursive IEKF-TI+CMN scheme. As on the AURORA 5 task at high global broadband RNR values, the IEKF-α-AUG+CMN yields the highest word error rate. Opposed to the IEKF-TI+CMN and the IEKF-α-TV+CMN schemes, the recursive counterpart of the IEKF-α-AUG+CMN scheme thereby always outperforms the non-recursive variant.

An increase in the number M of dynamic states in the MSLDM a priori model thereby in general leads to reduced error rates for all recursive and non-recursive variants of the considered BFE schemes. The non-recursive observation models thereby majorly give the best results for reverberation time estimates in the range of $\hat{T}_{60} = 550 - 650$ ms and $L_C \in \{4, 5\}$. While the recursive counterparts also give the best results for $L_C \in \{4, 5\}$, they seem to favor a slightly higher estimate of the reverberation time, i.e., $\hat{T}_{60} = 700 - 800$ ms.

The overall best result with 37.76 % and 55.90 % on the evaluation 1 and evaluation 2 test set, respectively, are obtained with the non-recursive IEKF-α-TV+CMN scheme at $M = 32$, $L_C = 4$ and $\hat{T}_{60} = 600$ ms. This amounts to a relative improvement of about 55 % and 40 % on the evaluation 1 and evaluation 2 test set, respectively. Note that the results thereby only slightly vary for L_C in the range $4 - 6$ and $\hat{T}_{60} = 550 - 650$ ms. The same insensitivity w.r.t. the estimated reverberation time can be observed for its recursive counterpart and the non-recursive and recursive IEKF-α-TV+CMN scheme, which, as expected, performs almost as good as the IEKF-TI+CMN scheme.

This may best be explained by the fact the noise conditions encountered in the MC-WSJ-AV task are, with the global broadband RNR in the range of 20-15 dB, very moderate and that the MC-WSJ-AV task is a continuous speech recognition task where pauses between words are rather rare. Thus, the relative occurrence of low IRNR values may be considered low, too. Further, the still large error rates may be considered an indicator for a potentially high uncertainty in the prediction of the clean speech feature vector. Since the covariance of the observation error in the IEKF-α-TV+CMN scheme, however, converges to the time-invariant covariance matrix of the observation error employed in the IEKF-TI+CMN scheme under these circumstances, the two schemes may be considered to be equivalent.

Before turning to the UD results for the best performing BFE scheme, it is worth looking at how the individual speakers contribute to the presented average recognition errors. Therefore, a closer look is taken at the results per speaker obtained with the non-recursive IEKF-TI+CMN scheme with $M = 32$, $L_C = 4$ and $\hat{T}_{60} = 600$ ms. The word error rates per speaker are listed in Tab. 5.45. For comparison purposes, the correpsonding speaker-specific word error rates obtained with the SFE+CMN on the headset data er listed, too.

Table 5.39: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the non-recursive observation model in the IEKF-TI+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400 - 800\,\text{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						evaluation 2 – array1-1					
M	$\hat{T}_{60}$			L_C						L_C			
	[ms]	1	2	3	4	5	6	1	2	3	4	5	6
1	400	76.40	54.03	53.86	55.74	57.97	60.69	87.07	70.90	70.00	71.43	74.37	76.60
	450	77.72	51.37	49.40	51.76	53.93	55.74	88.23	70.20	66.63	68.40	70.40	72.27
	500	83.16	51.96	**47.75**	48.50	49.89	52.09	89.67	72.47	67.03	65.53	68.17	69.97
	550	87.20	55.67	48.66	47.66	47.98	51.05	91.90	74.20	70.17	**65.30**	66.97	69.30
	600	89.33	58.13	49.47	**47.75**	48.17	50.34	93.37	76.50	70.17	67.20	67.60	69.20
	650	92.40	62.17	52.02	49.05	48.63	49.56	94.93	79.47	72.20	69.67	69.00	70.77
	700	95.02	67.86	54.38	51.21	50.34	51.28	94.40	83.23	74.47	71.57	71.00	72.70
	750	95.28	71.81	58.94	52.21	50.70	51.60	95.23	84.83	77.27	73.97	72.67	73.97
	800	96.51	77.50	63.37	56.09	53.25	54.64	95.07	87.33	82.53	76.87	74.57	75.47
2	400	77.79	54.48	53.99	56.16	58.68	61.40	89.17	69.83	69.60	71.93	72.97	76.30
	450	78.79	50.99	49.40	50.66	53.64	55.32	90.00	69.07	66.23	67.20	68.80	70.90
	500	82.90	51.99	48.08	47.75	48.79	51.83	90.60	71.53	66.03	**65.33**	66.70	68.40
	550	85.10	54.57	48.56	46.40	47.82	50.40	91.97	73.43	67.50	65.47	65.50	67.70
	600	88.75	57.10	48.92	46.33	46.62	49.24	92.83	76.40	69.13	66.17	65.57	68.10
	650	89.36	60.65	49.50	46.78	**45.88**	47.98	93.60	78.57	70.13	67.87	67.13	69.40
	700	92.21	65.34	54.32	49.27	48.30	50.37	94.43	81.00	72.97	70.20	71.10	70.90
	750	92.89	69.67	57.10	51.86	49.01	50.31	93.80	84.17	76.20	72.47	72.87	72.93
	800	93.82	74.14	60.14	55.29	51.50	52.44	95.07	85.50	79.67	74.63	74.47	75.47
4	400	74.65	53.41	52.83	54.74	57.61	60.30	88.37	69.33	68.73	69.90	72.63	75.73
	450	76.43	51.18	47.91	49.37	53.09	54.45	89.63	68.07	65.63	65.40	67.03	70.03
	500	78.89	51.28	46.23	47.53	48.50	50.99	91.23	70.67	65.37	64.27	64.77	67.27
	550	81.67	53.60	46.85	45.30	47.07	49.37	92.57	72.37	65.50	63.57	64.93	67.10
	600	85.19	55.54	47.30	45.20	**44.97**	48.33	93.83	74.60	66.10	**62.47**	63.87	66.07
	650	86.45	58.52	50.15	46.27	45.94	48.76	92.97	76.33	68.77	65.73	65.00	67.90
	700	91.11	62.82	53.31	48.95	47.30	48.76	94.63	79.87	70.87	68.53	67.47	68.80
	750	91.88	67.22	55.42	50.37	48.59	48.98	95.13	83.03	74.00	71.03	70.00	71.10
	800	92.85	72.55	60.33	53.70	51.44	51.37	94.43	85.37	76.63	72.30	72.37	73.23
8	400	74.36	51.54	50.11	52.31	54.83	57.32	90.17	69.33	66.10	68.07	71.63	74.17
	450	76.27	50.21	45.36	46.98	49.18	50.73	91.10	67.50	62.43	63.43	65.70	68.67
	500	77.72	48.79	43.26	43.52	45.36	47.37	90.43	68.27	61.50	63.33	64.07	66.37
	550	79.70	51.57	44.00	**41.90**	44.16	45.68	92.07	69.33	62.00	**61.37**	63.00	63.57
	600	82.54	51.31	45.30	42.97	42.58	45.00	93.57	71.90	63.60	61.53	62.17	63.83
	650	84.19	53.83	45.94	43.36	42.81	44.68	93.57	73.80	65.67	61.67	62.73	64.37
	700	84.74	58.26	47.37	44.36	43.97	44.68	95.03	77.43	68.97	65.13	65.07	65.27
	750	87.81	61.40	51.28	46.98	46.01	47.30	95.30	81.17	72.50	67.20	65.77	67.27
	800	91.37	65.60	54.22	48.69	48.43	48.95	96.27	83.10	75.50	68.33	68.00	70.03
16	400	73.10	49.76	49.30	51.18	54.32	57.06	86.87	66.07	65.07	66.93	69.87	71.77
	450	71.23	48.14	44.49	45.85	48.14	50.11	85.57	64.17	61.93	62.47	64.37	68.00
	500	73.65	47.59	42.84	42.94	45.20	46.69	88.53	67.00	60.07	61.53	62.13	64.13
	550	75.40	47.88	43.10	**41.29**	43.00	44.94	90.70	66.30	60.43	**58.93**	60.97	62.93
	600	78.18	50.11	44.16	**41.29**	42.68	42.84	91.70	69.53	61.30	59.30	60.77	62.33
	650	78.95	52.21	45.36	42.48	41.55	43.10	93.03	71.87	63.83	60.37	60.77	62.37
	700	81.05	53.25	46.75	42.84	42.71	42.71	93.30	75.13	65.73	63.57	62.93	64.23
	750	83.38	58.33	48.56	44.94	43.71	44.62	94.73	77.20	69.07	65.17	63.93	65.20
	800	84.84	60.56	52.18	47.82	46.85	46.49	95.37	79.43	70.37	67.13	66.40	68.03
32	400	67.64	47.79	46.88	47.20	50.99	52.80	84.20	64.37	62.37	64.97	68.53	69.97
	450	67.60	45.62	42.26	42.77	45.20	47.01	84.43	64.90	59.93	60.57	63.07	64.87
	500	68.12	44.84	41.03	40.38	42.13	43.52	84.70	64.30	59.70	57.50	60.00	61.63
	550	68.77	46.78	40.80	38.12	39.73	41.19	86.73	65.00	59.33	58.10	58.60	60.23
	600	70.16	46.85	41.64	**37.76**	38.51	39.51	86.97	66.60	58.67	**55.90**	57.20	59.17
	650	73.13	48.33	41.64	38.93	38.96	39.96	89.00	68.67	61.07	56.67	57.90	59.77
	700	74.75	49.56	43.81	40.35	39.48	40.87	91.33	73.03	62.73	59.17	59.73	60.97
	750	75.65	54.09	45.30	41.84	40.77	41.16	92.83	74.17	65.33	62.10	60.50	61.70
	800	80.15	55.74	47.30	44.55	42.87	43.26	91.87	76.60	67.50	63.87	62.97	64.60

Table 5.40: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the recursive observation model in the IEKF-TI+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400-800\,\text{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						evaluation 2 – array1-1					
M	$\hat{T}_{60}$			L_C						L_C			
	[ms]	1	2	3	4	5	6	1	2	3	4	5	6
	400	75.56	64.02	58.91	58.49	59.84	60.88	85.73	77.53	75.27	73.80	75.30	77.17
	450	75.56	61.59	55.45	53.67	54.64	55.03	86.57	75.27	71.43	71.13	71.73	72.30
	500	76.53	59.49	52.63	51.70	50.50	52.25	87.13	75.67	70.30	68.63	69.73	69.07
	550	77.21	58.97	51.34	49.69	48.72	50.34	87.87	74.57	68.43	67.77	66.97	67.23
1	**600**	78.27	58.91	49.60	48.92	46.94	49.11	88.20	73.30	67.53	66.77	65.30	65.47
	650	78.89	59.26	50.15	47.37	46.30	47.98	88.50	73.73	67.87	65.63	64.70	65.37
	700	79.57	59.33	50.50	**45.78**	45.97	47.91	88.80	74.53	67.23	65.10	**64.10**	65.47
	750	79.70	58.42	50.92	46.23	46.49	47.69	89.63	74.57	67.97	65.63	64.87	66.83
	800	80.76	58.71	50.70	47.11	46.72	48.01	89.60	75.50	68.80	66.47	64.37	67.17
	400	75.04	63.40	58.78	58.58	61.20	61.43	84.90	76.70	74.80	73.73	75.70	77.73
	450	76.20	60.33	54.87	54.12	56.29	56.71	85.37	74.93	71.13	70.83	71.63	72.53
	500	76.82	58.58	52.31	50.70	51.21	52.57	86.47	74.13	69.17	68.50	68.50	68.53
	550	77.85	58.29	50.50	48.76	48.76	50.18	86.57	72.87	67.73	66.67	65.67	67.27
2	**600**	79.70	58.36	49.89	47.40	46.91	48.50	87.70	73.77	64.83	64.25	63.77	66.20
	650	80.44	57.87	49.21	47.07	46.75	48.21	88.30	73.63	64.93	63.90	64.67	66.00
	700	81.02	58.13	48.98	**46.46**	46.52	47.33	88.53	72.60	64.97	63.60	**62.80**	66.00
	750	80.70	58.75	48.85	**46.46**	45.55	47.27	89.10	73.30	65.63	63.20	64.10	66.10
	800	81.28	58.81	48.82	45.23	44.97	47.07	89.43	73.50	66.57	64.37	64.73	67.47
	400	73.68	61.62	58.68	58.87	59.39	60.94	83.40	75.17	72.60	72.37	74.53	75.50
	450	74.10	58.81	54.45	54.38	54.15	56.39	84.10	74.60	70.13	69.03	69.93	71.63
	500	74.01	57.23	51.41	50.82	50.76	51.76	84.07	72.97	68.53	66.77	66.80	68.20
	550	75.82	55.71	49.50	48.40	49.01	49.47	84.50	70.80	66.07	65.43	64.03	66.13
4	**600**	78.21	55.90	49.21	46.91	47.27	47.30	85.40	70.93	65.10	65.10	64.47	64.53
	650	78.82	54.83	47.40	45.81	46.27	47.72	86.43	71.67	63.90	64.47	63.17	64.87
	700	80.47	56.55	47.17	45.75	45.10	46.98	88.13	71.60	64.37	63.73	**62.43**	64.70
	750	80.70	56.87	48.56	44.91	44.84	46.33	88.37	71.63	64.37	62.93	63.27	65.13
	800	82.15	57.58	48.08	45.75	**44.71**	46.30	88.87	72.07	64.67	62.97	63.17	66.40
	400	70.35	58.58	55.51	56.45	57.81	59.39	82.33	73.80	70.80	72.00	74.17	75.43
	450	72.00	55.51	51.50	51.18	52.21	53.35	83.50	71.10	68.07	69.03	69.73	71.00
	500	74.81	53.44	48.95	48.14	48.63	50.02	85.20	70.43	66.50	65.83	66.80	69.03
	550	76.46	53.12	46.91	46.98	46.62	47.66	85.73	70.27	63.70	63.33	64.03	65.13
8	**600**	78.56	52.93	46.20	44.88	45.00	45.39	87.23	70.03	62.90	62.27	63.43	63.80
	650	80.50	53.60	45.39	43.29	43.10	44.62	86.80	71.40	63.23	60.67	61.63	63.17
	700	82.15	53.41	45.43	42.42	42.61	43.74	87.67	71.40	62.07	**60.10**	61.03	63.33
	750	84.84	55.80	47.40	42.64	**42.10**	42.68	89.13	71.90	63.33	61.13	62.50	63.33
	800	85.58	58.16	48.04	43.03	42.48	43.68	89.63	72.80	63.67	61.30	62.27	63.90
	400	67.60	55.77	55.35	54.64	57.00	58.23	81.30	71.13	69.33	69.03	73.27	75.60
	450	70.22	52.93	49.27	49.69	51.28	51.92	82.53	68.93	65.47	66.13	68.43	70.93
	500	73.20	51.28	47.11	46.98	47.27	48.79	83.47	66.47	63.77	63.27	65.23	66.73
	550	76.11	50.76	45.43	45.39	45.17	45.75	84.23	67.00	61.47	62.67	62.30	64.23
16	**600**	78.98	51.34	45.46	42.64	43.13	43.84	86.00	66.53	60.40	60.57	61.57	62.47
	650	79.76	52.15	44.97	42.90	42.39	43.42	86.53	66.83	59.80	**59.40**	60.20	62.63
	700	81.77	52.41	44.78	42.29	**41.48**	42.52	87.30	67.63	60.80	59.43	59.93	61.20
	750	83.96	54.09	46.49	43.26	42.84	43.71	88.50	68.97	60.77	59.33	61.00	62.53
	800	85.16	55.77	47.24	43.61	43.23	44.46	88.87	70.13	61.67	59.43	62.23	63.17
	400	62.72	52.83	51.70	52.93	53.80	56.42	76.93	66.93	66.43	68.30	70.50	72.23
	450	64.05	48.76	47.27	48.59	48.88	49.50	78.93	65.70	64.57	64.40	66.40	67.93
	500	67.77	48.08	44.13	44.68	45.68	46.43	80.07	62.87	63.00	62.53	63.07	64.57
	550	69.54	47.37	42.94	42.90	42.42	44.13	81.10	62.20	60.40	59.87	60.77	61.33
32	**600**	73.46	47.53	41.51	40.25	39.54	40.41	81.97	62.70	58.13	57.97	58.80	60.77
	650	75.53	47.66	41.77	39.86	39.31	38.80	83.27	64.07	58.43	56.70	57.80	60.03
	700	76.69	49.30	41.13	38.89	**38.57**	39.15	84.47	63.63	58.70	**54.87**	58.10	59.23
	750	79.50	50.31	42.84	39.83	40.32	40.12	84.50	64.97	58.17	56.90	58.00	59.60
	800	80.86	51.63	42.97	40.32	40.87	40.90	85.63	65.97	59.27	57.50	58.50	60.07

Table 5.41: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the non-recursive observation model in the IEKF-α-AUG+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400-800\,\text{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						evaluation 2 – array1-1					
M	$\hat{\mathbf{T}}_{60}$			$\mathbf{L}_\mathbf{C}$						$\mathbf{L}_\mathbf{C}$			
	[ms]	1	2	3	4	5	6	1	2	3	4	5	6
1	400	78.95	58.81	58.97	63.89	65.24	67.12	87.27	76.37	75.83	78.17	79.73	81.13
	450	80.67	56.42	55.67	58.91	60.62	61.91	89.53	73.60	73.97	74.43	75.37	76.90
	500	83.22	56.29	53.25	57.42	58.23	59.81	92.20	73.47	72.90	72.40	74.60	75.27
	550	88.07	57.36	**53.02**	53.99	56.00	57.48	93.43	77.20	73.03	72.87	73.00	73.70
	600	90.79	60.39	54.83	53.90	55.71	56.16	93.80	79.40	73.07	**71.33**	72.60	74.15
	650	94.05	63.21	54.93	53.12	55.16	55.90	94.10	80.60	75.27	74.40	73.13	74.17
	700	96.25	70.51	58.52	55.22	54.96	56.42	95.50	85.20	78.50	75.30	75.33	74.73
	750	95.93	75.14	60.78	56.94	56.61	57.03	96.63	87.70	82.50	78.17	77.77	76.90
	800	96.38	78.69	66.76	60.04	59.04	58.00	96.10	88.47	84.40	81.67	80.73	79.33
2	400	80.83	58.00	59.59	62.85	65.13	66.63	89.70	74.47	74.90	75.99	77.50	81.23
	450	79.88	55.20	55.67	57.66	59.49	61.98	92.13	72.36	71.37	73.73	75.07	76.93
	500	82.44	54.81	53.09	54.38	55.65	58.45	92.03	72.93	**70.13**	71.53	73.70	73.67
	550	84.93	56.51	53.12	**51.08**	54.12	56.32	93.73	76.03	71.77	72.23	71.73	73.47
	600	88.62	58.00	53.29	51.70	52.34	54.12	92.73	77.40	72.33	71.50	71.27	73.70
	650	91.08	62.79	54.38	52.99	51.24	54.11	93.47	79.50	74.77	72.75	71.70	73.50
	700	92.24	68.38	56.75	55.00	53.44	55.42	94.40	83.53	77.03	75.83	74.35	74.70
	750	94.87	72.13	58.45	55.06	55.03	55.16	94.50	85.50	79.97	77.17	76.50	77.43
	800	94.70	75.27	63.01	58.33	56.58	56.55	94.57	87.90	80.80	78.83	78.40	79.63
4	400	76.68	59.26	59.76	63.89	64.61	66.55	90.53	74.40	73.78	76.31	78.69	81.31
	450	78.11	56.99	55.66	58.44	61.17	63.36	91.77	73.53	71.39	73.44	75.01	77.52
	500	80.31	55.16	53.30	54.72	58.63	57.81	92.20	73.38	72.40	71.67	72.43	75.16
	550	83.39	56.89	54.09	52.67	55.22	54.78	93.26	75.36	71.86	71.01	72.08	73.22
	600	85.58	58.73	53.51	**51.69**	52.20	55.75	93.57	78.61	72.13	**70.52**	71.41	73.19
	650	87.75	61.32	54.89	53.28	51.31	55.46	94.83	80.13	72.93	72.47	72.19	74.22
	700	91.30	66.03	57.11	54.79	53.13	55.42	95.47	84.17	75.36	74.75	73.20	75.13
	750	93.37	71.23	60.93	55.98	54.71	55.45	96.77	85.13	77.63	76.20	75.98	76.93
	800	92.89	74.72	63.24	58.54	56.57	57.62	96.73	88.50	81.97	78.93	76.83	78.03
8	400	79.26	59.43	57.69	58.97	62.02	64.61	93.66	75.88	73.72	76.47	79.50	80.63
	450	78.79	59.76	55.41	55.38	56.34	58.14	94.14	75.51	72.85	73.98	75.26	75.67
	500	80.15	57.57	53.96	54.29	55.19	55.58	95.28	76.33	72.77	73.17	74.30	74.24
	550	82.99	56.78	52.60	53.48	55.54	54.41	94.87	77.82	72.96	**71.42**	72.91	74.20
	600	85.00	58.81	53.70	**51.29**	51.87	52.96	96.75	78.59	73.83	72.00	71.83	73.10
	650	87.04	60.79	54.92	53.36	52.19	53.15	96.50	81.50	74.61	72.73	72.61	71.47
	700	88.72	64.10	56.51	52.89	53.25	53.54	98.50	85.97	77.09	74.00	74.08	73.45
	750	90.33	67.18	59.50	56.53	54.50	55.36	99.07	87.07	79.27	75.84	74.82	75.97
	800	92.11	72.30	61.14	57.17	56.22	55.75	99.20	88.80	82.41	78.06	77.33	77.27
16	400	77.59	57.65	54.64	57.36	59.68	61.68	89.83	73.07	72.10	73.03	75.26	78.60
	450	76.37	56.71	51.08	53.70	53.39	56.29	90.63	72.67	69.97	70.57	71.62	74.81
	500	78.37	55.00	49.89	51.31	50.29	52.64	92.57	71.07	68.67	69.76	70.18	72.10
	550	78.63	55.80	50.57	**48.72**	49.69	52.51	93.63	73.27	68.37	68.69	69.49	70.59
	600	78.90	56.94	50.92	49.79	50.37	50.18	94.63	76.11	69.27	67.67	68.48	70.22
	650	81.02	57.90	52.10	50.66	51.18	50.99	96.50	77.45	72.00	**67.14**	69.25	70.53
	700	82.86	59.62	53.83	50.99	49.17	49.79	96.17	80.39	73.73	70.37	70.53	71.70
	750	85.81	63.12	56.03	51.77	50.33	51.50	97.23	83.14	76.50	73.31	72.01	74.00
	800	88.36	65.63	57.68	54.57	53.45	52.46	98.27	84.22	78.03	73.24	73.86	74.83
32	400	71.97	52.51	51.76	53.60	56.81	57.39	88.67	72.00	68.40	71.60	74.00	75.80
	450	70.87	49.66	47.59	49.18	50.76	52.99	89.57	69.63	67.60	69.30	68.80	71.20
	500	72.26	51.15	47.33	46.59	47.01	48.95	88.77	70.80	66.47	66.33	66.33	67.47
	550	72.84	50.99	47.01	45.33	46.36	47.04	89.77	72.80	66.27	65.70	66.33	66.97
	600	74.39	51.92	46.85	44.81	45.65	45.97	91.97	73.27	66.53	**65.00**	66.23	67.37
	650	76.30	53.60	48.01	46.01	45.68	46.30	90.97	74.93	68.27	65.23	65.77	68.20
	700	77.14	54.90	49.47	46.07	**44.65**	47.14	92.60	77.87	70.30	67.87	66.77	68.40
	750	78.21	57.84	49.98	48.01	46.69	48.27	94.33	79.60	72.20	70.43	68.67	69.03
	800	80.12	61.40	52.99	50.53	48.50	49.56	94.27	81.37	74.90	72.90	72.40	71.43

Table 5.42: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the recursive observation model in the IEKF-α-AUG+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400 - 800\,\mathrm{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						**evaluation 2 – array1-1**					
M	$\hat{\mathbf{T}}_{60}$			$\mathbf{L_C}$						$\mathbf{L_C}$			
	[ms]	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	**400**	73.91	63.43	60.27	60.81	61.53	60.98	84.67	76.63	75.60	75.03	77.13	77.73
	450	73.65	60.94	56.39	56.55	57.45	57.03	85.03	75.33	73.00	72.63	73.07	73.57
	500	73.59	59.81	54.03	53.70	53.12	53.31	84.67	73.33	71.37	70.83	72.03	70.87
	550	73.88	58.42	52.18	51.96	51.12	51.96	85.40	74.13	69.60	69.57	68.33	69.03
	600	74.36	58.36	50.15	50.60	49.50	49.18	84.83	73.57	68.77	69.20	68.80	68.17
	650	75.23	57.58	50.57	49.08	47.82	49.56	85.63	74.33	67.80	68.13	67.20	67.90
	700	76.04	57.84	49.34	48.08	47.04	48.79	86.00	73.90	68.70	66.67	66.70	67.37
	750	76.85	57.36	49.79	46.91	47.40	48.92	86.83	74.60	68.07	**66.63**	67.77	68.47
	800	77.92	58.29	50.21	48.11	**46.94**	49.08	87.53	74.07	67.60	68.40	67.67	69.00
2	**400**	72.94	62.72	60.23	61.46	62.30	61.91	83.30	76.90	73.63	75.30	76.80	78.30
	450	73.13	59.84	56.09	56.35	57.32	57.97	83.30	74.87	71.40	71.50	72.60	73.93
	500	74.49	58.94	53.09	53.18	52.73	53.51	83.57	72.60	69.97	70.40	69.33	71.43
	550	74.49	57.36	50.73	51.86	50.99	51.89	84.33	72.50	68.40	68.23	67.83	69.27
	600	75.75	57.23	49.95	49.27	49.24	49.82	84.73	72.07	66.70	66.30	66.90	68.10
	650	76.56	56.87	48.79	48.66	48.11	48.98	85.87	72.47	66.00	66.10	65.67	68.43
	700	77.50	56.58	48.17	47.75	47.82	48.08	86.17	72.33	65.27	65.70	**64.93**	67.83
	750	79.02	57.00	48.76	46.91	**46.75**	49.05	86.50	72.40	65.80	65.57	66.37	67.70
	800	79.66	57.55	49.40	46.85	47.01	48.79	86.87	72.73	66.33	65.53	66.70	68.70
4	**400**	71.13	60.14	58.58	59.62	61.24	61.91	82.40	75.53	72.23	74.70	76.67	76.83
	450	71.52	58.20	54.64	55.06	56.06	56.51	82.80	73.63	70.37	71.17	72.23	73.67
	500	71.32	55.87	51.86	51.24	53.44	52.15	82.33	72.53	66.80	68.40	68.57	71.03
	550	72.45	54.77	49.98	49.73	50.63	50.79	82.07	70.83	67.03	66.53	67.53	68.77
	600	74.94	54.32	49.50	47.82	48.88	48.59	83.00	70.57	65.03	65.70	66.33	67.67
	650	76.01	54.80	48.63	46.91	47.69	48.63	82.87	69.87	64.13	65.00	64.57	67.63
	700	77.30	53.93	47.91	46.78	46.69	47.43	84.73	71.00	64.00	64.27	63.80	66.47
	750	77.72	55.16	48.14	46.20	46.72	47.43	84.43	70.77	63.40	63.83	64.47	67.60
	800	79.44	55.74	48.30	**45.26**	46.30	48.21	85.17	71.07	65.00	**63.23**	65.93	67.37
8	**400**	69.06	56.35	56.22	57.19	59.36	60.98	80.37	72.93	72.13	72.33	76.03	78.00
	450	70.48	54.83	51.92	51.96	54.51	55.06	81.07	71.23	68.50	69.87	70.57	74.13
	500	71.61	53.28	48.98	48.95	50.60	51.83	83.17	70.97	66.73	67.20	67.43	71.20
	550	75.17	53.02	47.30	47.82	48.08	49.14	84.00	70.43	65.43	64.13	65.93	69.37
	600	77.63	52.12	46.88	44.94	46.56	48.27	84.77	70.17	63.30	62.60	64.90	66.63
	650	78.79	53.38	47.33	44.71	45.68	47.85	85.60	71.23	64.07	62.77	64.67	65.97
	700	80.18	53.90	47.62	**43.81**	45.00	46.17	86.73	70.60	63.10	**61.97**	62.83	64.90
	750	82.80	56.61	48.95	43.84	43.87	43.97	87.80	72.40	64.10	62.70	64.20	66.37
	800	83.87	57.71	49.11	44.65	44.23	45.33	88.40	73.50	64.63	63.67	63.97	66.17
16	**400**	66.96	55.58	55.64	56.94	58.49	59.81	80.80	70.90	70.83	72.67	75.37	77.47
	450	69.32	52.96	50.63	51.76	52.83	54.83	81.53	68.80	67.57	68.07	69.77	72.47
	500	72.36	51.96	47.95	48.40	50.18	50.44	81.83	67.13	65.17	65.37	67.13	70.60
	550	73.84	50.60	47.17	46.56	46.62	47.37	83.10	66.80	64.23	63.50	64.70	67.87
	600	77.43	51.02	46.20	45.10	45.62	46.01	83.83	67.60	61.60	62.63	63.97	67.27
	650	78.73	52.15	47.33	43.61	44.46	45.23	84.97	67.53	61.57	61.50	63.67	65.30
	700	80.80	53.02	46.62	43.58	**43.49**	45.62	85.33	68.93	61.00	**60.37**	62.43	64.77
	750	82.51	54.48	47.49	45.07	44.81	46.14	85.57	69.57	62.43	61.33	63.27	64.90
	800	83.74	55.87	47.85	45.33	45.26	46.94	87.00	69.70	63.50	62.03	63.50	65.27
32	**400**	61.82	52.76	52.70	54.45	56.58	57.23	77.23	69.27	68.80	71.30	73.77	76.20
	450	63.95	49.30	47.49	49.43	51.05	52.02	76.60	67.10	65.80	67.27	68.63	70.90
	500	66.83	48.04	44.84	44.91	46.88	48.40	77.50	65.03	63.80	64.57	65.87	68.70
	550	69.61	47.20	43.29	43.58	44.07	45.55	80.00	64.50	61.63	61.60	63.27	65.43
	600	72.52	47.91	43.10	42.06	41.71	43.32	82.97	64.63	59.73	60.50	63.33	64.07
	650	74.98	48.40	43.10	41.35	42.03	42.71	83.27	64.37	60.90	60.00	60.77	63.30
	700	76.50	48.56	43.16	**41.06**	41.90	42.03	82.97	63.03	60.43	**58.90**	61.30	62.80
	750	77.56	50.66	44.07	43.10	42.26	43.68	82.93	64.67	60.90	60.07	61.67	63.37
	800	79.92	51.60	44.55	42.06	42.84	44.13	83.97	65.03	61.17	61.77	61.73	64.03

Table 5.43: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the non-recursive observation model in the IEKF-α-TV+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400 - 800\,\text{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						evaluation 2 – array1-1					
M	$\hat{T}_{60}$			L_C						L_C			
	[ms]	1	2	3	4	5	6	1	2	3	4	5	6
1	400	76.82	54.51	54.41	56.16	58.23	61.20	87.63	70.90	69.93	71.90	74.80	76.27
	450	79.02	51.70	50.99	52.47	54.32	56.26	88.17	70.30	66.90	68.47	70.30	72.47
	500	82.67	52.73	48.27	48.82	50.47	52.86	90.20	72.90	66.17	65.97	68.83	69.50
	550	86.78	56.00	48.92	**47.66**	49.05	51.89	91.73	74.37	69.03	**64.63**	66.20	68.97
	600	90.24	57.90	50.50	47.88	48.01	51.12	93.40	76.23	69.13	66.63	66.57	70.10
	650	92.43	61.98	51.37	49.01	48.69	49.40	93.43	79.57	71.73	68.53	67.73	69.57
	700	95.83	66.93	55.06	50.66	49.30	50.89	93.93	83.00	75.37	71.43	70.33	71.57
	750	95.47	72.03	58.52	53.15	50.08	51.37	94.50	84.87	77.07	72.60	72.43	72.37
	800	96.22	75.43	61.66	55.09	53.83	54.90	94.63	86.53	81.97	75.70	75.20	75.53
2	400	78.40	55.03	53.48	57.06	60.14	61.78	88.23	70.77	70.37	72.50	75.07	77.80
	450	78.98	52.28	50.18	52.31	54.38	56.64	89.83	68.73	66.37	67.53	70.00	72.50
	500	81.89	52.34	48.43	48.95	50.82	54.15	90.73	70.77	65.47	**65.23**	68.07	69.43
	550	85.84	55.38	48.88	47.62	49.11	51.41	92.10	73.27	66.47	65.40	66.57	68.87
	600	88.68	56.81	48.92	**47.40**	47.69	49.92	92.73	75.80	67.53	66.00	67.07	68.67
	650	89.56	60.30	50.57	48.08	47.11	49.34	93.50	78.57	69.97	68.03	67.77	69.67
	700	91.43	66.93	53.73	49.60	48.82	49.89	93.73	81.47	73.77	69.80	68.97	70.63
	750	92.43	69.35	56.32	51.28	49.76	51.47	95.07	83.50	75.57	72.43	71.23	72.77
	800	94.41	74.43	61.66	54.54	52.44	52.18	94.57	84.97	79.13	73.70	73.10	74.93
4	400	74.85	53.99	54.32	56.94	59.49	61.43	88.50	69.70	68.97	71.13	73.60	76.53
	450	76.14	51.34	49.47	51.31	54.03	56.32	89.83	68.63	65.67	65.67	68.60	71.00
	500	79.86	51.21	47.95	48.04	50.21	52.21	91.13	70.03	64.63	63.97	65.40	68.07
	550	81.25	54.06	46.75	47.43	48.33	49.53	92.70	71.50	65.63	**63.63**	65.07	67.13
	600	85.26	56.00	48.08	46.30	**46.17**	48.82	93.23	73.87	64.97	64.23	64.80	66.07
	650	87.58	58.71	51.12	46.69	46.36	48.63	93.67	75.73	68.80	65.90	64.60	68.10
	700	90.24	63.40	53.25	48.82	48.21	49.69	95.07	81.23	70.43	67.83	67.37	68.33
	750	91.66	67.41	55.22	50.63	49.21	49.24	95.40	83.47	73.67	70.33	68.93	70.10
	800	92.47	74.17	60.88	53.70	51.60	52.25	93.97	85.67	77.10	72.00	71.53	72.37
8	400	75.46	51.86	49.53	51.67	54.77	57.81	89.37	68.77	66.27	68.20	72.23	74.73
	450	76.75	50.95	45.59	47.53	49.27	51.21	90.83	68.20	63.80	63.77	66.10	68.70
	500	79.28	50.08	44.13	44.26	46.30	48.24	92.47	69.00	61.93	62.77	64.93	66.33
	550	80.86	51.24	43.42	**42.16**	44.26	45.65	92.67	70.50	63.47	**61.77**	63.97	65.03
	600	83.51	52.54	45.30	42.29	43.78	44.94	94.03	71.47	64.47	62.33	62.77	63.93
	650	85.74	55.12	46.14	43.78	43.16	43.94	94.47	74.77	65.80	62.10	63.73	63.97
	700	86.10	58.58	47.72	44.88	45.00	44.62	94.90	77.27	69.13	65.67	64.63	65.43
	750	87.46	62.37	51.08	46.91	46.69	45.75	95.33	80.53	71.50	67.00	67.30	67.50
	800	91.37	66.57	54.41	49.69	49.82	48.76	96.27	84.03	74.23	68.87	68.23	70.13
16	400	72.81	50.86	49.47	50.63	53.83	56.97	86.70	67.30	64.80	67.40	71.47	72.57
	450	71.74	48.69	44.91	47.82	48.04	50.44	86.30	65.40	62.83	63.13	65.00	68.60
	500	74.10	48.37	43.16	43.71	45.33	47.11	89.27	66.90	60.77	62.20	63.83	64.87
	550	76.14	48.85	43.39	41.48	43.10	44.58	90.70	66.90	61.70	61.67	63.67	63.80
	600	78.05	50.79	44.55	**41.32**	42.94	43.32	91.37	69.43	61.43	60.87	62.00	64.33
	650	79.57	52.44	45.20	42.74	40.90	43.19	92.80	72.60	63.27	**60.70**	62.07	63.43
	700	81.44	54.51	46.56	43.84	41.97	43.68	93.67	75.60	66.00	63.43	64.10	65.80
	750	84.03	58.36	49.34	45.04	43.49	44.29	94.80	77.23	69.60	64.60	64.97	66.93
	800	86.13	61.14	52.09	47.75	47.62	46.52	95.80	80.43	72.20	68.40	67.13	68.63
32	400	69.06	48.56	47.59	47.82	51.50	54.45	83.83	65.33	63.80	66.07	69.07	70.73
	450	68.61	46.65	42.06	43.29	45.97	48.08	84.47	65.53	61.07	61.63	63.23	64.97
	500	68.77	45.46	41.00	41.29	42.03	43.61	84.87	65.63	60.03	59.73	61.90	61.90
	550	69.32	47.07	41.32	38.96	40.61	41.84	86.13	66.93	58.67	58.07	59.77	61.20
	600	70.84	47.43	42.32	**38.76**	39.61	40.67	86.60	67.97	58.93	**57.87**	58.97	60.80
	650	73.36	49.85	42.29	39.57	38.89	40.45	90.07	69.97	60.73	58.87	60.00	61.20
	700	76.17	50.70	43.74	41.61	40.19	41.00	91.17	73.17	65.03	60.47	61.27	63.30
	750	75.91	54.48	46.27	42.52	42.19	42.64	91.57	74.37	66.10	62.20	63.23	63.97
	800	80.60	56.77	48.43	44.84	44.20	44.84	93.13	75.73	68.23	65.53	64.73	65.27

Table 5.44: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the recursive observation model in the IEKF-α-TV+CMN model-specific inference employing an MSLDM with $M \in \{1,2,4,8,16,32\}$ and $L_C \in \{1,2,3,4,5,6\}$ for reverberation time estimates $\hat{T}_{60} = 400 - 800\,\mathrm{ms}$. The recognizer utilizes a* clean *acoustic model.*

		evaluation 1 – array1-1						evaluation 2 – array1-1					
M	$\hat{T}_{60}$			L_C						L_C			
	[ms]	1	2	3	4	5	6	1	2	3	4	5	6
	400	78.05	63.37	59.84	59.23	60.36	62.01	86.30	78.33	75.77	74.80	76.17	77.67
	450	76.79	62.56	56.77	55.64	56.35	56.29	85.80	76.17	73.23	71.63	72.83	73.83
	500	75.95	60.78	54.22	53.09	51.60	53.25	85.50	75.43	70.23	69.20	69.83	70.00
	550	76.95	59.62	52.63	50.50	48.98	51.73	86.10	74.40	69.07	67.77	67.77	68.50
1	**600**	78.21	58.84	50.34	49.43	48.04	49.60	86.50	74.00	67.07	66.70	66.80	67.07
	650	78.47	58.36	49.34	47.88	47.01	48.21	87.23	74.13	65.97	65.43	64.63	65.83
	700	79.24	59.17	49.30	46.88	46.65	48.50	88.43	73.57	66.97	64.47	**64.30**	66.00
	750	79.34	59.07	49.89	46.75	**45.65**	47.75	88.10	73.70	66.67	64.83	65.03	66.00
	800	80.28	58.58	50.27	46.43	46.07	48.50	88.23	73.97	67.23	64.47	64.40	67.00
	400	77.66	64.53	60.52	59.84	62.40	62.92	85.23	77.17	75.00	75.60	77.53	79.53
	450	77.40	62.92	57.87	56.90	57.42	58.07	85.67	75.40	72.67	72.07	73.53	74.47
	500	77.37	60.91	54.12	52.63	52.47	54.03	85.30	74.40	70.03	70.10	70.60	70.93
	550	77.63	59.13	52.60	49.92	50.44	52.44	85.67	74.70	68.57	67.73	67.50	68.50
2	**600**	77.53	58.78	50.53	48.08	48.24	50.79	86.60	73.87	66.93	66.07	65.83	66.50
	650	78.05	58.36	49.18	46.65	47.37	48.50	86.93	73.37	65.10	65.20	64.17	65.70
	700	78.56	58.03	49.66	46.27	46.69	48.50	87.30	73.73	65.17	65.10	64.30	66.47
	750	79.37	58.65	49.92	46.52	45.91	48.24	88.07	73.43	64.97	64.00	64.70	66.80
	800	79.89	58.65	50.05	46.07	**45.62**	47.20	88.37	73.67	66.10	**63.63**	64.50	68.13
	400	74.81	63.17	60.14	59.94	61.17	62.43	85.13	77.30	74.07	74.77	77.10	76.60
	450	74.98	61.56	56.61	56.74	56.84	57.90	84.60	75.13	72.63	72.00	72.67	72.70
	500	75.69	58.49	54.96	52.70	53.35	53.93	84.97	74.67	69.90	74.13	69.70	70.43
	550	76.43	57.32	52.15	50.89	51.44	51.99	84.77	72.77	67.70	67.20	67.63	67.47
4	**600**	76.95	56.51	50.66	48.21	48.95	50.11	84.47	71.60	66.33	66.43	65.30	67.17
	650	77.98	56.87	49.14	47.56	47.46	48.63	84.70	71.57	66.20	65.23	64.70	66.20
	700	79.86	57.68	49.01	46.85	47.59	47.85	85.73	71.13	65.50	64.67	64.30	66.60
	750	79.70	57.39	47.75	46.17	46.62	47.49	86.43	70.90	64.43	**63.90**	64.03	66.17
	800	81.15	57.78	48.63	46.43	**45.97**	47.27	87.57	71.10	65.07	65.23	64.13	66.20
	400	71.58	59.49	57.23	58.52	58.81	60.98	82.17	75.10	72.30	74.17	75.17	76.60
	450	73.07	57.48	53.18	53.70	54.35	55.48	82.13	72.37	69.47	69.10	72.17	73.07
	500	75.40	54.90	50.27	50.24	49.85	51.96	83.87	71.33	66.33	67.40	68.60	70.60
	550	76.33	53.86	47.88	47.95	48.04	49.08	85.20	71.17	65.33	65.67	65.83	67.03
8	**600**	79.11	54.28	46.75	46.04	47.36	47.56	86.87	70.73	64.17	64.10	65.03	66.13
	650	79.53	54.12	45.49	44.75	44.52	45.91	86.90	71.83	64.03	62.67	63.57	65.23
	700	81.47	54.61	45.65	44.10	44.46	45.20	87.30	72.80	63.07	**63.03**	63.07	64.83
	750	83.22	55.06	46.56	43.78	44.23	44.33	87.47	73.30	64.03	62.37	62.23	66.67
	800	83.51	56.51	46.75	44.23	**43.55**	44.03	88.10	74.00	63.90	63.20	63.70	66.40
	400	70.16	57.90	56.48	57.39	58.36	60.91	81.57	73.07	70.97	72.27	75.90	76.07
	450	72.07	54.64	51.76	51.89	54.32	54.87	82.53	71.67	68.60	68.30	71.00	71.80
	500	73.97	52.57	48.82	48.30	48.92	50.31	84.13	68.60	64.93	65.80	68.00	69.17
	550	74.65	51.34	46.65	46.14	46.30	47.66	85.10	68.53	63.73	63.43	65.33	66.03
16	**600**	77.53	51.92	46.43	44.33	44.39	45.26	85.53	68.27	63.47	61.37	62.77	65.23
	650	80.08	52.05	45.88	43.84	44.13	43.06	85.87	68.93	63.10	62.00	61.87	64.73
	700	81.60	52.34	45.52	**41.90**	42.90	43.23	86.30	69.00	61.67	61.60	**61.07**	64.77
	750	82.22	53.90	45.88	42.22	42.32	43.58	87.03	69.47	61.53	61.40	62.93	64.33
	800	82.70	54.74	46.23	43.10	42.13	43.65	87.73	69.37	62.93	62.30	63.23	64.40
	400	63.17	53.80	53.70	55.06	56.42	58.07	77.50	70.53	70.10	70.83	72.87	73.93
	450	65.83	50.47	48.85	50.50	50.57	52.28	79.03	66.50	66.27	67.37	69.37	69.97
	500	67.22	48.98	45.94	46.17	47.14	47.75	80.10	65.23	63.33	64.17	65.87	66.43
	550	69.03	48.14	43.45	43.36	44.81	45.59	81.17	64.27	62.27	61.37	63.13	64.50
32	**600**	71.81	47.98	42.19	41.51	41.61	42.03	82.43	64.00	60.97	59.90	60.87	62.20
	650	74.07	48.30	42.19	40.41	41.00	40.83	82.87	64.87	59.73	59.47	58.73	60.73
	700	76.53	49.47	41.97	**39.51**	40.16	40.54	84.07	65.33	59.03	**57.77**	58.53	60.97
	750	77.79	49.98	42.71	39.70	40.64	41.00	85.30	66.00	59.60	58.77	58.73	61.13
	800	79.44	51.54	44.23	40.54	41.97	42.48	85.53	66.50	59.90	59.83	59.13	62.30

Table 5.45: *Speaker-specific word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the SFE+CMN on the headset data and the non-recursive observation model in the IEKF-TI+CMN scheme on the array1-1 data. The IEKF-TI+CMN employs an MSLDM with $M = 32$ dynamic states and $L_C = 4$ LMPSC feature vectors of the clean speech signal in the state vector. The estimated reverberation time is $\hat{T}_{60} = 600$ ms. Both times the* clean *acoustic model is employed.*

test set	speaker id	SFE+CMN – headset	IEKF-TI+CMN – array1-1
evaluation 1	**21**	10.59	30.48
	22	8.14	22.89
	23	11.94	28.55
	24	16.30	45.88
	25	25.72	57.05
evaluation 2	**36**	31.56	61.56
	37	28.55	54.14
	38	30.12	53.87
	39	19.49	57.03
	40	23.44	52.30

It can be seen, that the results highly vary from speaker to speaker, especially on the evaluation 1 test set. With speaker number 22 yielding a word error rate as low as 22.89 % and speaker number 25 a word error rate as high as 57 %, there seems to be a potential in improving the recognition performance by making (at least) the decoder more insensitive w.r.t. different speakers. This conclusion is also supported by the recognition results on the headset data with the SFE+CMN on the same test set.

For the evaluation 2 test set, the word error rates are already much worse with the SFE+CMN on the headset data, also indicating the need to adapt the decoder to different speaker characteristics, as already identified in [110].

Though UD decoding is not designed to resolve the speaker dependency in the recognition task, results obtained with IEKF-TI+CMN and additional UD are presented in Tab. 5.46 for the best performing non-recursive and recursive variant of the BFE scheme. The non-causal UD-m scheme can be found to be superior to the UD-n scheme (as observed on the AURORA 5 taks), improving the recognition results obtained with the non-recursive BFE

Table 5.46: *Word error rates λ_{WER} [%] on the MC-WSJ-AV task obtained with the IEKF-TI+CMN employing an MSLDM with $M = 32$ dynamic states and $L_C = 4$ LMPSC feature vectors of the clean speech signal in the state vector. Both the non-recursive and the recursive observation model are additionally employed with the non-causal MMSE estimate in the UD-m and UD-n scheme.*

	non-recursive observation model $\hat{\mathbf{T}}_{60} = 600$ ms		**recursive observation model $\hat{\mathbf{T}}_{60} = 700$ ms**	
processing scheme	**evaluation 1 – array1-1**	**evaluation 2 – array1-1**	**evaluation 1 – array1-1**	**evaluation 2 – array1-1**
IEKF-TI+CMN	37.76	55.90	38.57	54.87
IEKF-TI+CMN+UD-m	36.34	54.03	38.18	52.90
IEKF-TI+CMN+UD-n	36.53	51.33	37.25	51.47

scheme to 36.34 % and 54.03 % on the evaluation 1 and evaluation 2 test set, respectively.

6 Conclusion

This thesis has investigated different statistical observation models describing the relationship between the LMPSC feature vector of either the noisy, the reverberant or the noisy reverberant observation and the LMPSC vectors of the underlying clean speech signal and that of the noise. In particular, this work has introduced a new stochastic observation model for noisy reverberant speech.

Special focus has thereby been laid on a sound statistical formulation of the observation models and, in particular, the occurring observation errors. While the observation error in the presence of reverberation and the absence of noise has been considered a realization of a white, stationary and ergodic Gaussian process, this assumption has been shown to no longer be valid in the presence of noise, where the observation errors turn into functions of the ISNR and the IRNR in the observation models for noisy speech and noisy reverberant speech, respectively. In addition, the observation errors in the observation models for the presence of noise have also been found to become functions of the vector of phase factors, a term that arises from the cross-term in the computation of the power spectrum carried out during the front-end feature extraction. As a consequence, both observation models may be considered phase-sensitive. Since the derivation of the observation model for noisy reverberant speech has been based on the (existing) observation model for reverberant-only speech, the observation error in the former has also been shown to be a function of the observation error in the latter, eventually rendering the observation error highly non-stationary. However, a Gaussian approximation of the observation error has been found by considering the mean vector and the covariance matrix as a function of the IRNR. The new observation model for noisy reverberant speech has thereby been shown to be a generalization of the observation models targeting either noise or reverberation as the only distortion affecting the clean speech signal.

All observation models have been investigated in the context of Bayesian feature enhancement with subsequent speech recognition. As such, the Bayesian estimation of the a posteriori PDF of the clean speech feature vector has been motivated and soundly embedded into the statistical framework of ASR in Ch. 3. There, the a posteriori PDF of the clean speech feature vector sequence has been shown to be the key component to environmental robust speech recognition if recognition has to be carried out with acoustic models trained on clean speech signals. Since the theoretically optimal solution to the robustness problem has been identified as practically infeasible, approximate solutions have been presented in terms of (partly novel) UD schemes.

The conceptually optimal solution to the estimation of the a posteriori PDF of the clean speech LMPSC feature vector has then been outlined in the starting section of Ch. 4. The identified key components, namely the a priori model statistically describing the trajectory of the state vector consisting of some most recent clean speech LMPSC feature vectors plus the LMPSC feature vector of the noise and the observation model statistically relating the corrupted (either noisy, reverberant or noisy reverberant) observation to the LMPSC vectors in the state vector have been presented next. While the considered GMM and MSLDM a priori models have only briefly been reviewed, special stress has been put on the derivation of the observation models. Starting with the observation model for reverberant speech, which has previously been derived in [67], it has been shown how this observation model can be extended to the additional presence of noise. Thereby, the occurring observation error has been shown to depend on two random quantities, namely the observation error driving the observation model for reverberant-only speech and the vector of phase factors arising during the front-end feature extraction, and, moreover, the IRNR. Especially the latter renders the observation error highly time-variant, a fact that has been accounted for by approximating the PDF of the observation error by a GAUSSIAN distribution with time-variant (i.e., variant w.r.t. the IRNR) mean vector and covariance matrix. This complex model has further been shown to generalize not only the observation model for reverberant speech but also that for noisy speech, which has briefly been reviewed afterwards.

Since the observation models for reverberant and noisy reverberant speech call for a representation of the AIR in the LMPSC domain, a simplified model thereof has been described in the adjoined section to circumvent a sensitive blind estimation of the complete AIR. The employed model of the AIR requires only two parameters to be estimated, namely the reverberation time T_{60} and the energy $\sigma_{\check{h}}$ of the AIR. However, it not only simplifies the parameter estimation but also allows for a recursive formulation of the presented observation models for reverberant and noisy reverberant speech, which have been outlined next. The recursive observation model for noisy reverberant speech requires the LMPSC feature vector of the reverberant speech some L_R time instants earlier. Since it is not directly observable, an approximate MMSE estimate thereof has been derived employing the noisy reverberant observation L_R time instants earlier, the LMPSC feature vector of the noise at the very same time instant and the variance vector associated with the vector of phase factors.

The vector of phase factors arises in all (non-recursive and recursive) observation models involving additive background noise. Its properties have therefor been discussed in full detail and a parametric approximation to its PDF has been derived together with an analytic Solution to its central moments. Assuming independent components of the vector of phase factors, both have been shown to be fully characterized by the corresponding vector of variances.

The validity of the presented observation models has then been investigated by looking at the distributions of the corresponding observation errors. The dependency on the ISNR and the vector of phase factors for the observation model for noisy speech and the dependency on the IRNR and the vector of phase factors for the observation model for noisy reverberant speech have thereby been highlighted and confirmed.

Since the optimal inference of the a posteriori PDF of the LMPSC feature vector of the clean speech signal under the chosen a priori model and the presented observation model is computationally intractable, sub-optimal multi-model and model-specific inference schemes have been reviewed at the end of Ch. 4. The multi-model inference schemes thereby comprise the GPB1 and the IMM inference scheme. While the multi-model inference schemes have been presented in a generalized formulation, the model-specific inference schemes have been tailored to the considered observation models. However, they all have been based on a truncated VTS expansion of the observation function/mapping and a proper modeling of the observation error.

In Ch. 5, all proposed inference schemes have been experimentally assessed in joint Bayesian feature enhancement and recognition tasks on appropriate data. The inference schemes are thereby applied on both small and large vocabulary recognition tasks featuring both artificially distorted data and recordings in a real noisy reverberant environment. First, the observation model for noisy speech has been considered on the small vocabulary AURORA 2 task and the large vocabulary AURORA 4 tasks. On both tasks, the phase-sensitive observation models have been shown to be superior to the phase-insensitive ones, especially at low SNR values, which nicely reflects the theoretical considerations that have been made in Ch. 4. Thereby a GMM a priori model has been found to be superior to an MSLDM a priori model on the small vocabulary AURORA 2 task. However, on the large vocabulary AURORA 4 task an MSLDM has been found to perform equivalent or superior to a GMM model, especially at a comparable number of dynamic states. The additional application of the (partly novel) UD rules thereby resulted in comparable but rather inconsistent improvements on both databases, eventually reflecting the approximate nature of the practically realizable solution to the theoretically optimal decoding rule for robustness speech recognition. The best inference schemes have been able to increase the baseline recognition accuracy of 66.48 % and 72.06 % to 89.42 % and 90.15 % on the two test sets of the AURORA 2 database and from 38.94 % to 26.94 % on the AURORA 4 database.

Second, the observation models for reverberant and noisy reverberant speech have been considered on the reverberant data employed in the small vocabulary AURORA 5 task. The experiments with the observation models for reverberant speech have been carried out to validate the required modeling assumptions during its derivation. The recognition results have thereby been found to only slightly depend on the number of dynamic states in the a priori model, pointing out the modeling power of the observation model for reverberant only speech. The non-recursive observation model and the recursive observation model could further be observed to perform approximately equally well, although the latter comes with a reduced computational effort and lower memory requirements. This is also reflected by the obtained word error rates, which, even without UD have been found to reduce from 7.00 % and 16.25 % in the office and living room environment to 1.74 % and 3.75 % for the best setup when compared to the baseline. This corresponds to a relative improvement of 75 % and 80 % that has further slightly be improved by the application of additional UD.

Third, the observation model for noisy reverberant speech has been considered on the noisy reverberant data employed in the small vocabulary AURORA 5 task. The novel observation model with IRNR dependent modeling of the mean vector and covariance matrix of the observation error thereby has been found to be superior or equivalent to the variant with a fixed mean vector and covariance matrix of the observation error. The superiority has especially been observed at mid- and low-level values of the global broadband RNR, which also nicely reflects the theoretical considerations that have been made in Ch. 4. However, the performance of the inference schemes highly depends on the employed a priori model. While this may partly be attributed to the criteria employed for the training of the MSLDM, the major cause may be found in the sub-optimality of the employed multi-model and model-specific inference scheme.

Finally, the observation model for noisy reverberant speech has been considered on the noisy reverberant data employed in the large vocabulary MC-WSJ-AV task. In contrast to the aforementioned three AURORA databases, the MC-WSJ-AV data are not artificially distorted by reverberation and/or noise but are taken from real recordings in a noisy reverberant environment. Further, the MC-WSJ-AV database differs in the fact that the reverberation time is not known a priori, as opposed to the AURORA 5 database. Since the global broadband RNR is quite moderate on this database, the recognition results have been found to not vary much between the time-variant error modeling and the time-invariant counterpart, where the former appeared to be slightly more sensitive to the estimated reverberation time. The word error rate without additional UD could thereby be reduced from the baseline of 82.35 % and 91.07 % to 37.76 % and 55.90 % for the two considered evaluation sets – a relative improvement of 55 % and 40 %, respectively. However, a further analysis has shown that the results are highly varying from speaker to speaker and as such call for measures to adapt (at least) the acoustic model to the speaker. Additional application of UD has only partly been able to reduce the word error rates.

As a final remark it may be noted that all experiments with the observation models for noisy and noisy reverberant speech have been carried out with an oversimplified a priori model for the LMPSC feature vector of the noise, only. Considerable improvements of the performance of the proposed inference schemes may thus be expected by addressing this issue in future work.

A Appendix

A.1 Properties of Gaussian distributions

The following section summarizes some commonly employed properties of Gaussian distributions.

A.1.1 Quotient of Two Gaussian Distributions

When moving from (3.84) to (3.85), the following identity for the quotient of two Gaussian distributions has been employed:

$$\begin{aligned}&\frac{\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|a}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\right)}{\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|b}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)} \\ &= \frac{\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|a}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\right)}{\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}|b}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\breve{\mathbf{x}}}^{(\mathsf{eq})}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}}^{(\mathsf{eq})}\right)}\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\breve{\mathbf{x}}}^{(\mathsf{eq})}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}}^{(\mathsf{eq})}\right).\end{aligned} \tag{A.1}$$

Thereby, $\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} \in \mathbb{R}^{D\times D}$ denote the mean vector and the covariance matrix of Gaussian "a" and $\boldsymbol{\mu}_{\breve{\mathbf{x}}|b} \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b} \in \mathbb{R}^{D\times D}$ the corresponding moments of Gaussian "b". The *equivalent covariance matrix* $\boldsymbol{\Sigma}_{\breve{\mathbf{x}}}^{(\mathsf{eq})} \in \mathbb{R}^{D\times D}$ and the *equivalent mean vector* $\boldsymbol{\mu}_{\breve{\mathbf{x}}}^{(\mathsf{eq})} \in \mathbb{R}^{D\times 1}$ are, equivalent to (3.95) and (3.94), given by

$$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}}^{(\mathsf{eq})} = \left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}^{-1} - \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}^{-1}\right)^{-1}, \tag{A.2}$$

$$\boldsymbol{\mu}_{\breve{\mathbf{x}}}^{(\mathsf{eq})} = \boldsymbol{\Sigma}_{\breve{\mathbf{x}}}^{(\mathsf{eq})}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} - \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|b}\right). \tag{A.3}$$

The equality (A.1) will now be proven by comparing the exponents and the determinants of the left- and right-hand side.

The exponents of the quotient of the two Gaussians on the left-hand side of (A.1) may be combined to give

$$\begin{aligned}
&\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\right)^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\right)-\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)\\
&=\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}-\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\mathbf{x}_t-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\mathbf{x}_t-\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}-\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)\\
&\quad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b} && \text{(A.4)}\\
&=\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\mathbf{x}_t-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\mathbf{x}_t-\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\\
&\quad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\\
&\quad\underbrace{+\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}}_{=0} && \text{(A.5)}\\
&=\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)\\
&\quad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}, && \text{(A.6)}
\end{aligned}$$

which already are the desired exponents of the right-hand side. By expanding the fraction of the determinants of the left-hand side of (A.1) by the determinant of the equivalent covariance matrix, e.g.,

$$\frac{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right|}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right|}=\frac{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right|\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right|}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right|}\frac{1}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right|}, \tag{A.7}$$

the equality of the left- and right-hand side of (A.1) becomes apparent.

A.1.2 Product of Two Gaussian Distributions

When moving from (3.85) to (3.86), the following identity for the product of two Gaussian distributions "a" and "b" has been employed:

$$\begin{aligned}
&\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\check{\mathbf{x}}|a},\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right)\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\check{\mathbf{x}}|b},\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right)\\
&\quad=\frac{\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\check{\mathbf{x}}|a},\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right)\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\check{\mathbf{x}}|b},\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right)}{\mathcal{N}\left(\mathbf{0};\ \boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})},\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)}\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})},\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right) && \text{(A.8)}\\
&\quad=\mathcal{N}\left(\boldsymbol{\mu}_{\check{\mathbf{x}}|b};\ \boldsymbol{\mu}_{\check{\mathbf{x}}|a},\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right)\mathcal{N}\left(\mathbf{x}_t;\ \boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})},\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right). && \text{(A.9)}
\end{aligned}$$

Thereby, $\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\in\mathbb{R}^{D\times 1}$ and $\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\in\mathbb{R}^{D\times D}$ denote the mean vector and the covariance matrix of Gaussian "a" and $\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\in\mathbb{R}^{D\times 1}$ and $\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\in\mathbb{R}^{D\times D}$ the corresponding moments of Gaussian "b". The *equivalent covariance matrix* $\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\in\mathbb{R}^{D\times D}$ and the *equivalent mean vector* $\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\in\mathbb{R}^{D\times 1}$ are given by

$$\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}=\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}, \tag{A.10}$$

$$\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}=\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right). \tag{A.11}$$

The equality (A.8) and further (A.9) will now be proven by comparing the exponents and the respective determinants of the left- and right-hand side.

The exponents of the product of the two GAUSSIANs on the left-hand side of (A.8) may be combined to give

$$\begin{aligned}
&\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\right)^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\right)+\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)\\
&\quad=\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\mathbf{x}_t-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\mathbf{x}_t-\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right)\\
&\qquad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b} &\text{(A.12)}\\
&\quad=\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\mathbf{x}_t-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\mathbf{x}_t-\mathbf{x}_t^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\\
&\qquad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\\
&\qquad+\underbrace{\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}}_{=0} &\text{(A.13)}\\
&\quad=\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)\\
&\qquad+\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}, &\text{(A.14)}
\end{aligned}$$

which already are the desired exponents of the right-hand side of (A.8). By expanding the fraction of the determinants of the left-hand side of (A.8) by the determinant of the equivalent covariance matrix, e.g.,

$$\frac{1}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right|\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right|}=\frac{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right|}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}\right|\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}\right|}\frac{1}{\left|\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right|}, \tag{A.15}$$

the equality of the left- and right-hand side of (A.8) becomes apparent.

To further prove equality (A.9), the last three terms of (A.14) are rewritten by plugging in the definition of the equivalent mean vector (A.11) and the covariance matrix (A.10), which eventually yields

$$\begin{aligned}
&\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}}^{(\mathsf{eq})}\right)^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}}^{(\mathsf{eq})} &\text{(A.16)}\\
&=\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\\
&\quad-\left(\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\right) &\text{(A.17)}\\
&=\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}-\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\right)\boldsymbol{\mu}_{\check{\mathbf{x}}|a}\\
&\quad+\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}-\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)\boldsymbol{\mu}_{\check{\mathbf{x}}|b}\\
&\quad-\boldsymbol{\mu}_{\check{\mathbf{x}}|b}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|a}-\boldsymbol{\mu}_{\check{\mathbf{x}}|a}^{\dagger}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}\left(\boldsymbol{\Sigma}_{\check{\mathbf{x}}|a}^{-1}+\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\check{\mathbf{x}}|b}^{-1}\boldsymbol{\mu}_{\check{\mathbf{x}}|b}. &\text{(A.18)}
\end{aligned}$$

Employing the matrix inversion lemma [112, (144)]

$$\mathbf{A}^{-1}+\mathbf{B}^{-1}=\mathbf{A}^{-1}\left(\mathbf{A}+\mathbf{B}\right)\mathbf{B}^{-1} \tag{A.19}$$

and the matrix inversion lemma [112, (137)]

$$\mathbf{A}^{-1} - \mathbf{A}^{-1}\left(\mathbf{B}^{-1} + \mathbf{A}^{-1}\right)^{-1}\mathbf{A}^{-1} = (\mathbf{A}+\mathbf{B})^{-1} \tag{A.20}$$

for non-singular matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{D\times D}$ with

$$\mathbf{A} := \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} \tag{A.21}$$

$$\mathbf{B} := \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b} \tag{A.22}$$

eventually gives the desired exponent

$$\begin{aligned}
&\boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|a}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|a}\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} + \boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|b}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|b}\boldsymbol{\mu}_{\breve{\mathbf{x}}|b} - \left(\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}}\right)^{\dagger}\left(\boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}}\right)^{-1}\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}}\\
&= \boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|a}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} + \boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|b}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|b}\\
&\quad - \boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|a}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|b} - \boldsymbol{\mu}^{\dagger}_{\breve{\mathbf{x}}|b}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} &&\text{(A.23)}\\
&= \left(\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} - \boldsymbol{\mu}_{\breve{\mathbf{x}}|b}\right)^{\dagger}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right)^{-1}\left(\boldsymbol{\mu}_{\breve{\mathbf{x}}|a} - \boldsymbol{\mu}_{\breve{\mathbf{x}}|b}\right). &&\text{(A.24)}
\end{aligned}$$

and further the desired determinant

$$\begin{aligned}
\frac{\left|\boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}}\right|}{\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\right|\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right|} &= \frac{\left|\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|b}\right)^{-1}\right|}{\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\right|\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right|} &&\text{(A.25)}\\
&= \frac{1}{\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\right|\left|\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|b}\right)\right|\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right|} &&\text{(A.26)}\\
&= \frac{1}{\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a}\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}|b}\right)\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right|} &&\text{(A.27)}\\
&= \frac{1}{\left|\boldsymbol{\Sigma}_{\breve{\mathbf{x}}|a} + \boldsymbol{\Sigma}_{\breve{\mathbf{x}}|b}\right|}. &&\text{(A.28)}
\end{aligned}$$

A.2 Alternative Formulation of the Equivalent Mean

Plugging the definition of the equivalent covariance matrix (3.94) into the definition of the equivalent mean (3.95) gives

$$\begin{aligned}
\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} &= \boldsymbol{\Sigma}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right) &&\text{(A.29)}\\
&= \left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)\\
& &&\text{(A.30)}\\
&= \left(\mathbf{I} - \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\\
&\quad - \left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} - \mathbf{I}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}. &&\text{(A.31)}
\end{aligned}$$

The matrix inversion lemma [112, (145)]

$$(\mathbf{I}+\mathbf{A}\mathbf{B})^{-1}=\mathbf{I}-\mathbf{A}(\mathbf{I}+\mathbf{B}\mathbf{A})^{-1}\mathbf{B} \tag{A.32}$$

for non-singular matrices $\mathbf{A},\mathbf{B}\in\mathbb{R}^{D\times D}$ with

$$\begin{aligned}\mathbf{A} &:= \boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m} && \text{(A.33)}\\ \mathbf{B} &:= -\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}} && \text{(A.34)}\end{aligned}$$

then gives

$$\begin{aligned}&\left(\mathbf{I}-\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\right)^{-1}\\
&=\mathbf{I}+\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\left(\mathbf{I}-\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)^{-1}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}} && \text{(A.35)}\\
&=\mathbf{I}+\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)^{-1}\left(\mathbf{I}-\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)^{-1}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}} && \text{(A.36)}\\
&=\mathbf{I}+\left(\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)^{-1}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\left[\mathbf{I}-\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right]\right)^{-1} && \text{(A.37)}\\
&=\mathbf{I}+\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\left[\mathbf{I}-\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right]\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\right)^{-1} && \text{(A.38)}\\
&=\mathbf{I}+\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}-\mathbf{I}\right)^{-1} && \text{(A.39)}\end{aligned}$$

and hence (A.31) turns into (3.90), which is repeated here for convenience:

$$\begin{aligned}\boldsymbol{\mu}^{(\mathsf{eq})}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}=&\left(\mathbf{I}+\left[\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}-\mathbf{I}\right]^{-1}\right)\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}\\
&-\left(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}\boldsymbol{\Sigma}^{-1}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t},m}-\mathbf{I}\right)^{-1}\boldsymbol{\mu}_{\breve{\mathbf{x}}_t|\mathbf{o}_{1:t-1}}. && \text{(A.40)}\end{aligned}$$

A.3 Derivation of (4.115)

The conditional PDF $p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}$ may be expressed in terms of a conditional PDF of $\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}$ and that of $\breve{\boldsymbol{\alpha}}_t$ by writing

$$\begin{aligned}
&p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{v}_{o_t}^{(\mathsf{l})}\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\right.\right)\\
&=\int\limits_{\mathbb{R}^Q}\left[\quad\int\limits_{[-1,+1]^Q} p_{\breve{\mathbf{v}}_{o_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})},\breve{\boldsymbol{\alpha}}_t}\left(\mathbf{v}_{o_t}^{(\mathsf{l})}\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t}^{(\mathsf{l})},\boldsymbol{\alpha}_t\right.\right)\right.\\
&\qquad\qquad\left.p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\boldsymbol{\alpha}_t\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})}\mathbf{v}_{s_t}^{(\mathsf{l})}\right.\right)\mathrm{d}\boldsymbol{\alpha}_t\right]\\
&\qquad p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{v}_{s_t}^{(\mathsf{l})}\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\right.\right)\mathrm{d}\mathbf{v}_{s_t}^{(\mathsf{l})}
\end{aligned}\tag{A.41}$$

$$\begin{aligned}
&=\int\limits_{\mathbb{R}^Q}\left[\quad\int\limits_{[-1,+1]^Q}\delta\left(\mathbf{v}_{o_t}^{(\mathsf{l})}-\ln\left(1+\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}}-1\right)\circ\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)+2\boldsymbol{\alpha}_t\circ\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}}\circ\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right)\right)\right.\\
&\qquad\qquad\left.p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\boldsymbol{\alpha}_t\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t}^{(\mathsf{l})}\right.\right)\mathrm{d}\boldsymbol{\alpha}_t\right]\\
&\qquad p_{\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})}}\left(\mathbf{v}_{s_t}^{(\mathsf{l})}\left|\mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\right.\right)\mathrm{d}\mathbf{v}_{s_t}^{(\mathsf{l})}.
\end{aligned}\tag{A.42}$$

To solve the inner integral in (A.42), the bijective and continuously differentiable substitution function $\Phi_o:[-1,+1]^Q\to\mathbb{R}^Q$ and its *inverse* Φ_o^{-1} are introduced. They are defined by

$$\Phi_o\left(\boldsymbol{\alpha}_t\right):=\mathbf{v}_{o_t}^{(\mathsf{l})}-\ln\left(1+\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}}-1\right)\circ\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)+2\boldsymbol{\alpha}_t\circ\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}}\circ\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)\right)\tag{A.43}$$

$$\Phi_o^{-1}\left(\Phi_o\left(\boldsymbol{\alpha}_t\right)\right):=\frac{\mathrm{e}^{\mathbf{v}_{o_t}^{(\mathsf{l})}-\Phi_o(\boldsymbol{\alpha}_t)}-1-\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}}-1\right)\circ\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}}\circ\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)}\tag{A.44}$$

with associated (diagonal[1]) Jacobian matrix

$$J_{\Phi_o,\boldsymbol{\alpha}_t} = \begin{bmatrix} \frac{\partial \Phi_{o,0}(\boldsymbol{\alpha}_t)}{\partial \alpha_t(0)} & \cdots & \frac{\partial \Phi_{o,0}(\boldsymbol{\alpha}_t)}{\partial \alpha_t(Q-1)} \\ \vdots & \ddots & \vdots \\ \frac{\partial \Phi_{o,Q-1}(\boldsymbol{\alpha}_t)}{\partial \alpha_t(0)} & \cdots & \frac{\partial \Phi_{o,Q-1}(\boldsymbol{\alpha}_t)}{\partial \alpha_t(Q-1)} \end{bmatrix} \tag{A.45}$$

$$= -\operatorname{diag}\left(\begin{bmatrix} \dfrac{2\mathrm{e}^{\frac{v_{s_t}^{(\mathsf{l})}(0)}{2}}\zeta\left(r_t^{(\mathsf{l})}(0)\right)}{1+\left(\mathrm{e}^{v_{s_t}^{(\mathsf{l})}(0)}-1\right)\xi\left(r_t^{(\mathsf{l})}(0)\right)+2\alpha_t(0)\mathrm{e}^{\frac{v_{s_t}^{(\mathsf{l})}(0)}{2}}\zeta\left(r_t^{(\mathsf{l})}(0)\right)} \\ \vdots \\ \dfrac{2\mathrm{e}^{\frac{v_{s_t}^{(\mathsf{l})}(Q-1)}{2}}\zeta\left(r_t^{(\mathsf{l})}(Q-1)\right)}{1+\left(\mathrm{e}^{v_{s_t}^{(\mathsf{l})}(Q-1)}-1\right)\xi\left(r_t^{(\mathsf{l})}(Q-1)\right)+2\alpha_t(Q-1)\mathrm{e}^{\frac{v_{s_t}^{(\mathsf{l})}(Q-1)}{2}}\zeta\left(r_t^{(\mathsf{l})}(Q-1)\right)} \end{bmatrix}\right). \tag{A.46}$$

By further defining the volume

$$V_{\Phi_o} := \left\{ \Phi_o(\boldsymbol{\alpha}_t) \in \mathbb{R}^Q \middle| \Phi_{o,q}(-1) \leq \Phi_{o,q}(\alpha_t(q)) \leq \Phi_{o,q}(+1), \forall q \in \{0,\ldots,Q-1\} \right\} \tag{A.47}$$

the inner integral in (A.42) may eventually be expressed as [85, p. 244, Eq. (7-8)]

$$\int\limits_{V_{\Phi_o}} \delta(\Phi_o(\boldsymbol{\alpha}_t)) \frac{p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\Phi_o^{-1}(\Phi_o(\boldsymbol{\alpha}_t)) \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t}^{(\mathsf{l})}\right)}{\left|J_{\Phi_o,\boldsymbol{\alpha}_t}\right|} \mathrm{d}\Phi_o(\boldsymbol{\alpha}_t) \tag{A.48}$$

$$= \left.\frac{p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\Phi_o^{-1}(\Phi_o(\boldsymbol{\alpha}_t)) \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t}^{(\mathsf{l})}\right)}{\left|J_{\Phi_o,\boldsymbol{\alpha}_t}\right|}\right|_{\boldsymbol{\alpha}_t=\Phi_o^{-1}(\mathbf{0})} \tag{A.49}$$

$$= \left(\prod_{q=0}^{Q-1} \frac{\mathrm{e}^{v_{o_t}^{(\mathsf{l})}(q)}}{2\zeta\left(r_t^{(\mathsf{l})}(q)\right)}\right) \cdot p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_{t-L_H:t}^{(\mathsf{l})},\breve{\mathbf{n}}_t^{(\mathsf{l})},\breve{\mathbf{o}}_{1:t-1}^{(\mathsf{l})},\breve{\mathbf{v}}_{s_t}^{(\mathsf{l})}}\left(\frac{\mathrm{e}^{\mathbf{v}_{o_t}^{(\mathsf{l})}}-1-\left(\mathrm{e}^{\mathbf{v}_{s_t}^{(\mathsf{l})}}-1\right)\circ\xi\left(\mathbf{r}_t^{(\mathsf{l})}\right)}{2\mathrm{e}^{\frac{\mathbf{v}_{s_t}^{(\mathsf{l})}}{2}}\circ\zeta\left(\mathbf{r}_t^{(\mathsf{l})}\right)} \middle| \mathbf{x}_{t-L_H:t}^{(\mathsf{l})},\mathbf{n}_t^{(\mathsf{l})},\mathbf{o}_{1:t-1}^{(\mathsf{l})},\mathbf{v}_{s_t}^{(\mathsf{l})}\right). \tag{A.50}$$

[1] For readability purposes, the $\operatorname{diag}(\cdot)$ operator is introduced here. It either takes a vector argument and builds up a matrix with the vector components on its diagonal and zeros on all other positions or takes a matrix argument and builds up a vector containing the matrix' main diagonal components.

With (A.50), the conditional PDF $p_{\breve{\mathbf{v}}^{(l)}_{o_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}$ of the observation error in the presence of reverberation and noise is eventually given by

$$
\begin{aligned}
&p_{\breve{\mathbf{v}}^{(l)}_{o_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{o_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1}\right)\\
&\quad=\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v^{(l)}_{o_t}(q)}}{2\zeta\left(r^{(l)}_t(q)\right)}\right)\\
&\qquad\cdot\int\limits_{\mathbb{R}^Q}p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1},\breve{\mathbf{v}}^{(l)}_{s_t}}\left(\frac{\mathrm{e}^{\mathbf{v}^{(l)}_{o_t}}-1-\left(\mathrm{e}^{\mathbf{v}^{(l)}_{s_t}}-1\right)\circ\xi\left(\mathbf{r}^{(l)}_t\right)}{2\mathrm{e}^{\frac{\mathbf{v}^{(l)}_{s_t}}{2}}\circ\zeta\left(\mathbf{r}^{(l)}_t\right)}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1},\mathbf{v}^{(l)}_{s_t}\right)\\
&\qquad\quad p_{\breve{\mathbf{v}}^{(l)}_{s_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{s_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1},\right)\mathrm{d}\mathbf{v}^{(l)}_{s_t}.
\end{aligned}
\tag{A.51}
$$

By additionally introducing the sequence of the past LMPSCs $\mathbf{s}^{(l)}_{1:t-1}$ of the reverberant speech signal as a realization of the stochastic process $\breve{\mathbf{s}}^{(l)}_{1:t-1}$, the PDF $p_{\breve{\mathbf{v}}^{(l)}_{s_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}$ may be expressed as

$$
\begin{aligned}
&p_{\breve{\mathbf{v}}^{(l)}_{s_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{s_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1}\right)\\
&\quad=\int\limits_{\mathbb{R}^{(T-1)Q}}p_{\breve{\mathbf{v}}^{(l)}_{s_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1},\breve{\mathbf{s}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{s_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1},\mathbf{s}^{(l)}_{1:t-1}\right)\\
&\qquad\qquad p_{\breve{\mathbf{s}}^{(l)}_{1:t-1}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{s}^{(l)}_{1:t-1}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1}\right)\mathrm{d}\mathbf{s}^{(l)}_{1:t-1}
\end{aligned}
\tag{A.52}
$$

$$
\begin{aligned}
&\quad=\int\limits_{\mathbb{R}^{(T-1)Q}}p_{\breve{\mathbf{v}}^{(l)}_{s_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{s}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{s_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{s}^{(l)}_{1:t-1}\right)\\
&\qquad\qquad p_{\breve{\mathbf{s}}^{(l)}_{1:t-1}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{s}^{(l)}_{1:t-1}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1}\right)\mathrm{d}\mathbf{s}^{(l)}_{1:t-1}.
\end{aligned}
\tag{A.53}
$$

If the approximation (4.95) is now applied to (A.53), the conditional PDF $p_{\breve{\mathbf{s}}^{(l)}_{1:t-1}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1},\breve{\boldsymbol{\alpha}}_t}$ does not have to be determined explicitly and (A.41) turns into

$$
\begin{aligned}
&p_{\breve{\mathbf{v}}^{(l)}_{o_t}|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{o_t}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1}\right)\\
&\quad\approx\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v^{(l)}_{o_t}(q)}}{2\zeta\left(r^{(l)}_t(q)\right)}\right)\\
&\qquad\cdot\int\limits_{\mathbb{R}^Q}p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}^{(l)}_{t-L_H:t},\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{o}}^{(l)}_{1:t-1},\breve{\mathbf{v}}^{(l)}_{s_t}}\left(\frac{\mathrm{e}^{\mathbf{v}^{(l)}_{o_t}}-1-\left(\mathrm{e}^{\mathbf{v}^{(l)}_{s_t}}-1\right)\circ\xi\left(\mathbf{r}^{(l)}_t\right)}{2\mathrm{e}^{\frac{\mathbf{v}^{(l)}_{s_t}}{2}}\circ\zeta\left(\mathbf{r}^{(l)}_t\right)}\middle|\mathbf{x}^{(l)}_{t-L_H:t},\mathbf{n}^{(l)}_t,\mathbf{o}^{(l)}_{1:t-1},\mathbf{v}^{(l)}_{s_t}\right)\\
&\qquad\quad p_{\breve{\mathbf{v}}^{(l)}_{s_t}}\left(\mathbf{v}^{(l)}_{s_t}\right)\mathrm{d}\mathbf{v}^{(l)}_{s_t}
\end{aligned}
\tag{A.54}
$$

$$
\approx\left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v^{(l)}_{o_t}(q)}}{2\zeta\left(r^{(l)}_t(q)\right)}\right)\int\limits_{\mathbb{R}^Q}p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{v}^{(l)}_{o_t}}-1-\left(\mathrm{e}^{\mathbf{v}^{(l)}_{s_t}}-1\right)\circ\xi\left(\mathbf{r}^{(l)}_t\right)}{2\mathrm{e}^{\frac{\mathbf{v}^{(l)}_{s_t}}{2}}\circ\zeta\left(\mathbf{r}^{(l)}_t\right)}\right)p_{\breve{\mathbf{v}}^{(l)}_{s_t}}\left(\mathbf{v}^{(l)}_{s_t}\right)\mathrm{d}\mathbf{v}^{(l)}_{s_t}.
\tag{A.55}
$$

The last approximation thereby assumes the phase factor $\boldsymbol{\alpha}_t$ to be a realization of the RV $\breve{\boldsymbol{\alpha}}_t$ that is independent of $\breve{\mathbf{x}}^{(l)}_{t-L_H:t}$, $\breve{\mathbf{n}}^{(l)}_t$, $\breve{\mathbf{o}}^{(l)}_{1:t-1}$ and $\breve{\mathbf{v}}^{(l)}_{st}$. The validity of this approximation is discussed in more detail in Sec. 4.6.

A.4 Derivation of (4.139)

With (4.135), the conditional PDF $p_{\breve{\mathbf{v}}^{(l)}_{yt}|\breve{\mathbf{x}}^{(l)}_t,\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{y}}^{(l)}_{1:t-1}}$ may first be written as

$$\begin{aligned}
&p_{\breve{\mathbf{v}}^{(l)}_{yt}|\breve{\mathbf{x}}^{(l)}_t,\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{y}}^{(l)}_{1:t-1}}\left(\mathbf{v}^{(l)}_{yt}\middle|\mathbf{x}^{(l)}_t,\mathbf{n}^{(l)}_t,\mathbf{y}^{(l)}_{1:t-1}\right)\\
&=\int\limits_{[-1,+1]^Q} p_{\breve{\mathbf{v}}^{(l)}_{yt}|\breve{\mathbf{x}}^{(l)}_t,\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{y}}^{(l)}_{1:t-1},\breve{\boldsymbol{\alpha}}_t}\left(\mathbf{v}^{(l)}_{yt}\middle|\mathbf{x}^{(l)}_t,\mathbf{n}^{(l)}_t,\mathbf{y}^{(l)}_{1:t-1},\boldsymbol{\alpha}_t\right)\\
&\qquad\qquad p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}^{(l)}_t,\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{y}}^{(l)}_{1:t-1}}\left(\boldsymbol{\alpha}_t\middle|\mathbf{x}^{(l)}_t,\mathbf{n}^{(l)}_t,\mathbf{y}^{(l)}_{1:t-1}\right)\mathrm{d}\boldsymbol{\alpha}_t && \text{(A.56)}\\
&=\int\limits_{[-1,+1]^Q} \delta\left(\mathbf{v}^{(l)}_{yt}-\ln\left(1+2\boldsymbol{\alpha}_t\circ\zeta\left(\mathbf{r}^{(l)}_t\right)\right)\right)p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}^{(l)}_t,\breve{\mathbf{n}}^{(l)}_t,\breve{\mathbf{y}}^{(l)}_{1:t-1}}\left(\boldsymbol{\alpha}_t\middle|\mathbf{x}^{(l)}_t,\mathbf{n}^{(l)}_t,\mathbf{y}^{(l)}_{1:t-1}\right)\mathrm{d}\boldsymbol{\alpha}_t. && \text{(A.57)}
\end{aligned}$$

To solve the integral (A.57), the bijective and continuously differentiable substitution function $\Phi_y:[-1,+1]^Q\to\mathbb{R}^Q$ and its *inverse* Φ_y^{-1} defined by

$$\Phi_y\left(\boldsymbol{\alpha}_t\right):=\mathbf{v}^{(l)}_{yt}-\ln\left(1+2\boldsymbol{\alpha}_t\circ\zeta\left(\mathbf{r}^{(l)}_t\right)\right) \tag{A.58}$$

$$\Phi_y^{-1}\left(\Phi_y\left(\boldsymbol{\alpha}_t\right)\right):=\frac{\mathrm{e}^{\mathbf{v}^{(l)}_{yt}-\Phi_y(\boldsymbol{\alpha}_t)}-1}{2\zeta\left(\mathbf{r}^{(l)}_t\right)} \tag{A.59}$$

with the (diagonal) Jacobian matrix

$$\begin{aligned}
J_{\Phi_y,\boldsymbol{\alpha}_t}&=\begin{bmatrix}\frac{\partial\Phi_{y,0}(\boldsymbol{\alpha}_t)}{\partial\alpha_t(0)} & \cdots & \frac{\partial\Phi_{y,0}(\boldsymbol{\alpha}_t)}{\partial\alpha_t(Q-1)}\\ \vdots & \ddots & \vdots\\ \frac{\partial\Phi_{y,Q-1}(\boldsymbol{\alpha}_t)}{\partial\alpha_t(0)} & \cdots & \frac{\partial\Phi_{y,Q-1}(\boldsymbol{\alpha}_t)}{\partial\alpha_t(Q-1)}\end{bmatrix} && \text{(A.60)}\\
&=-\operatorname{diag}\left(\left[\frac{2\zeta\left(r^{(l)}_t(0)\right)}{1+2\alpha_t(0)\zeta\left(r^{(l)}_t(0)\right)},\ldots,\frac{2\zeta\left(r^{(l)}_t(Q-1)\right)}{1+2\alpha_t(Q-1)\zeta\left(r^{(l)}_t(Q-1)\right)}\right]\right) && \text{(A.61)}
\end{aligned}$$

are introduced. Defining the volume

$$V_{\Phi_y}:=\left\{\Phi_y\left(\boldsymbol{\alpha}_t\right)\in\mathbb{R}^Q\middle|\Phi_{y,q}(-1)\le\Phi_{y,q}\left(\alpha_t(q)\right)\le\Phi_{y,q}(+1),\forall q\in\{0,\ldots,Q-1\}\right\} \tag{A.62}$$

now allows the integral (A.57) to be expressed as [85, p. 244, Eq. (7-8)]

$$
\begin{aligned}
&p_{\breve{\mathbf{v}}_{yt}^{\text{(l)}}|\breve{\mathbf{x}}_t^{\text{(l)}},\breve{\mathbf{n}}_t^{\text{(l)}},\breve{\mathbf{y}}_{1:t-1}^{\text{(l)}}}\left(\mathbf{v}_{yt}^{\text{(l)}}\middle|\mathbf{x}_t^{\text{(l)}},\mathbf{n}_t^{\text{(l)}},\mathbf{y}_{1:t-1}^{\text{(l)}}\right)\\
&= \int\limits_{V_{\Phi_y}} \delta\left(\Phi_y\left(\boldsymbol{\alpha}_t\right)\right) \frac{p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_t^{\text{(l)}},\breve{\mathbf{n}}_t^{\text{(l)}},\breve{\mathbf{y}}_{1:t-1}^{\text{(l)}}}\left(\Phi_y^{-1}\left(\Phi_y\left(\boldsymbol{\alpha}_t\right)\right)\middle|\mathbf{x}_t^{\text{(l)}},\mathbf{n}_t^{\text{(l)}},\mathbf{y}_{1:t-1}^{\text{(l)}}\right)}{\left|J_{\Phi_y,\boldsymbol{\alpha}_t}\right|}\mathrm{d}\Phi_y\left(\boldsymbol{\alpha}_t\right) \qquad \text{(A.63)}\\
&= \left.\frac{p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_t^{\text{(l)}},\breve{\mathbf{n}}_t^{\text{(l)}},\breve{\mathbf{y}}_{1:t-1}^{\text{(l)}}}\left(\Phi_y^{-1}\left(\Phi_y\left(\boldsymbol{\alpha}_t\right)\right)\middle|\mathbf{x}_t^{\text{(l)}},\mathbf{n}_t^{\text{(l)}},\mathbf{y}_{1:t-1}^{\text{(l)}}\right)}{\left|J_{\Phi_y,\boldsymbol{\alpha}_t}\right|}\right|_{\boldsymbol{\alpha}_t=\Phi_y^{-1}(\mathbf{0})} \qquad \text{(A.64)}\\
&= \left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v_{yt}^{\text{(l)}}(q)}}{2\zeta\left(r_t^{\text{(l)}}(q)\right)}\right) p_{\breve{\boldsymbol{\alpha}}_t|\breve{\mathbf{x}}_t^{\text{(l)}},\breve{\mathbf{n}}_t^{\text{(l)}},\breve{\mathbf{y}}_{1:t-1}^{\text{(l)}}}\left(\frac{\mathrm{e}^{\mathbf{v}_{yt}^{\text{(l)}}}-1}{2\zeta\left(\mathbf{r}_t^{\text{(l)}}\right)}\middle|\mathbf{x}_t^{\text{(l)}},\mathbf{n}_t^{\text{(l)}},\mathbf{y}_{1:t-1}^{\text{(l)}}\right) \qquad \text{(A.65)}\\
&\approx \left(\prod_{q=0}^{Q-1}\frac{\mathrm{e}^{v_{yt}^{\text{(l)}}(q)}}{2\zeta\left(r_t^{\text{(l)}}(q)\right)}\right) p_{\breve{\boldsymbol{\alpha}}_t}\left(\frac{\mathrm{e}^{\mathbf{v}_{yt}^{\text{(l)}}}-1}{2\zeta\left(\mathbf{r}_t^{\text{(l)}}\right)}\right). \qquad \text{(A.66)}
\end{aligned}
$$

Note that the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ has again been considered to be independent of $\breve{\mathbf{x}}_t^{\text{(l)}}$, $\breve{\mathbf{n}}_t^{\text{(l)}}$ and $\breve{\mathbf{y}}_{1:t-1}^{\text{(l)}}$ while moving from (A.65) to (A.66).

A.5 Mean Vector and Covariance Matrix of the Observation Error in the Presence of Reverberation and Noise

If the observation error $\breve{\mathbf{v}}_{ot}^{\text{(l)}}$ is assumed to be distributed according to a Gaussian, the RV $\breve{\mathbf{v}}_{ot}^{\text{(m)}} = \mathrm{e}^{\breve{\mathbf{v}}_{ot}^{\text{(l)}}}$ may likewise be assumed to be log-normally distributed. However, as shown in Appendix A.11, the mean vector and the covariance matrix of a Gaussian distributed RV can be expressed in terms of the mean vector and the covariance matrix of the corresponding log-normally distributed RV and vice versa.

The components of the mean vector

$$\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right) := E\left[\breve{\mathbf{v}}_{ot}^{\text{(l)}}\middle|\mathcal{C}\right] \tag{A.67}$$

$$\tag{A.68}$$

and the covariance matrix

$$\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right) := E\left[\left(\breve{\mathbf{v}}_{ot}^{\text{(l)}} - E\left[\breve{\mathbf{v}}_{ot}^{\text{(l)}}\middle|\mathcal{C}\right]\right)\left(\breve{\mathbf{v}}_{ot}^{\text{(l)}} - E\left[\breve{\mathbf{v}}_{ot}^{\text{(l)}}\middle|\mathcal{C}\right]\right)^\dagger\middle|\mathcal{C}\right] \tag{A.69}$$

of the RV $\breve{\mathbf{v}}_{ot}^{\text{(l)}}$, where the short-hand notation $\mathcal{C} := \breve{\mathbf{x}}_{t-L_H:t}^{\text{(l)}}, \breve{\mathbf{n}}_t^{\text{(l)}}, \breve{\mathbf{o}}_{1:t-1}^{\text{(l)}}$ has been employed for readability purposes, are thus related to the conditional mean vector $\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(m)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)$ of the log-normally distributed RV $\breve{\mathbf{v}}_{ot}^{\text{(m)}}$ by (repeated here from (4.122) and (4.123) for convenience)

$$\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right) = \ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(m)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)\right) - \frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)\right) \tag{A.70}$$

$$\boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{\text{(l)}}}\left(\mathbf{r}_t^{\text{(l)}}\right) = \ln\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(m)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)\left(\boldsymbol{\mu}_{\breve{\mathbf{v}}_o^{\text{(m)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)\right)^\dagger + \boldsymbol{\Sigma}_{\breve{\mathbf{v}}_o^{\text{(m)}}}\left(\mathbf{r}_t^{\text{(l)}}\right)\right). \tag{A.71}$$

With (4.112), the RV $\breve{\mathbf{v}}_{o_t}^{\mathsf{(m)}}$ is, for a given IRNR $\mathbf{r}_t^{\mathsf{(l)}}$, related to the vector of phase factors $\breve{\boldsymbol{\alpha}}_t$ and the observation error in the reverberant but noise-free case $\breve{\mathbf{v}}_{s_t}^{\mathsf{(l)}}$ by

$$\breve{\mathbf{v}}_{o_t}^{\mathsf{(m)}} := 1+\left(\mathrm{e}^{\breve{\mathbf{v}}_{s_t}^{\mathsf{(l)}}}-1\right)\xi\left(\mathbf{r}_t^{\mathsf{(l)}}\right)+2\breve{\boldsymbol{\alpha}}_t\mathrm{e}^{\frac{\breve{\mathbf{v}}_{s_t}^{\mathsf{(l)}}}{2}}\zeta\left(\mathbf{r}_t^{\mathsf{(l)}}\right) \tag{A.72}$$

and hence it can immediately be seen that

$$\begin{aligned}
\mu_{\breve{v}_o^{\mathsf{(m)}}(q)}\left(\mathbf{r}_t^{\mathsf{(l)}}\right) &:= E\left[\breve{v}_{o_t}^{\mathsf{(m)}}(q)\middle|\mathcal{C}\right] && \text{(A.73)}\\
&= 1+\left(E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}\middle|\mathcal{C}\right]-1\right)\xi\left(r_t^{\mathsf{(l)}}(q)\right)+2E\left[\breve{\alpha}_t(q)|\mathcal{C}\right]E\left[\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}{2}}\middle|\mathcal{C}\right]\zeta\left(\mathbf{r}_t^{\mathsf{(l)}}\right) && \text{(A.74)}\\
&= 1+\left(E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}\right]-1\right)\xi\left(r_t^{\mathsf{(l)}}(q)\right). && \text{(A.75)}
\end{aligned}$$

For the last step, the independence of the observation error $\breve{v}_{s_t}^{\mathsf{(l)}}(q)$ and the phase factor $\breve{\alpha}_t(q)$ on the context $\mathcal{C}$ as well as the zero-mean property of the phase factor $\breve{\alpha}_t(q)$ have been employed.

For the covariance $\sigma_{\breve{v}_o^{\mathsf{(m)}}(q),\breve{v}_o^{\mathsf{(m)}}(q')}\left(\mathbf{r}_t^{\mathsf{(l)}}\right)$ it may further be written

$$\begin{aligned}
&\sigma_{\breve{v}_o^{\mathsf{(m)}}(q),\breve{v}_o^{\mathsf{(m)}}(q')}\left(\mathbf{r}_t^{\mathsf{(l)}}\right)\\
&:= E\left[\left(\breve{v}_{o_t}^{\mathsf{(m)}}(q)-E\left[\breve{v}_{o_t}^{\mathsf{(m)}}(q)\middle|\mathcal{C}\right]\right)\left(\breve{v}_{o_t}^{\mathsf{(m)}}\left(q'\right)-E\left[\breve{v}_{o_t}^{\mathsf{(m)}}\left(q'\right)\middle|\mathcal{C}\right]\right)\middle|\mathcal{C}\right] && \text{(A.76)}\\
&= E\Bigg[\left\{\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}\right]\right)\xi\left(r_t^{\mathsf{(l)}}(q)\right)+2\breve{\alpha}_t(q)\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}{2}}\zeta\left(r_t^{\mathsf{(l)}}(q)\right)\right\}\\
&\qquad\left\{\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}\right]\right)\xi\left(r_t^{\mathsf{(l)}}\left(q'\right)\right)+2\breve{\alpha}_t\left(q'\right)\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}{2}}\zeta\left(r_t^{\mathsf{(l)}}\left(q'\right)\right)\right\}\Bigg|\mathcal{C}\Bigg] && \text{(A.77)}\\
&= E\left[\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}\right]\right)\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}\right]\right)\middle|\mathcal{C}\right]\xi\left(r_t^{\mathsf{(l)}}(q)\right)\xi\left(r_t^{\mathsf{(l)}}\left(q'\right)\right)\\
&\quad+4E\left[\breve{\alpha}_t(q)\breve{\alpha}_t\left(q'\right)\middle|\mathcal{C}\right]E\left[\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q)+\breve{v}_{s_t}^{\mathsf{(l)}}(q')}{2}}\middle|\mathcal{C}\right]\zeta\left(r_t^{\mathsf{(l)}}(q)\right)\zeta\left(r_t^{\mathsf{(l)}}\left(q'\right)\right)\\
&\quad+2E\left[\breve{\alpha}_t(q)|\mathcal{C}\right]E\left[\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}{2}}\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}\right]\right)\middle|\mathcal{C}\right]\\
&\quad+2E\left[\breve{\alpha}_t\left(q'\right)\middle|\mathcal{C}\right]E\left[\mathrm{e}^{\frac{\breve{v}_{s_t}^{\mathsf{(l)}}(q')}{2}}\left(\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}-E\left[\mathrm{e}^{\breve{v}_{s_t}^{\mathsf{(l)}}(q)}\right]\right)\middle|\mathcal{C}\right]. && \text{(A.78)}
\end{aligned}$$

With the same reasoning as before, the above simplifies to

$$\begin{aligned}
&\sigma_{\breve{v}_o^{(\mathsf{m})}(q),\breve{v}_o^{(\mathsf{m})}(q')}\left(\mathbf{r}_t^{(\mathsf{l})}\right)\\
&= E\left[\left(\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)} - E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)}\right]\right)\left(\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q')} - E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q')}\right]\right)\right]\xi\left(r_t^{(\mathsf{l})}(q)\right)\xi\left(r_t^{(\mathsf{l})}\left(q'\right)\right)\\
&\quad + 4E\left[\breve{\alpha}_t(q)\,\breve{\alpha}_t\left(q'\right)\right]E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}_{s_t}^{(\mathsf{l})}(q)+\breve{v}_{s_t}^{(\mathsf{l})}(q')\right)}\right]\zeta\left(r_t^{(\mathsf{l})}(q)\right)\zeta\left(r_t^{(\mathsf{l})}\left(q'\right)\right) \qquad \text{(A.79)}\\
&= \left(E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)+\breve{v}_{s_t}^{(\mathsf{l})}(q')}\right] - E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)}\right]E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q')}\right]\right)\xi\left(r_t^{(\mathsf{l})}(q)\right)\xi\left(r_t^{(\mathsf{l})}\left(q'\right)\right)\\
&\quad + 4\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}_{s_t}^{(\mathsf{l})}(q)+\breve{v}_{s_t}^{(\mathsf{l})}(q')\right)}\right]\zeta\left(r_t^{(\mathsf{l})}(q)\right)\zeta\left(r_t^{(\mathsf{l})}\left(q'\right)\right), \qquad \text{(A.80)}
\end{aligned}$$

where $\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}$ denotes the covariance between the phase factors $\breve{\alpha}_t(q)$ and $\breve{\alpha}_t(q')$.

The remaining expectation values are, since $\breve{v}_{s_t}^{(\mathsf{l})}(q)$ and $\breve{v}_{s_t}^{(\mathsf{l})}(q')$ are assumed to be jointly GAUSSIAN distributed, given by

$$E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)}\right] = \mathrm{e}^{\mu_{\breve{v}_s^{(\mathsf{l})}(q)}+\frac{1}{2}\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)}} \qquad \text{(A.81)}$$

$$E\left[\mathrm{e}^{\breve{v}_{s_t}^{(\mathsf{l})}(q)+\breve{v}_{s_t}^{(\mathsf{l})}(q')}\right] = \mathrm{e}^{\mu_{\breve{v}_s^{(\mathsf{l})}(q)}+\mu_{\breve{v}_s^{(\mathsf{l})}(q')}+\frac{1}{2}\left(\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)}+\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q')}+2\sigma_{\breve{v}_s^{(\mathsf{l})}(q),\breve{v}_s^{(\mathsf{l})}(q')}\right)} \qquad \text{(A.82)}$$

$$E\left[\mathrm{e}^{\frac{1}{2}\left(\breve{v}_{s_t}^{(\mathsf{l})}(q)+\breve{v}_{s_t}^{(\mathsf{l})}(q')\right)}\right] = \mathrm{e}^{\frac{1}{2}\left(\mu_{\breve{v}_s^{(\mathsf{l})}(q)}+\mu_{\breve{v}_s^{(\mathsf{l})}(q')}\right)+\frac{1}{8}\left(\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q)}+\sigma^2_{\breve{v}_s^{(\mathsf{l})}(q')}+2\sigma_{\breve{v}_s^{(\mathsf{l})}(q),\breve{v}_s^{(\mathsf{l})}(q')}\right)}. \qquad \text{(A.83)}$$

A.6 Frequency Dependent Power Compensation Constant

The frequency dependent power compensation constant $C_P(k)$ is chosen such that the error $E_t(k)$ in (4.69) is of zero mean, which leads to postulation

$$C_P(k) = \frac{E\left[\left|\sum\limits_{t'=-L_{H,\ell}}^{L_H}\sum\limits_{k'=0}^{K-1}\breve{X}_{t-t'}(k')\,h_{t'}(k,k')\right|^2\right]}{E\left[\sum\limits_{t'=0}^{L_H}\left|\breve{X}_{t-t'}(k)\right|^2\left|h_{t'}(k,k)\right|^2\right]}. \qquad \text{(A.84)}$$

Assuming the speech signal to be a realization of a real-valued white GAUSSIAN random process with auto-correlation function

$$E\left[x(l)\,x\left(l'\right)\right] = \sigma_{\breve{x}}^2\delta\left(l-l'\right), \qquad \text{(A.85)}$$

where $\sigma_{\breve{x}}^2$ denotes the power of the involved RV, allows the denominator in (A.84) to be written as

$$E\left[\sum_{t'=0}^{L_H}\left|\breve{X}_{t-t'}\left(k\right)\right|^2\left|h_{t'}\left(k,k\right)\right|^2\right] \tag{A.86}$$

$$=\sum_{t'=0}^{L_H}E\left[\left|\breve{X}_{t-t'}\left(k\right)\right|^2\right]\left|h_{t'}\left(k,k\right)\right|^2 \tag{A.87}$$

$$=\sum_{t'=0}^{L_H}\sum_{l,l'=0}^{L_w-1}w_{\mathsf{A}}\left(l\right)w_{\mathsf{A}}\left(l'\right)E\left[x\left(l+\left(t-t'\right)B\right)x\left(l'+\left(t-t'\right)B\right)\right]\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(l-l'\right)k}\left|h_t\left(k,k\right)\right|^2 \tag{A.88}$$

$$=\sigma_{\breve{x}}^2\left(\sum_{l=0}^{L_w-1}w_{\mathsf{A}}^2\left(l\right)\right)\sum_{t'=0}^{L_H}\left|h_{t'}\left(k,k\right)\right|^2 \tag{A.89}$$

$$=\sigma_{\breve{x}}^2C_D(k), \tag{A.90}$$

where

$$C_D(k):=\left(\sum_{l=0}^{L_w-1}w_{\mathsf{A}}^2\left(l\right)\right)\sum_{t'=0}^{L_H}\left|h_{t'}\left(k,k\right)\right|^2. \tag{A.91}$$

Equivalently, the numerator in (A.84) may be expressed as

$$E\left[\left|\sum_{t'=-L_{H,\ell}}^{L_H}\sum_{k'=0}^{K-1}\breve{X}_{t-t'}\left(k'\right)h_{t'}\left(k,k'\right)\right|^2\right]$$

$$=\sum_{t',t''=-L_{H,\ell}}^{L_H}\sum_{k',k''=0}^{K-1}E\left[\breve{X}_{t-t'}\left(k'\right)\breve{X}_{t-t''}^{*}\left(k''\right)\right]h_{t'}\left(k,k'\right)h_{t''}^{*}\left(k,k''\right) \tag{A.92}$$

$$=\sum_{t',t''=-L_{H,\ell}}^{L_H}\sum_{k',k''=0}^{K-1}\sum_{l,l'=0}^{L_w-1}w_{\mathsf{A}}\left(l\right)w_{\mathsf{A}}\left(l'\right)E\left[x\left(l+\left(t-t'\right)B\right)x\left(l'+\left(t-t''\right)B\right)\right]$$
$$\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-l'k''\right)}h_{t'}\left(k,k'\right)h_{t''}^{*}\left(k,k''\right) \tag{A.93}$$

$$=\sigma_{\breve{x}}^2\sum_{t',t''=-L_{H,\ell}}^{L_H}\sum_{l=0}^{L_w-1}w_{\mathsf{A}}\left(l\right)w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right)$$
$$\sum_{k',k''=0}^{K-1}h_{t'}\left(k,k'\right)h_{t''}^{*}\left(k,k''\right)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-\left(l+\left[t''-t'\right]B\right)k''\right)} \tag{A.94}$$

$$=\sigma_{\breve{x}}^2C_N(k) \tag{A.95}$$

with

$$C_N(k):=\sum_{t',t''=-L_{H,\ell}}^{L_H}\sum_{l=0}^{L_w-1}w_{\mathsf{A}}\left(l\right)w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right)$$
$$\sum_{k',k''=0}^{K-1}h_{t'}\left(k,k'\right)h_{t''}^{*}\left(k,k''\right)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-\left(l+\left[t''-t'\right]B\right)k''\right)}. \tag{A.96}$$

Thus, the frequency dependent power compensation constant for a given AIR may be computed as

$$C_P(k) = \frac{C_N(k)}{C_D(k)}, \tag{A.97}$$

which is independent of the power $\sigma^2_{\breve{x}}$ of the clean speech RVs. Note, that if the AIR is not known, but rather modeled as a realization of a random process, the expectation in (A.86) and (A.92) also have to be taken w.r.t. the cross-band/band-to-band filters, i.e., any occurrence of $h_{t'}(k,k')\,h^*_{t''}(k,k'')$ in (A.91) and (A.96) has to be replaced by $E\left[h_{t'}(k,k')\,h^*_{t''}(k,k'')\right]$.

A.6.1 Applying the Model of the AIR

With the definition of the cross-band and band-to-band filters given in (4.51), these expectation value may be written as

$$\begin{aligned}
&E\left[\breve{h}_{t'}\left(k,k'\right)\breve{h}^*_{t''}\left(k,k''\right)\right]\\
&= E\left[\left(\sum_{p=0}^{L_h-1}\breve{h}(p)\,\varPhi_{t'B-p}\left(k,k'\right)\right)\left(\sum_{p'=0}^{L_h-1}\breve{h}\left(p'\right)\varPhi_{t''B-p'}\left(k,k''\right)\right)^*\right] &&\text{(A.98)}\\
&= \sum_{p,p'=0}^{L_h-1} E\left[\breve{h}(p)\,\breve{h}\left(p'\right)\right]\varPhi_{t'B-p}\left(k,k'\right)\varPhi_{t''B-p'}\left(k,k''\right). &&\text{(A.99)}
\end{aligned}$$

With the stochastic model of the AIR given by (4.147), Eq. (A.99) may be rewritten as

$$\begin{aligned}
&E\left[\breve{h}_{t'}\left(k,k'\right)\breve{h}^*_{t''}\left(k,k''\right)\right]\\
&= \sum_{p=0}^{L_h-1}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2p}{\tau_h}}\varPhi_{t'B-p}\left(k,k'\right)\varPhi^*_{t''B-p}\left(k,k''\right) &&\text{(A.100)}\\
&= \sum_{p'=t'B-L_h+1}^{t'B}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\varPhi_{p'}\left(k,k'\right)\varPhi^*_{(t''-t')B+p'}\left(k,k''\right) &&\text{(A.101)}\\
&= \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\varPhi_{p'}\left(k,k'\right)\varPhi^*_{(t''-t')B+p'}\left(k,k''\right), &&\text{(A.102)}
\end{aligned}$$

where the change of the summation limits to $\mathcal{L}(t')$ and $\mathcal{U}(t')$ (specified in (4.55) and (4.56)) in the last conversion is again due to the limited support of the auxiliary function $\varPhi_{p'}(k,k')$, which is defined for $-L_w+1 \leq p' \leq L_w+1$. The definition of the auxiliary function given in (4.53) may now be employed to further expand (A.102) to

$$\begin{aligned}
&E\left[\breve{h}_{t'}\left(k,k'\right)\breve{h}^*_{t''}\left(k,k''\right)\right]\\
&= \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\sum_{l',l''=0}^{L_w-1} w_{\mathsf{A}}\left(l'\right)w_{\mathsf{A}}\left(l''\right)w_{\mathsf{S}}\left(l'+p'\right)w_{\mathsf{S}}\left(l''+\left(t''-t'\right)B+p'\right)\\
&\qquad \mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}(l'+p')k'}\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}(l''+(t''-t')B+p')k''}\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}(l'-l'')k}. &&\text{(A.103)}
\end{aligned}$$

Expected value of denominator (A.91) For the calculation of the expectation value of (A.91), the special case $t' = t''$, $k' = k'' = k$ is of interest, for which (A.103) turns into

$$E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] = E\left[\breve{h}_{t'}(k,k)\,\breve{h}^*_{t'}(k,k)\right] \tag{A.104}$$

$$= \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \sigma^2_{\breve{h}} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{A.105}$$

with

$$w\left(p'\right) = \sum_{l'=0}^{L_w-1} w_{\mathsf{A}}\left(l'\right) w_{\mathsf{S}}\left(l'+p'\right) \tag{A.106}$$

and renders the expected value of the denominator term (A.91) to be independent of the actual frequency index k, i.e.,

$$E\left[\breve{C}_D(k)\right] = \left(\sum_{l=0}^{L_w-1} w^2_{\mathsf{A}}(l)\right) \sum_{t'=0}^{L_H} E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right] \tag{A.107}$$

$$= \sigma^2_{\breve{h}} C_D. \tag{A.108}$$

where

$$C_D := \left(\sum_{l=0}^{L_w-1} w^2_{\mathsf{A}}(l)\right) \sum_{t'=0}^{L_H} \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w\left(p'\right). \tag{A.109}$$

Expected value of numerator (A.96) For the calculation of the expectation value of (A.96), the expectation value (A.103) is required for all $t', t'' \in \left\{-L_{H,\ell}, \ldots, L_H\right\}$ and all $k', k'' \in \{0, \ldots, K\}$. However, since (A.96) involves a weighted summation of over all $k', k'' \in \{0, \ldots, K\}$, the expression

$$\begin{aligned}
&\sum_{k',k''=0}^{K-1} E\left[\breve{h}_{t'}\left(k,k'\right) \breve{h}^*_{t''}\left(k,k''\right)\right] \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-\left(l+\left[t''-t'\right]B\right)k''\right)} \\
&= \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \sigma^2_{\breve{h}} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} \sum_{l',l''=0}^{L_w-1} w_{\mathsf{A}}\left(l'\right) w_{\mathsf{A}}\left(l''\right) w_{\mathsf{S}}\left(l'+p'\right) w_{\mathsf{S}}\left(l''+\left(t''-t'\right)B+p'\right) \\
&\qquad \left(\sum_{k'=0}^{K-1} \mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}\left(l'+p'-l\right)k'}\right) \left(\sum_{k''=0}^{K-1} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(l''+p'-l\right)k''}\right) \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(l'-l''\right)k}
\end{aligned} \tag{A.110}$$

is of particular interest. Due to the sum orthogonality of complex exponentials, i.e.,

$$\frac{1}{K}\sum_{k=0}^{K} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}kp} = \sum_{\vartheta=-\infty}^{\infty} \delta\left(p-\vartheta K\right), \tag{A.111}$$

Eq. (A.110) may be rewritten as

$$E\left[\sum_{k',k''=0}^{K-1}\breve{h}_{t'}\left(k,k'\right)\breve{h}^{*}_{t''}\left(k,k''\right)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-\left(l+\left[t''-t'\right]B\right)k''\right)}\right] \tag{A.112}$$

$$= K^2\sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\sum_{l',l''=0}^{L_w-1}w_{\mathsf{A}}\left(l'\right)w_{\mathsf{A}}\left(l''\right)w_{\mathsf{S}}\left(l'+p'\right)w_{\mathsf{S}}\left(l''+\left(t''-t'\right)B+p'\right)$$
$$\sum_{\vartheta,\vartheta'=-\infty}^{\infty}\delta\left(l'+p'-l-\vartheta K\right)\delta\left(l''+p'-l-\vartheta' K\right)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(l'-l''\right)k} \tag{A.113}$$

$$= K^2\sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\sum_{l'=0}^{L_w-1}w^2_{\mathsf{A}}\left(l'\right)w_{\mathsf{S}}\left(l'+p'\right)w_{\mathsf{S}}\left(l'+\left(t''-t'\right)B+p'\right)$$
$$\sum_{\vartheta=-\infty}^{\infty}\delta\left(l'+p'-l-\vartheta K\right) \tag{A.114}$$

$$= K^2\sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}\sum_{\vartheta=-\infty}^{\infty}w^2_{\mathsf{A}}\left(\vartheta K+l-p'\right)w_{\mathsf{S}}\left(\vartheta K+l\right)w_{\mathsf{S}}\left(\vartheta K+l+\left(t''-t'\right)B\right) \tag{A.115}$$

$$= K^2\sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')}\sigma^2_{\breve{h}}\mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}}w^2_{\mathsf{A}}\left(l-p'\right)w_{\mathsf{S}}\left(l\right)w_{\mathsf{S}}\left(l+\left(t''-t'\right)B\right). \tag{A.116}$$

Equality (A.114) thereby holds, since the two trains of Dirac-delta pulses only "overlap" if $l' = \mu K + l''$, where $\mu \in \mathbb{N}$. However, since $l', l'' \in \{0,\ldots,L_w-1\}$ and $L_w \leq K$, only $\mu = 0$ and thus $l' = l''$ contributes to the double summation – which in turn reduces to a single summation – and further renders the respective expression to be independent of the actual frequency index k. Moreover, since the synthesis window $w_{\mathsf{S}}(p)$ is only non-zero for $p \in \{0,\ldots,L_w-1\}$, only $\vartheta = 0$ contributes to (A.115), eventually resulting in (A.116).

The final expectation value of the numerator term (A.96) is thus given by

$$E\left[\breve{C}_N(k)\right] = \sum_{t',t''=-L_{H,\ell}}^{L_H}\sum_{l=0}^{L_w-1}w_{\mathsf{A}}\left(l\right)w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right)$$
$$\sum_{k',k''=0}^{K-1}E\left[\breve{h}_{t'}\left(k,k'\right)\breve{h}^{*}_{t''}\left(k,k''\right)\right]\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(lk'-\left(l+\left[t''-t'\right]B\right)k''\right)} \tag{A.117}$$

$$= \sigma^2_{\breve{h}}C_N. \tag{A.118}$$

with

$$\begin{aligned}C_N &:= K^2 \sum_{t',t''=-L_{H,\ell}}^{L_H} \sum_{l=0}^{L_w-1} w_{\mathsf{A}}(l)\, w_{\mathsf{S}}(l)\, w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right) w_{\mathsf{S}}\left(l+\left(t''-t'\right)B\right) \\ &\qquad \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w_{\mathsf{A}}^2\left(l-p'\right) \qquad \text{(A.119)} \\ &= K^2 \sum_{t',t''=-L_{H,\ell}}^{L_H} \sum_{l=l_{\mathsf{min}}(t''-t')}^{l_{\mathsf{max}}\left(t''-t'\right)} w_{\mathsf{A}}(l)\, w_{\mathsf{S}}(l)\, w_{\mathsf{A}}\left(l+\left(t''-t'\right)B\right) w_{\mathsf{S}}\left(l+\left(t''-t'\right)B\right) \\ &\qquad \sum_{p'=\max\{\mathcal{L}(t'),l-(L_w-1)\}}^{\min\left\{\mathcal{U}\left(t'\right),l\right\}} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w_{\mathsf{A}}^2\left(l-p'\right) \qquad \text{(A.120)}\end{aligned}$$

The last equality with

$$l_{\mathsf{min}}\left(t''-t'\right) := \max\left\{-\left(t''-t'\right)B, 0\right\} \qquad \text{(A.121)}$$

$$l_{\mathsf{max}}\left(t''-t'\right) := \min\left\{L_w-1-\left(t''-t'\right)B, L_w-1\right\} \qquad \text{(A.122)}$$

thereby employs the limited support of the analysis and synthesis windows $w_{\mathsf{A}}(l)$ and $w_{\mathsf{S}}(l)$, which are non-zero only for $0 \leq l \leq L_w - 1$.

Since both (A.107) and (A.117) are independent of the actual frequency index k, the final power compensation constant under the employed AIR model is also independent of k and is given by

$$\begin{aligned}C_P &:= \frac{E\left[\breve{C}_N(k)\right]}{E\left[\breve{C}_D(k)\right]} \qquad \text{(A.123)} \\ &= \frac{C_N}{C_D}. \qquad \text{(A.124)}\end{aligned}$$

Note that this frequency independent power compensation constant is also independent of the energy of the AIR.

A.7 Derivation of the AIR Representation in the Mel Power Spectral Domain

The AIR representation $\mathcal{H}_{t'}(q)$ in (4.81) is chosen such that the error $\epsilon_{t'}^{(\mathsf{m})}(q)$ in (4.82) is zero-mean. This, however, is equivalent to postulating

$$E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|\breve{X}_{t-t'}(k)\right|^2 C_P(k) \left|h_{t'}(k,k)\right|^2\right] \overset{!}{=} \mathcal{H}_{t'}(q)\, E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \left|\breve{X}_{t-t'}(k)\right|^2\right]. \qquad \text{(A.125)}$$

Assuming the speech signal to be a realization of a real-valued white GAUSSIAN random process of variance $\sigma_{\breve{x}}^2$ (compare (A.85)) the expectation on the right hand side of (A.125) is given by

$$
\begin{aligned}
&E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\left|\breve{X}_{t-t'}(k)\right|^2\right] \\
&= \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, E\left[\left|\breve{X}_{t-t'}(k)\right|^2\right] \qquad \text{(A.126)}\\
&= \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k) \sum_{l,l'=0}^{L_w-1} w_{\mathsf{A}}(l)\, w_{\mathsf{A}}\left(l'\right) E\left[x\left(l+\left(t-t'\right)B\right)x\left(l'+\left(t-t'\right)B\right)\right]\mathrm{e}^{-\mathrm{j}\frac{2\pi}{K}\left(l-l'\right)k} \qquad \text{(A.127)}\\
&= \sigma_{\breve{x}}^2\left(\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\right)\left(\sum_{l=0}^{L_w-1} w_{\mathsf{A}}^2(l)\right). \qquad \text{(A.128)}
\end{aligned}
$$

In an analogous manner, the left hand side of (A.125) may then be written as

$$
\begin{aligned}
&E\left[\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\left|\breve{X}_{t-t'}(k)\right|^2 C_P(k)\left|h_{t'}(k,k)\right|^2\right] \\
&= \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, E\left[\left|\breve{X}_{t-t'}(k)\right|^2\right] C_P(k)\left|h_{t'}(k,k)\right|^2 \qquad \text{(A.129)}\\
&= \sigma_{\breve{x}}^2\left(\sum_{l=0}^{L_w-1} w_{\mathsf{A}}^2(l)\right)\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, C_P(k)\left|h_{t'}(k,k)\right|^2 \qquad \text{(A.130)}
\end{aligned}
$$

resulting in $\mathcal{H}_{t'}(q)$ to be given by

$$
\mathcal{H}_{t'}(q) = \frac{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, C_P(k)\left|h_{t'}(k,k)\right|^2}{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)}. \qquad \text{(A.131)}
$$

Note, that if the AIR is not known, but rather modeled as a realization of a random process according to (4.147), this has first to be considered in the computation of the frequency dependent power compensation constant $C_P(k)$ – leading to the frequency independent power compensation constant C_P given in (A.123). The expectation in (A.129) then also has to be taken w.r.t. the band-to-band filters, i.e., $\left|h_{t'}(k,k)\right|^2$ has to be replaced by $E\left[\left|h_{t'}(k,k)\right|^2\right]$, which is given in (A.105).

A.7.1 Applying the model of the AIR

Under the AIR model (4.147), the expected value of the AIR representation $\mathcal{H}_{t'}(q)$ is, employing (A.105), eventually given by

$$E\left[\breve{\mathcal{H}}_{t'}(q)\right] := C_P \frac{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)\, E\left[\left|\breve{h}_{t'}(k,k)\right|^2\right]}{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \Lambda_q(k)} \tag{A.132}$$

$$= C_P \sigma_{\breve{h}}^2 \sum_{p'=\mathcal{L}(t')}^{\mathcal{U}(t')} \mathrm{e}^{-\frac{2\left(t'B-p'\right)}{\tau_h}} w^2\left(p'\right) \tag{A.133}$$

Thereby note has to be taken of the fact that the AIR has already been applied to calculate the (then frequency independent) power compensation constant. As a consequence, it may be considered a deterministic quantity when taking the expectation of $\breve{\mathcal{H}}_{t'}(q)$.

The computation of the covariances of the AIR representations with different (or the same) mel frequency bin indices, i.e.,

$$\begin{aligned} &E\left[\left(\breve{\mathcal{H}}_{t'}(q) - E\left[\breve{\mathcal{H}}_{t'}(q)\right]\right)\left(\breve{\mathcal{H}}_{t'}\left(q'\right) - E\left[\breve{\mathcal{H}}_{t'}\left(q'\right)\right]\right)\right] \\ &\quad = E\left[\breve{\mathcal{H}}_{t'}(q)\,\breve{\mathcal{H}}_{t'}\left(q'\right)\right] - E\left[\breve{\mathcal{H}}_{t'}(q)\right] E\left[\breve{\mathcal{H}}_{t'}\left(q'\right)\right] \end{aligned} \tag{A.134}$$

requires additional knowledge of $E\left[\breve{\mathcal{H}}_{t'}(q)\,\breve{\mathcal{H}}_{t'}(q')\right]$, which is given by

$$E\left[\breve{\mathcal{H}}_{t'}(q)\,\breve{\mathcal{H}}_{t'}\left(q'\right)\right] := C_P^2 \frac{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \sum\limits_{k'=K_{q'}^{(\mathsf{low})}}^{K_{q'}^{(\mathsf{up})}} \Lambda_q(k)\,\Lambda_q(k')\, E\left[\left|\breve{h}_{t'}(k,k)\right|^2 \left|\breve{h}_{t'}(k',k')\right|^2\right]}{\sum\limits_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \sum\limits_{k'=K_{q'}^{(\mathsf{low})}}^{K_{q'}^{(\mathsf{up})}} \Lambda_q(k)\,\Lambda_{q'}(k')}. \tag{A.135}$$

With property (4.152) and the definition of the band-to-band filters given in (4.51), the occurring expectation value may further be expressed by

$$
\begin{aligned}
&E\left[\left|\breve{h}_{t'}\left(k,k\right)\right|^2\left|\breve{h}_{t'}\left(k',k'\right)\right|^2\right]\\
&=\sum_{p,p',p'',p'''=0}^{L_h-1} E\left[\breve{h}\left(p\right)\breve{h}\left(p'\right)\breve{h}\left(p''\right)\breve{h}\left(p'''\right)\right]\varPhi_{t'B-p}\left(k,k\right)\varPhi^{*}_{t'B-p'}\left(k,k\right)\\
&\qquad\qquad\varPhi_{t'B-p''}\left(k',k'\right)\varPhi^{*}_{t'B-p'''}\left(k',k'\right) \qquad \text{(A.136)}\\
&=\sigma_{\breve{h}}^4\left(\sum_{p,p''=0}^{L_h-1}\mathrm{e}^{-\frac{2\left(p+p''\right)}{\tau_h}}\left|\varPhi_{t'B-p}\left(k,k\right)\right|^2\left|\varPhi_{t'B-p''}\left(k',k'\right)\right|^2\right.\\
&\quad+\sum_{p,p'=0}^{L_h-1}\mathrm{e}^{-\frac{2\left(p+p'\right)}{\tau_h}}\varPhi_{t'B-p}\left(k,k\right)\varPhi_{t'B-p}\left(k',k'\right)\varPhi^{*}_{t'B-p'}\left(k,k\right)\varPhi^{*}_{t'B-p'}\left(k',k'\right)\\
&\quad\left.+\sum_{p,p'=0}^{L_h-1}\mathrm{e}^{-\frac{2\left(p+p'\right)}{\tau_h}}\varPhi_{t'B-p}\left(k,k\right)\varPhi^{*}_{t'B-p}\left(k',k'\right)\varPhi_{t'B-p'}\left(k,k\right)\varPhi^{*}_{t'B-p'}\left(k',k'\right)\right). \qquad \text{(A.137)}
\end{aligned}
$$

Since

$$\varPhi_{p}\left(k,k\right)\varPhi_{p}\left(k',k'\right)=w^2\left(p\right)\mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}p\left(k+k'\right)}, \qquad \text{(A.138)}$$

$$\varPhi_{p}\left(k,k\right)\varPhi^{*}_{p}\left(k',k'\right)=w^2\left(p\right)\mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}p\left(k-k'\right)} \qquad \text{(A.139)}$$

with $w\left(\cdot\right)$ defined in (4.169), (A.137) may now be simplified to

$$
\begin{aligned}
&E\left[\left|\breve{h}_{t'}\left(k,k\right)\right|^2\left|\breve{h}_{t'}\left(k',k'\right)\right|^2\right]\\
&=\sigma_{\breve{h}}^4\left[\left|\mathcal{V}_{t'}\left(0\right)\right|^2+\mathcal{V}_{t'}\left(k+k'\right)\left(\mathcal{V}_{t'}\left(k+k'\right)\right)^{*}+\mathcal{V}_{t'}\left(k-k'\right)\left(\mathcal{V}_{t'}\left(k-k'\right)\right)^{*}\right] \qquad \text{(A.140)}
\end{aligned}
$$

where

$$\mathcal{V}_{t'}\left(k\right)=\sum_{p=\mathcal{L}\left(t'\right)}^{\mathcal{U}\left(t'\right)}\mathrm{e}^{-\frac{2\left(t'B-p\right)}{\tau_h}}w^2\left(p\right)\mathrm{e}^{+\mathrm{j}\frac{2\pi}{K}pk} \qquad \text{(A.141)}$$

with $\mathcal{L}\left(t'\right)$ and $\mathcal{U}\left(t'\right)$ defined in (4.55) and (4.56), respectively.

Noting that the first term in (A.140) (including $\sigma_{\breve{h}}^4$) is nothing but the product of $E\left[\left|\breve{h}_{t'}\left(k,k\right)\right|^2\right]$ and $E\left[\left|\breve{h}_{t'}\left(k',k'\right)\right|^2\right]$ eventually gives the final covariance to be

$$
\begin{aligned}
&E\left[\left(\breve{\mathcal{H}}_{t'}\left(q\right)-E\left[\breve{\mathcal{H}}_{t'}\left(q\right)\right]\right)\left(\breve{\mathcal{H}}_{t'}\left(q'\right)-E\left[\breve{\mathcal{H}}_{t'}\left(q'\right)\right]\right)\right]\\
&=\sigma_{\breve{h}}^4C_P^2\frac{\sum\limits_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}}\sum\limits_{k'=K_{q'}^{(\mathrm{low})}}^{K_{q'}^{(\mathrm{up})}}\varLambda_q\left(k\right)\varLambda_{q'}\left(k'\right)\left[\left|\mathcal{V}_{t'}\left(k+k'\right)\right|^2+\left|\mathcal{V}_{t'}\left(k-k'\right)\right|^2\right]}{\sum\limits_{k=K_q^{(\mathrm{low})}}^{K_q^{(\mathrm{up})}}\sum\limits_{k'=K_{q'}^{(\mathrm{low})}}^{K_{q'}^{(\mathrm{up})}}\varLambda_q\left(k\right)\varLambda_{q'}\left(k'\right)}. \qquad \text{(A.142)}
\end{aligned}
$$

A.8 MMSE Estimate of the MPSC Feature Vector of Reverberant Speech

The MMSE estimate of the MPSC feature vector $\mathbf{s}^{\mathsf{(m)}}_{t-L_R}$ of the reverberant speech signal at time instant t is defined as

$$
\begin{aligned}
&E\left[\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}\middle|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1}\right]\\
&\quad=\int\limits_{\mathbb{R}^D_{\geq 0}}\mathbf{s}^{\mathsf{(m)}}_{t-L_R}p_{\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1}}\left(\mathbf{s}^{\mathsf{(m)}}_{t-L_R}\middle|\mathbf{x}^{\mathsf{(m)}}_{t-L_R+1:t},\mathbf{n}^{\mathsf{(m)}}_{t-L_R,t},\mathbf{o}^{\mathsf{(m)}}_{1:t-1}\right)\mathrm{d}\mathbf{s}^{\mathsf{(m)}}_{t-L_R}.
\end{aligned}
\tag{A.143}
$$

The PDF $p_{\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1}}$ may, employing the short-hand notation

$$\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t}:=\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t} \tag{A.144}$$

$$\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}:=\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1-L_R-1},\breve{\mathbf{o}}^{\mathsf{(m)}}_{t-L_R+1:t-1}, \tag{A.145}$$

now be written as

$$
\begin{aligned}
&p_{\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1}}\left(\mathbf{s}^{\mathsf{(m)}}_{t-L_R}\middle|\mathbf{x}^{\mathsf{(m)}}_{t-L_R+1:t},\mathbf{n}^{\mathsf{(m)}}_{t-L_R,t},\mathbf{o}^{\mathsf{(m)}}_{1:t-1}\right)\\
&\quad\propto\int\limits_{[-1,+1]^Q}p_{\breve{\mathbf{o}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R},\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R},\breve{\boldsymbol{\alpha}}_{t-L_R}}\left(\mathbf{o}^{\mathsf{(m)}}_{t-L_R}\middle|\mathbf{n}^{\mathsf{(m)}}_{t-L_R},\mathbf{s}^{\mathsf{(m)}}_{t-L_R},\boldsymbol{\alpha}_{t-L_R}\right)p_{\breve{\boldsymbol{\alpha}}_{t-L_R}}\left(\boldsymbol{\alpha}_{t-L_R}\right)\mathrm{d}\boldsymbol{\alpha}_{t-L_R}\\
&\qquad p_{\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}}\left(\mathbf{s}^{\mathsf{(m)}}_{t-L_R}\middle|\mathbf{x}^{\mathsf{(m)}}_{t-L_R+1:t},\mathbf{n}^{\mathsf{(m)}}_{t-L_R,t},\mathbf{o}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}\right).
\end{aligned}
\tag{A.146}
$$

Thereby, the vector of phase factors $\breve{\boldsymbol{\alpha}}_{t-L_R}$ has been assumed to be independent of $\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1,t}$, $\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}$, $\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t}$ and $\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}$. From (4.100), $p_{\breve{\mathbf{o}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R},\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R},\breve{\boldsymbol{\alpha}}_{t-L_R}}$ may be found to be a Dirac-delta distribution, i.e.,

$$p_{\breve{\mathbf{o}}^{\mathsf{(m)}}_{t-L_R}|\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R},\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R},\breve{\boldsymbol{\alpha}}_{t-L_R}}\left(\mathbf{o}^{\mathsf{(m)}}_{t-L_R}\middle|\mathbf{n}^{\mathsf{(m)}}_{t-L_R},\mathbf{s}^{\mathsf{(m)}}_{t-L_R},\boldsymbol{\alpha}_{t-L_R}\right) \tag{A.147}$$

$$=\delta\left(\mathbf{o}^{\mathsf{(m)}}_{t-L_R}-\left(\mathbf{s}^{\mathsf{(m)}}_{t-L_R}+\mathbf{n}^{\mathsf{(m)}}_{t-L_R}+2\boldsymbol{\alpha}_{t-L_R}\sqrt{\mathbf{n}^{\mathsf{(m)}}_{t-L_R}}\sqrt{\mathbf{s}^{\mathsf{(m)}}_{t-L_R}}\right)\right). \tag{A.148}$$

Once $\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R}$, $\breve{\mathbf{s}}^{\mathsf{(m)}}_{t-L_R}$, and $\breve{\boldsymbol{\alpha}}_{t-L_R}$ are given, the components of $\breve{\mathbf{o}}^{\mathsf{(m)}}_{t-L_R}$ are statistically independent and the q-th component of the MMSE estimate is given by

$$
\begin{aligned}
&E\left[s^{\mathsf{(m)}}_{t-L_R}(q)\middle|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1}\right]\\
&\propto\int\limits_{-1}^{+1}\Bigg[\int\limits_{\mathbb{R}_{\geq 0}}\delta\left(o^{\mathsf{(m)}}_{t-L_R}(q)-\left(s^{\mathsf{(m)}}_{t-L_R}(q)+n^{\mathsf{(m)}}_{t-L_R}(q)+2\alpha_{t-L_R}(q)\sqrt{n^{\mathsf{(m)}}_{t-L_R}(q)}\sqrt{s^{\mathsf{(m)}}_{t-L_R}(q)}\right)\right)\\
&\qquad p_{\breve{s}^{\mathsf{(m)}}_{t-L_R}(q)|\breve{\mathbf{x}}^{\mathsf{(m)}}_{t-L_R+1:t},\breve{\mathbf{n}}^{\mathsf{(m)}}_{t-L_R,t},\breve{\mathbf{o}}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}}\left(s^{\mathsf{(m)}}_{t-L_R}(q)\middle|\mathbf{x}^{\mathsf{(m)}}_{t-L_R+1:t},\mathbf{n}^{\mathsf{(m)}}_{t-L_R,t},\mathbf{o}^{\mathsf{(m)}}_{1:t-1\backslash t-L_R}\right)\\
&\qquad s^{\mathsf{(m)}}_{t-L_R}(q)\,\mathrm{d}s^{\mathsf{(m)}}_{t-L_R}(q)\Bigg]p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q).
\end{aligned}
\tag{A.149}
$$

The constant of proportionality is simply given by

$$\int_{-1}^{+1} \Bigg[\int_{\mathbb{R}_{\geq 0}} \delta\left(o^{(\mathsf{m})}_{t-L_R}(q) - \left(s^{(\mathsf{m})}_{t-L_R}(q) + n^{(\mathsf{m})}_{t-L_R}(q) + 2\alpha_{t-L_R}(q)\sqrt{n^{(\mathsf{m})}_{t-L_R}(q)}\sqrt{s^{(\mathsf{m})}_{t-L_R}(q)}\right)\right)$$
$$p_{\breve{s}^{(\mathsf{m})}_{t-L_R}(q)|\breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1\backslash t-L_R}}\left(s^{(\mathsf{m})}_{t-L_R}(q)\,\middle|\,\mathbf{x}^{(\mathsf{m})}_{t-L_R+1:t},\mathbf{n}^{(\mathsf{m})}_{t-L_R,t},\mathbf{o}^{(\mathsf{m})}_{1:t-1\backslash t-L_R}\right)$$
$$\mathrm{d}s^{(\mathsf{m})}_{t-L_R}(q)\Bigg]\, p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q) \tag{A.150}$$

The argument of the DIRAC-delta distribution may now be considered a polynomial in $\sqrt{s^{(\mathsf{m})}_{t-L_R}(q)}$ with its roots given by

$$\sqrt{s^{(\mathsf{m}),\pm}_{t-L_R}(q)} = -\alpha_{t-L_R}(q)\sqrt{n^{(\mathsf{m})}_{t-L_R}(q)} \pm \sqrt{\alpha^2_{t-L_R}(q)\,n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)}. \tag{A.151}$$

Since $\breve{s}^{(\mathsf{m})}_{t-L_R}(q)$ is a non-negative, real-valued random variable (and thus its square root), the inner integral is only non-zero if (A.151) is real-valued and non-negative. The MMSE estimate is thus characterized by those phase factors $\alpha_{t-L_R}(q)$ that ensure (A.151) to be real-valued and non-negative.

The requirement for a real-valued solution is met if

$$\alpha^2_{t-L_R}(q) \geq 1 - \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}. \tag{A.152}$$

Hence, the complex-valued roots may only exist for certain $o^{(\mathsf{m})}_{t-L_R}(q) < n^{(\mathsf{m})}_{t-L_R}(q)$. For the non-negativeness, the two solutions given in (A.151) are looked at separately. Thereby, it is assumed that (A.152) holds.

Calling for the $\sqrt{s^{(\mathsf{m}),+}_{t-L_R}(q)}$ solution to be non-negative leads to

$$\sqrt{\alpha^2_{t-L_R}(q)\,n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)} \overset{!}{\geq} \alpha_{t-L_R}(q)\sqrt{n^{(\mathsf{m})}_{t-L_R}(q)}. \tag{A.153}$$

For negative phase factors $\alpha_{t-L_R}(q)$, (A.153) is always true (recall: (A.152) is assumed to hold). For non-negative phase factors $\alpha_{t-L_R}(q)$, (A.153) is true only if $o^{(\mathsf{m})}_{t-L_R}(q) \geq n^{(\mathsf{m})}_{t-L_R}(q)$.

Calling for the $\sqrt{s^{(\mathsf{m}),-}_{t}(q)}$ solution to be non-negative leads to

$$-\sqrt{\alpha^2_{t-L_R}(q)\,n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)} \overset{!}{\geq} \alpha_{t-L_R}(q)\sqrt{n^{(\mathsf{m})}_{t-L_R}(q)}. \tag{A.154}$$

For positive phase factors $\alpha_{t-L_R}(q)$, (A.154) is never true (recall: (A.152) is assumed to hold). For non-positive phase factors $\alpha_{t-L_R}(q)$, (A.154) is true only if $o^{(\mathsf{m})}_{t-L_R}(q) \leq n^{(\mathsf{m})}_{t-L_R}(q)$.

Hence, two cases are of special interest.

Case $\mathbf{o}^{(\mathsf{m})}_{\mathbf{t-L_R}}(\mathbf{q}) \geq \mathbf{n}^{(\mathsf{m})}_{\mathbf{t-L_R}}(\mathbf{q})$: For $o^{(\mathsf{m})}_{t-L_R}(q) \geq n^{(\mathsf{m})}_{t}(q)$, (A.152) holds for an arbitrary phase factor in the range $-1 \leq \alpha_{t-L_R}(q) \leq 1$. However, only $\sqrt{s^{(\mathsf{m}),+}_{t-L_R}(q)}$ is a valid solution and as such contributes to the integrals (A.149) and (A.150), where the latter simply reduces to 1. The MMSE estimate is thus given by

$$E\left[s^{(\mathsf{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1}; \breve{o}^{(\mathsf{m})}_{t-L_R}(q) \geq \breve{n}^{(\mathsf{m})}_{t-L_R}(q) \right]$$

$$= \int_{-1}^{+1} s^{(\mathsf{m}),+}_{t-L_R}(q)\, p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q) \tag{A.155}$$

$$\begin{aligned} &= \int_{-1}^{+1} \alpha^2_{t-L_R}(q)\, n^{(\mathsf{m})}_{t-L_R}(q)\, p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q) \\ &\quad - 2\int_{-1}^{+1} \alpha_{t-L_R}(q) \sqrt{n^{(\mathsf{m})}_{t-L_R}(q)} \sqrt{\alpha^2_{t-L_R}(q)\, n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)} \\ &\qquad\quad p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q) \\ &\quad + \int_{-1}^{+1} \left(\alpha^2_{t-L_R}(q)\, n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)\right) p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q). \end{aligned} \tag{A.156}$$

Since the phase factor distribution is an even function, the second integral in (A.156) is zero and the MMSE estimate turns into

$$\begin{aligned} &E\left[s^{(\mathsf{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1}; \breve{o}^{(\mathsf{m})}_{t-L_R}(q) \geq \breve{n}^{(\mathsf{m})}_{t-L_R}(q) \right] \\ &\quad = \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q) \end{aligned} \tag{A.157}$$

Thereby $\sigma^2_{\breve{\alpha}_q}$ denotes the variance of the phase factor in mel frequency bin q. Since $o^{(\mathsf{m})}_{t-L_R}(q) \geq n^{(\mathsf{m})}_{t-L_R}(q)$, the MMSE estimate is always positive.

Case $\mathbf{o}^{(\mathsf{m})}_{\mathbf{t-L_R}}(\mathbf{q}) < \mathbf{n}^{(\mathsf{m})}_{\mathbf{t-L_R}}(\mathbf{q})$: For $o^{(\mathsf{m})}_{t-L_R}(q) < n^{(\mathsf{m})}_{t-L_R}(q)$, (A.152) poses the first constraint on possible values for the phase factor $\alpha_{t-L_R}(q)$, i.e.,

$$\left|\alpha_{t-L_R}(q)\right| \geq \sqrt{1 - \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}. \tag{A.158}$$

However, for $o^{(\mathsf{m})}_{t-L_R}(q) < n^{(\mathsf{m})}_{t-L_R}(q)$ solution $s^{(\mathsf{m}),+}_{t-L_R}(q)$ is only valid for $\alpha_{t-L_R}(q) < 0$ and solution $s^{(\mathsf{m}),-}_{t-L_R}(q)$ is only valid for $\alpha_{t-L_R}(q) \leq 0$. Hence, only those phase factors $\alpha_{t-L_R}(q)$ in the range $-1 \leq \alpha_{t-L_R}(q) \leq -\sqrt{1 - \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}$ have to be considered for the integration. Introducing the short-hand notation

$$\beta_{t-L_R}(q) := \sqrt{1 - \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}} \tag{A.159}$$

thus allows the the MMSE estimate to be written as

$$
\begin{aligned}
&E\left[s^{(\mathsf{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1}; \breve{o}^{(\mathsf{m})}_{t-L_R}(q) < \breve{n}^{(\mathsf{m})}_{t-L_R}(q) \right] \\
&= \frac{\int\limits_{-1}^{-\beta_{t-L_R}(q)} \left[s^{(\mathsf{m}),+}_{t-L_R}(q)\, P^{+}_{t-L_R}(q) + s^{(\mathsf{m}),-}_{t-L_R}(q)\, P^{-}_{t-L_R}(q) \right] p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)}{\left(P^{+}_{t-L_R}(q) + P^{-}_{t-L_R}(q) \right) \int\limits_{-1}^{-\beta_{t-L_R}(q)} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)},
\end{aligned}
\tag{A.160}
$$

where $P^{+}_{t-L_R}(q)$ an $P^{-}_{t-L_R}(q)$ are the likelihoods of the solution $s^{(\mathsf{m}),+}_{t-L_R}(q)$ and the solution $s^{(\mathsf{m}),-}_{t-L_R}(q)$ under the conditional PDF $p_{\breve{s}^{(\mathsf{m})}_{t-L_R}(q) | \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1 \backslash t-L_R}}$. By plugging the solutions $s^{(\mathsf{m}),\pm}_{t-L_R}(q)$ into (A.160), the MMSE estimate may be rewritten as

$$
\begin{aligned}
&E\left[s^{(\mathsf{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1}; \breve{o}^{(\mathsf{m})}_{t-L_R}(q) < \breve{n}^{(\mathsf{m})}_{t-L_R}(q) \right] \\
&= \frac{\int\limits_{-1}^{-\beta_{t-L_R}(q)} \left[\left(2\alpha^{2}_{t-L_R}(q) - 1 \right) n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q) \right] p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)}{\int\limits_{-1}^{-\beta_{t-L_R}(q)} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)} \\
&\quad + \int\limits_{-1}^{-\beta_{t-L_R}(q)} 2\alpha_{t-L_R}(q) \sqrt{n^{(\mathsf{m})}_{t-L_R}(q)} \sqrt{\alpha^{2}_{t-L_R}(q)\, n^{(\mathsf{m})}_{t-L_R}(q) - n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q)} \\
&\qquad\qquad p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q) \\
&\quad \cdot \frac{P^{-}_{t-L_R}(q) - P^{+}_{t-L_R}(q)}{\left[P^{+}_{t-L_R}(q) + P^{-}_{t-L_R}(q) \right] \int\limits_{-1}^{-\beta_{t-L_R}(q)} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)}.
\end{aligned}
\tag{A.161}
$$

Assuming the likelihoods of the two solutions to be approximately equal, i.e., $P^{+}_{t-L_R}(q) \approx P^{-}_{t-L_R}(q)$, finally yields

$$
\begin{aligned}
&E\left[s^{(\mathsf{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1}; \breve{o}^{(\mathsf{m})}_{t-L_R}(q) < \breve{n}^{(\mathsf{m})}_{t-L_R}(q) \right] \\
&\approx \left(2 \frac{\int\limits_{-1}^{-\beta_{t-L_R}(q)} \alpha^{2}_{t-L_R}(q)\, p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)}{\int\limits_{-1}^{-\beta_{t-L_R}(q)} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q) \right) \mathrm{d}\alpha_{t-L_R}(q)} - 1 \right) n^{(\mathsf{m})}_{t-L_R}(q) + o^{(\mathsf{m})}_{t-L_R}(q).
\end{aligned}
\tag{A.162}
$$

Note that approximation (A.162) is exact for $o^{(\mathsf{m})}_{t-L_R}(q) = 0$. In this case, only the phase factor $\alpha_{t-L_R}(q) = -1$ can ensure (A.151) to be real-valued and non-negative. Hence, $s^{(\mathsf{m}),+}_{t-L_R}(q) = s^{(\mathsf{m}),-}_{t-L_R}(q) = n^{(\mathsf{m})}_{t-L_R}(q)$ and $P^{-}_{t-L_R}(q) = P^{+}_{t-L_R}(q)$.

Since the PDF of the phase factor is an even function, (A.162) may also be rewritten as

$$E\left[s^{(\mathsf{m})}_{t-L_R}(q)\middle|\breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1};\breve{o}^{(\mathsf{m})}_{t-L_R}(q)<\breve{n}^{(\mathsf{m})}_{t-L_R}(q)\right]$$
$$\approx\left(2\frac{\sigma^2_{\breve{\alpha}_q}-\int\limits_{-\beta_{t-L_R}(q)}^{\beta_{t-L_R}(q)}\alpha^2_{t-L_R}(q)\,p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q)}{1-\int\limits_{-\beta_{t-L_R}(q)}^{\beta_{t-L_R}(q)}p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q)}-1\right)n^{(\mathsf{m})}_{t-L_R}(q)$$
$$+\,o^{(\mathsf{m})}_{t-L_R}(q)\,. \tag{A.163}$$

Unfortunately, neither the integrals in (A.162) nor those in (A.163) have known closed-form solutions. Employing the parametric approximation to the PDF of the phase factor derived in Sec. 4.6.3, these integrals may be solved numerically, however, here, the parametric approximation to the PDF will only be used to support the derivation of the following lower and upper bounds of the ratio of integrals

$$I\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right):=\frac{\int\limits_{-1}^{-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}}\alpha^2_{t-L_R}(q)\,p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q)}{\int\limits_{-1}^{-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}}p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q)} \tag{A.164}$$

occurring in (A.162). In particular, it holds that

$$\sigma^2_{\breve{\alpha}_q}\le I\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)\le\left(\sigma^2_{\breve{\alpha}_q}-1\right)\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}+1\le 1. \tag{A.165}$$

The lower bound $\sigma^2_{\breve{\alpha}_q}$ and the upper bound 1 arise as a consequence of (A.164) being a strictly monotonically decreasing function in $0\le\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\le 1$ (see Sec. A.8.1 for a proof) and the limiting cases

$$\lim_{\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\to 0} I\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)=1 \tag{A.166}$$

$$\lim_{\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\to 1} I\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)=\sigma^2_{\breve{\alpha}_q}. \tag{A.167}$$

The tighter upper bound

$$I_{\mathsf{U}}\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right):=\left(\sigma^2_{\breve{\alpha}_q}-1\right)\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}+1 \tag{A.168}$$

is obtained from the limiting cases by a linear interpolation, i.e., it is assumed that (A.164) is a convex function. Instead of a formal mathematical proof thereof (which additionally requires to prove that the second derivative of (A.164) is always positive), the convexity of (A.164) is inferred from Fig. A.1, where (A.164) and its bounds are displayed as functions of $\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}$ for different mel frequency indices. While (A.164) is clearly upper bounded by 1 and lower bounded by $\sigma^2_{\breve{\alpha}_q}$ for all (displayed) mel frequency indices, (A.164) can also be seen to be a strictly monotonically decreasing and convex function, eventually showing $I_{\mathsf{U}}(\cdot)$ to be a tighter upper bound than the bound given in (A.166).

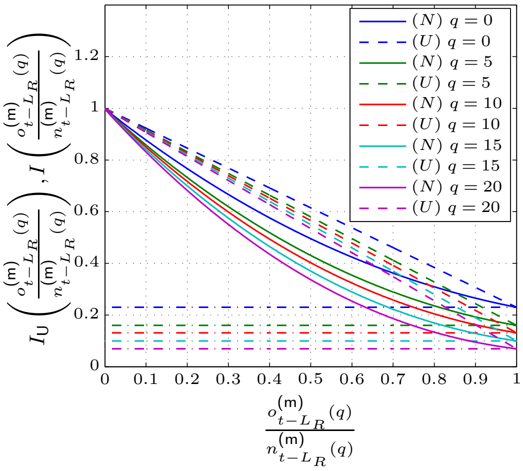


Figure A.1: *(N)umerical approximation to $I(\cdot)$ defined in* (A.164) *(solid lines) and the (U)pper bound $I_U(\cdot)$ defined in* (A.168) *(dashed lines) for $0 \leq \frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)} \leq 1$ and $q \in \{0, 5, 10, 15, 20\}$. The horizontal (dash-dotted) lines show the variance $\sigma^2_{\breve{\alpha}_q}$ of the phase factor at the respective mel frequency indices.*

With the above considerations, the final MMSE estimate in the case where $o^{(\mathsf{m})}_{t-L_R}(q) < n^{(\mathsf{m})}_{t-L_R}(q)$ is thus upper bounded by

$$\begin{aligned}
&E\left[s^{(\mathsf{m})}_{t-L_R}(q)\middle|\breve{\mathbf{x}}^{(\mathsf{m})}_{t-L_R+1:t},\breve{\mathbf{n}}^{(\mathsf{m})}_{t-L_R,t},\breve{\mathbf{o}}^{(\mathsf{m})}_{1:t-1};\breve{o}^{(\mathsf{m})}_{t-L_R}(q)<\breve{n}^{(\mathsf{m})}_{t-L_R}(q)\right]\\
&\leq\left(2\left[\left\{\sigma^2_{\breve{\alpha}_q}-1\right\}\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}+1\right]-1\right)n^{(\mathsf{m})}_{t-L_R}(q)+o^{(\mathsf{m})}_{t-L_R}(q) \qquad \text{(A.169)}\\
&=\left(2\sigma^2_{\breve{\alpha}_q}-1\right)o^{(\mathsf{m})}_{t-L_R}(q)+n^{(\mathsf{m})}_{t-L_R}(q). \qquad \text{(A.170)}
\end{aligned}$$

Since (A.170) also converges to (A.157) as $\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)} \to 1$, in summary, the following

MMSE estimate of the MPSC of the reverberant speech signal will be used in this work:

$$
\begin{aligned}
\hat{s}^{(\mathrm{m,R})}_{t-L_R}(q) &:= E\left[s^{(\mathrm{m})}_{t-L_R}(q) \middle| \breve{\mathbf{x}}^{(\mathrm{m})}_{t-L_R+1:t}, \breve{\mathbf{n}}^{(\mathrm{m})}_{t-L_R,t}, \breve{\mathbf{o}}^{(\mathrm{m})}_{1:t-1} \right] && \text{(A.171)} \\
&\approx \begin{cases} \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) n^{(\mathrm{m})}_{t-L_R}(q) + o^{(\mathrm{m})}_{t}(q), & \text{if } o^{(\mathrm{m})}_{t-L_R}(q) \geq n^{(\mathrm{m})}_{t-L_R}(q) \\ \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) o^{(\mathrm{m})}_{t-L_R}(q) + n^{(\mathrm{m})}_{t-L_R}(q), & \text{if } o^{(\mathrm{m})}_{t-L_R}(q) < n^{(\mathrm{m})}_{t-L_R}(q) \end{cases} && \text{(A.172)} \\
&= \max\left\{ \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) n^{(\mathrm{m})}_{t-L_R}(q) + o^{(\mathrm{m})}_{t-L_R}(q),\ \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) o^{(\mathrm{m})}_{t-L_R}(q) + n^{(\mathrm{m})}_{t-L_R}(q) \right\} && \text{(A.173)} \\
&= \left(2\sigma^2_{\breve{\alpha}_q} - 1\right) \min\left\{ n^{(\mathrm{m})}_{t-L_R}(q), o^{(\mathrm{m})}_{t-L_R}(q) \right\} + \max\left\{ n^{(\mathrm{m})}_{t-L_R}(q), o^{(\mathrm{m})}_{t-L_R}(q) \right\}. && \text{(A.174)}
\end{aligned}
$$

A.8.1 Proof of Strict Monotonicity of (A.164)

Lemma 1. *Defining (repeated here from* (A.164) *for convenience)*

$$
I\left(\frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)}\right) := \frac{\displaystyle\int_{-1}^{-\sqrt{1-\frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)}}} \alpha^2_{t-L_R}(q)\, p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q)}{\displaystyle\int_{-1}^{-\sqrt{1-\frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)}}} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right) \mathrm{d}\alpha_{t-L_R}(q)} \qquad \text{(A.175)}
$$

it holds that

$$
I\left(\frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)}\right) > I\left(\frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)} + \epsilon\right), \ \textit{for}\ 0 < \epsilon < 1 - \frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)}, \qquad \text{(A.176)}
$$

i.e., that $I(\cdot)$ *is a strictly monotonically decreasing function on the interval* $0 \leq \frac{o^{(m)}_{t-L_R}(q)}{n^{(m)}_{t-L_R}(q)} \leq 1$.

Proof of Lemma 1.
The derivative of (A.175) is given by

$$
\frac{\mathrm{d}I\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)} = \frac{\frac{\mathrm{d}I_N\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)} I_D\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right) - \frac{\mathrm{d}I_D\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)} I_N\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)}{I_D^2\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)}, \qquad \text{(A.177)}
$$

where $I_N\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)$ and $I_D\left(\frac{o^{(\mathrm{m})}_{t-L_R}(q)}{n^{(\mathrm{m})}_{t-L_R}(q)}\right)$ denote the (N)umerator and the (D)enominator of (A.175), respectively.

The derivative of either of the two functions may now be found by applying LEIBNIZ' integral rule [113, p. 11, Eq. (3.3.7)]

$$\frac{\mathrm{d}}{\mathrm{d}c}\left(\int\limits_{a(c)}^{b(c)} f(x,c)\mathrm{d}x\right) = \int\limits_{a(c)}^{b(c)} \frac{\mathrm{d}f(x,c)}{\mathrm{d}c}\mathrm{d}x + f(b(c),c)\frac{\mathrm{d}b(c)}{\mathrm{d}c} - f(a(c),c)\frac{\mathrm{d}a(c)}{\mathrm{d}c} \tag{A.178}$$

and noting that only the upper limit of the involved integral is a function of the variable to be differentiated w.r.t.. In particular, the two derivatives are given by

$$\frac{\mathrm{d}I_N\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)} = \frac{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}{2\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}} p_{\breve{\alpha}_{t-L_R}(q)}\left(-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}\right), \tag{A.179}$$

$$\frac{\mathrm{d}I_D\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)} = \frac{1}{2\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}} p_{\breve{\alpha}_{t-L_R}(q)}\left(-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}\right). \tag{A.180}$$

The derivative of (A.175) w.r.t. $\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}$ is thus given by

$$\begin{aligned}&\frac{\mathrm{d}I\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)}{\mathrm{d}\left(\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)}\\ &= \frac{1}{2}\frac{p_{\breve{\alpha}_{t-L_R}(q)}\left(-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}\right)}{\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}\left(\int\limits_{-1}^{-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}} p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q)\right)^2}\\ &\quad \int\limits_{-1}^{-\sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}} \left[\left(1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right) - \alpha^2_{t-L_R}(q)\right] p_{\breve{\alpha}_{t-L_R}(q)}\left(\alpha_{t-L_R}(q)\right)\mathrm{d}\alpha_{t-L_R}(q).\end{aligned} \tag{A.181}$$

The quotient term is always positive since $0 < \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)} < 1$. Since further $\alpha^2_{t-L_R}(q) \geq$

$\left(1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}\right)$ for $-1 \le \alpha_{t-L_R}(q) \le \sqrt{1-\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)}}$, the derivative is negative for $0 \le \frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)} < 1$ and 0 only for $\frac{o^{(\mathsf{m})}_{t-L_R}(q)}{n^{(\mathsf{m})}_{t-L_R}(q)} = 1$. □

A.9 Moments of the Phase Factor

With the characteristic function $\Phi_{\breve{\alpha}_t(q)}(\tau)$ of the phase factor RV $\breve{\alpha}_t(q)$ given by

$$\Phi_{\breve{\alpha}_t(q)}(\tau) := \prod_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \mathrm{J}_0(c_q(k)\tau), \tag{A.182}$$

where $c_q(k)$ are defined by (4.225), the *raw* moments of the phase factor RV $\breve{\alpha}_t(q)$ are given by

$$E[\breve{\alpha}_t^n(q)] = (-\mathrm{j})^n \left.\frac{\mathrm{d}^n \Phi_{\alpha_t(q)}(\tau)}{\mathrm{d}\tau^n}\right|_{\tau=0}, \quad \forall n \in \mathbb{N}. \tag{A.183}$$

Hence, the n-th derivative of (A.182) w.r.t. τ evaluated at $\tau = 0$ is required.

A differentiation of (A.182) by re-factoring the occurring product and recursively employing LEIBNIZ' theorem [113, p. 12, Eq. (3.3.8)], i.e.,

$$f(\tau) = u(\tau)v(\tau) \Rightarrow \frac{\mathrm{d}^n f(\tau)}{\mathrm{d}\tau^n} = \sum_{k=0}^{n} \binom{n}{k} \frac{\mathrm{d}^k u(\tau)}{\mathrm{d}\tau^k} \frac{\mathrm{d}^{n-k} v(\tau)}{\mathrm{d}\tau^{n-k}}, \tag{A.184}$$

will, however, result in rather bulky expressions. A more elegant solution may be obtained, if the natural logarithm of (A.182) is taken first, resulting in the *second characteristic function* [85, p. 153, Eq. (5-97)] $\Upsilon_{\breve{\alpha}_t(q)}(\tau)$ given by

$$\begin{aligned}\Upsilon_{\breve{\alpha}_t(q)}(\tau) &:= \ln\left(\Phi_{\breve{\alpha}_t(q)}(\tau)\right) && \text{(A.185)}\\ &= \sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \ln(\mathrm{J}_0(c_q(k)\tau)) && \text{(A.186)}\end{aligned}$$

with

$$\frac{\mathrm{d}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau} = \frac{\frac{\mathrm{d}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau}}{\Phi_{\breve{\alpha}_t(q)}(\tau)}. \tag{A.187}$$

Solving (A.184) for $\frac{\mathrm{d}^n u(\tau)}{\mathrm{d}x^n}$ yields

$$\frac{\mathrm{d}^n u(\tau)}{\mathrm{d}\tau^n} = \frac{\frac{\mathrm{d}^n f(\tau)}{\mathrm{d}\tau^n} - \sum_{k=0}^{n-1} \binom{n}{k} \frac{\mathrm{d}^k u(\tau)}{\mathrm{d}\tau^k} \frac{\mathrm{d}^{n-k} v(\tau)}{\mathrm{d}\tau^{n-k}}}{v(\tau)} \tag{A.188}$$

which, by further setting

$$u(\tau) := \frac{\mathrm{d}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau}, \tag{A.189}$$

$$v(\tau) := \Phi_{\breve{\alpha}_t(q)}(\tau), \tag{A.190}$$

$$f(\tau) := \frac{\mathrm{d}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau}, \tag{A.191}$$

first gives

$$\frac{\mathrm{d}^{n+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n+1}} = \frac{\frac{\mathrm{d}^{n+1}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n+1}} - \sum\limits_{k=0}^{n-1}\binom{n}{k}\frac{\mathrm{d}^{k+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{k+1}}\frac{\mathrm{d}^{n-k}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n-k}}}{\Phi_{\breve{\alpha}_t(q)}(\tau)} \tag{A.192}$$

and eventually results in

$$\frac{\mathrm{d}^{n+1}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n+1}} = \Phi_{\breve{\alpha}_t(q)}(\tau)\frac{\mathrm{d}^{n+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n+1}} + \sum_{k=0}^{n-1}\binom{n}{k}\frac{\mathrm{d}^{k+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{k+1}}\frac{\mathrm{d}^{n-k}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{n-k}}. \tag{A.193}$$

In Appendix A.9.1 it will be proven that

$$\left.\frac{\mathrm{d}^{2m+1}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m+1}}\right|_{\tau=0} = 0, \quad \forall m \in \mathbb{N}. \tag{A.194}$$

Hence, it immediately follows that all raw moments of *uneven* order are zero, i.e.,

$$E\left[\breve{\alpha}_t^{2m+1}(q)\right] = (-\mathrm{j})^{2m+1}\left.\frac{\mathrm{d}^{2m+1}\Phi_{\alpha_t(q)}(\tau)}{\mathrm{d}\tau^{2m+1}}\right|_{\tau=0} = 0, \quad \forall m \in \mathbb{N}. \tag{A.195}$$

For all raw moments of *even* order, (A.193) has to be evaluated at $\tau = 0$ only for $n = 2m-1$, $m \in \mathbb{N}_{>0}$, resulting in

$$\begin{aligned}\left.\frac{\mathrm{d}^{2m}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} &= \left.\Phi_{\breve{\alpha}_t(q)}(\tau)\frac{\mathrm{d}^{2m}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} \\ &\quad + \left.\sum_{k=0}^{2m-2}\binom{2m-1}{k}\frac{\mathrm{d}^{k+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{k+1}}\frac{\mathrm{d}^{2m-1-k}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m-1-k}}\right|_{\tau=0} \qquad \text{(A.196)} \\ &= \Phi_{\breve{\alpha}_t(q)}(\tau)\Big|_{\tau=0}\left.\frac{\mathrm{d}^{2m}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} \\ &\quad + \sum_{k=0}^{2m-2}\binom{2m-1}{k}\left.\frac{\mathrm{d}^{k+1}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{k+1}}\right|_{\tau=0}\left.\frac{\mathrm{d}^{2m-1-k}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m-1-k}}\right|_{\tau=0}. \qquad \text{(A.197)}\end{aligned}$$

With the same reasoning only those k for which $k = 2l-1$, $l \in \{1, \ldots, m-1\}$, holds have to be considered for the summation, leading to

$$\begin{aligned}\left.\frac{\mathrm{d}^{2m}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} &= \Phi_{\breve{\alpha}_t(q)}(\tau)\Big|_{\tau=0}\left.\frac{\mathrm{d}^{2m}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} \\ &\quad + \sum_{l=1}^{m-1}\binom{2m-1}{2l-1}\left.\frac{\mathrm{d}^{2l}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0}\left.\frac{\mathrm{d}^{2(m-l)}\Phi_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2(m-l)}}\right|_{\tau=0}. \qquad \text{(A.198)}\end{aligned}$$

Recalling (4.262) while utilizing (A.214) to find $\Phi_{\breve{\alpha}_t(q)}(0) = 1$ and employing the identity

$$(-\mathrm{j})^{2k} = \mathrm{e}^{-\mathrm{j}\frac{\pi}{2}2k} = \mathrm{e}^{-\mathrm{j}\pi k} = (-1)^k, \quad \forall k \in \mathbb{Z} \tag{A.199}$$

finally leads to the recurrence formula for the raw moments of order $2m$ to be given by

$$\begin{aligned} E\left[\breve{\alpha}_t^{2m}(q)\right] &= (-1)^m \left.\frac{\mathrm{d}^{2m}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2m}}\right|_{\tau=0} \\ &\quad + \sum_{l=1}^{m-1} (-1)^l \binom{2m-1}{2l-1} \left.\frac{\mathrm{d}^{2l}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} E\left[\breve{\alpha}_t^{2(m-l)}(q)\right] \end{aligned} \tag{A.200}$$

which, employing

$$E\left[\breve{\alpha}_t^{0}(q)\right] = 1, \tag{A.201}$$

may also be rewritten as (note the change of the upper summation limit)

$$E\left[\breve{\alpha}_t^{2m}(q)\right] = \sum_{l=1}^{m} (-1)^l \binom{2m-1}{2l-1} \left.\frac{\mathrm{d}^{2l}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} E\left[\breve{\alpha}_t^{2(m-l)}(q)\right]. \tag{A.202}$$

The $2l$-th derivative of the second characteristic function $\Upsilon_{\breve{\alpha}_t(q)}(\tau)$ w.r.t. τ evaluated at $\tau = 0$ is now given by

$$\left.\frac{\mathrm{d}^{2l}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} = \sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} \left.\frac{\mathrm{d}^{2l}\ln\left(\mathrm{J}_0\left(c_q(k)\,\tau\right)\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0}. \tag{A.203}$$

With

$$\frac{\mathrm{d}^n f(c_q(k)\,x)}{\mathrm{d}x^n} = c_q(k)^n \left.\frac{d^n f(y)}{\mathrm{d}y^n}\right|_{y=c_q(k)x}, \tag{A.204}$$

which arises as a special case of the univariate Faà di Bruno's formula for the higher order derivatives of a composition of functions (see [114] for an overview and detailed discussions on the univariate case and [115] for the multivariate extension), (A.203) may be simplified to

$$\left.\frac{\mathrm{d}^{2l}\Upsilon_{\breve{\alpha}_t(q)}(\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} = \left.\frac{\mathrm{d}^{2l}\ln\left(\mathrm{J}_0(\tau)\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} \left(\sum_{k=K_q^{(\text{low})}}^{K_q^{(\text{up})}} c_q(k)^{2l}\right). \tag{A.205}$$

The pre-factor $\left.\frac{\mathrm{d}^{2l}\ln(\mathrm{J}_0(\tau))}{\mathrm{d}\tau^{2l}}\right|_{\tau=0}$ in (A.205) is independent of the particular mel filter bank index q. Though Faà di Bruno's formula may be employed to express this pre-factor for arbitrary l in an explicit way, an implicit, recursive solution to it will be considered next.

By noting that

$$\frac{\mathrm{d}\ln\left(\mathrm{J}_0(\tau)\right)}{\mathrm{d}\tau} = \frac{\frac{\mathrm{d}\,\mathrm{J}_0(\tau)}{\mathrm{d}\tau}}{\mathrm{J}_0(\tau)} \tag{A.206}$$

Eq. (A.188) may be employed with

$$u(\tau) := \frac{\mathrm{d}\ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau}, \tag{A.207}$$

$$v(\tau) := \mathrm{J}_0\left(\tau\right), \tag{A.208}$$

$$f(\tau) := \frac{\mathrm{d}\,\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau}, \tag{A.209}$$

to find

$$\left.\frac{\mathrm{d}^{n+1}\ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{n+1}}\right|_{\tau=0} = \left.\frac{\frac{\mathrm{d}^{n+1}\mathrm{J}_0(\tau)}{\mathrm{d}\tau^{n+1}} - \sum\limits_{k=0}^{n-1}\binom{n}{k}\frac{\mathrm{d}^{k+1}\ln(\mathrm{J}_0(\tau))}{\mathrm{d}\tau^{k+1}}\frac{\mathrm{d}^{n-k}\mathrm{J}_0(\tau)}{\mathrm{d}\tau^{n-k}}}{\mathrm{J}_0\left(\tau\right)}\right|_{\tau=0} \tag{A.210}$$

$$= \left.\frac{\mathrm{d}^{n+1}\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{n+1}}\right|_{\tau=0} - \sum_{k=0}^{n-1}\binom{n}{k}\left.\frac{\mathrm{d}^{k+1}\ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{k+1}}\right|_{\tau=0}\left.\frac{\mathrm{d}^{n-k}\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{n-k}}\right|_{\tau=0}, \tag{A.211}$$

where $\mathrm{J}_0\left(0\right)=1$ has been employed. For $n=2l-1$, $l\in\mathbb{N}_{>0}$, (A.211) turns into

$$\left.\frac{\mathrm{d}^{2l}\ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} = \left.\frac{\mathrm{d}^{2l}\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} - \sum_{k=0}^{2l-2}\binom{2l-1}{k}\left.\frac{\mathrm{d}^{k+1}\ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{k+1}}\right|_{\tau=0}\left.\frac{\mathrm{d}^{2l-1-k}\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{2l-1-k}}\right|_{\tau=0}. \tag{A.212}$$

By using the differentiation rule [113, p. 361, Eq. (9.1.31)]

$$\frac{\mathrm{d}^n\mathrm{J}_\nu\left(\tau\right)}{\mathrm{d}\tau^n} = \frac{1}{2^n}\sum_{k=0}^{n}(-1)^k\,\mathrm{J}_{\nu-n+2k}\left(x\right)\binom{n}{k}, \quad \forall n\in\mathbb{N} \tag{A.213}$$

and

$$\mathrm{J}_\nu\left(0\right) = \begin{cases} 1 & \text{, if } \nu=0 \\ 0 & \text{, else} \end{cases}, \tag{A.214}$$

which may best be seen by looking at the TAYLOR series expansion of the BESSEL function at $\tau=0$ given by [113, p. 360, Eq. (9.1.10)]

$$\mathrm{J}_\nu\left(\tau\right) = \sum_{k=0}^{\infty}\frac{(-1)^k\left(\frac{\tau}{2}\right)^{2k+\nu}}{\Gamma(k+\nu+1)k!}, \quad \forall\nu\in\mathbb{C}, \tag{A.215}$$

the identity

$$\left.\frac{\mathrm{d}^n\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^n}\right|_{\tau=0} = \begin{cases} \frac{(-1)^{\frac{n}{2}}}{2^n}\binom{n}{\frac{n}{2}} & \text{, if } n \bmod 2 = 0 \\ 0 & \text{, if } n \bmod 2 = 1 \end{cases} \tag{A.216}$$

may be inferred. Hence, the last term in the sum of (A.212) is non-zero only for those k for which $k = 2i - 1$, $i \in \{1, \ldots, l-1\}$, holds. Employing (A.216) while changing the summation variable accordingly thus gives the recursive formulation

$$\begin{aligned} \left.\frac{\mathrm{d}^{2l} \ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} &= \left.\frac{\mathrm{d}^{2l}\,\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0} \\ &\quad - \sum_{i=1}^{l-1} \binom{2l-1}{2i-1} \left.\frac{\mathrm{d}^{2i} \ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{2i}}\right|_{\tau=0} \left.\frac{\mathrm{d}^{2(l-i)}\,\mathrm{J}_0\left(\tau\right)}{\mathrm{d}\tau^{2(l-i)}}\right|_{\tau=0} && \text{(A.217)} \\ &= \frac{(-1)^l}{2^{2l}} \binom{2l}{l} \\ &\quad - \sum_{i=1}^{l-1} \binom{2l-1}{2i-1} \left.\frac{\mathrm{d}^{2i} \ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{2i}}\right|_{\tau=0} \frac{(-1)^{l-i}}{2^{2(l-i)}} \binom{2(l-i)}{l-i}. && \text{(A.218)} \end{aligned}$$

Note that only derivatives of $\ln\left(\mathrm{J}_0\left(\tau\right)\right)$ with *even* order occur in (A.218).

By defining

$$\varsigma_l := (-1)^l 2^{2l} \left.\frac{\mathrm{d}^{2l} \ln\left(\mathrm{J}_0\left(\tau\right)\right)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0}, \tag{A.219}$$

for which with (A.218)

$$\varsigma_l = \binom{2l}{l} - \sum_{i=1}^{l-1} \binom{2l-1}{2i-1} \binom{2(l-i)}{l-i} \varsigma_i \tag{A.220}$$

holds, and employing (A.205) now allows the raw moments of the phase factor RV $\breve{\alpha}_t\left(q\right)$ to eventually be expressed as

$$\begin{aligned} E\left[\breve{\alpha}_t^{2m-1}\left(q\right)\right] &= 0, && \text{(A.221)} \\ E\left[\breve{\alpha}_t^{2m}\left(q\right)\right] &= \sum_{l=1}^{m} \binom{2m-1}{2l-1} \left(\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \left(\frac{c_q\left(k\right)}{2}\right)^{2l} \right) \varsigma_l E\left[\breve{\alpha}_t^{2(m-l)}\left(q\right)\right], && \text{(A.222)} \end{aligned}$$

where $m \in \mathbb{N}_{>0}$.

An alternative to (A.222) may be obtained by defining

$$\beta_{m,l} := \binom{2m-1}{2l-1} \varsigma_l \tag{A.223}$$

and employing the identity

$$\binom{n}{m}\binom{m}{l} = \binom{n-l}{n-m}\binom{n}{l} \tag{A.224}$$

to find

$$E\left[\breve{\alpha}_t^{2m}\left(q\right)\right] = \sum_{l=1}^{m} \left(\sum_{k=K_q^{(\mathsf{low})}}^{K_q^{(\mathsf{up})}} \left(\frac{c_q\left(k\right)}{2}\right)^{2l} \right) \beta_{m,l} E\left[\breve{\alpha}_t^{2(m-l)}\left(q\right)\right], \tag{A.225}$$

where again $m \in \mathbb{N}_{>0}$. The determination of $\beta_{m,l}$ is now subject to the recursion

$$\beta_{m,l} = \binom{2m-1}{2l-1}\binom{2l}{l} - \sum_{i=1}^{l-1}\binom{2(l-i)}{l-i}\binom{2(m-i)}{2(m-l)}\beta_{m,i} \tag{A.226}$$

which, opposed to recursion (A.220), now also depends on the index m. However, according to (A.225) its values only need to be determined for $l \leq m$.

A.9.1 Proof of (A.194)

Lemma 2. *Let $\Phi_I(\tau)$ be the product of $I \in \mathbb{N}_{>0}$ Bessel functions of the first kind and order 0 according to*

$$\Phi_I(\tau) := \prod_{i=1}^{I} \mathrm{J}_0(c_i\tau), \tag{A.227}$$

where $\tau, c_i \in \mathbb{R}$. For the $(2n+1)$-th derivative of (A.227) it then holds that

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_I(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = 0, \quad \forall I \in \mathbb{N}_{>0} \tag{A.228}$$

and arbitrary $n \in \mathbb{N}$.

Proof of Lemma 2.
Basis Let $I = 1$. For $I = 1$, the $(2n+1)$-th derivative of (A.227) is given by

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_1(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = \left.\frac{\mathrm{d}^{2n+1}\,\mathrm{J}_0(c_1\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} \tag{A.229}$$

$$= c_1^{2n+1}\sum_{k=0}^{2n+1}\left.\frac{(-1)^k\,\mathrm{J}_{2(k-n)+1}(c_1\tau)\binom{2n+1}{k}}{2^{2n+1}}\right|_{\tau=0}, \tag{A.230}$$

where for the last equality (A.213) and (A.204) have been employed.

Since with (A.214) only the (not occurring) half-integer $k = n - \frac{1}{2}$ would contribute to the sum (A.230), eventually

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_1(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = 0, \quad \forall n \in \mathbb{N}. \tag{A.231}$$

Inductive step on I With

$$\Phi_I(\tau) = \Phi_{I-1}(\tau)\,\mathrm{J}_0(c_I\tau) \tag{A.232}$$

and LEIBNIZ' theorem [113, p. 12, Eq. (3.3.8)] (outlined in (A.184)), the $(2n+1)$-th derivative of $\Phi_I(\tau)$ may be written as

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_I(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = \left.\frac{\mathrm{d}^{2n+1}(\Phi_{I-1}(\tau)\,\mathrm{J}_0(c_I\tau))}{\mathrm{d}\tau^{2n+1}}\right|_{v=0} \tag{A.233}$$

$$= \left.\left(\sum_{k=0}^{2n+1}\binom{2n+1}{k}\frac{\mathrm{d}^k\,\mathrm{J}_0(c_I\tau)}{\mathrm{d}\tau^k}\frac{\mathrm{d}^{2n+1-k}\Phi_{I-1}(\tau)}{\mathrm{d}\tau^{2n+1-k}}\right)\right|_{\tau=0} \tag{A.234}$$

$$= \sum_{k=0}^{2n+1}\binom{2n+1}{k}\left.\frac{\mathrm{d}^k\,\mathrm{J}_0(c_I\tau)}{\mathrm{d}\tau^k}\right|_{\tau=0}\left.\frac{\mathrm{d}^{2n+1-k}\Phi_{I-1}(\tau)}{\mathrm{d}\tau^{2n+1-k}}\right|_{\tau=0}. \tag{A.235}$$

Since with (A.204) and (A.216)

$$\left.\frac{\mathrm{d}^k\,\mathrm{J}_0(c_I\tau)}{\mathrm{d}\tau^k}\right|_{\tau=0} = \begin{cases}(-1)^{\frac{k}{2}}\binom{k}{\frac{k}{2}}\left(\frac{c_I}{2}\right)^k & \text{,if } k \bmod 2 = 0\\ 0 & \text{,if } k \bmod 2 = 1\end{cases} \tag{A.236}$$

only those summands in (A.235) where $k=2l$, $l\in\{0,\ldots,n\}$, have to be considered further. Hence,

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_I(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = \sum_{l=0}^{n}\binom{2n+1}{2l}\left.\frac{\mathrm{d}^{2l}\,\mathrm{J}_0(c_I\tau)}{\mathrm{d}\tau^{2l}}\right|_{\tau=0}\left.\frac{\mathrm{d}^{2(n-l)+1}\Phi_{I-1}(\tau)}{\mathrm{d}\tau^{2(n-l)+1}}\right|_{\tau=0}. \tag{A.237}$$

Now, the induction hypothesis (A.228) may be employed for the last term under the sum to eventually find

$$\left.\frac{\mathrm{d}^{2n+1}\Phi_I(\tau)}{\mathrm{d}\tau^{2n+1}}\right|_{\tau=0} = 0, \quad \forall n\in\mathbb{N}, I\in\mathbb{N}_{>0}. \tag{A.238}$$

□

A.10 Moments of the Transformed Phase Factor

The RVs $\breve{\boldsymbol{\alpha}}_t$ are assumed to be i.i.d. with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$. The inverse error function $\mathrm{erf}^{-1}(\cdot)$ has been observed to transform the RV $\breve{\boldsymbol{\alpha}}_t$ into a RV $\breve{\boldsymbol{\gamma}}_t$ that is approximately GAUSSIAN distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$, i.e.,

$$\breve{\boldsymbol{\alpha}}_t \sim p_{\breve{\boldsymbol{\alpha}}_t}(\boldsymbol{\alpha}_t) \text{ with } E[\breve{\boldsymbol{\alpha}}_t]=\mathbf{0},\ E\left[\breve{\boldsymbol{\alpha}}_t\breve{\boldsymbol{\alpha}}_t^\dagger\right]=\boldsymbol{\Sigma}_{\breve{\alpha}} \tag{A.239}$$

$$\Downarrow \breve{\boldsymbol{\gamma}}_t = \mathrm{erf}^{-1}(\breve{\boldsymbol{\alpha}}_t)$$

$$\breve{\boldsymbol{\gamma}}_t \sim p_{\breve{\boldsymbol{\gamma}}_t}(\boldsymbol{\gamma}_t) \approx \mathcal{N}\left(\boldsymbol{\gamma}_t;\mathbf{0},\boldsymbol{\Sigma}_{\breve{\gamma}}\right). \tag{A.240}$$

The covariance matrix $\boldsymbol{\Sigma}_{\breve{\gamma}}$ will now be obtained from the covariance matrix $\boldsymbol{\Sigma}_{\breve{\alpha}}$ by means of matching the *true* $\boldsymbol{\Sigma}_{\breve{\alpha}}$ to that according to the inverse transformation, i.e.,

$$\boldsymbol{\Sigma}_{\breve{\alpha}} \overset{!}{=} \int_{\mathbb{R}^Q}\mathrm{erf}(\boldsymbol{\gamma}_t)(\mathrm{erf}(\boldsymbol{\gamma}_t))^\dagger p_{\breve{\boldsymbol{\gamma}}_t}(\boldsymbol{\gamma}_t)\,\mathrm{d}\boldsymbol{\gamma}_t \tag{A.241}$$

$$= \int_{\mathbb{R}^Q}\mathrm{erf}(\boldsymbol{\gamma}_t)(\mathrm{erf}(\boldsymbol{\gamma}_t))^\dagger\mathcal{N}\left(\boldsymbol{\gamma}_t;\mathbf{0},\boldsymbol{\Sigma}_{\breve{\gamma}}\right)\mathrm{d}\boldsymbol{\gamma}_t. \tag{A.242}$$

The solution to $\mathbf{\Sigma}_{\breve{\gamma}}$ will be obtained by first solving for the diagonal entries and then for the off-diagonal ones.

The diagonal elements – the variances Looking at the diagonal entries first, the variance $\sigma^2_{\breve{\gamma}_q}$ at mel index q may be obtained by postulating

$$\sigma^2_{\breve{\alpha}_q} \overset{!}{=} \int\limits_{-\infty}^{\infty} \mathrm{erf}^2\left(\gamma_t\left(q\right)\right) p_{\breve{\gamma}_t(q)}\left(\gamma_t\left(q\right)\right) \mathrm{d}\gamma_t\left(q\right) \tag{A.243}$$

$$= \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{\breve{\gamma}_q}}} \mathrm{e}^{-\frac{1}{2\sigma^2_{\breve{\gamma}_q}}\gamma_t(q)^2} \mathrm{erf}^2\left(\gamma_t\left(q\right)\right) \mathrm{d}\gamma_t\left(q\right) \tag{A.244}$$

$$= \frac{2}{\pi}\tan^{-1}\left(\sqrt{\frac{4\sigma^4_{\breve{\gamma}_q}}{1+4\sigma^2_{\breve{\gamma}_q}}}\right), \tag{A.245}$$

where the last equality has been obtained by employing the definite integral [116, p. 137, (52)]

$$\int\limits_{0}^{\infty} \mathrm{e}^{-a^2x^2} \mathrm{erf}\left(bx\right)\mathrm{erf}\left(cx\right)\mathrm{d}x = \frac{1}{a\sqrt{\pi}}\tan^{-1}\left(\frac{bc}{a\left(a^2+b^2+c^2\right)^{\frac{1}{2}}}\right) \tag{A.246}$$

with $b = c = 1$ and $a = \frac{1}{\sqrt{2}\sigma_{\breve{\gamma}_q}}$. Note that $\tan^{-1}(\cdot)$ denotes the inverse function of the tangent (also denoted by $\arctan(\cdot)$).

Calling for (A.245) to hold is equivalent to

$$\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) \overset{!}{=} \frac{4\sigma^4_{\breve{\gamma}_q}}{1+4\sigma^2_{\breve{\gamma}_q}} \tag{A.247}$$

$$4\sigma^4_{\breve{\gamma}_q} - 4\sigma^2_{\breve{\gamma}_q}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) - \tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) \overset{!}{=} 0 \tag{A.248}$$

$$\sigma^4_{\breve{\gamma}_q} - \sigma^2_{\breve{\gamma}_q}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) - \frac{1}{4}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) \overset{!}{=} 0. \tag{A.249}$$

The solution to (A.249) is given by

$$\sigma^2_{\breve{\gamma}_q} = \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) \pm \sqrt{\frac{1}{4}\tan^4\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) + \frac{1}{4}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)} \tag{A.250}$$

$$= \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) \pm \left(\frac{1}{2}\tan\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)\sqrt{\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)+1}\right). \tag{A.251}$$

Since $\sigma^2_{\breve{\gamma}_q} \in \mathbb{R}_{>0}$, only the "+" solution is valid. A more convenient solution may be obtained by employing in turn the trigonometric identities

$$\tan^2\left(x\right) + 1 = \frac{1}{\cos^2\left(x\right)} \tag{A.252}$$

$$\tan\left(x\right) = 2\frac{\tan\left(\frac{x}{2}\right)}{1-\tan^2\left(\frac{x}{2}\right)} \tag{A.253}$$

$$\sin\left(x\right) = 2\frac{\tan\left(\frac{x}{2}\right)}{1+\tan^2\left(\frac{x}{2}\right)}. \tag{A.254}$$

Since the phase factor $\alpha_t(q)$ is limited to the range $[-1,+1]$, the variance $\sigma^2_{\breve{\alpha}_q}$ is bounded by $0 \leq \sigma^2_{\breve{\alpha}_q} \leq 1$. Consequently, with $x = \frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}$, (A.253) and (A.254) are always non-negative.

Application of (A.252) first leads to

$$\begin{aligned}
\sigma^2_{\breve{\gamma}_q} &= \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) + \left(\frac{1}{2}\tan\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)\frac{1}{\cos\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}\right) && \text{(A.255)}\\
&= \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) + \left(\frac{1}{2}\tan\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)\frac{\sin\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}{\cos\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}\frac{1}{\sin\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}\right) && \text{(A.256)}\\
&= \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right) + \left(\frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)\frac{1}{\sin\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}\right) && \text{(A.257)}\\
&= \frac{1}{2}\tan^2\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)\left(1+\frac{1}{\sin\left(\frac{\pi}{2}\sigma^2_{\breve{\alpha}_q}\right)}\right). && \text{(A.258)}
\end{aligned}$$

Further application of (A.253) and (A.254) results in

$$\sigma^2_{\breve{\gamma}_q} = 2\frac{\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}\left(1+\frac{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{2\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}\right) \quad \text{(A.259)}$$

which may finally be written as

$$\begin{aligned}
\sigma^2_{\breve{\gamma}_q} &= 2\frac{\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}\left(1+\frac{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{2\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}\right) && \text{(A.260)}\\
&= 2\frac{\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}\frac{2\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)+1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{2\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)} && \text{(A.261)}\\
&= 2\frac{\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2\left(1+\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}\frac{\left(1+\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}{2\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)} && \text{(A.262)}\\
&= \frac{\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}{\left(1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)\right)^2}. && \text{(A.263)}
\end{aligned}$$

The off-diagonal elements – the covariances To determine the cross-terms of the covariance matrix $\mathbf{\Sigma}_{\breve{\gamma}}$ from the entries of the covariance matrix $\mathbf{\Sigma}_{\breve{\alpha}}$, the joint PDF of two random variables $\breve{\gamma}_t(q)$ and $\breve{\gamma}_t(q')$ is required. Since $\breve{\gamma}_t$ is assumed to be GAUSSIAN distributed, this joint PDF is also GAUSSIAN and given by

$$p_{\breve{\gamma}_t(q),\breve{\gamma}_t(q')}\left(\gamma_t(q),\gamma_t\left(q'\right)\right) = \mathcal{N}\left(\begin{bmatrix}\gamma_t(q)\\ \gamma_t(q')\end{bmatrix};\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma^2_{\breve{\gamma}_q} & \sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}\\ \sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} & \sigma^2_{\breve{\gamma}_{q'}}\end{bmatrix}\right), q \neq q'. \quad \text{(A.264)}$$

Assuming the variances $\sigma^2_{\breve{\gamma}_q}$ and $\sigma^2_{\breve{\gamma}_{q'}}$ to be computed according to (A.263), only $\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}$ has to be determined. For its derivation, it is now postulated that

$$\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}} \overset{!}{=} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \mathrm{erf}\left(\gamma_t\left(q\right)\right)\mathrm{erf}\left(\gamma_t\left(q'\right)\right) p_{\breve{\gamma}_t(q),\breve{\gamma}_t(q')}\left(\gamma_t\left(q\right),\gamma_t\left(q'\right)\right)\mathrm{d}\gamma_t\left(q\right)\mathrm{d}\gamma_t\left(q'\right) \tag{A.265}$$

$$= \int\limits_{-\infty}^{\infty}\left(\int\limits_{-\infty}^{\infty} \mathrm{erf}\left(\gamma_t\left(q\right)\right) p_{\breve{\gamma}_t(q)|\breve{\gamma}_t(q')}\left(\gamma_t\left(q\right)\middle|\gamma_t\left(q'\right)\right)\mathrm{d}\gamma_t\left(q\right)\right) \mathrm{erf}\left(\gamma_t\left(q'\right)\right) p_{\breve{\gamma}_t(q')}\left(\gamma_t\left(q'\right)\right)\mathrm{d}\gamma_t\left(q'\right). \tag{A.266}$$

Since the joint PDF $p_{\breve{\gamma}_t(q),\breve{\gamma}_t(q')}$ given in (A.264) is zero-mean, the conditional PDF $p_{\breve{\gamma}_t(q)|\breve{\gamma}_t(q')}$ is given by [112, (276)]

$$p_{\breve{\gamma}_t(q)|\breve{\gamma}_t(q')}\left(\gamma_t\left(q\right)\middle|\gamma_t\left(q'\right)\right) = \mathcal{N}\left(\gamma_t\left(q\right);\nu_{q|q'}\gamma_t\left(q'\right),\sigma^2_{q|q'}\right), \tag{A.267}$$

where

$$\nu_{q|q'} = \frac{\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}{\sigma^2_{\breve{\gamma}_{q'}}} \tag{A.268}$$

$$\sigma^2_{q|q'} = \sigma^2_{\breve{\gamma}_q} - \frac{\sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}{\sigma^2_{\breve{\gamma}_{q'}}} = \frac{\sigma^2_{\breve{\gamma}_q}\sigma^2_{\breve{\gamma}_{q'}} - \sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}{\sigma^2_{\breve{\gamma}_{q'}}}. \tag{A.269}$$

By employing the definite integral [116, p. 136, (45)]

$$\int\limits_{-\infty}^{\infty} \mathrm{e}^{-(ax+b)^2}\,\mathrm{erf}\left(x\right)\mathrm{d}x = \frac{\sqrt{\pi}}{a}\,\mathrm{erf}\left(\frac{b}{\left(a^2+1\right)^{\frac{1}{2}}}\right) \tag{A.270}$$

with $a = \frac{1}{\sqrt{2}\sigma_{q|q'}}$ and $b = \frac{\nu_{q|q'}\gamma_t\left(q'\right)}{\sqrt{2}\sigma_{q|q'}}$, the inner integral in (A.266) can be found to be

$$\int\limits_{-\infty}^{\infty} \mathrm{erf}\left(\gamma_t\left(q\right)\right) p_{\breve{\gamma}_t(q)|\breve{\gamma}_t(q')}\left(\gamma_t\left(q\right)\middle|\gamma_t\left(q'\right)\right)\mathrm{d}\gamma_t\left(q\right) \tag{A.271}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{q|q'}}\int\limits_{-\infty}^{\infty} \mathrm{erf}\left(\gamma_t\left(q\right)\right)\mathrm{e}^{-\frac{1}{2}\frac{\left(\gamma_t(q)-\nu_{q|q'}\gamma_t\left(q'\right)\right)^2}{\sigma^2_{q|q'}}}\mathrm{d}\gamma_t\left(q\right) \tag{A.272}$$

$$= \mathrm{erf}\left(c_{q|q'}\gamma_t\left(q'\right)\right), \tag{A.273}$$

with

$$c_{q|q'} = \frac{\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}{\sigma_{\breve{\gamma}_{q'}}\sqrt{\sigma^2_{\breve{\gamma}_{q'}} + 2\left(\sigma^2_{\breve{\gamma}_q}\sigma^2_{\breve{\gamma}_{q'}} - \sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}}\right)}}. \tag{A.274}$$

Hence, (A.266) turns into

$$\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}} \overset{!}{=} \int\limits_{-\infty}^{\infty} \operatorname{erf}\left(\gamma_t\left(q'\right)\right) \operatorname{erf}\left(c\gamma_t\left(q'\right)\right) p_{\breve{\gamma}_t(q')}\left(\gamma_t\left(q'\right)\right) \mathrm{d}\gamma_t\left(q'\right) \tag{A.275}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{\breve{\gamma}_q}} \int\limits_{-\infty}^{\infty} \operatorname{erf}\left(\gamma_t\left(q'\right)\right) \operatorname{erf}\left(c_{q|q'}\gamma_t\left(q'\right)\right) \mathrm{e}^{-\frac{1}{2\sigma_{\breve{\gamma}_q}}\gamma_t(q)^2} \mathrm{d}\gamma_t\left(q'\right) \tag{A.276}$$

and, employing (A.246) with $a = \frac{1}{\sqrt{2}\sigma_{\breve{\gamma}_q}}$, $b = 1$ and $c = c_{q|q'}$, eventually yields

$$\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}} \overset{!}{=} \frac{2}{\pi} \tan^{-1}\left(\frac{2\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}{\sqrt{\left(1+2\sigma^2_{\breve{\gamma}_q}\right)\left(1+2\sigma^2_{\breve{\gamma}_{q'}}\right) - 4\sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}}}}\right). \tag{A.277}$$

Solving for $\sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}}$ then yields

$$\sigma^2_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \frac{\tan^2\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)\left(1+2\sigma^2_{\breve{\gamma}_q}\right)\left(1+2\sigma^2_{\breve{\gamma}_{q'}}\right)}{2+4\tan^2\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)} \tag{A.278}$$

and, recalling that the solution to $\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}}$ has to fulfill (A.277), eventually gives

$$\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \frac{1}{2}\tan\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)\sqrt{\frac{\left(1+2\sigma^2_{\breve{\gamma}_q}\right)\left(1+2\sigma^2_{\breve{\gamma}_{q'}}\right)}{1+\tan^2\left(\frac{\pi}{2}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)}}. \tag{A.279}$$

With the variances $\sigma^2_{\breve{\gamma}_q}$ and $\sigma^2_{\breve{\gamma}_{q'}}$ characterized by (A.263) in terms of $\sigma^2_{\breve{\alpha}_q}$ and $\sigma^2_{\breve{\alpha}_{q'}}$, respectively, the trigonometric identity (A.253) may be employed to find the compact representation

$$\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \frac{\tan\left(\frac{\pi}{4}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)}{1+\tan^2\left(\frac{\pi}{4}\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}}\right)} \cdot \frac{\sqrt{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)}}{1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_q}\right)} \cdot \frac{\sqrt{1+\tan^2\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_{q'}}\right)}}{1-\tan\left(\frac{\pi}{4}\sigma^2_{\breve{\alpha}_{q'}}\right)}. \tag{A.280}$$

Note that $\sigma_{\breve{\alpha}_q,\breve{\alpha}_{q'}} = \sigma^2_{\breve{\alpha}_q}$ and $\sigma_{\breve{\gamma}_q,\breve{\gamma}_{q'}} = \sigma^2_{\breve{\gamma}_q}$ for $q = q'$. Hence, (A.280) turns into (A.263) and can eventually be considered the *generalized solution* to the problem of determining the elements of the covariance matrix $\mathbf{\Sigma}_{\breve{\gamma}}$ from those of $\mathbf{\Sigma}_{\breve{\alpha}}$.

A.11 Multivariate Normal and Log-Normal Distribution

If the RV $\breve{\mathbf{x}} \in \mathbb{R}^D$ is distributed according to a Normal distribution with mean vector $\boldsymbol{\mu}_{\breve{\mathbf{x}}}$ and covariance matrix $\mathbf{\Sigma}_{\breve{\mathbf{x}}}$, i.e.,

$$p_{\breve{\mathbf{x}}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{\breve{\mathbf{x}}}, \mathbf{\Sigma}_{\breve{\mathbf{x}}}\right), \tag{A.281}$$

then the RV $\breve{\mathbf{y}} \in \mathbb{R}^D_{>0}$ obtained from the transformation

$$\breve{\mathbf{y}} = \mathrm{e}^{\breve{\mathbf{x}}} \tag{A.282}$$

follows a multivariate Log-Normal distribution, i.e.,

$$\begin{aligned}
p_{\breve{\mathbf{y}}}(\mathbf{y}) &= \frac{p_{\breve{\mathbf{x}}}(\ln(\mathbf{y}))}{|J_{\mathrm{e},\mathbf{x}}|_{\mathbf{x}=\ln(\mathbf{y})}} && \text{(A.283)}\\
&= \frac{1}{\prod\limits_{d=0}^{D-1} y_d} \mathcal{N}(\ln(\mathbf{y}); \boldsymbol{\mu}_{\breve{\mathbf{x}}}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}}) && \text{(A.284)}\\
&:= \mathcal{LN}(\mathbf{y}; \boldsymbol{\mu}_{\breve{\mathbf{x}}}, \boldsymbol{\Sigma}_{\breve{\mathbf{x}}}), && \text{(A.285)}
\end{aligned}$$

which is characterized by the mean vector and the covariance matrix of the RV $\breve{\mathbf{x}}$ rather than the mean vector and the covariance matrix of the RV $\breve{\mathbf{y}}$ since interpretation of the former is usually more intuitive. However, the mean vector and covariance matrix of the normally distributed RV $\breve{\mathbf{y}}$ may be related to the mean vector and covariance matrix of the log-normally distributed RV $\breve{\mathbf{y}}$ by

$$\boldsymbol{\mu}_{\breve{\mathbf{x}}} = \ln\left(\boldsymbol{\mu}_{\breve{\mathbf{y}}}\right) - \frac{1}{2}\operatorname{diag}(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}}), \tag{A.286}$$

where

$$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}} = \ln\left(\boldsymbol{\mu}_{\breve{\mathbf{y}}}\left(\boldsymbol{\mu}_{\breve{\mathbf{y}}}\right)^{\dagger} + \boldsymbol{\Sigma}_{\breve{\mathbf{y}}}\right) - \ln\left(\boldsymbol{\mu}_{\breve{\mathbf{y}}}\left(\boldsymbol{\mu}_{\breve{\mathbf{y}}}\right)^{\dagger}\right). \tag{A.287}$$

Note that the logarithm thereby has to be understood to be applied to the vectors and matrices component-wise. Equation (A.286) and (A.287) will now be derived by looking at the individual elements of the mean vector and the covariance matrix.

The Mean Vector The d-th component of the mean vector is given by

$$\begin{aligned}
\mu_{y_d} &= E\left[\mathrm{e}^{x_d}\right] && \text{(A.288)}\\
&= \int\limits_{-\infty}^{\infty} \mathrm{e}^{x_d} p_{\breve{x}_d}(x_d)\,\mathrm{d}x_d && \text{(A.289)}\\
&= \frac{1}{\sqrt{2\pi}\sigma_{x_d}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{x_d} \mathrm{e}^{-\frac{1}{2\sigma_{x_d}^2}x_d^2 + \frac{\mu_{x_d}}{\sigma_{x_d}^2}x_d - \frac{\mu_{x_d}^2}{2\sigma_{x_d}^2}}\,\mathrm{d}x_d && \text{(A.290)}\\
&= \frac{1}{\sqrt{2\pi}\sigma_{x_d}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma_{x_d}^2}x_d^2 + \frac{\mu_{x_d}+\sigma_{x_d}^2}{\sigma_{x_d}^2}x_d - \frac{\mu_{x_d}^2}{2\sigma_{x_d}^2}}\,\mathrm{d}x_d && \text{(A.291)}
\end{aligned}$$

The closed form solution to the integral may be obtain from [117, p. 307, 3.323(2)]

$$\int\limits_{-\infty}^{\infty} \mathrm{e}^{-p^2x^2 \pm qx}\,\mathrm{d}x = \frac{\sqrt{\pi}}{p}\mathrm{e}^{\frac{q^2}{4p^2}}, \qquad p > 0 \tag{A.292}$$

with $p = \frac{1}{\sqrt{2}\sigma_{x_d}}$ and $q = \frac{\mu_{x_d} + \sigma_{x_d}^2}{\sigma_{x_d}^2}$. Hence,

$$\begin{aligned}\mu_{y_d} &= \mathrm{e}^{-\frac{\mu_{x_d}^2}{2\sigma_{x_d}^2}} \mathrm{e}^{\frac{\left(\mu_{x_d}+\sigma_{x_d}^2\right)^2}{2\sigma_{x_d}^2}} && \text{(A.293)}\\ &= \mathrm{e}^{\mu_{x_d}+\frac{1}{2}\sigma_{x_d}^2} && \text{(A.294)}\end{aligned}$$

and consequently

$$\mu_{x_d} = \ln\left(\mu_{y_d}\right) - \frac{1}{2}\sigma_{x_d}^2. \qquad \text{(A.295)}$$

The Covariance Matrix To find the elements $\sigma_{y_d,y_{d'}}$ of the covariance matrix $\mathbf{\Sigma}_{\breve{\mathbf{y}}}$ it is sufficient to look at the expectation value

$$E\left[y_d y_{d'}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{x_d + x_{d'}} p_{\breve{x}_d,\breve{x}_{d'}}\left(x_d, x_{d'}\right) \mathrm{d}x_d \mathrm{d}x_{d'} \qquad \text{(A.296)}$$

since

$$\begin{aligned}\sigma_{y_d,y_{d'}} &= E\left[\left(y_d - E\left[y_d\right]\right)\left(y_{d'} - E\left[y_{d'}\right]\right)\right] && \text{(A.297)}\\ &= E\left[y_d y_{d'}\right] - E\left[y_d\right] E\left[y_{d'}\right]. && \text{(A.298)}\end{aligned}$$

An elegant solution to the integrals may be found by considering the sum $x_d + x_{d'}$ in the integral (A.296) to form a new RV. Thus, (A.296) may also be written as

$$\begin{aligned}E\left[y_d y_{d'}\right] &= E\left[\mathrm{e}^{x_d + x_{d'}}\right] && \text{(A.299)}\\ &= \int_{-\infty}^{\infty} \mathrm{e}^{x_d + x_{d'}} p_{\breve{x}_d + \breve{x}_{d'}}\left(x_d + x_{d'}\right) \mathrm{d}\left(x_d + x_{d'}\right). && \text{(A.300)}\end{aligned}$$

Since $\breve{x}_d$ and $\breve{x}_{d'}$ are jointly Gaussian distributed, their sum also follows a Gaussian distribution with mean $\mu_{x_d} + \mu_{x_{d'}}$ and variance $\sigma_{x_d}^2 + \sigma_{x_{d'}}^2 + 2\sigma_{x_d,x_{d'}}$.

Hence, following the steps taken in the derivation of the mean in the previous paragraph, it immediately follows that for $d = d'$ the variance $\sigma_{y_d}^2$ is given by

$$\begin{aligned}\sigma_{y_d}^2 &= \mathrm{e}^{2\mu_{x_d} + 2\sigma_{x_d}^2} - \mathrm{e}^{2\mu_{x_d} + \sigma_{x_d}^2} && \text{(A.301)}\\ &= \mathrm{e}^{2\mu_{x_d} + \sigma_{x_d}^2}\left(\mathrm{e}^{\sigma_{x_d}^2} - 1\right). && \text{(A.302)}\end{aligned}$$

Noting that $\mathrm{e}^{2\mu_{x_d} + \sigma_{x_d}^2} = \mu_{y_d}^2$, it also holds that

$$\begin{aligned}\sigma_{x_d}^2 &= \ln\left(1 + \frac{\sigma_{y_d}^2}{\mu_{y_d}^2}\right) && \text{(A.303)}\\ &= \ln\left(\mu_{y_d}^2 + \sigma_{y_d}^2\right) - \ln\left(\mu_{y_d}^2\right). && \text{(A.304)}\end{aligned}$$

Equivalently, for $d \neq d'$, the covariance $\sigma_{y_d,y_{d'}}$ is, with

$$E\left[y_d y_{d'}\right] = \mathrm{e}^{\mu_{x_d}+\mu_{x_{d'}}+\frac{1}{2}\left(\sigma^2_{x_d}+\sigma^2_{x_{d'}}+2\sigma_{x_d,x_{d'}}\right)}, \tag{A.305}$$

given by

$$\sigma_{y_d,y_{d'}} = \mathrm{e}^{\mu_{x_d}+\mu_{x_{d'}}+\frac{1}{2}\left(\sigma^2_{x_d}+\sigma^2_{x_{d'}}\right)}\left(\mathrm{e}^{\sigma_{x_d,x_{d'}}}-1\right). \tag{A.306}$$

Noting that $\mathrm{e}^{\mu_{x_d}+\frac{1}{2}\sigma^2_{x_d}} = \mu_{y_d}$ and $\mathrm{e}^{\mu_{x_{d'}}+\frac{1}{2}\sigma^2_{x_{d'}}} = \mu_{y_{d'}}$, it immediately follows that

$$\begin{aligned}\sigma_{x_d,x_{d'}} &= \ln\left(1+\frac{\sigma_{y_d,y_{d'}}}{\mu_{y_d}\mu_{y_{d'}}}\right) && \text{(A.307)}\\ &= \ln\left(\mu_{y_d}\mu_{y_{d'}}+\sigma_{y_d,y_{d'}}\right)-\ln\left(\mu_{y_d}\mu_{y_{d'}}\right). && \text{(A.308)}\end{aligned}$$

A.12 Vector-Taylor Series Expansion

The VTS expansion of a vector-valued function $g : \mathbb{R}^{N_{\mathbf{z}}} \to \mathbb{R}^{N_{\mathbf{o}}}$ mapping the vector $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$ to the vector $\mathbf{o}$ around the expansion point $\mathbf{z}_0 \in \mathbb{R}^{N_{\mathbf{z}}}$ up to second-order terms may be written as [93]

$$\mathbf{o} = g\left(\mathbf{z}_0\right) + J_{g,\mathbf{z}_0}\left[\mathbf{z}-\mathbf{z}_0\right] + \frac{1}{2}\sum_{n=1}^{N_{\mathbf{o}}} \mathbf{e}_n[\mathbf{z}-\mathbf{z}_0]^{\dagger} H^n_{g,\mathbf{z}_0}\left[\mathbf{z}-\mathbf{z}_0\right] + \mathit{HOT}. \tag{A.309}$$

Thereby, $J_{g,\mathbf{z}_0}$ denotes the JACOBIAN matrix of the function g and $H^n_{g,\mathbf{z}_0}$ the HESSIAN matrix of the n-th component of the function g, defined by

$$J_{g,\mathbf{z}} = \begin{bmatrix} \frac{\partial g_1(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial g_1(\mathbf{z})}{\partial z_{N_{\mathbf{z}}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{N_{\mathbf{o}}}(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial g_{N_{\mathbf{o}}}(\mathbf{z})}{\partial z_{N_{\mathbf{z}}}} \end{bmatrix}, \quad H^n_{g,\mathbf{z}} = \begin{bmatrix} \frac{\partial^2 g_n(\mathbf{z})}{\partial z_1 \partial z_1} & \cdots & \frac{\partial^2 g_n(\mathbf{z})}{\partial z_1 \partial z_{N_{\mathbf{z}}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g_n(\mathbf{z})}{\partial z_{N_{\mathbf{z}}} \partial z_1} & \cdots & \frac{\partial^2 g_n(\mathbf{z})}{\partial z_{N_{\mathbf{z}}} \partial z_{N_{\mathbf{z}}}} \end{bmatrix}, \tag{A.310}$$

both times evaluated at $\mathbf{z} = \mathbf{z}_0$. Further, $\mathbf{e}_n \in \mathbb{R}^{N_{\mathbf{o}}}$ is the n-th CARTESIAN basis vector.

Considering $\mathbf{z}$ to be a realization of a RV $\breve{\mathbf{z}}$, the mean vector $\boldsymbol{\mu}_{\breve{\mathbf{o}}}$ of the resulting RV $\breve{\mathbf{o}}$ may be computed (ignoring *HOT*) by

$$\begin{aligned}\boldsymbol{\mu}_{\breve{\mathbf{o}}} &:= E\left[\breve{\mathbf{o}}\right] && \text{(A.311)}\\ &\approx g\left(\mathbf{z}_0\right) + J_{g,\mathbf{z}_0}E\left[\breve{\mathbf{z}}-\mathbf{z}_0\right] + \frac{1}{2}\sum_{n=1}^{N_{\mathbf{o}}} \mathbf{e}_n E\left[\left(\breve{\mathbf{z}}-\mathbf{z}_0\right)^{\dagger} H^n_{g,\mathbf{z}_0}\left(\breve{\mathbf{z}}-\mathbf{z}_0\right)\right] && \text{(A.312)}\\ &= g\left(\mathbf{z}_0\right) + J_{g,\mathbf{z}_0}E\left[\breve{\mathbf{z}}-\mathbf{z}_0\right] + \frac{1}{2}\sum_{n=1}^{N_{\mathbf{o}}} \mathbf{e}_n \operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[\left(\breve{\mathbf{z}}-\mathbf{z}_0\right)\left(\breve{\mathbf{z}}-\mathbf{z}_0\right)^{\dagger}\right]\right) && \text{(A.313)}\end{aligned}$$

and the associated covariance matrix $\boldsymbol{\Sigma}_{\breve{\mathbf{o}}}$ by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\breve{\mathbf{o}}} &:= E\left[(\breve{\mathbf{o}}-\boldsymbol{\mu}_{\breve{\mathbf{o}}})(\breve{\mathbf{o}}-\boldsymbol{\mu}_{\breve{\mathbf{o}}})^{\dagger}\right] && \text{(A.314)}\\
&= J_{g,\mathbf{z}_0} E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^{\dagger}\right](J_{g,\mathbf{z}_0})^{\dagger}\\
&\quad + \frac{1}{2} J_{g,\mathbf{z}_0} \sum_{n=1}^{N_{\mathbf{o}}} \mathbf{e}_n \Big\{ E\left[[\breve{\mathbf{z}}-\mathbf{z}_0][\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\left(H^{n}_{g,\mathbf{z}_0}\right)^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\\
&\qquad\qquad - E[\breve{\mathbf{z}}-\mathbf{z}_0]\, E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\left(H^{n}_{g,\mathbf{z}_0}\right)^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\Big\}\\
&\quad + \frac{1}{2}\sum_{n=1}^{N_{\mathbf{o}}} \Big\{ E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0][\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\right]\\
&\qquad\qquad - E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\Big\} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\right]\mathbf{e}_n^{\dagger}(J_{g,\mathbf{z}_0})^{\dagger}\\
&\quad + \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0][\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\left(H^{m}_{g,\mathbf{z}_0}\right)^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\\
&\quad - \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\left(E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{m}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)^{\dagger}. && \text{(A.315)}
\end{aligned}$$

If the expansion point is now chosen to be $\mathbf{z}_0 = E[\breve{\mathbf{z}}]$ and the third-order moments are approximately zero, the above central moments of the observation $\breve{\mathbf{o}}$ simplify to

$$\begin{aligned}
\boldsymbol{\mu}_{\breve{\mathbf{o}}} &= g(\mathbf{z}_0) + \frac{1}{2}\sum_{n=1}^{N_{\mathbf{o}}} \mathbf{e}_n \operatorname{tr}\left(H^{n}_{g,\mathbf{z}_0} E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^{\dagger}\right]\right) && \text{(A.316)}\\
\boldsymbol{\Sigma}_{\breve{\mathbf{o}}} &= J_{g,\mathbf{z}_0} E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^{\dagger}\right](J_{g,\mathbf{z}_0})^{\dagger}\\
&\quad + \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0][\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\left(H^{m}_{g,\mathbf{z}_0}\right)^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\\
&\quad - \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\left(E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{m}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)^{\dagger} && \text{(A.317)}\\
&= J_{g,\mathbf{z}_0} E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^{\dagger}\right](J_{g,\mathbf{z}_0})^{\dagger}\\
&\quad + \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger} H^{n}_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0][\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}\left(H^{m}_{g,\mathbf{z}_0}\right)^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\\
&\quad - \frac{1}{4}\sum_{n,m=1}^{N_{\mathbf{o}}} \mathbf{e}_n\mathbf{e}_m^{\dagger} \operatorname{tr}\left(H^{n}_{g,\mathbf{z}_0} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)\\
&\qquad\qquad \operatorname{tr}\left(\left(H^{m}_{g,\mathbf{z}_0}\right)^{\dagger} E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^{\dagger}[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right) && \text{(A.318)}
\end{aligned}$$

For a Gaussian distributed RV $\breve{\mathbf{z}}$ the third-order moments are exactly zero and it can further be shown that [93, p. 55, (1.4.15-6)]

$$\begin{aligned}&E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger H^n_{g,\mathbf{z}_0}[\breve{\mathbf{z}}-\mathbf{z}_0]\,[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger\left(H^m_{g,\mathbf{z}_0}\right)^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\\ &\quad=\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)\operatorname{tr}\left(\left(H^m_{g,\mathbf{z}_0}\right)^\dagger E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)\\ &\qquad+2\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\left(H^m_{g,\mathbf{z}_0}\right)^\dagger E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right).\end{aligned}\tag{A.319}$$

Hence, the mean vector and the covariance matrix of the vector $\breve{\mathbf{o}}$ are given by

$$\boldsymbol{\mu}_{\breve{\mathbf{o}}}=g(\mathbf{z}_0)+\frac{1}{2}\sum_{n=1}^{N_\mathbf{o}}\mathbf{e}_n\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^\dagger\right]\right)\tag{A.320}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\breve{\mathbf{o}}}&=J_{g,\mathbf{z}_0}E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^\dagger\right](J_{g,\mathbf{z}_0})^\dagger\\ &\quad+\frac{1}{2}\sum_{n,m=1}^{N_\mathbf{o}}\mathbf{e}_n{\mathbf{e}_m}^\dagger\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\left(H^m_{g,\mathbf{z}_0}\right)^\dagger E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right)\end{aligned}\tag{A.321}$$

and the PDF $p_{\breve{\mathbf{o}}}$ may be approximated by a Gaussian as

$$p_{\breve{\mathbf{o}}}(\mathbf{o})\approx\mathcal{N}(\mathbf{o};\boldsymbol{\mu}_{\breve{\mathbf{o}}},\boldsymbol{\Sigma}_{\breve{\mathbf{o}}}).\tag{A.322}$$

In an equivalent manner, the conditional PDF $p_{\breve{\mathbf{o}}|\breve{\mathbf{z}}}$ may be obtained by truncating the VTS expansion to the linear terms and approximating the error

$$\begin{aligned}\boldsymbol{\varepsilon}&=\frac{1}{2}\sum_{n=1}^{N_\mathbf{o}}\mathbf{e}_n[\mathbf{z}-\mathbf{z}_0]^\dagger H^n_{g,\mathbf{z}_0}[\mathbf{z}-\mathbf{z}_0]+\mathit{HOT} &&\text{(A.323)}\\ &\approx\frac{1}{2}\sum_{n=1}^{N_\mathbf{o}}\mathbf{e}_n[\mathbf{z}-\mathbf{z}_0]^\dagger H^n_{g,\mathbf{z}_0}[\mathbf{z}-\mathbf{z}_0], &&\text{(A.324)}\end{aligned}$$

where the *HOT* have again be dropped, by a Gaussian as

$$p_{\breve{\boldsymbol{\varepsilon}}}(\boldsymbol{\varepsilon})\approx\mathcal{N}(\boldsymbol{\varepsilon};\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}},\boldsymbol{\Sigma}_{\breve{\boldsymbol{\varepsilon}}}).\tag{A.325}$$

With the above considerations, the mean vector $\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\breve{\boldsymbol{\varepsilon}}}$ can directly be given by

$$\begin{aligned}\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}}&:=E[\breve{\boldsymbol{\varepsilon}}] &&\text{(A.326)}\\ &=\frac{1}{2}\sum_{n=1}^{N_\mathbf{o}}\mathbf{e}_n\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[(\breve{\mathbf{z}}-\mathbf{z}_0)(\breve{\mathbf{z}}-\mathbf{z}_0)^\dagger\right]\right) &&\text{(A.327)}\\ \boldsymbol{\Sigma}_{\breve{\boldsymbol{\varepsilon}}}&:=E\left[(\breve{\boldsymbol{\varepsilon}}-\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}})(\breve{\boldsymbol{\varepsilon}}-\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}})^\dagger\right] &&\text{(A.328)}\\ &=\frac{1}{2}\sum_{n,m=1}^{N_\mathbf{o}}\mathbf{e}_n{\mathbf{e}_m}^\dagger\operatorname{tr}\left(H^n_{g,\mathbf{z}_0}E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\left(H^m_{g,\mathbf{z}_0}\right)^\dagger E\left[[\breve{\mathbf{z}}-\mathbf{z}_0]^\dagger[\breve{\mathbf{z}}-\mathbf{z}_0]\right]\right) &&\text{(A.329)}\end{aligned}$$

and the conditional PDF $p_{\breve{\mathbf{o}}|\breve{\mathbf{z}}}$ turns into a Gaussian distribution as

$$p_{\breve{\mathbf{o}}|\breve{\mathbf{z}}}(\mathbf{o}|\mathbf{z})\approx\mathcal{N}(\mathbf{o};g(\mathbf{z}_0)+J_{g,\mathbf{z}_0}[\mathbf{z}-\mathbf{z}_0]+\boldsymbol{\mu}_{\breve{\boldsymbol{\varepsilon}}},\boldsymbol{\Sigma}_{\breve{\boldsymbol{\varepsilon}}}).\tag{A.330}$$

Notation

This chapter summarizes the terms and notation used throughout this work.

Special Operators/Symbols

$\mathrm{tr}\left(\cdot\right)$	Trace
$(\cdot)!$	Factorial
$\lvert\cdot\rvert$	Absolute value/determinant
$(\cdot)^{-1}$	Inverse of a matrix, inverse function or inverse transform
$(\cdot)^{*}$	Conjugation
$(\cdot)^{\dagger}$	Transposition
$(\cdot)^{H}$	Conjugation and Transposition (Hermitian)
$\mathbf{I}$	Identity matrix
$\mathbf{0}$	Zero vector/matrix
$\mathbf{1}$	One vector/matrix
$a \bmod b$	Modulo operator, "a modulo b"
$\binom{a}{b}$	Binomial coefficient, "a over b"
$\mathrm{Re}\left\{\cdot\right\}$	Real part
$\mathrm{Im}\left\{\cdot\right\}$	Imaginary part
j	Imaginary unit
$\mathcal{F}\left\{\cdot\right\}$	FOURIER transform
$\pounds\left(\cdot\right)$	Length operator
$*$	Linear convolution
$\mathrm{erf}\left(\cdot\right)$	Error function
$\mathrm{sech}\left(\cdot\right)$	Hyperbolic secant
$\tanh\left(\cdot\right)$	Hyperbolic tangent
$\circ$	SCHUR/HADAMARD product operator
$\breve{\cdot}$	Superscript indicating a RV
$\hat{\cdot}$	Superscript indicating an estimate

Roman Symbols

$a_{i|h}$ HMM state transition probability (see (3.40)/(4.15))

$\mathbf{a}_t$ Auxiliary vector subsuming all variables in the observation model that are not part of the state vector $\mathbf{z}_t^{(\mathsf{l})}$, i.e., the vector of phase factor $\boldsymbol{\alpha}_t$

$\mathbf{A}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ State transition matrix

b_t Generic GMM mixture

$\mathbf{b}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ State prediction bias vector

B Shift of the analysis window

$c_{j|i}$ GMM mixture weight (defined in (3.42))

$c_k(q)$ Normalized coefficient of the q-th mel filter (defined in (4.225))

$\mathbf{C}_{\mathsf{DCT}}$ DCT matrix

$C_P(k)$ Frequency-dependent power compensation constant at frequency bin k (defined in (4.73))

C_P, $C_P^{(\mathsf{opt})}$ Frequency-independent power compensation constant (defined in (4.170))

d, d' Component indices

D # of components in a feature vector

$D_{\mathsf{KL}}(p_{\breve{x}} \| \hat{p}_{\breve{x}})$ Kulback-Leibler divergence between the PDF $p_{\breve{x}}(x)$ and the approximation to/estimate of it $\hat{p}_{\breve{x}}(x)$ (defined in (4.255))

$\mathbf{e}_i$ i-th Cartesian basis vector

$E[\breve{x}]$ Expectation of the RV $\breve{x}$

$E[\breve{x}|y]$ Conditional expectation of the RV $\breve{x}$ given that the RV $\breve{y}$ takes a value of y

$e_t^{(\mathsf{m})}(q)$ Error term in the derivation of the non-recursive observation model for reverberant-only speech in the MPSC domain (see (4.79))

$E_t(k)$ Error term in the derivation of the non-recursive observation model for reverberant-only speech in the PSC domain (see (4.69))

$\tilde{E}_t(k)$ Error term in the derivation of the non-recursive observation model for reverberant-only speech in the PSC domain after introduction of the AIR model (see (4.179))

$\tilde{\tilde{E}}_t(k), \tilde{\tilde{\tilde{E}}}_t(k)$ Error terms in the derivation of the recursive observation model for reverberant-only speech in the PSC domain (see (4.181) and (4.182))

f_S Sampling rate

$f_o^{(\mathsf{l})}(\cdot)$ Non-recursive observation mapping for noisy reverberant speech (defined in (4.111))

$f_{o,L_R}^{(\mathsf{l},\mathsf{R})}(\cdot)$ Recursive observation mapping for noisy reverberant speech (defined in (4.199))

$f_s^{(\mathsf{l})}(\cdot)$ Non-recursive observation mapping for reverberant-only speech (defined in (4.93))

$f_{s,L_R}^{(\mathsf{l},\mathsf{R})}(\cdot)$ Recursive observation mapping for reverberant-only speech (defined in (4.185))

$f_y^{(\mathsf{l})}(\cdot)$	Observation mapping for noisy speech (defined in (4.134))
$F_{\breve{x}}(a)$	Short form of $P_{\breve{x}}(x \leq a)$
$F_{w_{\mathsf{A}}}$	Correction factor accounting for the neglected correlations in the derivation of the moments of the vector of phase factors (defined in (4.273))
$g(\cdot)$	Generic observation/transformation function
$g_n(\cdot)$	n-th component of the generic observation function (defined in (4.323))
$g_o^{(\mathsf{l})}(\cdot)$	Non-recursive observation function for noisy reverberant speech (defined in (4.316))
$g_{o,L_R}^{(\mathsf{l},\mathsf{R})}(\cdot)$	Recursive observation function for noisy reverberant speech
$g_s^{(\mathsf{l})}(\cdot)$	Non-recursive observation function for reverberant-only speech (defined in (4.314))
$g_{s,L_R}^{(\mathsf{l},\mathsf{R})}(\cdot)$	Recursive observation function for reverberant-only speech (defined in (4.321))
$g_y^{(\mathsf{l})}(\cdot)$	Observation function for noisy speech (defined in (4.319))
h	HMM state index
$h(p)$	(Time-invariant) AIR
$h_{\breve{x}}(x)$	Histogram approximation to PDF of $\breve{x}$ evaluated at x
$h_t(k,k')$	Cross-band/band-to-band filters (defined in (4.50))
$\bar{\mathbf{h}}_t^{(\mathsf{l})}$, $\bar{h}_t^{(\mathsf{l})}(q)$	Representation of the AIR in the LMPSC domain (defined in (4.89))
$\mathcal{H}_t(q)$	AIR representation in the MPSC domain that is independent of the clean speech signal (defined in (4.85))
$\mathcal{H}_{t,t'}^{X}(q)$	AIR representation in the MPSC domain depending on the PSC of the clean speech signal (defined in (4.80))
$H_{f(\mathbf{x}),\mathbf{x}_0}^{n}$	Hessian matrix of the n-th component of the vector-valued function $f(\mathbf{x})$ evaluated at $\mathbf{x} = \mathbf{x}_0$
i	HMM state index
I	# of HMM states
$\mathcal{I}_{i\|\mathbf{o}_{1:t}}$, $\mathcal{I}_{i\|\mathbf{o}_{1:T}}$	Integrals to be solved for the UD rule to be applicable (defined in (3.78) and (3.79))
j	GMM mixture index/a priori model state index
J	# of mixture components in a GMM
$\mathrm{J}_n(\cdot)$	Bessel function of the n-th kind
$J_{f(\mathbf{x}),\mathbf{x}_0}$	Jacobian matrix of the vector-valued function $f(\mathbf{x})$ evaluated at $\mathbf{x} = \mathbf{x}_0$
k, k', k''	Discrete frequency indices
K	# of frequency bins, DFT length
$K_q^{(\mathsf{low})}$, $K_q^{(\mathsf{up})}$	Lower and upper cutoff frequency indices of the q mel filter
l, l', l''	Discrete time indices within an analysis frame
L_{Δ}, $L_{\Delta\Delta}$	Parameters controlling the # of MFCC feature vectors employed for calculation of the dynamic features
L_h	Length of the (truncated) AIR
$L_{w_{\mathsf{A}}}$, $L_{w_{\mathsf{S}}}$	Length of the analysis/synthesis window
L_w	Length of the analysis and synthesis window if $L_{w_{\mathsf{A}}} = L_{w_{\mathsf{S}}}$

$L_{H,\ell}$, L_H	Summation limits in the derivation of the observation model for reverberant speech in the SC domain (defined in (4.59) and (4.60))
L_H	Length of the AIR representation in the LMPSC domain
L_C	# of LMPSC vectors of the clean speech signal in the state vector
L_R	Recursion length, # of LMPSC vectors of the clean speech signal in the state vector if the recursive observation model is employed
$\mathcal{L}(t)$	Lower summation limits used to express the cross-band filters $h_t(k,k')$ in terms of the AIR $h(p)$ and the auxiliary function $\Phi_p(k,k')$ (defined in (4.55))
$\mathcal{LN}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	Log-Normal PDF of the RV $\breve{\mathbf{x}}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ associated with the underlying Normal PDF and evaluated at $\mathbf{x}$
m	GMM mixture index
m_t	Generic MSLDM/MSGMM state
M	# of mixture components in a GMM/# of dynamic states in the a priori model of the clean speech feature vector trajectory
n	Current word index
$n(p)$	Noise signal after pre-emphasis
$\mathbf{n}_t^{(\mathsf{l})}$, $n_t^{(\mathsf{l})}(q)$	LMPSC (vector) of the noise signal
$\mathbf{n}_t^{(\mathsf{m})}$, $n_t^{(\mathsf{m})}(q)$	MPSC (vector) of the noise signal
$N_t(k)$	SC of the noise signal
$\mathbf{N}_{t,q}$	Vector of SCs of the noise signal falling into mel band q
N_{DEL}, N_{INS}, N_{SUB}	# of required edit operations for string alignment: deletions, insertions and substitutions
N_w	# of words in a sentence $\mathcal{S}$
$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	Normal PDF of the RV $\breve{\mathbf{x}}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$
$o(p)$	Observed microphone signal after pre-emphasis
$o_t(l)$	Observed signal after application of the analysis window
$\mathbf{o}_t$	Complete MFCC feature vector of the observed (possibly corrupted) speech signal, i.e., including dynamic features
$\mathbf{o}_t^{(\mathsf{c})}$, $o_t^{(\mathsf{c})}(\kappa)$	MFCC (vector) of the noisy reverberant speech signal
$\mathbf{o}_t^{(\mathsf{l})}$, $o_t^{(\mathsf{l})}(q)$	LMPSC (vector) of the noisy reverberant speech signal
$\mathbf{o}_t^{(\mathsf{m})}$, $o_t^{(\mathsf{m})}(q)$	MPSC (vector) of the noisy reverberant speech signal
$O_t(k)$	SC of the noisy reverberant speech signal
p, p', p'', p'''	Discrete time indices
$p_{\breve{x}}(x)$	PDF of $\breve{x}$ evaluated at x
$p_{\breve{x}\vert\breve{y}}(x\vert y)$	Conditional PDF of RV $\breve{x}$ evaluated at x given that the RV $\breve{y}$ takes a value of y
$P_{\breve{x}}(x \leq a)$	CDF; probability of the RV $\breve{x}$ taking a value x that is lower or equal to a
$P_{\breve{x}}(x)$	PMF

$P_{\breve{x}|\breve{y}}(x\,|y)$ .... Conditional PMF of RV $\breve{x}$ evaluated at x given that the RV $\breve{y}$ takes a value of y

q, q' .... Mel frequency indices

q_t .... Generic HMM state

Q .... # of triangular-shaped mel filters

$\mathbf{r}_t^{\mathsf{(l)}}$, $r_t^{\mathsf{(l)}}(q)$ .... IRNR (defined in (4.108))

$\mathbf{r}_t^{\mathsf{(l,N)}}$, $r_t^{\mathsf{(l,N)}}(q)$ .... IRNR in the non-recursive observation model after introduction of the AIR model (defined in (4.108))

$\mathbf{r}_{t,L_R}^{\mathsf{(l,R)}}$, $r_{t,L_R}^{\mathsf{(l,R)}}(q)$ .... IRNR in the recursive observation model (defined in (4.201))

$s_t(l)$ .... Reverberant signal after application of the analysis window

$s(p)$ .... Reverberant signal after pre-emphasis

$\mathbf{s}_t^{\mathsf{(l)}}$, $s_t^{\mathsf{(l)}}(q)$ .... LMPSC (vector) of the reverberant speech signal

$\hat{\mathbf{s}}_t^{\mathsf{(l,R)}}$, $\hat{s}_t^{\mathsf{(l,R)}}(q)$ .... Estimate of the LMPSC (vector) of the reverberant speech signal in the recursive observation mode for noisy reverberant speech

$\mathbf{s}_t^{\mathsf{(m)}}$, $s_t^{\mathsf{(m)}}(q)$ .... MPSC (vector) of the reverberant speech signal

$\hat{\mathbf{s}}_t^{\mathsf{(m,R)}}$, $\hat{s}_t^{\mathsf{(m,R)}}(q)$ .... Estimate of the MPSC (vector) of the reverberant speech signal in the recursive observation mode for noisy reverberant speech (defined in (4.190))

$S_t(k)$ .... SC of the reverberant speech signal

$\mathbf{S}_{t,q}$ .... Vector of SCs of the reverberant speech signal falling into mel band q

$\mathcal{S}$ .... Sentence

t,t', t'' .... Frame indices

T .... # of observations

$\tilde{T}$ .... # of samples of the acoustic signal

T_S .... Sampling period

T_{60} .... Reverberation time

$\mathcal{T}_{L_\Delta,L_{\Delta\Delta}}^{\mathsf{(l)/(c)}}$ .... Transformation matrices transforming a sequence of $2(L_\Delta+L_{\Delta\Delta})$ LMPSC/MFCC vectors to the final features used by the recognizer, i.e., static MFCCs with δ and $\delta\delta$ features

U .... # of utterance used for training of the a priori model for speech

u .... Utterance ID

$\mathcal{U}(t)$ .... Upper summation limits used to express the cross-band filters $h_t(k,k')$ in terms of the AIR $h(p)$ and the auxiliary function $\Phi_p(k,k')$ (defined in (4.56))

$\mathbf{v}_{o_t}^{\mathsf{(l)}}$, $v_{o_t}^{\mathsf{(l)}}(q)$ .... Observation error in the observation model for noisy reverberant speech in the LMPSC domain (defined in (4.112))

$\mathbf{v}_{o_t}^{\mathsf{(l,N)}}$, $v_{o_t}^{\mathsf{(l,N)}}(q)$ .... Observation error in the non-recursive observation model for noisy reverberant speech in the LMPSC domain after introduction of the AIR model

$\mathbf{v}_{o_t,L_R}^{\mathsf{(l,R)}}$, $v_{o_t,L_R}^{\mathsf{(l,R)}}(q)$ .... Observation error in the recursive observation model for noisy reverberant speech in the LMPSC domain (defined in (4.200))

$\mathbf{v}_{s_t}^{\mathsf{(l)}}$, $v_{s_t}^{\mathsf{(l)}}(q)$ .... Observation error in the observation model for reverberant-only speech in the LMPSC domain (defined in (4.90))

$\mathbf{v}_{s_t}^{(\mathsf{l,N})}$, $v_{s_t}^{(\mathsf{l,N})}(q)$ .... Observation error in the non-recursive observation model for reverberant speech in the LMPSC domain after introduction of the AIR model

$\mathbf{v}_{s_t,L_R}^{(\mathsf{l,R})}$, $v_{s_t,L_R}^{(\mathsf{l,R})}(q)$ .... Observation error in the recursive observation model for reverberant speech in the LMPSC domain

$\mathbf{v}_{y_t}^{(\mathsf{l})}$, $v_{y_t}^{(\mathsf{l})}(q)$, $\mathbf{v}_{y_t}^{(\mathsf{l,N})}$, $v_{y_t}^{(\mathsf{l,N})}(q)$. Observation error in the observation model for noisy speech in the LMPSC domain (defined in (4.135))

$\mathbf{V}_{\breve{\mathbf{x}}^{(\mathsf{l})}|i}$ .... State prediction error covariance matrix

w_n .... Word at position n of sentence $\mathcal{S}$

$w(p)$ .... Window function resulting from the convolution of the synthesis window and the time-reversed analysis window (defined in (4.169))

$w_{\mathsf{A}}(l)$, $w_{\mathsf{S}}(l)$ .... Analysis/Synthesis window

$\mathbf{w}_{o_t,L_R}^{(\mathsf{l,R})}$ .... Error in the recursive observation model for noisy reverberant speech due to employing only an estimate of the LMPSC of the reverberant speech signal (defined in (4.195))

$x(p)$ .... Speech signal after pre-emphasis

$\mathbf{x}_t$ .... Complete MFCC feature vector of the clean speech signal, i.e., including dynamic features

$\mathbf{x}_t^{(\mathsf{c})}$, $x_t^{(\mathsf{c})}(\kappa)$ .... MFCC (vector) of the clean speech signal

$\mathbf{x}_t^{(\mathsf{l})}$, $x_t^{(\mathsf{l})}(q)$ .... LMPSC (vector) of the clean speech signal

$\mathbf{x}_t^{(\mathsf{m})}$, $x_t^{(\mathsf{m})}(q)$ .... MPSC (vector) of the clean speech signal

$X_t(k)$ .... SC of the clean speech signal

$\mathbf{y}_t^{(\mathsf{l})}$, $y_t^{(\mathsf{l})}(q)$ .... MPSC (vector) of the noisy speech signal

$\mathbf{z}$, z_n .... Generic state vector (component)

$\mathbf{z}_0$ .... Vector the VTS expansion of the observation functions/mappings is carried out at

$\mathbf{z}_t^{(\mathsf{l})}$ .... State vector employed for the inference; it comprises L_C LMPSC vectors of the clean speech signal and, if a noisy scenario is considered, the LMPSC vector of the noise (defined in (4.1))

Greek Symbols

$\boldsymbol{\alpha}_t$, $\alpha_t(q)$ .... (Vector of) phase factor(s) (defined in (4.217))

α_{LMS} .... Language model scale factor

α_{WIP} .... Word insertion penalty

$\beta_{m,l}$ .... Recursion parameter in the derivation of the moments of the phase factor RV $\breve{\alpha}_t(q)$ (defined in (A.223))

$\beta_{t-L_R}(q)$ .... Short-hand notation for the integration limit employed in the estimation of the MPSC (vector) of the reverberant speech signal in the recursive observation mode for noisy reverberant speech (defined in (A.159))

$\boldsymbol{\gamma}_t$, $\gamma_t(q)$ .... Transformed (vector of) phase factors, $\boldsymbol{\gamma}_t = \mathrm{erf}(\boldsymbol{\alpha}_t)$

$\gamma_{m|\mathbf{o}_{1:t}}$, $\gamma_{m|\mathbf{o}_{1:T}}$ .... A posteriori mixture probability

$\gamma_{t,u}^{(\iota)}(i)$ .... A posteriori probability of GMM/MSLDM state i (defined in (4.27))

$\delta(\mathbf{x}-\boldsymbol{\mu})$ .... DIRAC-Delta distribution of the RV $\breve{\mathbf{x}}$, centered at $\boldsymbol{\mu}$ and evaluated at $\mathbf{x}$

δ_p .... KRONECKER-Delta; 1 if $p=0$, 0 otherwise

$\boldsymbol{\varepsilon}$ .... Generic error due to truncating the VTS to linear terms

$\epsilon_t^{(\mathsf{m})}(q)$, $\bar{\epsilon}_t^{(\mathsf{m})}(q)$ .... Error terms in the derivation of the non-recursive observation model for reverberant-only speech in the MPSC (see (4.82))

$\zeta(\cdot)$ .... Auxiliary function related to the IRNR (defined in (4.107))

$\eta_{t,u}^{(\iota)}(j,i)$ .... A posteriori probability of GMM/MSLDM state sequence j,i (defined in (4.28))

$\Theta_{\breve{\mathbf{x}}^{(\mathsf{l})}}^{\mathsf{HMM}}$ .... Set of model parameters for the model named in the superscript, trained on data identified by the subscrip (here an HMM trained on clean speech data)

$\nu_{\breve{x}}$ .... Skewness of the RV $\breve{x}$ (defined in (4.277))

ι .... Iteration counter, EM algorithm

κ .... Cepstral feature coefficient index

$\mathcal{K}$ .... $\#$ of cepstral feature vector coefficients

λ_{ACC} .... Word accuracy

λ_{DEL}, λ_{INS}, λ_{SUB} .... Deletion, Insertion and Substitution rate

λ_{WER} .... Word error rate

$\boldsymbol{\Lambda}_q$, $\Lambda_q(k)$ .... Mel filter coefficient (vector)

$\mu_{\breve{x}}$ .... Mean of the RV $\breve{x}$

$\boldsymbol{\mu}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}$ .... A priori/initial state prediction bias vector

$\nu_{t,k}(q)$ .... Mel weighted cosine of the phase difference between the SC of the reverberant signal and that of the noise (defined in (4.227))

$\xi(\cdot)$ .... Auxiliary function related to the IRNR (defined in (4.106))

$\xi_{\breve{x}}$ .... Excess kurtosis of the RV $\breve{x}$ (defined in (4.276))

π_i .... A priori/initial HMM state probabilities (see (3.39)/(4.14))

$\sigma_{\breve{h}}$ .... Energy parameter in the AIR model (defined in (4.157))

$\tilde{\sigma}_{\breve{h}}$ .... Energy parameter in the AIR model if signals are normalized (defined in (4.158))

$\sigma_{\breve{x}}^2$ .... Variance of the RV $\breve{x}$

$\boldsymbol{\sigma}_{\breve{\mathbf{x}}}^2$ .... Vector of variance of the RV $\breve{\mathbf{x}}$, $\mathrm{diag}(\boldsymbol{\Sigma}_{\breve{\mathbf{x}}})$

$\sigma_{\breve{x},\breve{y}}$ .... Covariance of the RVs $\breve{x}$ and $\breve{y}$

$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}}$ .... Covariance matrix of the RV $\breve{\mathbf{x}}$

$\boldsymbol{\Sigma}_{\breve{\mathbf{x}}_1^{(\mathsf{l})}|i}$ .... A priori/initial state prediction error covariance

ς_l .... Recursion parameter in the derivation of the moments of the phase factor RV $\breve{\alpha}_t(q)$ (defined in (A.219))

τ_h .... Decay constant in the AIR model (defined in (4.154))

Υ .... $\#$ of words in/size of vocabulary Ω

$\Upsilon_{\breve{x}}(\cdot)$ .... Second characteristic function of the RV $\breve{x}$, $\ln(\Phi_{\breve{x}}(\cdot))$

$\boldsymbol{\chi}_t^{(\mathsf{l})}$ .... Augmented state vector comprising the state vector $\mathbf{z}_t^{(\mathsf{l})}$ and the auxiliary vector $\mathbf{a}_t$

$\varphi_{S_t(k),N_t(k)}$ .... Phase difference between the SC of the reverberant signal and that of the noise

$\chi_{L_h}(p)$ .... Binary indicator function, 1 if $0 \leq p \leq L_h - 1$, 0 otherwise

$\Phi_{\breve{x}}(\cdot)$ .... Characteristic function of the RV $\breve{x}$

$\Phi_o(\boldsymbol{\alpha}_t)$, $\Phi_y(\boldsymbol{\alpha}_t)$ .... Substitution functions employed to solve the integrals in Ch. A.3 and Ch. A.4

$\Phi_{w_{1:N_w}}$ .... Set of possible HMM state sequence of length T for the word sequence $w_{1:N_w}$ (defined in (3.34))

$\mathit{\Phi}_p(k, k')$ .... Cross-band/band-to-band filter auxiliary function (defined in (4.53))

Ψ .... # of iterations in the IEKF

ψ .... IEKF iteration counter

ω_v .... Word v in vocabulary of size Υ

Ω .... Vocabulary

List of Figures

List of Tables

Bibliography

[1] M. Gürelli and C. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals", *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

[2] S. Gannot, "Multi-microphone speech dereverberation using eigen-decomposition", in P. A. N. und Nikolay D. Gaubitch, ed., *Speech Dereverberation*, chap. 5, Springer, 2010.

[3] S. Subramaniam, A. Petropulu and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 392–396, Sep. 1996.

[4] M. Triki and D. T. M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction", in *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 173–176, Eindhoven, The Netherlands, Sep. 2005.

[5] M. Delcroix, T. Hikichi and M. Miyoshi, "Precise dereverberation using multichannel linear prediction", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[6] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[7] B. Gillespie, H. Malvar and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3701–3704, Salt Lake City, UT, USA, May 2001.

[8] K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 17, no. 4, pp. 534–545, May 2009.

[9] T. Nakatani, K. Kinoshita and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[10] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, Jun. 2007.

[11] K. Lebart, J. Boucher and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation", *Acta Acustica united with Acustica*, vol. 87, pp. 359–366(8), 2001.

[12] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction", in *Proc. of Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*, pp. 250–254, Veldhoven, The Netherlands, Nov. 2004.

[13] T. Yoshioka, T. Nakatani and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[14] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[15] M. Delcroix, T. Hikichi and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

[16] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification", *Signal Processing*, vol. 82, pp. 1127–1138, Aug. 2002.

[17] A. E. Rosenberg, C.-H. Lee and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification", in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, pp. 1835–1838, 1994.

[18] C. Avendano, S. Tibrewala and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments", in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1107–1110, Rhodes, Greece, Sep. 1997.

[19] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant asr", in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 103–106, Madonna Di Campiglio, Italy, Dec. 2001.

[20] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer, Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] D. Gelbart and N. Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition", in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, pp. 2185–2188, Denver, CO, USA, Sep. 2002.

[22] E. Habets, S. Gannot and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model", *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[23] J. Erkelens and R. Heusdens, "Noise and late-reverberation suppression in time-varying acoustical environments", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4706–4709, Mar. 2010.

[24] M. Wölfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 17, no. 2, pp. 312–323, Feb. 2009.

[25] T. Takiguchi and M. Nishimura, "Acoustic model adaptation using first order prediction for reverberant speech", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 869–872, Montreal, Quebec, Canada, May 2004.

[26] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise", *Speech Communication*, vol. 50, no. 3, pp. 244–263, 2008.

[27] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments", in E. Hänsler and G. Schmidt, eds., *Speech and Audio Processing in Adverse Environments*, Signals and Communication Technology, pp. 679–728, Springer Berlin Heidelberg, 2008.

[28] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 18, no. 7, pp. 1692–1707, 2010.

[29] K. Kumar, *A Spectro-Temporal Framework for Compensation of Reverberation for Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Feb. 2011.

[30] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain", *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, May 2007.

[31] J. Polack, *La Transmission de l'énergie Sonore dans les Salles*, Dissertation, Université du Maine, 1988.

[32] M. Gales and Y.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments", in *Proc. of Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 121 –126, Jun. 2011.

[33] A. Krueger and R. Haeb-Umbach, "A model based approach to joint compensation of noise and reverberation for speech recognition", in R. Haeb-Umbach and D. Kolossa, eds., *Robust Speech Recognition of Uncertain or Missing Data*, chap. 10, Springer, 2011.

[34] L. Deng, J. Droppo and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–829–I–832, May 2002.

[35] S. Haykin, *Kalman Filtering and Neural Networks*, John Wiley & Sons, Inc., 2001.

[36] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, URL http://www.worldcat.org/isbn/0262100665, Jan. 1998.

[37] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, Bristol, PA, USA: Taylor & Francis, Inc., 2nd edn., 2002.

[38] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech", *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[39] H. Hermansky and N. Morgan, "Rasta processing of speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[40] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[41] ETSI, *ETSI standard document, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 V1.1.3 (2003-09).*

[42] ETSI, *ETSI standard document, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050 V1.1.5 (2007-01).*

[43] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., New York, NY., Jun. 2003.

[44] G. Fant, *Acoustic Theory of Speech Production (With Calculation based on X-Ray Studies of Russian Articulations)*, Mouton De Gruyter, 1970.

[45] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, Feb. 1986.

[46] J. A. Bilmes, "Graphical models and automatic speech recognition", in *Mathematical Foundations of Speech and Language Processing*, Springer-Verlag, 2003.

[47] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Information Science and Statistics, Springer, corrected edn., Jul. 2002.

[48] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[49] V. Digalakis, J. Rohlicek and M. Ostendorf, "Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[50] M. Ostendorf, V. Digalakis and O. Kimball, "From hmm's to segment models: a unified view of stochastic modeling for speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, Sep. 1996.

[51] M. Russell and W. Holmes, "Linear trajectory segmental hmms", *IEEE Signal Processing Letters*, vol. 4, no. 3, pp. 72–74, Mar. 1997.

[52] H. Zen, K. Tokuda and T. Kitamura, "Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences", *Computer, Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[53] D. Alspach and H. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations", *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, 1972.

[54] B.-H. Juang, S. Levinson and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.)", *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 307–309, Mar. 1986.

[55] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*, Prentice Hall, 2 edn., 2008.

[56] A. Ogawa, K. Takeda and F. Itakura, "Language modeling for robust balancing of acoustic and linguistic probabilities", in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 246 –253, Dec. 1997.

[57] X. Huang, A. Acero and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st edn., 2001.

[58] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, CSLI - Center for the Study of Language and Information, Stanford, CA,, 1999.

[59] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[60] R. Lippmann, E. Martin and D. Paul, "Multi-style training for robust isolated-word speech recognition", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, pp. 705–708, Apr. 1987.

[61] J. Benesty, M. M. Sondhi and Y. A. Huang, *Springer Handbook of Speech Processing*, Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[62] R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data*, chap. Uncertainty Decoding and Conditional Bayesian Estimation, Springer, 2011.

[63] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 16, no. 5, pp. 1047–1060, 2008.

[64] L. Deng, J. Droppo and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[65] V. Stouten, H. Van hamme and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR", *Speech Communication*, vol. 48, no. 11, pp. 1502–1514, robustness Issues for Conversational Interaction, 2006.

[66] S. Windmann, *Ausnutzung von Inter-Frame Korrelationen in der automatischen Spracherkennung*, Ph.D. thesis, University of Paderborn, 2008.

[67] A. Krüger, *Modellbasierte Merkmalsverbesserung zur robusten automatischen Spracherkennung in Gegenwart von Nachhall und Hintergrundstörungen*, Ph.D. thesis, University of Paderborn, Germany, Dec. 2011.

[68] K. P. Murphy, "Switching Kalman filters", *Tech. rep.*, U.C. Berkeley, 1998.

[69] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, Marcel Dekker, Inc., 2003.

[70] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding", in *Proc. of Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007.

[71] A. Krueger, V. Leutnant, R. Haeb-Umbach, M. Ackermann and J. Bloemer, "On the initialization of dynamic models for speech features", in *Proc. of Informationstechnische Gesellschaft (ITG) Fachtagung Sprachkommunikation*, Bochum, Oct. 2010.

[72] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[73] J. Deng, M. Bouchard and T. H. Yeap, "Noisy speech feature estimation on the Aurora2 database using a switching linear dynamic model", *Journal of Multimedia*, vol. 2, no. 2, pp. 47–52, Apr. 2007.

[74] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book V3.4*, Cambridge University Press, Cambridge, UK, 2006.

[75] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[76] M. R. Portnoff, *Time-scale modification of speech based on short-time FOURIER analysis*, Ph.D. thesis, Massachusetts Institute of Technology, 1978.

[77] S. Farkash and S. Raz, "Linear systems in gabor time-frequency space", *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 611–617, 1994.

[78] J. Droppo, A. Acero and L. Deng, "Uncertainty decoding with splice for noise robust speech recognition", in A. Acero, ed., *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–57–I–60, Orlando, Florida, May 2002.

[79] J. Droppo, A. Acero and L. Deng, “A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies”, in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, 2002.

[80] N. L. Johnson and S. Kotz, *Continuous univariate distributions - I , Distributions in statistics*, Wiley series in probability and mathematical statistics, John Wiley, 1970.

[81] L. Isserlis, “On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables”, *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918.

[82] V. Leutnant, A. Krueger and R. Haeb-Umbach, “A new observation model in the logarithmic mel power spectral domain for the automatic recognition of noisy reverberant speech”, *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 22, no. 1, pp. 95–109, 2014.

[83] V. Leutnant and R. Haeb-Umbach, “An analytic derivation of a phase sensitive observation model for noise robust speech recognition”, in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2395–2398, Sep. 2009.

[84] R. C. van Dalen, *Statistical Models for Noise-Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, 2011.

[85] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th edn., 2002.

[86] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag New York, Inc., 1986.

[87] R. A. Fisher, “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population”, *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.

[88] J. Kenney and E. Keeping, *Mathematics of Statistics II*, D. van Nostrand Company, Inc., 2nd edn., 1951.

[89] M. Paetzold, *Mobile Fading Channels*, John Wiley & Sons, Inc., 2002.

[90] D. R. Brillinger, *Time Series: Data Analysis and Theory*, Holt, Rinehart and Winston, Inc., 1975.

[91] J. B. Allen, “Image method for efficiently simulating small-room acoustics”, *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[92] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary and M. Brookes, “Performance comparison of algorithms for blind reverberation time estimation from speech”, *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 1–4, Sep. 2012.

[93] Y. Bar-Shalom, X. Rong Li and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, John Wiley & Sons, Inc., 2001.

[94] M. Evans and T. Swartz, *Approximating Integrals via Monte Carlo and Deterministic Methods*, OUP Catalogue, Oxford University Press, 2000.

[95] P. J. Moreno, B. Raj and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition", in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. On Conference Proceedings., 1996 IEEE International Conference - Volume 02*, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 733–736, Washington, DC, USA: IEEE Computer Society, 1996.

[96] M. Roth and F. Gustafsson, "An efficient implementation of the second order extended kalman filter", in *Proc. of International Conference on Information Fusion (FUSION)*, vol. 14, IEEE, 2011.

[97] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *The Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, Jun. 1974.

[98] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Oct. 2000.

[99] R. Leonard, "A database for speaker independent digit recognition", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9, pp. 328–331, San Diego, California, Mar. 1984.

[100] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge, UK: Cambridge University Engineering Department, 2006.

[101] J. Droppo, L. Deng and A. Alex, "A comparison of three non-linear observation models for noisy speech features", in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 681–684, Geneva, Switzerland: International Speech Communication Association, Sep. 2003.

[102] H. Liao and M. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition", *Speech Communication*, vol. 50, no. 4, pp. 265–277, 2008.

[103] H.-G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task AU/417/02", *Tech. rep.*, STQ AURORA DSR WORKING GROUP, Nov. 2002.

[104] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus", in *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pp. 357–362, Morristown, NJ, USA: Association for Computational Linguistics, 1992.

[105] H. Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments", *Tech. rep.*, Niederrhein University of Applied Sciences, 2007.

[106] H.-G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems.", in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2697–2700, ISCA, 2005.

[107] V. Leutnant, A. Krueger and R. Haeb-Umbach, "Bayesian feature enhancement for reverberation and noise robust speech recognition", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 21, no. 8, pp. 1640–1652, 2013.

[108] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAM0 A British English speech corpus for large vocabulary continuous speech recognition", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 81–84, IEEE, 1995.

[109] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant and B. Raj, "The reverb challenge: a common evaluation framework for dereverberation and recognition of reverberant speech", in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[110] M. Lincoln, "The multi-channel wall street journal audio-visual corpus (MC-WSJ-AV) Specification and initial experiments", in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 357–362, 2005.

[111] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. Garner and W. Li, "Beamforming with a maximum negentropy criterion", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 17, no. 5, pp. 994–1008, Jul. 2009.

[112] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, Petersen and Pedersen, 2008.

[113] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover, ninth dover printing, tenth gpo printing edn., 1964.

[114] W. P. Johnson, "The curious history of Faà di Bruno's formula", *The American Mathematical Monthly*, vol. 109, pp. 217–234, 2002.

[115] T. H. Savits, "Some statistical applications of Faà di Bruno", *The Journal of Multivariate Analysis*, vol. 97, no. 10, pp. 2131–2140, URL http://www.sciencedirect.com/science/article/pii/S0047259X06000340, 2006.

[116] A. Apelblat, *Table of Definite and Infinite Integrals*, vol. 13 of *Physical sciences data*, Amsterdam: Elsevier Scientific Publishing Company, 1983.

[117] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, Fifth Edition*, Academic Press, corrected and enlarged edition edn., 1979.

Own Publications

[118] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr and T. Yoshioka, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research", 2016.

[119] V. Leutnant, A. Krueger and R. Haeb-Umbach, "A new observation model in the logarithmic mel power spectral domain for the automatic recognition of noisy reverberant speech", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 22, no. 1, pp. 95–109, 2014.

[120] V. Leutnant, A. Krueger and R. Haeb-Umbach, "Bayesian feature enhancement for reverberation and noise robust speech recognition", *IEEE Transactions on Audio, Speech, and Language Processing (IEEE-T-ASL)*, vol. 21, no. 8, pp. 1640–1652, 2013.

[121] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant and B. Raj, "The reverb challenge: a common evaluation framework for dereverberation and recognition of reverberant speech", in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[122] A. H. Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnant and R. Haeb-Umbach, "GMM based significance decoding", in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[123] A. Krueger, O. Walter, V. Leutnant and R. Haeb-Umbach, "Bayesian feature enhancement for ASR of noisy reverberant real-world data", in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, USA, Sep. 2012.

[124] V. Leutnant, A. Krueger and R. Haeb-Umbach, "Investigations into a statistical observation model for logarithmic mel power spectral density features of noisy reverberant speech", in *Proc. of Informationstechnische Gesellschaft (ITG) Fachtagung Sprachkommunikation*, Braunschweig, Sep. 2012.

[125] V. Leutnant, A. Krueger and R. Haeb-Umbach, "A statistical observation model for noisy reverberant speech features and its application to robust ASR", in *Proc. of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Hong Kong, China, Aug. 2012.

[126] V. Leutnant and R. Haeb-Umbach, “Conditional Bayesian estimation employing a phase-sensitive observation model for noise robust speech recognition”, in R. Haeb-Umbach and D. Kolossa, eds., *Robust Speech Recognition of Uncertain or Missing Data*, Springer, 2011.

[127] V. Leutnant, A. Krueger and R. Haeb-Umbach, “A versatile Gaussian splitting approach to non-linear state estimation and its application to noise robust ASR”, in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Florence, Italy, Nominated for "Best Student Paper Award", Aug. 2011.

[128] A. Krueger, V. Leutnant, R. Haeb-Umbach, M. Ackermann and J. Bloemer, “On the initialization of dynamic models for speech features”, in *Proc. of Informationstechnische Gesellschaft (ITG) Fachtagung Sprachkommunikation*, Bochum, Oct. 2010.

[129] V. Leutnant and R. Haeb-Umbach, “On the exploitation of hidden Markov models and linear dynamic models in a hybrid decoder architecture for continuous speech recognition”, in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Tokyo, Japan, Sep. 2010.

[130] V. Leutnant and R. Haeb-Umbach, “Options for modelling temporal statistical dependencies in an acoustic model for ASR”, in *Proc. of Deutsche Jahrestagung für Akustik (DAGA)*, Berlin, Mar. 2010.

[131] V. Leutnant and R. Haeb-Umbach, “An analytic derivation of a phase-sensitive observation model for noise robust speech recognition”, in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, Sep. 2009.

[132] V. Leutnant and R. Haeb-Umbach, “On the estimation and use of feature reliability information for noise robust speech recognition”, in *Proc. of Deutsche Jahrestagung für Akustik (DAGA)*, Rotterdam, Netherlands, Mar. 2009.

[133] S. Windmann, R. Haeb-Umbach and V. Leutnant, “A segmental HMM based on a modified emission probability”, in *Proc. of Informationstechnische Gesellschaft (ITG) Fachtagung Sprachkommunikation*, Aachen, Oct. 2008.

[134] J. Schmalenstroeer, V. Leutnant and R. Haeb-Umbach, “Amigo context management service with applications in ambient communication scenarios”, in *European Conference on Ambient Intelligence (AmI)*, Darmstadt, Nov. 2007.