PADERBORN UNIVERSITY
The University for the Information Society

Faculty for
Computer Science,
Electrical Engineering
and Mathematics

Improving the Listening Experience for SSB-Modulated HF Transmissions Using Neural Networks

der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

M.Sc. Jens Heitkämper

Erster Gutachter: Prof. Dr.-Ing. Reinhold Häb-Umbach
Zweiter Gutachter: Prof. Dr.-Ing. Gerhard Schmidt
Tag der mündlichen Prüfung: 18.01.2023
Diss. EIM-E/370

Abstract

To this day, analog high frequency radio transmissions play a pivotal role in some vital sectors of our society, such as marine, aviation and police communications. However, the listening experience for such recordings is limited due to the highly distorted transmission channel and the dependency on accurate carrier frequency generation. In this thesis, neural network-based solutions to three of the main causes of poor listening quality are presented, using single-sideband modulated speech signals as an example of high frequency analog radio communication. First, a reliable speech activity estimation is presented to avoid the strenuous task of tracking the inactive, noisy high frequency channel. The developed architecture, which to this day achieves the best published results on the public Fearless Steps data, is shown to outperform comparable activity estimators. As a second challenge, a frequency shift in the recorded signal due to a mismatch between the modulation and demodulation frequency is identified. This is a challenge specific to analog transmission systems, which can lead to a highly reduced quality and intelligibility of the recorded speech signal. Here, two network architectures are designed and compared to a state-of-the-art statistical estimator. Third, a source separation network is adapted to extract the speech signal from the noisy recording, which leads to an improvement of more than 20 % in the speech intelligibility metric STOI. The models for the different tasks are combined into one system to improve the signal quality and intelligibility at the receiver in a single step. For all combinations of the presented models a large gain in both intelligibility and speech quality is reported.

Acknowledgments

First and foremost I like to thank my academic supervisor Prof. Dr.-Ing. Reinhold Haeb-Umbach for his support and his patients during all the times I tried to skip to possible solutions without taking the time to appreciate the research already done by the community. Without his support and guidance this thesis in its finale form would not have been possible.

Furthermore, I would like to state my gratitude to all the colleagues in our department who worked with me, even if only for a short time. All of them influenced my work with our (sometimes pointless) discussions about current research, programming and the world, whether at the coffee maker, over lunch or, any other time of the day. The possibility to work with so many brilliant minds helped me tremendously to grow as a researcher.

There are two colleagues whose influence on me and my work I would like to highlight in particular. First of all, thanks to Christoph Boedekker for all his insights into programming, speech processing and all our discussions at times when everybody else has already gone home. I will always treasure our time working together on the CHiME challenge and I am still convinced our ICASSP paper should have been accepted. A special thanks to Dr.-Ing Joerg Schmalenstroeer for his guidance during my later projects and the possibility to grow building on his original work. I am already missing our banter and morning discussions about research and the world. Your insights into signal processing and our competitions between statistical and neural network based processing have made my success possible, especially during the Fearless Steps Challenge.

Of course none of this work would have been possible without the love and support of my family. I owe everything I am and everything I will be to you. Thank you all for supporting me and believing in me.

Last but not least, I would like to thank all the people who make my life worth living. Due to space limitation I cannot give everybody the acknowledgment they deserve. However, rest assured that I am grateful to all of you for allowing me to be part of your life. In now specific order, thank you to the GaWlers for all the amazing years, thank you to every member of the A-Cup family for all the unforgettable moments, thank you to my VBC team for giving me the distractions I need and finally, thank you to everybody who stuck with me during my time at the University.

In short, thanks to everybody who made the last years to such an unforgettable time in my life.

Contents

1 Introduction

In today's communication systems, analog radio transmissions are increasingly replaced by their digital counterpart. However, analog modulation techniques are still of interest for aircraft, marine and police communication. Despite these important use cases, only a few studies in recent years dealt with the improvement of listening quality in recordings of high frequency transmissions, e.g., [1] and [2].

This thesis aims to offer modern approaches based on deep learning to solve some of the main challenges in analog high frequency (HF) communication. Most of the systems discussed in this thesis are based on neural networks (NN), as research in the last decades has shown the benefits of this data-driven approach for various tasks. A few examples for advances based on NNs are the design of an automatic speech recognition (ASR) model on-par in performance with human transcribers on some databases [3], a close to perfect separation of a mixture of multiple speakers [4] and the development of a speaker recognition systems with an error rate close to zero even for a large number of speakers [5]. Some of these algorithms have also shown impressive feats in real world applications like home assistants [6], chatbots [7] and more.

Three of the key unresolved challenges of analog HF communication are explored in this work, i.e., speech activity detection (SAD), noise reduction (NR) and carrier frequency offset (CFO) estimation and correction. A CFO is the frequency difference between the frequency

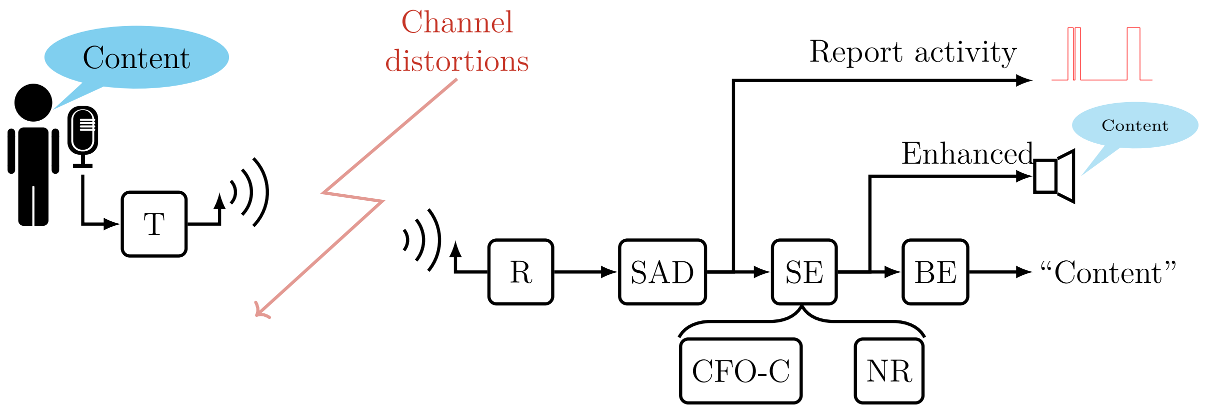


Figure 1.1: Illustration of a multi-stage speech signal processing system for analog HF recordings, where SE includes both the CFO correction (CFO-C) and noise reduction techniques. T represents the transmitter, R the receiver and BE stands for a back-end, e.g., an ASR system.

generated in the modulation oscillator in the transmitter, which is used to modulate the baseband signal to the transmission band, and the frequency generated in the demodulation oscillator of the receiver, which is used to demodulate the signal back to the baseband. These three tasks are chosen as example for challenges in HF signal processing because they are interdependent, and a reliable solution to each task provides a high impact with respect to the listening experience. Furthermore, a combination of these systems can be used as a front-end for subsequent processing like speaker recognition or ASR. An example of such a processing pipeline is displayed in Figure 1.1.

One possible application for the discussed system is an assistance system for air traffic control: The SAD component alerts the ground personnel at an airport of possible activity in a frequency band, while the speech enhancement systems could reduce intelligibility issues. Since most remaining applications of analog HF have strong safety requirements, only systems optimized for the use-case with a very low error rate can be considered. To achieve this goal with the help of neural networks, two steps must be taken. First, a database for training and evaluation has to be designed, which is as close as possible to the described use-case. The second step is the creation of network architectures specific to the tasks and data.

To lead through these steps and give an introduction into speech processing on analog HF data, the thesis is structured as follows. First, Chapter 2 provides a brief introduction to the basics necessary to understand the particularities of analog HF transmissions. Then, the current work in SAD is presented in Section 3.1 and in Section 3.2 an overview over possible approaches for CFO estimation is given. Afterwards, the state-of-the-art in speech enhancement (SE) is discussed in Section 3.3, focusing on single channel noise reduction as a special case of source separation for two sources. In **??**, a database of real analog HF recordings is described and a simulation tool is presented to supplement or replace the real data during training.

Chapter 5 and Chapter 6 introduce and evaluate the architectures and baselines for the SAD and CFO correction task, respectively. For both tasks statistical and NN-based systems are discussed. Although the focus of this work lies on NN for speech processing, a statistical baseline system can often be considered. A description of the SE system is given in Chapter 7, and the performance of the described system is evaluated for both source separation on simulated two speaker mixtures and noise reduction on HF transmitted signals. In Chapter 8, the individual systems for SAD, CFO correction and noise reduction are combined and evaluated. Finally, a conclusion and an outlook for future research on these topics are given in Chapter 9.

2 High Frequency Radio Transmission

Analog high frequency (HF) transmissions date back to the late 19th century when Tesla demonstrated the first wireless transmission system [8]. A few years later, the first high frequency communication system was presented by Marconi [9, p. 4]. Over the years more and more complex transmission systems were deployed leading to the first mobile telephone system in 1946 [9, p. 4]. The following years saw some breakthroughs with the publication of Shannon's work [10] and the development of the cellular principle [8]. Many of today's wireless communication devices can be attributed to techniques developed in the 1970s [11, p. 2]. Wireless communication stayed a focal point of research and in the 1990s the digital transmission techniques started to replace the analog systems [9, p. 6]. Digital transmissions offer many benefits like higher speech quality, higher flexibility, data encryption and more [12, p. 5]. Although these strengths of the digital systems lead to the mostly digital communication system of our days, there are still tasks for which analog transmission is an important tool, like marine and aircraft communication [2]. Analog transmission is still used for many of these tasks due to its low power demands for long distance communication [13].

During the last decades the design of both the analog and digital HF radio systems evolved by shifting the focus from hardware to software development [14, p. 3]. The resulting design idea is called software-defined radio and allows for more flexibility of these communication systems by controlling some or all of the physical layer functions via software [15, cha. 1].

In a radio communication system an information or baseband signal is modulated with a carrier frequency F_{M} which is very high compared to the baseband bandwidth [11, p. 255]. Afterwards, the modulated signal, i.e., the bandpass signal, is transmitted via an antenna. At the receiver the recorded signal is demodulated using a predefined carrier frequency F_{D}, which is as close as possible to the modulation frequency.

An illustration of a communication system is depicted in Figure 2.1. The next sections discuss each of the stages of the systems individually and introduce the signal properties of the resulting speech recordings. Finally, an overview over current databases for research on analog HF speech signals is given.

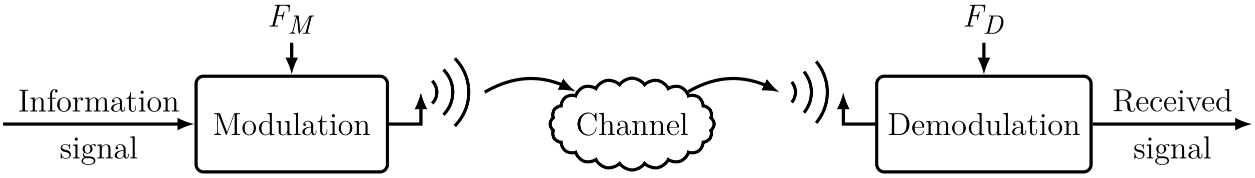


Figure 2.1: Illustration of a transmission system.

2.1 Modulation

Over the years, a host of different modulation techniques has been investigated. The two most common approaches for analog transmissions are frequency and amplitude modulation [11, p. 255]. While frequency modulation offers a higher noise immunity than amplitude modulation, it also requires a larger bandwidth [11, p. 256].

In this work, only single-sideband (SSB) modulation is considered, which is a specific form of amplitude modulation. The SSB method was derived in 1915 by John R. Carson [16]. Historically, SSB was first introduced to the broader public in 1923, when a trans-Atlantic radio telephone demonstrated the capability of SSB modulation [17, p. 1-2]. Although SSB profits from a lower bandwidth and power requirements than other amplitude modulation approaches like double sideband (DSB), the higher dependency on frequency stability and selectivity in the receiver delayed its rise to prominence in HF transmission until the 1950s [18, p. 3]. In 1990 SSB was declared the international standard for HF broadcasting and was recommended by the International Telecommunication Union (ITU) [19]. This declaration was withdrawn in 2012 but SSB is still a standard for maritime mobile service [20].

SSB is a form of amplitude modulation, where the hermitian symmetry of the Fourier transform of a real-valued signal is exploited so that the modulated signal occupies the same bandwidth as the original baseband signal [11, p. 260]. This is worth mentioning because other amplitude modulations like DSB require twice this bandwidth [21, p. 176].

The SSB signal can be mathematically expressed using the analytic signal s_t^+ of the baseband signal s_t [22] with

$$s_t^+ = s_t - \mathrm{j} \cdot \mathrm{H}\{s_t\} = s_t - \mathrm{j} \cdot s_t * \frac{1}{\pi t}, \tag{2.1}$$

where $\mathrm{H}\{\cdot\}$ is called the Hilbert transform [23] and $*$ is the convolutional operator. Applying the ideal Hilbert transform to a signal leads to a 90° phase shift between the real and imaginary part of the signal. The combination of the signal and its phase-shifted counterpart results in a spectrum of the analytic signal which is zero for all negative frequencies:

$$S_\omega^+ = S_\omega - \mathrm{j} \cdot (\mathrm{j} \cdot \mathrm{sgn}(\omega) \cdot S_\omega) = S_\omega + \mathrm{sgn}(\omega) \cdot S_\omega = \begin{cases} 2 \cdot S_\omega & \text{if } \omega \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.2}$$

Similarly, a signal $s_t^- = s_t + \mathrm{j} \cdot \mathrm{H}\{s_t\}$ is derived with a spectrum S_ω^- that is zero for all positive frequencies.

For the SSB modulation the analytical signal is shifted to a higher frequency resulting in the so called bandpass signal

$$\begin{aligned} x_t^{\mathrm{SSB}} &= \sqrt{2}\mathcal{R}\{\ s_t^+ \exp(\mathrm{j}2\pi \cdot F_{\mathrm{M}} \cdot t)\} \\ &= \sqrt{2}s_t \cos(2\pi \cdot F_{\mathrm{M}} \cdot t) - \sqrt{2}H\{s_t\} \sin(2\pi \cdot F_{\mathrm{M}} \cdot t) \end{aligned} \tag{2.3}$$

for a sinusoidal carrier, where $\mathcal{R}\{\cdot\}$ reduces a signal to its real part. The frequency spectrum of the bandpass signal for a modulation frequency F_{D} greater than the bandwidth of the baseband signal can be written as

$$\begin{aligned} X_\omega =& \frac{1}{\sqrt{2}} \left(S_\omega * \left(\delta(\omega - 2\pi F_{\mathrm{M}}) + \delta(\omega + 2\pi F_{\mathrm{M}}) \right) \right. \\ &\left. + \left(\operatorname{sgn}(\omega) \cdot S_\omega \right) * \left(\delta(\omega - 2\pi F_{\mathrm{M}}) - \delta(\omega + 2\pi F_{\mathrm{M}}) \right) \right), \qquad (2.4) \\ =& \sqrt{2} \begin{cases} S_\omega^+ * \delta(\omega - 2\pi F_{\mathrm{M}}) & \text{for } \omega \geq 0 \\ S_\omega^- * \delta(\omega + 2\pi F_{\mathrm{M}}) & \text{otherwise} \end{cases}. \qquad (2.5) \end{aligned}$$

Since S_ω^+ and S_ω^- only require half the bandwidth of S_ω, X occupies the same bandwidth as S_ω.

SSB modulation is a carrier suppressed modulation because the carrier does not influence the average power of the bandpass signal x_t^{SSB}. This is easily verified by calculating the power of the bandpass signals using Parseval's theorem [21, p. 42].

The equation above describes a specific form of the SSB modulation called upper sideband (USB) because only the spectrum for the positive (upper) frequencies of the baseband signal are modulated. From this analytical signal the negative frequencies of the spectrum can be calculated using the hermitian symmetry of the Fourier transform for real-valued signals. If only the spectrum for the negative (lower) frequencies of the baseband are modulated, the approach is called lower sideband (LSB) modulation [11, p. 260]. This modulation can be achieved by replacing s_t^+ with s_t^- in Equation (2.3), which leads to a modulated signal with a spectrum that only contains S_ω^- in the positive frequencies. An illustration of the described USB modulation steps for a sinusoidal carrier is depicted in Figure 2.2.

There are different approaches to generate a SSB signal. One is called balanced modulator and can be considered a direct implementation of Equation (2.3) [11, p. 261]. Another method sends a double sided amplitude modulation signal through a bandpass filter to remove the unwanted sideband [11, p. 260]. A third approach called Weaver method uses an intermediate frequency to remove the negative sideband with a lowpass filter [24].

After modulating the information signal, it is transmitted over a HF channel via an antenna.

2.2 Transmission

In this work all signals are transmitted over HF bands, where HF is a designation for frequencies between 3 and 30 MHz according to a recommendation by the ITU [25]. The considered HF transmissions are propagated either as ground or sky waves [26, p. 260]. Here, ground waves imply a transmission to the receiver by reflection at the surface without refraction in the ionosphere. These ground waves can travel a distance of up to 100 km for specific antennas and surfaces [27].

Sky waves travel to the ionosphere, where they are refracted by collision with free ions and diverted back to earth. Although this propagation allows the waves to travel farther distances,

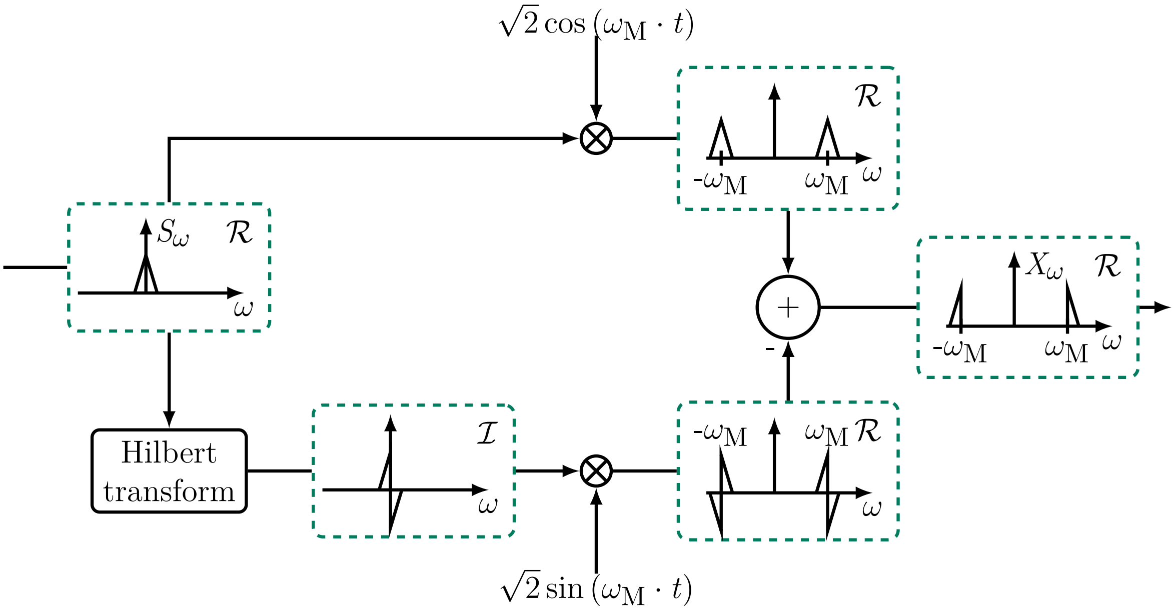


Figure 2.2: Illustration of USB modulation with a sinusoidal carrier for a baseband signal with a positive, real-valued triangle function as Fourier transform and $\omega_\mathrm{M} = 2\pi \cdot F_\mathrm{M}$. The green boxes show the Fourier transform of the signal. The symbol in the corner of each box symbolizes the displayed signal part (real: $\mathcal{R}$ or imaginary: $\mathcal{I}$)

it also leads to increased variation of the transmission conditions [26, p. 69]. These conditions are mainly influenced by the state of the ionosphere, which is dependent on solar activity [17, ch. 11]. For further information regarding the propagation of HF in the ionosphere refer to [26, ch. 4].

Both propagation modes can lead to the received signal consisting of multiple scaled and delayed reflections of the original transmission, which is called “multi-path” [21, p. 113]. These distortions can be represented by a finite impulse response filter [28, p. 246], similar to the distortion in recordings of reverberant environments [29]. After the transmission, the next step is the demodulation and signal reconstruction at the receiver.

2.3 Demodulation

The first stage of any HF receiver is an antenna with a subsequent amplifier [21, p. 288]. Afterwards a generic SSB receiver has to fulfill three tasks, which are frequency selection, demodulation and amplification [21, p. 288]. For most radio receivers these tasks are solved using the super heterodyne principle [21, p. 288]. However, with the rise of software defined radios, the zero-IF approach has gained attention [15, p. 11]. In contrast to the super heterodyne principle the received real-valued bandpass signal y_t^BP is not translated to an intermediate frequency (IF), but directly demodulated to the complex-valued analytic baseband signal y_t^+ with the in-phase and quadrature (IQ) components as the real (I) and imaginary part (Q) [15, p. 12]

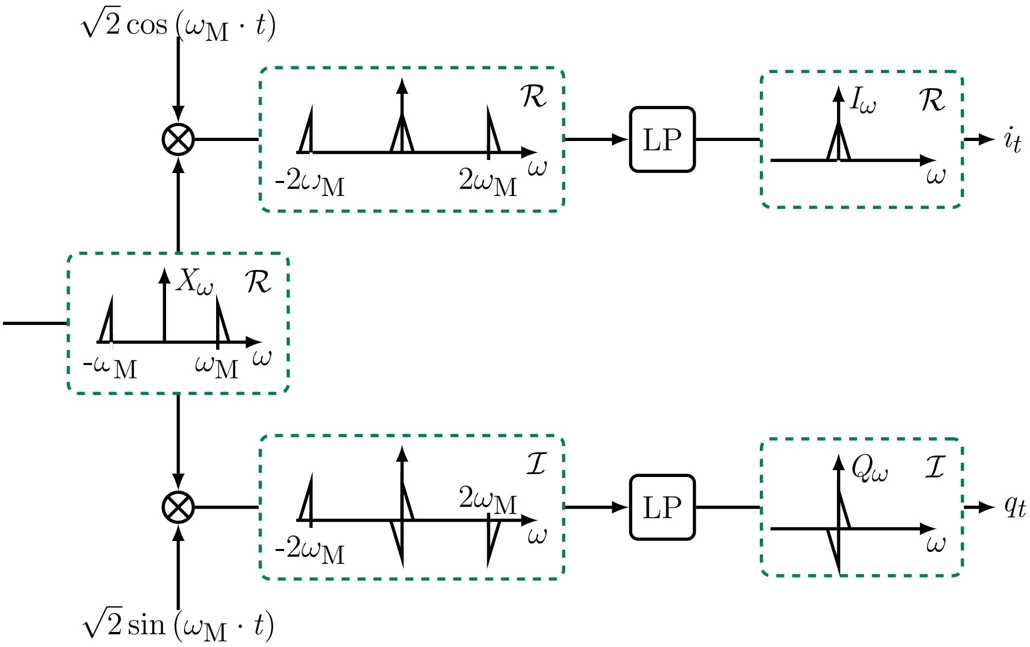


Figure 2.3: Illustration of an idealized zero-IF demodulation with a positive, real-valued triangular function as Fourier transform and $\omega_M = 2\pi \cdot F_\mathrm{M} = 2\pi \cdot F_\mathrm{D}$. The green boxes show the Fourier transform of the signal. The symbol in the corner of each box symbolizes the displayed signal part (real: $\mathcal{R}$ or imaginary: $\mathcal{I}$)

$$i_t = \mathrm{LP}\left\{y_t^{\mathrm{BP}} \cdot \cos\left(2\pi \cdot F_\mathrm{D} \cdot t\right)\right\}, \quad (2.6a) \qquad q_t = \mathrm{LP}\left\{y_t^{\mathrm{BP}} \cdot \sin\left(2\pi \cdot F_\mathrm{D} \cdot t\right)\right\}, \quad (2.6b)$$

where LP$\{\cdot\}$ is a low pass filter with a cut-off frequency below $F_\mathrm{D} - F_\mathrm{M}$ and F_D is the demodulation frequency. This approach is especially beneficial for software defined radios compared to the digitization of the signal at the IF, since digitizing the IQ signal reduces the maximum frequency of the signal and thus the necessary sampling rate, resulting in a reduced power consumption at the analog-to-digital converter [15, p. 12].

After demodulation the IQ-signals are further processed by an amplifier with automatic gain control (AGC) to reduce fluctuations in their amplitude [17, p. 3-7]. A generic AGC consists of a comparator that tracks the difference of the energy of the output signal with a reference level to adjust the gain applied to the input signal [14, p. 29]. Therefore, it can be represented by an infinite impulse response filter [14, p. 31]. In software defined radios, the AGC is especially tasked to maintain the signal amplitude at a level compatible with the analog-to-digital converter (ADC) to prevent non-linear distortions caused by signal clipping [14, p. 29].

The transmitted signal can be reconstructed by phase-shifting q_t by 90° and adding it to i_t in case of USB and by subtracting the phase-shifted quadrature from the in-phase component for LSB modulation [21, p. 293]:

$$y_t = \mathcal{R}\left\{i_t \mp \mathrm{j} q_t\right\}. \quad (2.7)$$

Reducing the complex-valued basisband signal to its real part forces the signals values for positive and negative frequency to be identical, thereby removing all additional information located in the negative frequencies. In case of perfect reconstruction no information is lost since previous steps already moved all information to the positive frequencies.

One reason for not digitizing the real-valued reconstructed signal but the IQ component is that different analog and digital modulation types can be recovered from the IQ signal (e.g. LSB, USB, quadrature modulation [28, p. 88]) and some errors in the demodulation frequency can be rectified. This is partly due to missing information in the negative frequencies which would be lost if only the real-valued signal is stored. An illustration of the zero-IF demodulation with perfect knowledge of the modulation frequency ($F_{\mathrm{D}} = F_{\mathrm{M}}$) and a perfect channel, i.e., $y_t^{\mathrm{BP}} = x_t^{\mathrm{SSB}}$, is displayed in Figure 2.3.

All prior discussions assume perfect knowledge of the modulation frequency and an idealized oscillator for the carrier signal generation for both modulation and demodulation. In such an idealized scenario the quadrature component q_t of an SSB signal is not required to reconstruct the transmitted signal but it can be used to reject interfering signals from the opposite sideband [21, p. 292].

After the signal is demodulated, it can now be further manipulated by subsequent signal processing. The next section gives an overview over the challenges of HF signals for further processing steps.

2.4 Signal properties

An analog HF recording is a challenging signal because of distortions due to multi-path propagation, concurrent transmissions and non-linearities in the transmission and receiver system. The non-linearity in the systems may lead to so called intermodulation distortions, which can be represented by scaled repetitions of the original signal with a shifted center frequency [21, p. 114]. In the demodulated signal these repetitions overlap with the original signal or appear as its harmonics [14, p. 38]. Further non-linear distortions are caused by the amplitude range compression of the AGC during demodulation.

Possible concurrent transmission in neighboring or overlapping HF channels can be modeled as an additive distortion. Similarly, the imperfect HF channel and the transmission as well as the receiver systems can be modeled as additive distortions. Another challenge for many further processing steps and the human listener is the limited frequency range of the transmitted signal, which is set at 2.7 kHz following ITU regulations for the considered HF links.

Another distortion is a shift in the signal frequency spectrum called carrier frequency offset (CFO), which is caused by imperfect knowledge about the modulation frequency or small deviations between the frequency oscillator during modulation and demodulation ($F_{\mathrm{D}} = F_{\mathrm{M}} - \mathit{\Delta}^{\mathrm{f}}$). Because of the high modulation frequencies, even small deviations from the original frequency can lead to a large frequency shift. For example an oscillator with an instability of 25 ppm in a transmission over a 7 MHz channel can lead to a CFO of $7 \cdot 10^6\,\mathrm{Hz} \cdot 25 \cdot 10^{-6} = 175\,\mathrm{Hz}$. Reasons for such a deviation in the oscillator frequency can be variations in the electrical characteristics of the receiver or transmission components due to mechanical vibrations or temperature changes [18, p. 92]. Even with perfect knowledge of the modulation frequency and perfectly stable oscillators small CFOs can be caused due to the high variability of the ionosphere [17, ch. 11] or a Doppler shift in the signal due to moving receiver or transmitter, which is a common problem for airplane communication [18, p. 91]. The frequency shift associated with a small CFO can already lead to reduced intelligibility of the demodulated signal [2]. Furthermore, in some applications, for example amateur radio, only rough information about the transmission frequency is known so that the CFO can get very large. For further discussion of CFOs refer to Section 3.2.

Even the general structure of the conversations over HF can be challenging for many signal processing algorithms, as there are long silence intervals between activities and speech segments, which can consist of only a few words [OC5]. These short activities are due to the limited frequency range shared with other transmitting parties. Adherence to the described transmission style is sometimes called radio discipline.

An advantage for further processing is that the transmitted signal is mostly recorded by a headset or similar microphone close to the speaker's mouth. Therefore, the final recording is not distorted by reverberation or masked by environmental additive noise prior to the modulation.

After discussing the HF signal in general, the next section deals with some of the HF data currently used for research on HF signals.

2.5 Databases

During the last decade, the release of two databases has led to an increased research interest in signal processing on analog HF transmission data: The Robust Automatic Transcription of Speech (RATS) program [30] and the database released for the Fearless Steps challenges [31]. For the RATS data a clean signal is transmitted over eight different channels with different bandwidths and modulations types, leading to a range of degradation in the recorded signal. The Fearless Steps database, on the other hand, contains recordings of the communication between the ground personnel and the astronauts from the Apollo 11 mission [31].

On these data multiple publications have presented statistical and neural network (NN)-based systems for a variety of tasks, including speech activity detection (SAD) [OC4], [32], [33], speaker identity detection (SID) [34], speech enhancement (SE) [2], language recognition [35], automatic speech recognition (ASR) [36], [37], and more. However, these databases cannot be used for most of the supervised NN-based speech enhancement systems discussed in the next chapter, since no paired data is provided. Paired data means the recording of both the degraded and the underlying clean speech signal, which is the target signal to be estimated by an enhancement system. For simulated signals, paired data often includes the distortion signal that is added to the clean speech to construct the observation.

Another problem not considered in these databases is a possible mismatch between the modulation and demodulation frequency. While the RATS database includes some data with differences in the modulation and demodulation frequencies, the Fearless Steps database does not explicitly include CFOs because the modulation frequency was always fixed and known during the Apollo mission. Nevertheless this mismatch is a common problem in HF transmissions, as discussed above.

In this thesis, a new database is developed which provides high frequency radio transmission data with paired clean signals. To evaluate the influence of a CFO on the introduced systems, both recordings with and without a CFO are included in the database. Furthermore, transmissions in both English and Russian language are recorded to investigate the impact of different languages on systems like NNs that are reliant on appropriate training data. In **??** the resulting database is described in more detail.

After describing the HF transmissions, the next chapter offers an overview of some possible subsequent signal processing systems. To be specific, current approaches for SAD, CFO estimation and SE with a focus on NN-based systems are discussed.

3 Related Work

Speech signal processing using neural networks (NN) has been a focus of research for the last decade. During this time, the field experienced several breakthroughs, most notably the effective application of neural networks in automatic speech recognition. This advancement led to NN-based automatic speech recognition (ASR) systems with a human level accuracy on some databases [3].

Although researchers have been working on NNs since the 1940s [38, p. 12] with renewed attention since the 1980s [38, p. 16] due to the invention of the back-propagation algorithm [39], NN were not a common tool for everyday applications. The breakthrough for a broader application only started in the years after the development of the greedy layer-wise pretraining strategy [40] in 2006, after which an increasing number of systems relied on NNs [41]. The possibilities of applying NNs to increasingly complex applications while outperforming many former state-of-the-art algorithms, grew with the amount of available data and the increasing computational capacity [38, p. 18-21].

This trend could also be witnessed in speech processing where by now a variety of tasks have seen NN-based approaches surpassing the statistical systems. Examples for this rise of neural networks in speech processing can be found in speech activity detection (SAD) [OC4], speech enhancement (SE) [42], diarization [43], source separation [44] and so on. However, ASR still is the most famous application of NNs in speech processing [41], [45]. Other well known NN-based speech processing tasks include speech synthesis [46] and speaker recognition [47].

One possibility to appreciate the importance of NNs in speech processing is the high number of recent speech processing challenges where the top systems are heavily utilizing NNs. Example for this trend are the CHiME [48], [49] and Reverb challenges [50] for robust speech recognition, the VOXCELEB challenge for speaker recognition [47], the Dihard challenges for diarization [51], the Deep Noise Supression Challenge for speech enhancement [42], the Fearless Steps challenges for SAD, speaker identification, diarization and ASR [31] and many more. A lesson learned from these challenges is that although NN-based systems can solve most of the presented tasks on clean speech signals, distortions like low signal to noise ratio (SNR) or reverberation still lead to a significant performance loss [OC3], [52], [53].

In this work the potential of neural networks for processing speech signals that have been transmitted over high frequency (HF) channels is shown on the example of SAD, carrier frequency offset (CFO) estimation and noise reduction (NR). Therefore, new architectures for SAD and CFO estimation are developed and some well known NR networks are extended. For all three tasks, the models are comprehensively evaluated and compared with appropriate baseline systems. All considered NNs consist of a combination of convolutional neural network (CNN) [54] and recurrent neural network (RNN) [39] layers, which are two of the most popular architectures for speech processing with NNs [45]. For the RNNs both gated recurrent units (GRUs) [55] and long short-term memory (LSTM) [56] layers are considered.

The following sections give an overview over the current research for SAD, CFO estimation and NR.

3.1 Speech activity detection

Speech activity detection (SAD) has been a subject of research for years, especially as a preprocessing step of other, more elaborated systems like ASR [57, p. 662], SE [58] or diarization [59]. SAD, also known as voice activity detection, describes a binary classifier that declares either speech activity or inactivity for a signal segment. A common approach for SAD systems consists of an initial feature extraction followed by a speech presence probability (SPP) estimator and a subsequent decision stage [60]. A block diagram of such a SAD system is displayed in Figure 3.1.

Features are a representation of the information in the input signal calculated by a transformation function. There is a multitude of possible features for SAD, ranging from the well-known short time Fourier transform (STFT) [61] to mel frequency cepstral coefficients and cochleagram-features [62]. Some systems use multiple feature extraction algorithms and combine the resulting features to allow the classification algorithm to exploit different views of the data [33], [63].

The currently popular SPP estimators can be split into two categories: statistical [61], [64], [65] and deep neural network (DNN)-based [33], [66] approaches. In recent years the DNN-

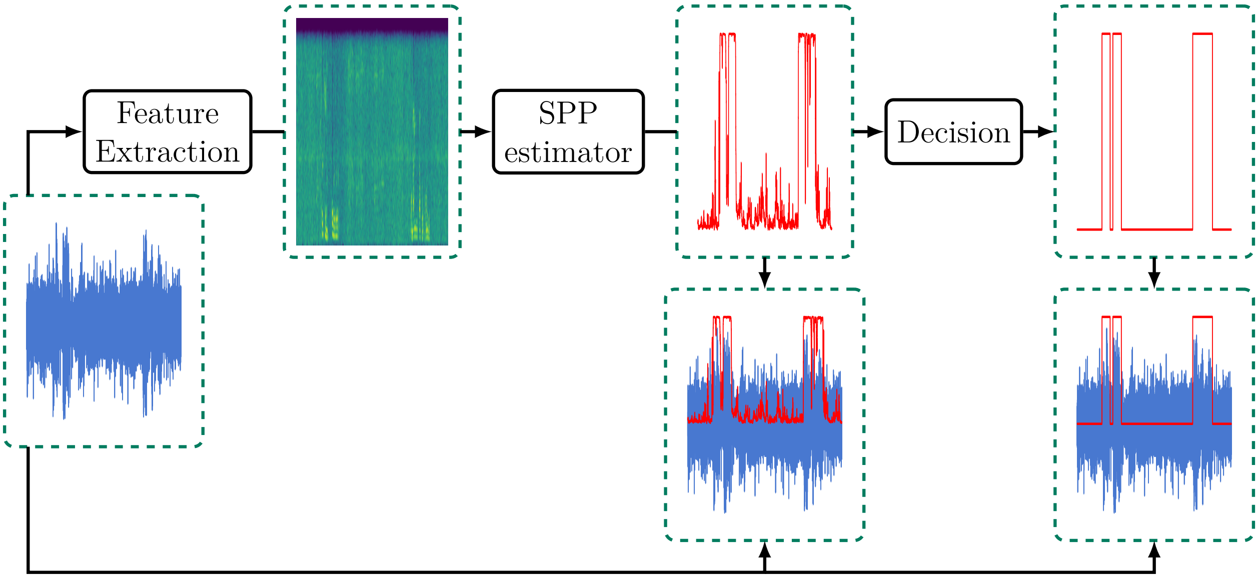


Figure 3.1: Block diagram of a SAD system with example output. The lower images show an overlay of the estimation over the input signal.

based approaches have consistently shown to outperform the statistical SAD if appropriate training data is available [OC4], [67], [68]. However, most statistical SAD systems are less computationally complex and are more robust to changes in the data statistics due to their unsupervised nature.

The decision stage for both NN-based and statistical SAD can be either a simple thresholding or a more complex system with automatic smoothing. Many SPP systems have a high variance in the output, which leads to a high variance in the estimated activity in case of a simple thresholding. To reduce the variance, which does not correspond to the reality of the activity in most speech signals, the decision may include smoothing the original estimate. A simple example for such a smoothing algorithm enforces a minimum duration of speech activity. Since a human requires around 0.1 s to produce a single phoneme [69, p. 298], the minimum length of activity segments detected by a SAD system should be at least that long. There are various ways to achieve a smoothed estimation using a simple median filter, a finite state transducer [70], smoothing the estimation by pooling over the time dimension in one of the NN layer [OC4], et cetera. For simplicity only median and NN internal smoothing will be considered in this work.

The statistical SAD systems may be divided into two groups. One focuses on a combination of multiple feature extraction algorithms to enable a back-end classification algorithm to exploit different views of the information [63]. Other systems use basic feature extractions schemes like the STFT and focus on a multi-step classification system [61]. The statistical baseline system used in this work for the SAD [OC4] can be assigned to the second group.

Although there are many competitive statistical SAD systems, this work focuses on NN-based SAD. Various NN-based SAD approaches have been published over the last decade [32], [66]–[68]. Depending on the database different architectures and input feature have led to highly accurate SAD results. A common NN architecture for SAD consists of a CNN followed by a RNN to capture temporal information, and a fully-connected (FC) classification layer [68], [71], [72]. Most of these networks rely on different variations of frequency domain features like the magnitude of the STFT [73, p. 94], Mel frequency cepstral coefficients (MFCC)- or cochleagram-features [62], [74]. Some systems rely on a combination of different frequency domain features to provide the system with alternative representations of the signal [33]. However, these approaches increase the pre-processing pipeline and require large input layers to map the feature combination to a latent space. Additionally, most of these feature combinations do not provide additional information but provide the network with different views of the same information [62]. Therefore, the network could learn to map one kind of input feature to the appropriate latent representation without a large number of different features. Some approaches rely on a CNN to extract the features directly from the time domain signal [75], although previous publications have shown such a feature extraction may result in less robust features [76], [77]. For this work, only the STFT features are considered to reduce the scope of the investigation. Additional information like visual features [78] or a-priori speaker embeddings [79] are not considered since such information can in general not be provided for high frequency radio transmissions.

The prior discussions are mostly independent of the specific data on which the systems are running. However, there are some publications regarding SAD on high frequency radio transmission as part of the Robust Automatic Transcription of Speech (RATS) program [30] and the Fearless Steps Challenge [31]. For these databases both statistical and NN-based SAD systems were presented [OC4], [32], [33]. However, the SAD results achieved on these databases are not directly transferable to the data considered in this work since they assume perfect knowledge of the carrier frequency during demodulation in case of the Fearless Steps database or only small deviations between modulation and demodulation frequency for RATS. Nevertheless, the SAD system developed as part of this thesis, achieved the lowest error rate of all contributions to the Fearless Steps 2020 SAD challenge [80], as published in [OC4]. The introduced system even outperforms the best performing contributions [81] to the follow-up Fearless Steps 2021 SAD challenge [82].

In the next section the state-of-the-art systems for CFO estimation on single-sideband (SSB) transmissions are discussed which are applied to speech regions identified by a previous SAD system.

3.2 Carrier frequency offset estimation and correction

As described in Section 2.4 one source of errors in SSB transmissions is a difference between the modulation and demodulation frequency, which is also called carrier frequency offset (CFO). A CFO leads to a shift in the frequency spectrum of the demodulated signal and therefore to a deterioration of the signal, reducing intelligibility and thus increasing listeners fatigue [83].

Figure 3.2 depicts an example for a recording with a 500 Hz difference between modulation and demodulation frequency. One can see that the speech spectrum is shifted to the higher frequencies, which leads to a change in pitch. This affects the gender and speaker information in the signal [84] and thereby also degrades many automatic speaker and language identification systems trained on signals without a CFO [2]. A more vivid description of the resulting effect is to characterize the recorded speech as “chipmunk-like” [85] or “like a duck speaking” [2]. This effect can already be witnessed at CFO greater than 5 Hz [85] and

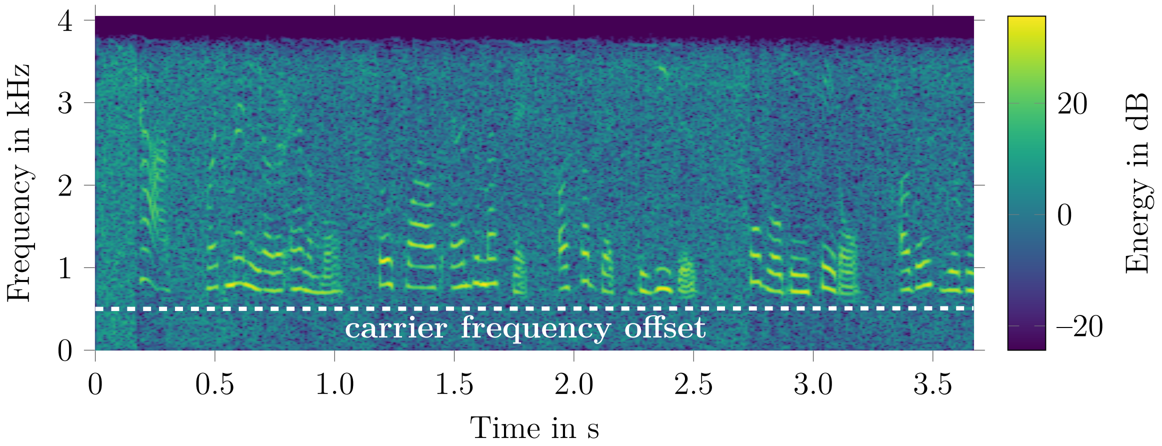


Figure 3.2: Spectrogram of an analog HF recording in case of a 500 Hz carrier frequency offset.

significantly decreases the speech quality for CFO above 10 Hz [86]. These effects are the result of a shift of the speech signal to higher frequencies or a positive CFO. For a negative CFO the energy in the lower frequencies of the speech signal is lost during demodulation and with it most of the speech energy, which is located in the lower frequencies [87]. Therefore, a reconstruction of the original transmitted signal from the demodulated signal in the case of a CFO is possible only for positive or very small negative CFOs.

The following algorithms for CFO estimation only consider positive frequency shifts to allow for a CFO correction after the estimation. There are various approaches to estimate a CFO $\varDelta^{\mathrm{f}}$ by exploiting the statistical properties of speech [88]. Many of those systems estimate the pitch and harmonic frequencies of the signal to calculate the CFO as in [83], [88]. Others use a third-order modulation spectral analysis [85] or a two-step approach combining a rough estimate of $\varDelta^{\mathrm{f}}$ using supervised learning algorithms with a subsequent fine-tuning step [2].

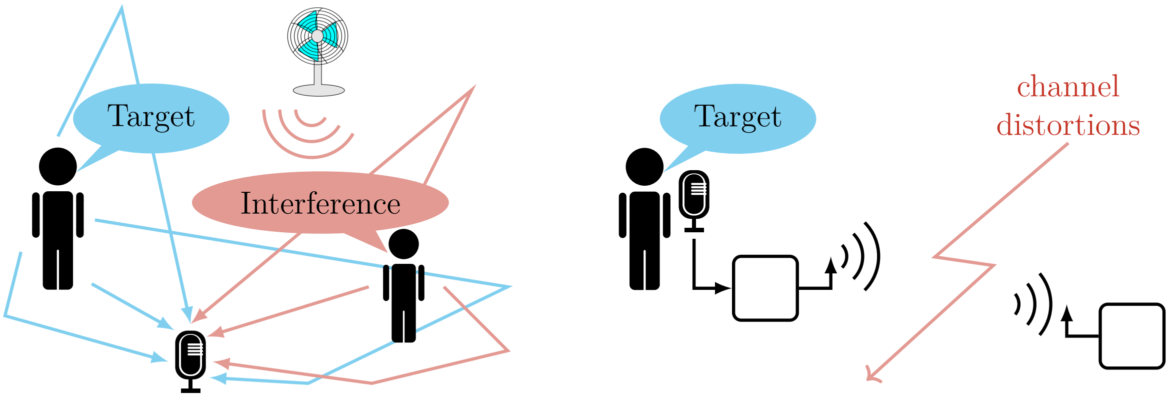
This work presents a statistical and two NN-based CFO estimation algorithms. All three algorithms are compared and the differences in performance are discussed. Parts of these investigations have been published in [OC15] and [OC6]. Both the statistical and NN-based approaches operate in the STFT domain with a high frequency resolution. The first NN-based CFO estimator segments the spectrum along the frequency dimension and processes each segment independently in NN layers with shared parameters. Afterwards the calculated signals are combined and further processed to classify the input signal regarding its CFO. The second NN architecture considered consists only of CNNs and estimates, for each frequency bin, whether the speech signal spectrum is shifted to this bin using the information from a limited number of adjacent frequency bins.

In comparison, the statistical algorithm, called RAKE, filters the power spectrum of the signal with a filterbank to emphasizes the harmonics in the speech signal for different CFOs and pitches. Afterwards, the pitch and CFO are chosen dependent on the filtered power spectrum. As a side note, the RAKE algorithm can also be used to track the pitch of the active speaker [OC15].

A more thorough explanation of both the RAKE and the NN-based CFO estimators is given in Chapter 6. After the CFO is estimated the signal spectrum can be shifted to correct the CFO as discussed in Section 6.3. Most of the enhancement systems discussed in the following sections are developed for signals without a frequency shift. Therefore, the CFO correction is a prerequisite for an application of these algorithms on SSB recordings.

3.3 Speech enhancement

Speech enhancement is a generic term, which includes research areas like noise reduction, source separation and dereverberation [89, ch. 1.2]. In all these areas, the enhancement system is tasked to improve the audio quality and/or intelligibility of a speech signal. There is a multitude of applications for speech enhancement including communication, hearing aids, voice-controlled human-machine interfaces and many speech processing tasks which profit from a high input signal quality [89, ch. 1]. One can classify the speech enhancement



(a) Standard scenario

(b) High frequency scenario

Figure 3.3: Illustration of two different speech enhancement scenarios.

task by the number of microphones used during recording as multi-channel, binaural and single-channel, also called monaural SE [89, ch. 1.2].

Monaural speech enhancement can be traced back more than 60 years and started gaining more attention in the 1970s [90]. Over the years, different techniques like spectral subtraction [91], Wiener filtering [92], [93], signal subspace decomposition [94], computational auditory scene analysis [95] or parametric [96] and statistical [97] model based enhancement were developed. Additionally, a host of multi-channel approaches like independent vector analysis [98], multi-channel non-negative matrix factorization [99] and others [29] were invented, which utilize spatial information to improve the enhancement results. In the 2010s, the sequential modeling capacity of RNNs was shown to improve NN-based enhancement systems [100], [101]. This led to an increased attention on NN-based enhancement systems, culminating in NN-based systems outperforming the common approaches on multiple databases [102]–[105].

In Figure 3.3 two distortion scenarios are depicted. The first is a scenario commonly associated with the enhancement task, where the distortions are introduced during propagation of the speech signal to the microphone [57, p. 844]. Therefore, the recorded signal includes reverberation, environmental noise and overlapping speaker interference. In the second image a single person is speaking directly into a microphone and the speech is transferred to the listener. Here, the distortions are introduced during transmissions and may include atmospheric distortions, multi-path fading, channel noise and distortion due to the receiver automatic gain control (AGC), as well as the modulation and demodulation. A possible third scenario is near-end listening enhancement, where a person is listening to clean speech on a mobile device in the presence of background noise [106]. In this work only the second scenario is discussed. Therefore, neither reverberation nor speaker overlap have to be considered.

As a possible solution for the NR tasks, current systems developed for source separation are used to reduce additive noise in the observed signal. Here, noise reduction is considered a special case of source separation for two sources. There are multiple, recent publications that show the benefits of NN architectures developed for source separation on the NR tasks [107]–

[109]. In [108], it is shown that training a network to calculate an estimate for both the speech and noise signal outperforms a network trained to estimate only the clean speech signal.

The following sections give a short overview of current systems for source separation and NR to offer an intuition for the network architectures discussed in Chapter 7.

3.3.1 Source separation

In this section the current state of the art in source separation or more specific blind source separation (BSS) is discussed. BSS is the task of extracting the speech signals of an unknown number of speakers from an audio recording without knowledge of the mixing situation [110, p. 47]. A common BSS scenario is called the “Cocktail Party Problem” [111], where a conversation between multiple participants in a noisy environment is recorded. During recent years NN-based approaches outperformed the conventional enhancement systems on most single channel databases with low reverberation [112], [113]. Especially for BSS, NNs achieved some remarkable results on data with simulated overlap [114]–[116].

Many databases like the WHAMR [117] database, the SMS-WSJ [118] database, the data recorded for the CHiME-5 challenge [49], and other databases [119]–[122] were published to further research in this area. For all of these databases there are NN-based [117] and/or statistical model based systems [OC8] which achieve strong separation results. However, many of these systems rely on spatial, i.e., multi-channel, information to improve the prediction. Some approaches directly use spatial features as input for a NN [121], others combine the modeling power of a NN with spatial statistical models [123], [124] or directly utilize the strength of spatial statistical models [OC8], [125]. In this work no spatial information is available.

Most BSS approaches suffer from the so called “permutation problem” [114] which describes the ambiguity of the mapping between the NN output and the target signals. For the network all output orders might be equally valid, if no further information about the desired order is provided. In other words, if a network outputs two separate signals $\hat{x_1}$ and $\hat{x_2}$ for a mixture of two sources x_1 and x_2, $\hat{x_1}$ is not always an estimate of x_1 because the order of the network outputs need not match the arbitrary order of the target sources. There is a similar issue for statistical models if statistical independence over the frequency dimension is assumed [126]. For this frequency permutation problem one solution uses a permutation alignment [126] to rearrange the estimated output. For BSS with neural networks the permutation of the output signals can be addressed by the design of the loss function. In recent years different loss functions were proposed like deep clustering [114], deep attractor networks [127] and utterance-level permutation invariant training (u-PIT) [115]. u-PIT describes a neural network with a loss function that is calculated for different permutations of the output and target signal. The permutation with the lowest loss is chosen to calculate the gradients and update the model. This approach is called utterance-wise because it is solved for a whole utterance and not frame-wise as in the original publication [113], which prevents a permutation of the speakers over the time dimension,. For a BSS system with a mixture of two overlapping speakers as input the loss can be written as

$$\mathcal{L}^{\mathrm{PIT}}\left(x_1, x_2, \hat{x}_1, \hat{x}_2\right) = \min\{\mathcal{L}\left(x_1, \hat{x}_1\right) + \mathcal{L}\left(x_2, \hat{x}_2\right), \mathcal{L}\left(x_1, \hat{x}_2\right) + \mathcal{L}\left(x_2, \hat{x}_1\right)\}, \tag{3.1}$$

with $\mathcal{L}(\cdot)$ as an arbitrary distance measure for the estimated and target signal.

Commonly, the NN-based BSS systems operate in the frequency domain with the STFT as input transformation. The network then estimates an activity mask for each source on the magnitude of the frequency spectrum. This mask is multiplied with the input STFT coefficients to calculate an estimate for each source signal in the input mixture. These estimates are computed in the frequency domain, and the separated signals are transformed back to the time domain using the inverse STFT to derive the final estimates. The networks are either trained as a classification network to estimate activity for each time frequency bin similar to a prior defined oracle mask or as a regression network by comparing the enhanced signals to the target signals in the frequency domain [115]. These systems have shown to achieve strong separation results for monaural data [115], [128]. However, such a system just enhances the magnitude of the complex-valued input signal and use the noisy phase for the reconstruction of the time-domain signal as suggested in [129]. Thereby, the possible enhancement is limited due to the distorted phase as shown in [130]. Some systems try to solve this problem by either directly estimating the complex-valued signal [131] or having an explicit network for phase estimation [132], [133].

The current state-of-the-art systems for BSS on non-reverberated data [4], [44], [134] are variants of the time-domain audio separation network (TasNet) architecture [116], which transforms the time domain signal into a real-valued, learned latent domain which is optimized for separation. During training the u-PIT method in combination with a time domain loss function is used to train the network. One reason for switching from the frequency to the latent domain is that the phase enhancement problem discussed above can be prevented since the latent domain signal is real valued.

There are some extensions to the original TasNet which have shown to further improve the separation results. Some of the extended networks use an additional layer, for example to estimate speaker embeddings as in [4], others change the separation network as in [44], [134]–[136]. For this work we assume that no speaker information is available during training and to reduce the scope of the investigation only the extensions introduced in [44] and [135] are considered.

In Figure 3.4 both frequency and latent domain separation are illustrated to show both the similarities and differences between the approaches. Both networks are trained to estimate a mask to separate the encoded signal and perform the separation in a transformed domain. The main difference between the networks is the domain and the loss function. For the frequency domain separation networks the loss is mostly calculated in the frequency domain [114], [115], [127] except for more recent publications [OC3], [137], whereas the latent domain separation mostly use time domain loss functions [116]. One exception is the so called two-step training presented in [138] where the separation network is trained with a loss in the latent domain.

Although, strong separation results can also be achieved in the frequency domain [115], [128], [139] the latent domain networks outperform other architectures for non-reverberated

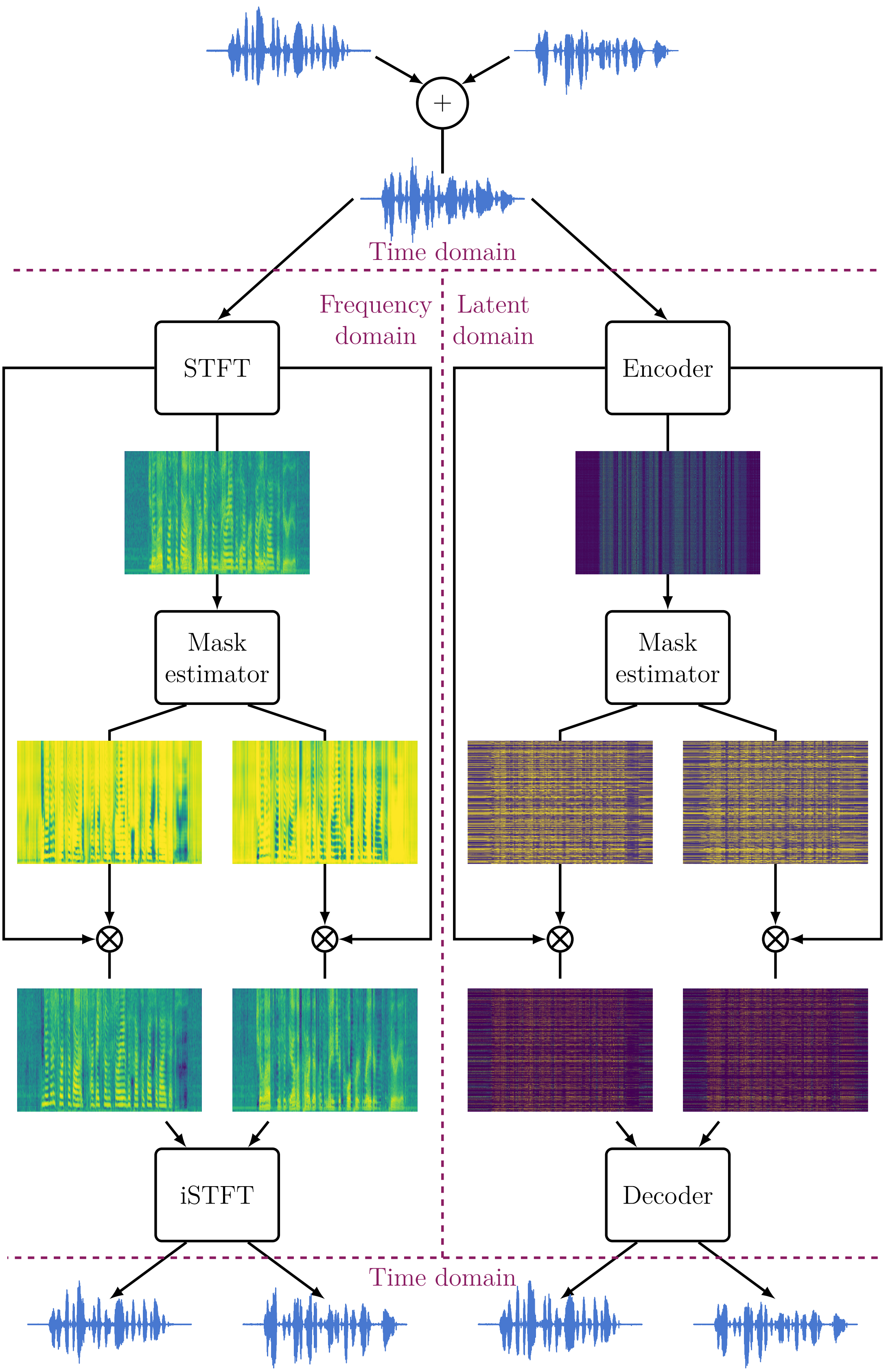


Figure 3.4: Illustration of source separation in case of two sources with a frequency domain network compared to a network operating in the latent domain.

monaural data [135]. As part of this thesis, a combination of frequency domain separation and the TasNet architecture, inspired by the experiments in [137], is introduced and evaluated. A subset of these experiments is published in [OC3], which show that the TasNet struggles in case of reverberation, where it is still outperformed by the frequency domain networks. This decline in performance can be mitigated using spatial information as in [137], [140]–[142]. However, this solution is not possible for the monaural HF data. In general, our results indicate that the TasNet is more susceptible to non-additive distortions than the frequency domain separation. Since the HF transmission includes many non-linear distortions from both the channel and the input AGC this work compares both frequency and latent domain networks.

However, two simultaneous speakers in one transmission are not a common scenario for HF signals. Therefore, this work focuses more on NR and in specific the application of NNs designed for BSS on NR, which will be discussed in the next section.

3.3.2 Noise reduction

Similar to BSS the NN-based approaches for NR started to gain attention in the early 2010s [143] and started to surpass the conventional systems in the following years. For example, in 2015 the strongest enhancement for the CHiME-3 challenge [144] was a statistical model [145]. One year later, the prior best result was surpassed by a neural network based enhancement [146] during the CHiME-4 challenge [48].

Some of the first NN-based NR systems tried to emulate conventional systems consisting of well tuned statistical models [OC7], [145] with subsequent filtering using beamforming vectors or a Wiener filter by replacing part of the systems with a neural network [105], [146]. Other systems rely on the conventional algorithms to guide the network as part of the loss function [147] or use the NN as initialization for a statistical model [148]. A third kind of network directly estimated the clean signal from the noisy observation with a regression model [149] without the aid of conventional approaches. In recent years, more and more NR systems especially for monaural NR rely on these single models with a regression loss to allow for direct training on the clean source signal [150], [151]. In [1] such a regression network is applied to simulated SSB recordings, where the results indicate the benefits of NN-based enhancement for speech signals transmitted over HF channels.

As discussed earlier, NR can be viewed as a special case of source separation with two sources, one of which is the desired speech signal and the other the noise signal. Therefore, NN architectures that have shown to be effective for source separation can be transferred to NR. However, NR does not have the “permutation problem” described in Section 3.3.1 because noise and speech statistics can be considered distinct for most noise types. Therefore, the network is able to learn a fixed mapping between the estimates and targets without requiring u-PIT or other BSS specific loss functions.

While most published NR approaches have shown to achieve strong noise suppression in the frequency domain [149], recent studies have shown the effectiveness of NR in a latent domain with a regression loss computed in the time domain [107], [108], [150]. In [152] different strengths and weaknesses are identified for both latent and frequency domain NR. Latent domain NR networks are shown to lead to higher intelligibility improvements, while the evaluated frequency domain models generalize better to previously unseen data.

In this work, the combination of frequency domain separation and the TasNet architecture we introduced for BSS is applied to the NR task. To illustrate the benefits of each component of the system the evaluation from [152] and [OC3] are performed on simulated and real HF recordings. First, the steps from frequency to latent domain NR are examined to allow a comparison of their importance for the NR results. Furthermore, the time and frequency domain systems are evaluated on their performance on previously unseen data. Whether a system is able to generalize well to unseen data is especially important for speech enhancement on HF recordings since the HF data may contain a high variety of noise types as discussed in Section 4.6. Therefore, the network has to generalize to noise signals previously unseen during the training. Furthermore, the systems are trained on simulated data so that there is a large mismatch between training and evaluation independent of the varying noise types.

To evaluate the performance of a NN in any of the discussed topics first the training and evaluation data have to be specified. Therefore, the next chapter gives an overview of the recorded HF database and the simulation tools used to generate additional training data.

4 High Frequency Radio Database

The objective of this work is to evaluate neural networks (NN) for processing speech signals transmitted over high frequency radio channels. Therefore, a first step is the selection of appropriate data reflecting the main challenges of the application, which are

1. non-stationary noise,
2. errors in the demodulation frequency,
3. long silence intervals between speech activity,
4. short speech segments with only a few words.

As part of this thesis, more than 23 hours of clean speech were transmitted, recorded and processed to create a database representing the challenges of high frequency transmissions. In [OC5] a subset of the recorded data is published to support research in high frequency (HF) signal processing.

This chapter gives an overview over the database, its creation and limitations. In the design of the database setup, one focus is the automatic synchronization of the recordings with the transmitted clean speech signal. The availability of both the synchronized transmitted clean speech and received noisy signal is also referred to as paired data and is important for the noise reduction (NR) performance evaluation as discussed in Section 7.5.2. Finally, a simulation framework is presented to augment the recorded data with simulated signals containing demodulation frequency errors.

4.1 Transmission

The recordings that make up the database are created by automatic transmission from an amateur radio station at Paderborn University in Germany. The HF transmissions were received in parallel from several Kiwi-SDR stations [153] in Germany and other European countries and transferred back to Paderborn as a data stream using web services. Kiwi-SDRs are software defined radios with a frequency range of 10-30 MHz that allow users around the world to download their received signals. The transmission and recording scheme is depicted in Figure 4.1.

All transmitted signals are single-sideband (SSB) modulated with a bandwidth of 2.7 kHz following the International Telecommunication Union (ITU) recommendation [19]. As discussed in Section 2.1 as well as Section 2.3, lower sideband (LSB) and upper sideband (USB) modulation are closely related, so that the signals and errors after demodulation

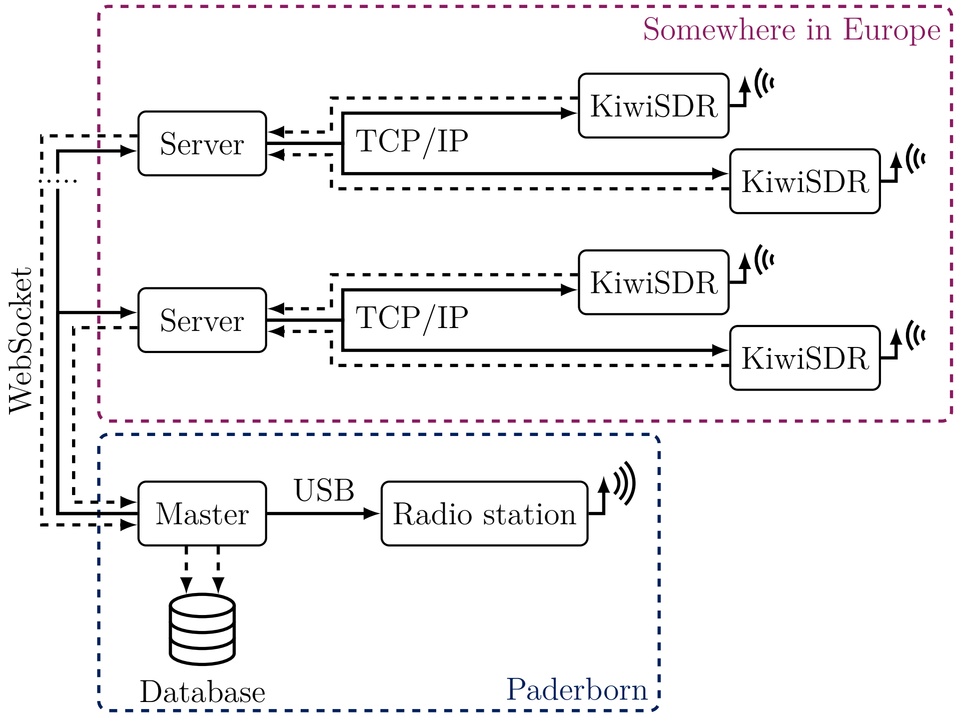


Figure 4.1: System for distributed recording of radio signals.

are fairly similar. Therefore, only USB modulation is considered to reduce the size of the recorded database without affecting the generalizability of the results.

In Figure 4.2 the signal processing steps from the original to the received signal and the bandwidth and sampling rate of the intermediate signals are depicted. First the clean 16 kHz signal s_n is downsampled to 8 kHz. Before modulation, the signal has to be further band-limited to comply with the regulations for amateur radio transmissions set by the ITU, which limits the bandwidth to 2.7 kHz per transmission. Afterwards, the band-limited signal $\tilde{s}_n$ is processed by a digital-to-analog converter (DAC) to generate the analog signal $\tilde{s}_t$ with t as the time index, which is then SSB-modulated at carrier frequency F_M in the range of 7.05 MHz - 7.053 MHz or 3.6 MHz - 3.62 MHz resulting in the bandpass signal x_t^SSB. These frequencies lie in the frequency band reserved for amateur radio transmissions [154].

During transmission additive noise, multi-path propagation, fading and other distortions lead to a deteriorated signal dependent on the channel b_t and the additive noise $\tilde{d}_t$.

The received bandpass signal y_t^BP is demodulated at frequency $F_\mathrm{D} = F_\mathrm{M} - \mathit{\Delta}^\mathrm{f}$ where $\mathit{\Delta}^\mathrm{f}$ symbolizes a deviation of the demodulation from the modulation frequency, i.e., a carrier frequency offset (CFO). As part of the demodulation block the reconstructed signal is processed by an automatic gain control (AGC) to reduce fluctuations in its amplitude as discussed in Section 2.3. Afterwards, the received baseband signal $\tilde{y}_t$ is sampled at 12 kHz resulting in $\tilde{y}_n$.

The demodulated signal can be further downsampled to 8 kHz without loosing speech information, since the transmitted signal is already band-limited to 2.7 kHz due to the ITU regulations. After these steps the final recording y_n is obtained.

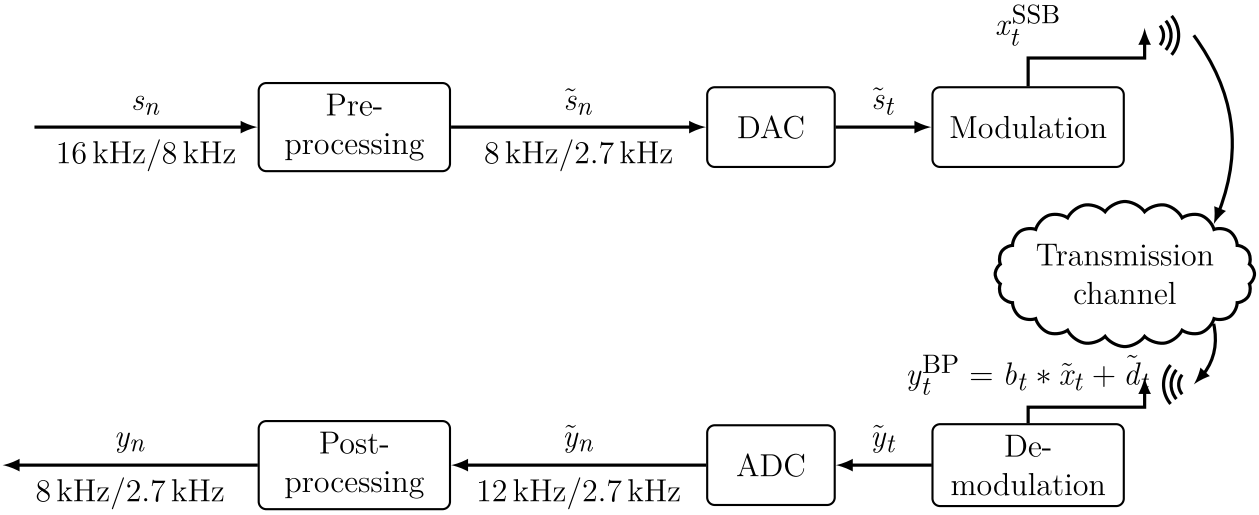


Figure 4.2: Block diagram of the signal processing steps from the clean signal s_n to the received signal y_n with sampling rate and maximum frequency.

Note, that different Kiwi-SDR stations may receive the transmitted signal with different degradation depending on the propagation conditions as discussed in Section 2.2.

4.2 Clean signal design

To model real communications between two parties on a HF link, the emitted signal has to reflect common speech patterns of such radio communications. As discussed in Section 2.4 many usage patterns over HF channels consist of long pauses between activities and speech segments that are only a few words long. Additionally, the used speech signals have to be recorded with a low amount of reverberation and a high signal to noise ratio (SNR). This ensures that the signal degradation observed at the receiver is exclusively due to the transmission channel. Since no freely available database currently meets these requirements, suitable speech signals are generated from the clean training subset of the LibriSpeech corpus [155].

The speech signal to be transmitted is compiled from five excerpts taken from different utterances from the LibriSpech data, each with random start and end points. All excerpts have a duration ranging from 1 to 8 s. To ensure that the random start and end points are in silence segments of the utterance, both are adjusted by taking phoneme alignments generated by an automatic speech recognition (ASR) system into consideration. If a phoneme label representing non-silence is aligned with either the start or the end sample of the excerpt, the duration of the excerpt is increased until both the start and end sample are aligned with a silence label. The phoneme alignments are calculated using a hidden Markov model (HMM)-Gaussian mixture model (GMM) system for ASR trained with the baseline script for LibriSpeech provided in the Kaldi toolkit [156], [157].

The excerpts are concatenated with larger silence segments in-between to reflect real recordings of high frequency transmission, which usually do not occupy a channel permanently to observe

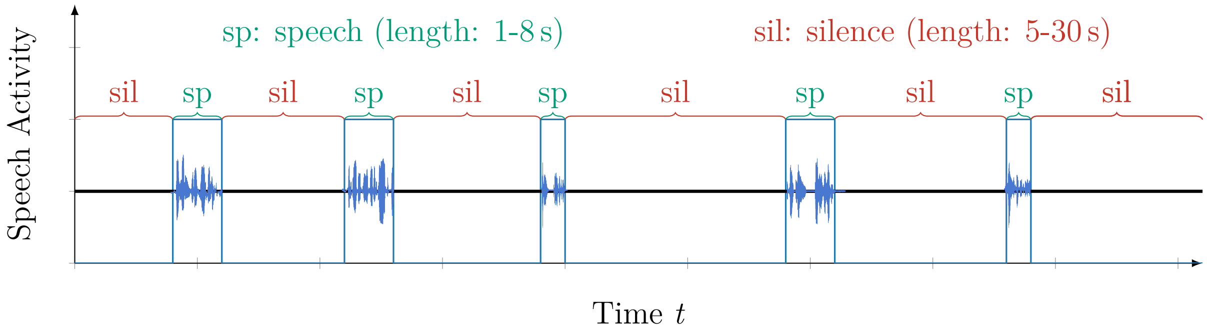


Figure 4.3: Exemplary structure of an audio signal sequence. Labels "sp" and "sil" represent speech and silence, respectively.

radio discipline. Each LibriSpeech utterance chosen for one audio sequence is spoken by a different speaker. In Figure 4.3 an example of the structure of an audio sequence is shown.

4.3 Data preparation

For the database design, solutions for three key tasks have to be devised. How to:

1. annotate speech activity in the received signal y_n,
2. synchronize the clean and recorded signals s_n and y_n to generate paired data,
3. evaluate automatically for each station whether the transmitted signal was received.

The first task is related to the second. If the clean and received signals are synchronized the speech activity on y_n can be annotated by estimating the activity on each of the clean sequences. Here, the activity information for the clean signal is derived by calculating forced alignments using a HMM-GMM acoustic model in Kaldi [157] and assuming speech activity for all samples, which are not aligned with the phoneme label representing silence. Thereby, highly accurate activity labels for 10 ms windows are generated, which are later used for neural network training and speech activity detection (SAD) evaluation.

To enable a robust fine synchronization of the transmitted and received signal a marker is added before and after each audio signal sequence. Each marker consists of an initial sine signal, followed by a 4 s sequence of 26 chirp symbols with different starting frequencies and orientations (ramp-up or ramp-down). To ensure orthogonality between the markers a gold code [158, p. 82] is used as encoding for the chirp symbol combination. The initial sine leads to a ramp-up of the AGC in the Kiwi-SDR, so that the following chirps are processed with a more constant gain.

5 s of silence are added between the marker and audio sequence to mitigate the effects of the marker on the Kiwi-SDR's AGC reaction to the following audio sequence. Common transmissions are not preceded by a warning of the upcoming transmission and therefore the gain change due to the marker would result in an unrealistic recording. The added silence allows for a readjustment of the gain to its original level prior to receiving the marker and

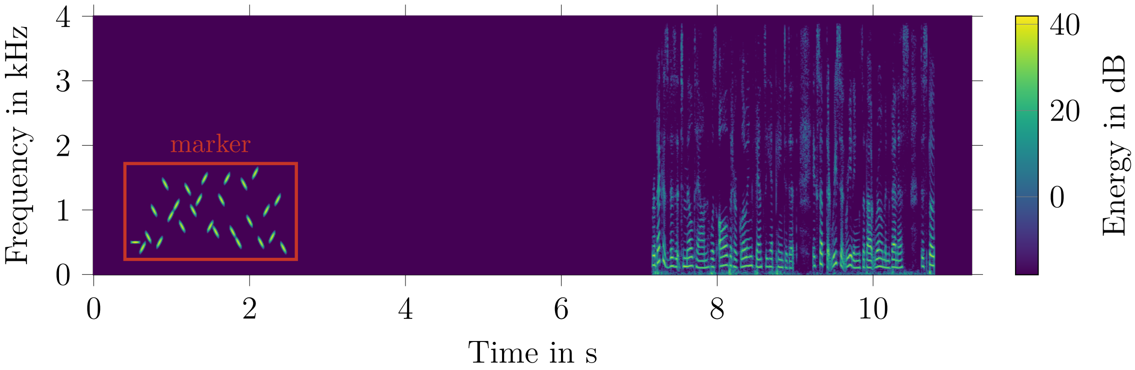


Figure 4.4: Example of a chirp sequence, followed by the start of an utterance from the LibriSpeech corpus.

therefore leads to a more realistic recording. An additional 1 s of silence is included after the audio sequence to ensure that no overlap occurs between the audio sequence and the end marker. Each transmission consists of $N_S = 5$ of the described audio sequences with preceding markers. The beginning of such a transmitted signal is depicted in Figure 4.4. The recording starts with a marker, shown between 0.5 s and 2.5 s.

The markers described above can also be used to automatically detect whether the transmitted signal was received at a specific Kiwi-SDR station. To this end, the order and temporal position of the detected markers are compared with the emitted signal. Only if both the order and temporal position of all markers coincide with their expected value do we assume that the signal is received by the Kiwi-SDR station.

However, all presented solutions to the three tasks posed at the beginning of this section depend on a robust detection of the markers in the received signals. This detection will be explained in more detail in the next section.

4.4 Marker detection

To detect the markers in the received signal, it is first transformed to the frequency domain using the short time Fourier transform (STFT) with a 16 ms shift and a window size of 40 ms. Here, the active time-frequency bins of the marker are used to define a binary mask which is shifted along the time dimension of the received signal spectrum. For all shifts the correlation between the mask and the received spectrum is calculated and the shift with the maximum correlation is chosen as temporal location of the marker. Therefore, the time resolution of the marker location is limited by the 16 ms shift between STFT windows. Both the length of 4 s and the orthogonality of the markers allow for a highly precise estimation even in low SNR conditions.

To keep only valid recordings both the number and order of markers in all recordings have to match the original markers. Furthermore, the time difference between markers is compared to the original transmission and the recording is discarded if they do not match. These sanity checks lead to a highly accurate time synchronization between the transmitted and

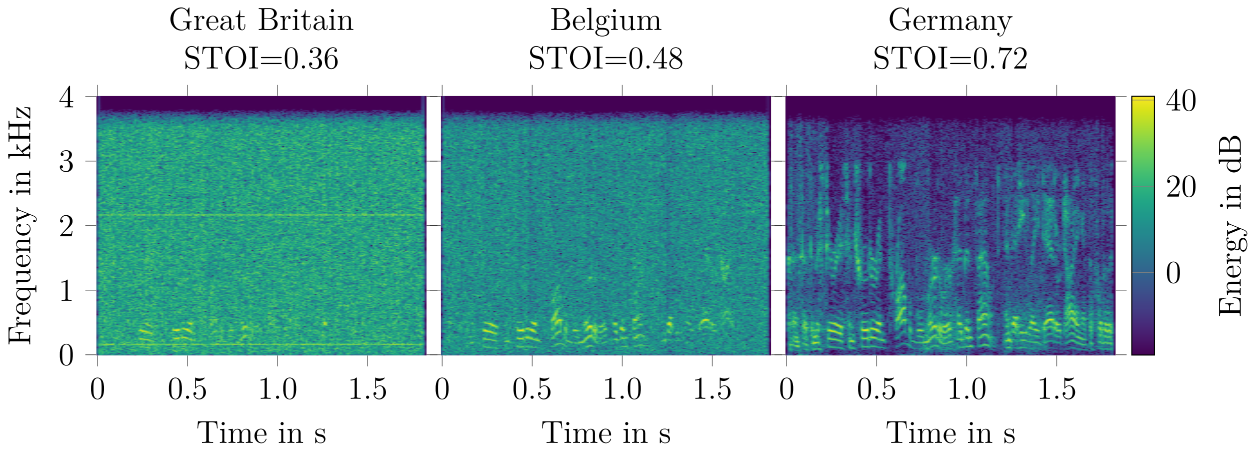


Figure 4.5: Example comparison of recordings of the same signal received at different locations in Europe.

received audio sequence. Especially, the last check forces the synchronization error to be less than 16 ms, which is the temporal resolution of the applied frequency transformation. From the temporal position of the marker the position of the transmitted audio sequence in the received signal can be inferred.

In Figure 4.5 exemplary segments of speech activity in the received audio signal are shown for different SNR conditions. All spectrograms depict the same emitted speech signal that were observed by different receiving stations under different propagation conditions. Here, short-time objective intelligibility (STOI) [159] is used as an objective measure of speech quality in the recorded signal. A more thorough discussion of the STOI metric for SSB recordings can be found in Section 6.4.2. The low STOI value and the high noise in the signal recorded by the station in Great Britain demonstrate the ability of the described marker detection algorithm to detect the signal even in such adverse conditions.

The next section introduces the post processing steps to remove some possible delays between the transmitted and recorded signal.

4.5 Post-processing

This post-processing step aims to reduce the possible misalignment of up to 16 ms between the original clean signal and the recording due to the temporal resolution of the STFT calculated during the marker detection. To this end, the recorded and clean signal are split into the five speech segments and a delay is estimated for each segment using generalized cross-correlation with phase transform (GCC-PHAT) [160]. A single delay estimation is obtained by calculating the median over the delays for the five segments per recorded sequence. To compensate for the delay the original clean signal is zero-padded from one side and equally many values discarded from the other side. While the markers can be detected in very low SNR recordings due to their high energy and specific design, the same is not true for the delay estimation with GCC-PHAT. To prevent the introduction of errors by the delay estimation, only delays lower than 16 ms are considered, which is the highest possible error in the marker detection. For all other recordings with a higher delay estimate, the signals are kept unchanged. More information about the difference between the recorded and the delay-corrected data after post processing is provided in Appendix A.1.

The next section discusses the recorded signals and the partitioning of the recordings into data sets.

4.6 Database

Using the recording process described above, more than 23 hours of speech were transmitted and the paired, distorted data was recorded in multiple KiwiSDR stations. Additional recordings with a CFO ($\Delta^{\mathrm{f}} > 0$) are created to extend the evaluation set to include examples of these problems common to analog HF transmissions [161, p. 8]. These recordings are performed as described above, however adding an error in the demodulation frequency from the following set $\{0\,\mathrm{Hz}, 100\,\mathrm{Hz}, 300\,\mathrm{Hz}, 500\,\mathrm{Hz}, 1000\,\mathrm{Hz}\}$. The resulting recordings have been made publicly available in [OC15] as evaluation data and are called “evaluation shift”.

Furthermore, with the clean speech signals taken from the LibriSpeech database all recordings consist of English speech. However, HF transmissions are used internationally as a communication method. Therefore, additional Russian data is recorded to test the presented systems against a different language. The same transmission and preparation steps as described above are taken for the Russian data. Only the part of the clean signal design, described in Section 4.2, is changed. Here, the Russian Open Speech To Text Dataset [162] is used as the basis for the clean data sequences instead of the Librispeech database and the activity information is calculated using a simple energy based SAD similar to [61] and not from ASR alignments as was done above. As for the English recordings, there is both an evaluation set without a CFO, called “Russian”, and a data set with recordings that include a CFO, called “Russian shift”.

The recordings are split into six data sets, the first for training, the second for development and four evaluation sets. All four evaluation sets represent a different challenge. The first set includes only signals with English speech without a CFO, the second consists of recordings of English speech with a CFO. For the third and fourth data set only Russian speech is used, where the third set only includes signals without a CFO and the fourth contains signals with a CFO. Note, that only English recordings with $\Delta^{\mathrm{f}} = 0$ are used as training and development sets.

For the database recordings from 56 different Kiwi-SDR stations are chosen. Signals from each station may appear in multiple data sets. All English speakers are uniformly distributed between female and male and are strictly disjoint among the data sets. Table 4.1 displays a comparison of the data sets.

Table 4.1: Various statistics for the different data sets.

Data set	Duration in h	Speech activity in %	# Speakers	# Stations
Training	121.72	13.91	705	35
Development	18.77	13.83	75	32
Evaluation	37.12	12.53	175	36
Evaluation shift	23.53	14.16	140	8
Russian	6.50	5.84	150	11
Russian shift	25.54	5.53	120	13

4.7 IQ Recordings

All previously discussed recorded signals are sampled at 8 kHz and therefore have a bandwidth of 4 kHz, which means that part of the signal energy may be lost in case of a large CFO. Therefore, additional transmissions were recorded by the PLATH GmbH & Co. KG with a high sampling rate and large CFOs. In these recordings, the complex-valued analytic baseband signal $y_n^+ = i_n + \mathrm{j} q_n$ is recorded instead of the reconstructed real-valued transmitted signal y_n. To only record real-valued signals the baseband signal is captured in form of its in-phase and quadrature (IQ) components as shown in Equations (2.6a) and (2.6b). This is not uncommon for software defined radio receivers as discussed in Section 2.3. Both the in-phase i_n and quadrature q_n component are saved with a high sampling rate of 64 kHz so that the signals have an increased bandwidth that retains the signal energy even for high CFOs. These recordings are combined in a data set referred to as "evaluation IQ" in the following, which contains 46 unique speakers in 4.5 hours of recordings with speech activity of 13.61 %.

To illustrate the benefits of working on the IQ components instead of the reconstructed signal,

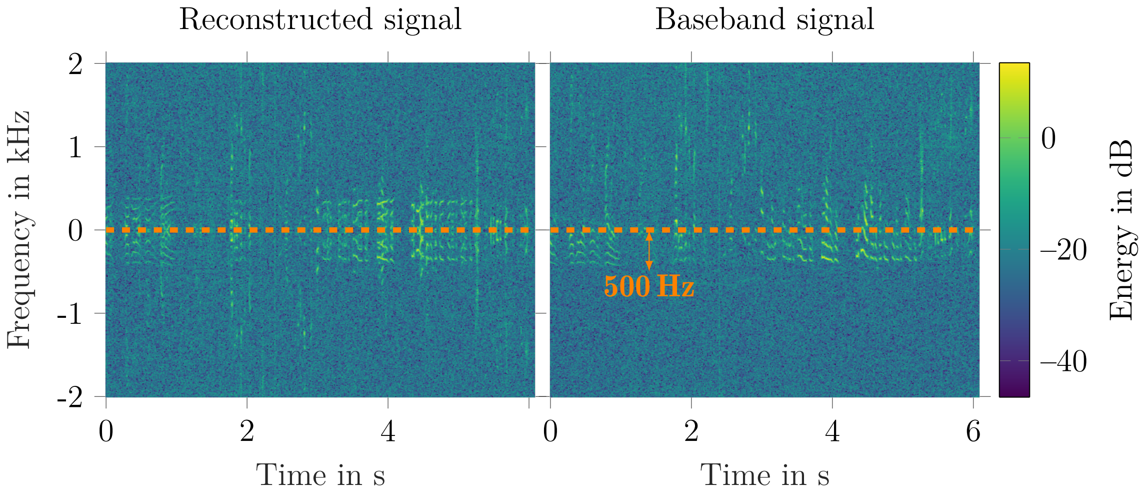


Figure 4.6: Excerpt from the frequency spectrum of the reconstructed real-valued transmitted signal (y_n) and the complex baseband signal (y_n^+) for a 8 kHz recording of a SSB transmission with a CFO of −500 Hz. The orange lines mark the frequency bin representing 0 Hz.

both the complex-valued baseband signal y_n^+ and the real-valued reconstructed signal y_n for a recording of a HF transmission with a CFO of –500 Hz are displayed in Figure 4.6, where the signals are downsampled to a rate of 8 kHz. Usually, the symmetry of the frequency spectrum for real-valued signals is used to reduce the size of the displayed spectrum to the unique bins. Here, the whole frequency spectrum is shown to compare the real-valued and complex-valued signal. For the real-valued reconstructed signal the spectrum is severely distorted due to the loss of the pitch and harmonics in the first 500 Hz of the signal. The complex-valued baseband however still includes the whole signal, which allows a reconstruction of the original transmitted signal even for negative CFOs.

Additionally, the larger bandwidth of the recording allows a realistic simulation of a CFO by considering the high CFO of $\left|\Delta^{\mathrm{f}}\right| \geq 2000\,\mathrm{Hz}$ introduced during recording as an intermediate frequency (IF). Then, baseband IQ components with a smaller CFO can be simulated by further demodulating the IQ signals similar to Equation (2.6a) and Equation (2.6b) with

$$\tilde{i}_n = \mathrm{LPF}\left\{ i_n \cdot \cos\left(2\pi \cdot \frac{\tilde{F}_D}{F_S} \cdot n \right) \right\}, \tag{4.1}$$

$$\tilde{q}_n = \mathrm{LPF}\left\{ q_n \cdot \sin\left(2\pi \cdot \frac{\tilde{F}_D}{F_S} \cdot n \right) \right\}, \tag{4.2}$$

where F_S represents the sampling rate and the demodulation frequency is set to $\tilde{F}_D = \Delta^{\mathrm{f}} - \tilde{\Delta}^{\mathrm{f}}$ with $\tilde{\Delta}^{\mathrm{f}}$ as the resulting smaller CFO. The real-valued reconstructed signal y_n can be calculated as the sum of $\tilde{i}_n$ and $\tilde{q}_n$ for a USB and as the difference between $\tilde{q}_n$ and $\tilde{i}_n$ for a LSB transmission following the steps of Weaver demodulation [24]. Finally, the demodulated signal is downsampled to 8 kHz to be comparable to other recordings with a CFO $\tilde{\Delta}^{\mathrm{f}}$.

Due to the unsupervised nature of all recordings considered, some of the signals in the data sets are distorted by concurrent speakers, which is discussed further in the next section.

4.8 Concurrent speakers

The discussed recordings are performed automatically in a HF channel reserved for amateur radio communication. Therefore, the recorded signals may accidentally be corrupted by adjacent channel interference caused by other HF transmissions [OC5]. In all data sets most examples include at least some activity of a concurrent speaker in the higher frequency and around 5-10 % of the signals include more than 1 s of activity of such an interfering speaker. An example of a recording with a concurrent speaker on a particularly close channel is displayed in Figure 4.7, where the transmission is demodulated without a CFO, while the concurrent speech signal has a CFO of about 1200 Hz and can bee seen above the red line.

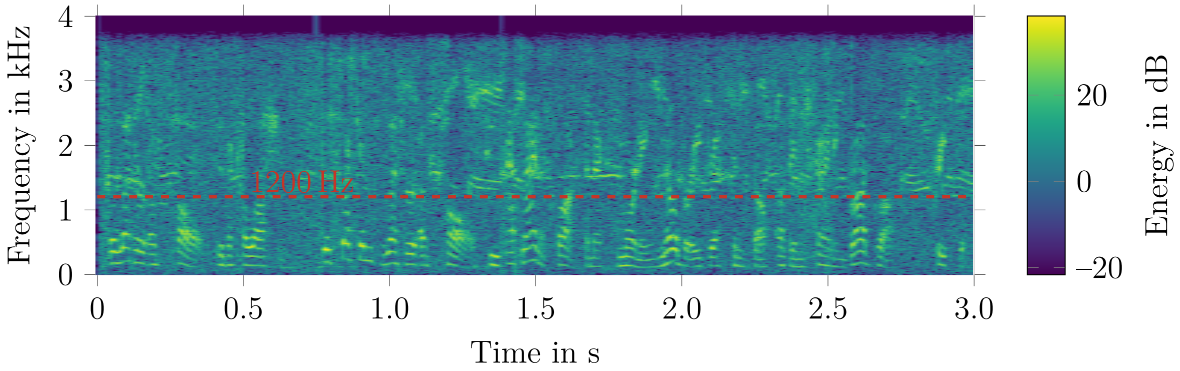


Figure 4.7: Example of a recording with a concurrent speaker, where the interfering transmission has a CFO of about 1200 Hz and can be seen above the red line.

For the SAD systems discussed in Chapter 5 the concurrent speakers active during the transmission of silence segments lead to false positives in the activity detection, as these signal regions are, obviously, not annotated with activity, since no speech has been transmitted. The labels for segments with speech activity are not affected by the concurrent speakers, because the additional activity by the concurrent speakers does not falsify the original annotation.

To examine the influence of concurrent speech on the SAD a manual annotation of the activity in the evaluation sets with English speakers is used. The annotation was performed by the PLATH GmbH & Co. KG. Here all speech activity is marked even if it is not part of the original transmission. Therefore, the difference between the SAD performance when evaluated with the automatic and manual annotation can be attributed to the concurrent speakers.

For both the CFO estimation and the NR systems discussed in Chapter 6 and Chapter 7, concurrent speakers active during the transmission of speech activity pose a unique challenge. This interference however is less frequent than the activity of concurrent speakers during silence transmissions, since other users are aware of the activity in the transmission channel.

While the database includes some evaluation sets with deviation between the modulation and demodulation frequency ($\varDelta^{\mathrm{f}} > 0$) the training set only includes data with the optimal demodulation frequency, i.e., $\varDelta^{\mathrm{f}} = 0$. A CFO greater than zero leads to a shift in the frequency spectrum of the speech signal as discussed in Section 3.3 and thereby introduces a large mismatch between the training and the evaluation data sets with a CFO. Therefore, the next section is dedicated to simulating these shifted signals to complement the real recordings during training.

4.9 Simulating high frequency radio signals

Creating a training database of real signals with representatives for all challenges in high frequency radio transmissions requires a large amount of data. To reduce the required

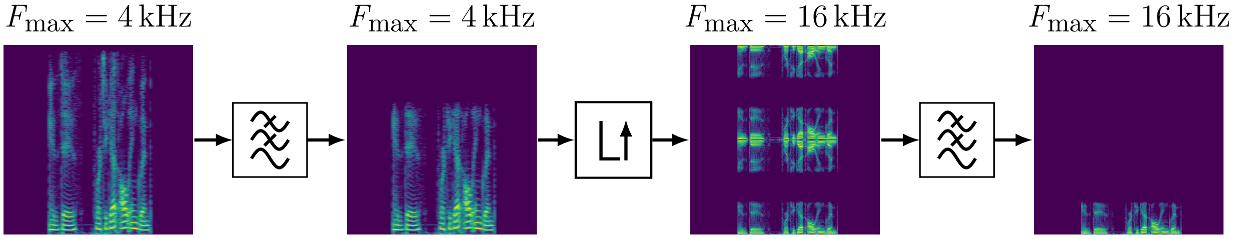

$F_{\text{max}} = 4\,\text{kHz}$
$F_{\text{max}} = 4\,\text{kHz}$
$F_{\text{max}} = 16\,\text{kHz}$
$F_{\text{max}} = 16\,\text{kHz}$

Figure 4.8: Pre-processing pipeline for simulating bandwidth limitations and preparing the signal for CFO effects. F_{max} represents the Nyquist frequency of the shown signal and the displayed frequency range is changed between the images to reflect the change in F_{max}.

number of transmissions some of the signals can be simulated. Therefore, a processing pipeline comparable to the steps taken during modulation and demodulation as described in Chapter 2 and Figure 4.2 is implemented to generate a simulated signal from the digital clean speech signal similar to the recording of a SSB modulated signal transmitted over a HF channel.

Here, the modulation approach described in [11, p. 260] is applied, where a double sideband (DSB) signal is filtered by a bandpass to remove one of the sidebands. To reduce the complexity only USB demodulation is simulated since all error patterns of LSB modulation also occur in USB modulated signals.

As a pre-processing step, the clean signal is low-pass filtered to simulate the limited bandwidth of the transmission channel. The signal is then interpolated to a four times higher sampling rate to provide bandwidth for shifting the signal along the frequency axis. Figure 4.8 depicts a visualization of the pre-processing pipeline.

The modulation is described as follows:

$$x_n^{\text{sim}} = \tilde{s}_n^{\text{sim}} \cdot \cos\left(2\pi \cdot \frac{\tilde{F}_M}{F_{\text{max}}} \cdot n\right), \tag{4.3}$$

with $\tilde{s}_n^{\text{sim}}$ as the output of the above described processing steps, n as the time index, $\tilde{F}_M$ as the simulated carrier frequency and F_{max} as the Nyquist frequency. For the simulation $\frac{\tilde{F}_M}{F_{\text{max}}}$ is chosen to be 0.5.

The modulated bandpass signal is reduced to the USB signal with a high pass. Subsequently, the sideband signal is filtered with a band pass to simulate the limited frequency range of the transmission channel.

The demodulation reverses the frequency shift applied during modulation:

$$\tilde{y}_n^{\text{sim}} = \tilde{y}_n \cdot \cos\left(2\pi \cdot \frac{\tilde{F}_D}{F_{\text{max}}} \cdot n\right) = (b * x)_n^{\text{sim}} \cdot \cos\left(2\pi \cdot \frac{\tilde{F}_D}{F_{\text{max}}} \cdot n\right), \tag{4.4}$$

where $\tilde{y}_n^{\text{sim}}$ is the simulated signal after the channel distortion is applied and $\tilde{F}_D = \tilde{F}_M - \Delta^{\text{f}}$ is the demodulation frequency with Δ^{f} as the CFO. These demodulation steps are independent of the chosen sideband for transmission. Note that no additional channel distortions other

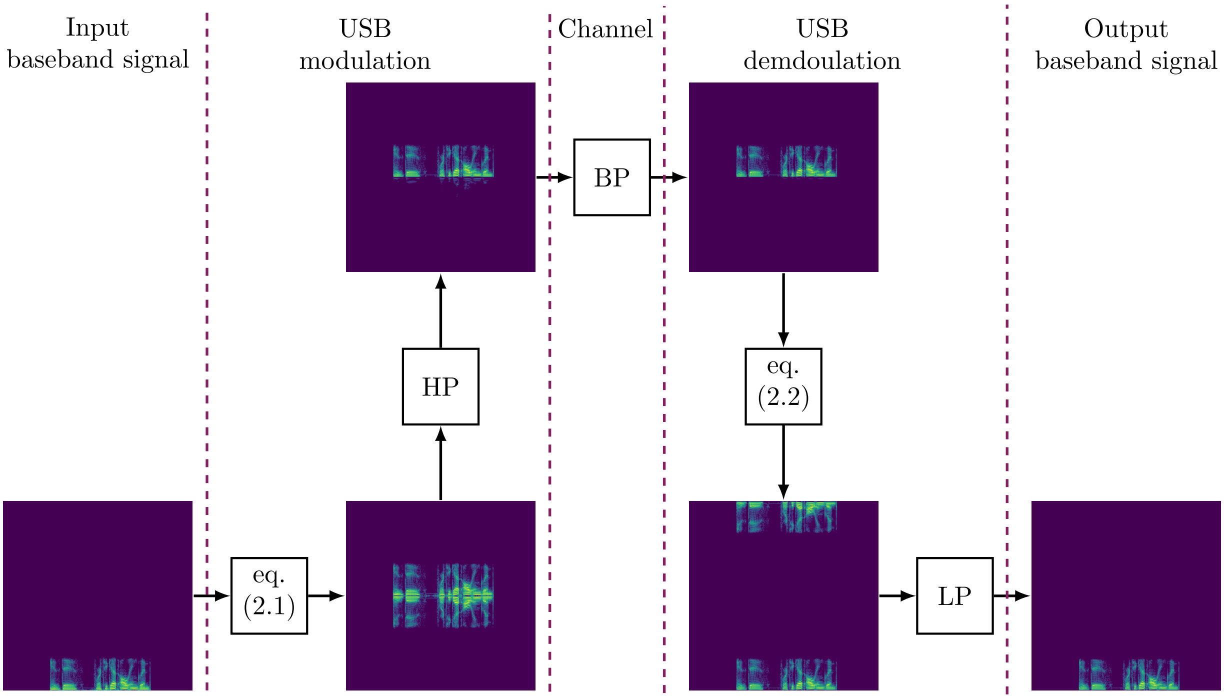


Figure 4.9: Example for the signal spectrum during the simulated transmission without a CFO. The low, high and band pass operations are abbreviated with LP, HP and BP, respectively.

than bandpass limitation is applied during this simulation. For more realistic simulations, other distortions such as multi-path propagation could be added.

Afterwards, a low pass filter is applied to reject the signal at image frequencies introduced by the demodulation. Furthermore, the signal is downsampled to the original sampling rate. The different signal spectra during transmission for a simulation with a CFO are shown in Figure 4.9.

To simulate a CFO only the demodulation frequency $\tilde{F}_D$ in Equation (4.4) has to be adjusted. Examples for the simulated signal spectra in case of negative and positive CFO are displayed in Figure 4.10. For the simulation framework, the CFOs are drawn from an uniform distribution in the range from $\mathit{\Delta}^{\mathrm{f}}_{\mathrm{min}}$ to $\mathit{\Delta}^{\mathrm{f}}_{\mathrm{max}}$, where $\mathit{\Delta}^{\mathrm{f}}_{\mathrm{max}}$ has to be smaller than the Nyquist frequency F_{max}.

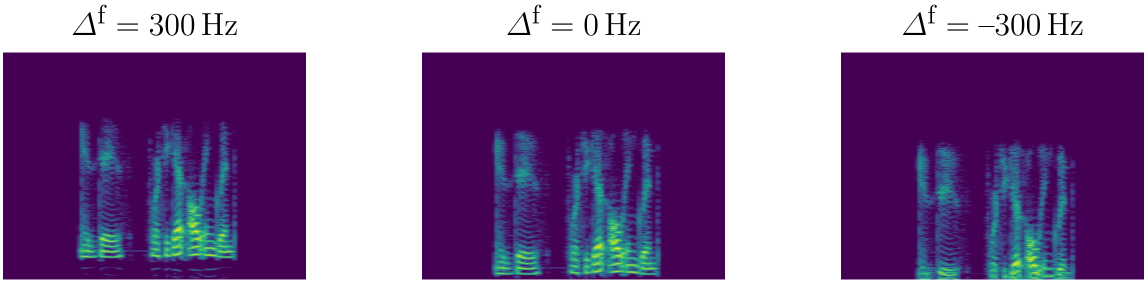


Figure 4.10: Example for the signal spectrum after demodulation with a CFO $\mathit{\Delta}^{\mathrm{f}}$.

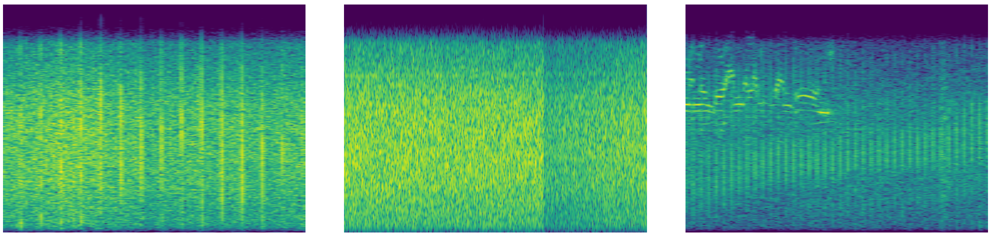


Figure 4.11: Examples for real recordings used as additive noise during simulation.

At this point, the simulated signal does not include channel noise. Therefore, the KiwiSDR receivers are used to record multiple hours of noise on non-occupied frequencies. A random chunk of recorded noise d_n is added to the simulated recordings at a random SNR $\in [\mathrm{SNR}_{\mathrm{min}}\ ,\ \mathrm{SNR}_{\mathrm{max}}]$ to generate a realistic high frequency transmission signal

$$y_n^{\mathrm{sim}} = \tilde{y}_n^{\mathrm{sim}} + d_n. \tag{4.5}$$

In Figure 4.11 a few examples of the recorded signals used as noise d_n are displayed to emphasize the variability of the encountered noise. The first image includes a periodically repeating interference, which could be caused by an overlapping radar transmission. In the second image the effect of the AGC on the recorded noise can be seen. After a few seconds, the amplitude of the recorded signal is reduced, resulting in a sudden change in the noise energy. The last image is an example of an interfering speech transmission on a neighboring frequency.

The presented simulation framework consists of multiple independent components which can be replaced to create training data suitable to adapt to other languages, different noise conditions or specific CFOs. To emphasize the modularity of the presented system a simple representation of the different components during simulation are shown in Figure 4.12.

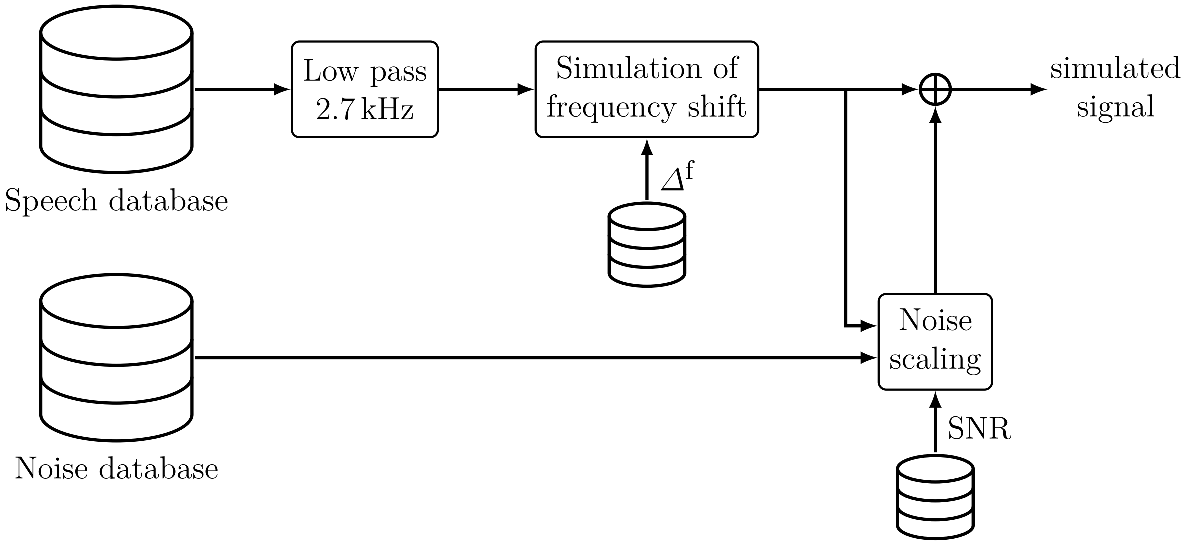


Figure 4.12: Overview of the presented simulation framework.

4.10 Summary

This chapter offered an in-depth discussion of database design and the automatic recording process for the HF recordings used for training and evaluation in the next chapters. Here, concurrent speakers during the transmission are identified as one of the main challenges for training and evaluation on the recorded signal. Additionally, a simulation framework is presented, which can either be used to generate signals to extend the recorded training set as in Chapter 5 or to create a new training set as in Chapter 6 and Chapter 7.

The next chapters will examine each component of the multi-stage system introduced in Chapter 1, starting with the SAD block.

5 Speech Activity Detection

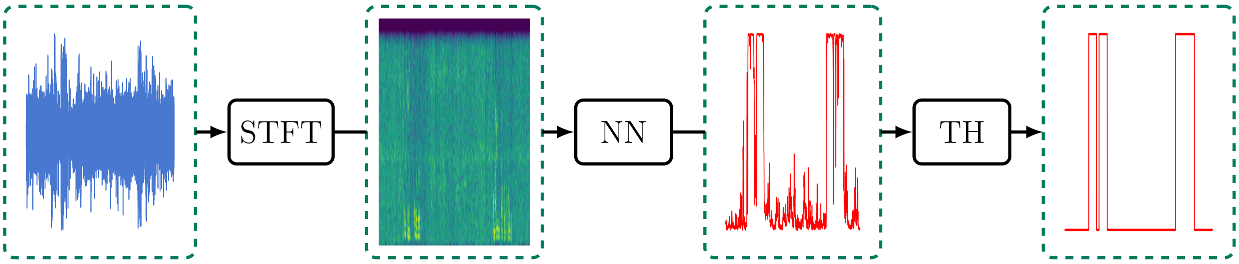


Figure 5.1: Block diagram of the SAD system, where TH represents thresholding.

As discussed in Section 2.4, high frequency channels are often inactive for a longer period of time while containing only short segments of speech activity. Therefore, the first step of most signal processing pipelines is to detect speech activity, so that subsequent processing steps can be performed only on speech-active segments.

There are both statistical and neural network (NN)-based speech activity detection (SAD) systems which achieve a strong detection performance. However, the NN-based systems outperform the statistical detection for scenarios for which appropriate training data is available as discussed in Section 3.1. Therefore, NN-based systems with the structure displayed in Figure 5.1 are the focus of this chapter. The block diagram of the system is the comparable to the one discussed in Section 3.1.

In the following section, an architecture for the NN-based speech presence probability (SPP) estimation is presented, which is a combination of convolutional neural network (CNN) and recurrent neural network (RNN) layers. Additionally, we propose a novel RNN architecture called segment recurrent neural network (SRNN) that leads to large improvements over a simple RNN layer on multiple data sets. Furthermore, the statistical approach we presented in [OC5] is discussed as a baseline for the NN-based SAD system. Finally, the network is evaluated for different scenarios of high frequency (HF) radio transmissions. We published a part of this evaluation in [OC4] as the winning submission to the 2020 Fearless Steps SAD challenge [80].

5.1 Architecture

As described in Section 3.1 there are many possible architectures for a NN-based SAD system. In this thesis an architecture is proposed that consist of CNNs and RNNs to combine the strength of both layer types.

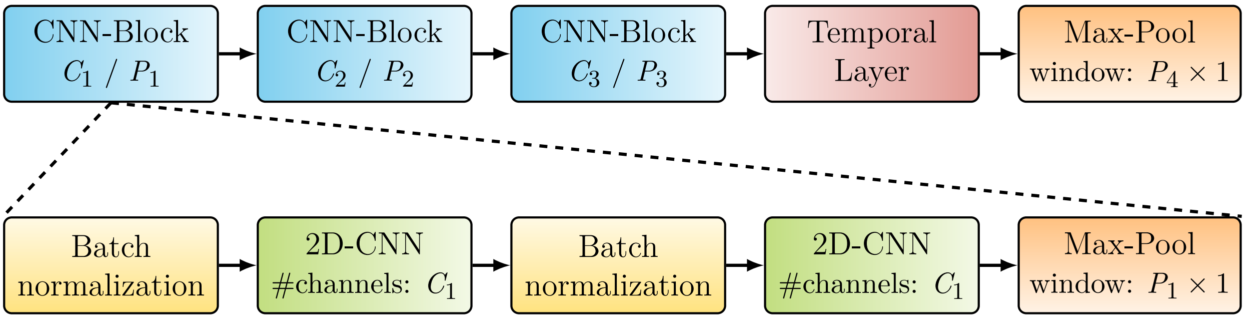


Figure 5.2: Block diagram of the SAD model, C_ν represents the number of channels in a layer and P_ν the pooling window size with $\nu \in [1, 2, 3]$. The window size of the output pooling P_4 is equal to the output size of the temporal layer $\tilde{F}$. All 2D-CNNs are designed with a 3×3 kernel.

The network uses the magnitude of the short time Fourier transform (STFT) of the audio signal as an input. A more detailed discussion about the STFT as feature extraction can be found in Section 7.2. To understand the following discussion, note that the STFT is a transformation of the input signal into the frequency domain, by segmenting the signal into multiple, overlapping windows and calculating the discrete Fourier transformation (DFT) for each window independently. For this SAD system a input signal with a sample rate of 8 kHz is assumed. Therefore, a STFT size of 256, a window length of $L_W = 25\,\text{ms}$ is chosen with a shift between the windows of $L_S = 10\,\text{ms}$, resulting in an overlap of 15 ms. Each segment is zero-padded to account for the STFT size which corresponds to a window size of $L_W = 32\,\text{ms}$. Since the DFT of a real valued signal leads to a symmetric frequency spectrum the feature vector has the size $F = 129$.

A block diagram of the architecture is shown in Figure 5.2. It consists of three initial CNN-blocks where each block consists of two 2D-CNN layers with 3x3 kernels, strides of one and rectified linear unit (ReLU) activation functions. Each CNN layer is preceded by a batch normalization [163]. The CNN blocks end with a maximum pooling with a pooling window of size $P_\nu \times 1$ with $\nu \in [1, 2, 3]$. To allow for a frame-wise activity estimation no pooling is applied along the time dimension.

The context per bin or receptive field of a CNN layer with a stride of one increases for each layer by the number of neighboring bins included in the kernel [38, p. 327]. Each pooling with a stride equal to the pooling window size P_ν further increases the context of the following layer P_ν times. In other words, the receptive field $R_{m,\nu}$ of the $\nu{-}th$ 2D-CNN layer for the dimension m can be calculated from the context of the previous layer with

$$R_{m,\nu} = R_{m,\nu-1} \cdot P_{m,\nu-1} + K_{m,\nu-1} - 1 \tag{5.1}$$

where m is in the range $[1, 2]$ with $m = 1$ representing the feature and $m = 2$ the time dimension. $K_{m,\nu}$ is the kernel size, which equals three for all dimensions m and layers ν for the given architecture. As described above the pooling window size for the time dimension is set to one, which leads to a simplified context calculation with $R_{2,\nu} = R_{2,\nu-1} + 2$.

Therefore, the configuration of the input CNN-blocks leads to a final temporal context of $1 + 3 \cdot (2 + 2) = 13$ frames for each considered frame, which is equal to 0.13 s for a sample rate

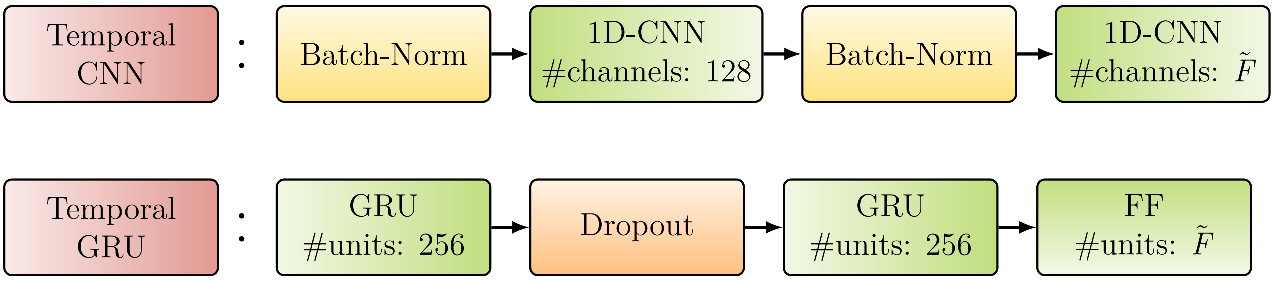


Figure 5.3: Block diagram of possible temporal layers, $\tilde{F}$ represents the output size of the temporal layer. The 1D-CNN layers are designed with a kernel of length 3 and 1 respectively.

of 8 kHz and the STFT configuration given above. This context is determined only by the kernel size of the CNN-layers, since no maximum pooling is applied to the time dimension. While this is only a small temporal context, it is increased by a subsequent temporal layer with a larger context, which has shown to improve the results in [68]. As in [OC4], three possible temporal layers are compared:

1. a 1D-CNN block,
2. a uni-directional gated recurrent unit (GRU) [55] followed by a fully-connected (FC) layer,
3. a segment recurrent neural network (SRNN), to be introduced here.

The configuration of the CNN and GRU temporal layers are displayed in Figure 5.3.

While the 1D-CNN layer only increases the temporal context by 2 frames to 0.15 s, given the receptive field of the initial CNN layer, the temporal layer with a GRU can gather information from the whole input signal. These architectures for the two temporal layers are chosen to offer two distinct networks as a comparison for the SRNN layer, a fast network with a small number of trainable parameters and a larger network with full temporal context. As a final step a maximum pooling over the feature dimension is applied to the output of the temporal layer with a window size P_4 equal to the feature size of the temporal layer output $\tilde{F}$. In the following section the SRNN is explained in more detail.

5.2 Segment recurrent neural network

Like many SPP systems, the two temporal layers described above suffer from a high variance in the output, which leads to a high variance in the estimated activity if the decision is performed by thresholding as discussed in Section 3.1. This can lead to a high dependency on the threshold applied to the network output, which, to the same extent, can be mitigated by external smoothing with a median filter or a hidden Markov model (HMM) as in the statistical SAD system. However, the smoothing can also be done inside of the network. There are already some publications on this subject, for example in [164] and [165]. However, most of these systems require additional trainable parameters to estimate a smoothed output.

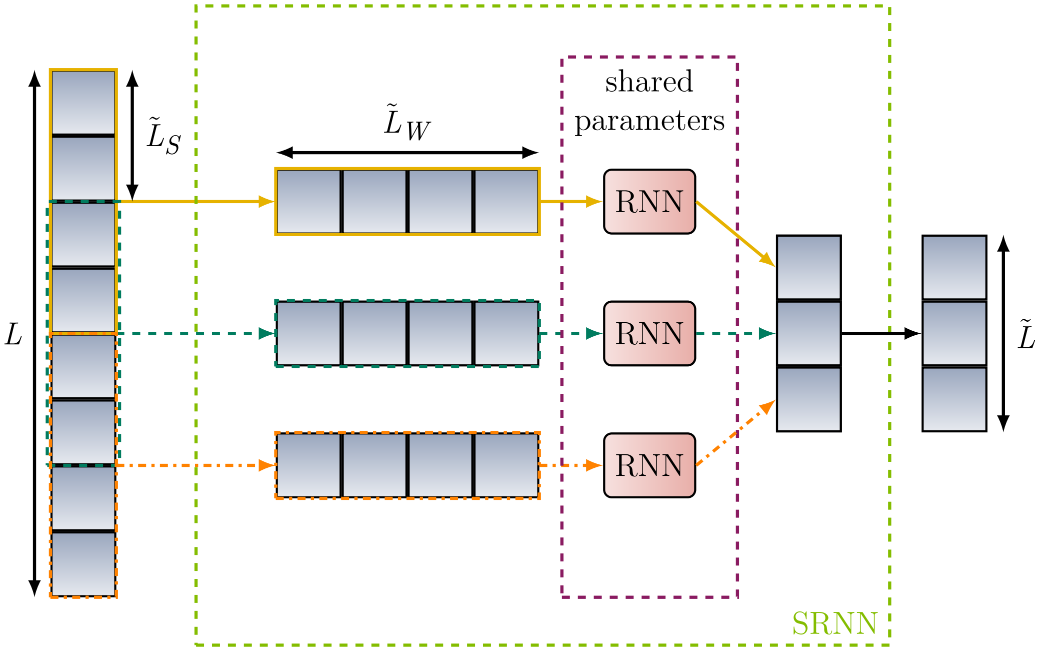


Figure 5.4: Block diagram of the SRNN. Here the RNN is a temporal GRU described above. The input is segmented in into $\tilde{L}$ chunks with length $\tilde{L}_W$ and an overlap of $\tilde{L}_W - \tilde{L}_S$. Each blue block represents a feature vector.

In this work, the SRNN layer is introduced, where a fixed segmentation of the input over the time dimension is used to restrict the temporal context of an RNN layer. This idea to restrict the temporal information exploited by the RNN is related to the dual-path recurrent neural network (DPRNN) [44] designed for speech enhancement and described in Section 7.3.2.

The L frames of the input are split into $\tilde{L}$ segments of length $\tilde{L}_W$ with a shift $\tilde{L}_S$ which results in an overlap $\tilde{L}_W - \tilde{L}_S$ between the segments. Each segment is processed independently by an RNN layer. In this work the temporal GRU defined in Figure 5.3 is chosen but could be replaced by other RNN configurations. Note, that only the time dimension is segmented. Each RNN processes the whole feature dimension for each segment.

For each RNN only the last frame of the output is passed on to force the RNN to aggregate the information of each segment. This value is the output of the SRNN layer. In case of the network architecture displayed in Figure 5.2 a global maximum pooling is applied to the feature dimension of the output before it is evaluated by the cost function. A block diagram of the SRNN is displayed in Figure 5.4

During the evaluation, the network output is further processed to obtain a frame-wise, binary estimate. If the estimated output $\hat{q}_i$, with $i \in [0, \tilde{L} - 1]$ as the segment index, exceeds a fixed threshold $\mathrm{th}^{\mathrm{nn,fixed}}$ the i-th segment is assumed to contain speech activity:

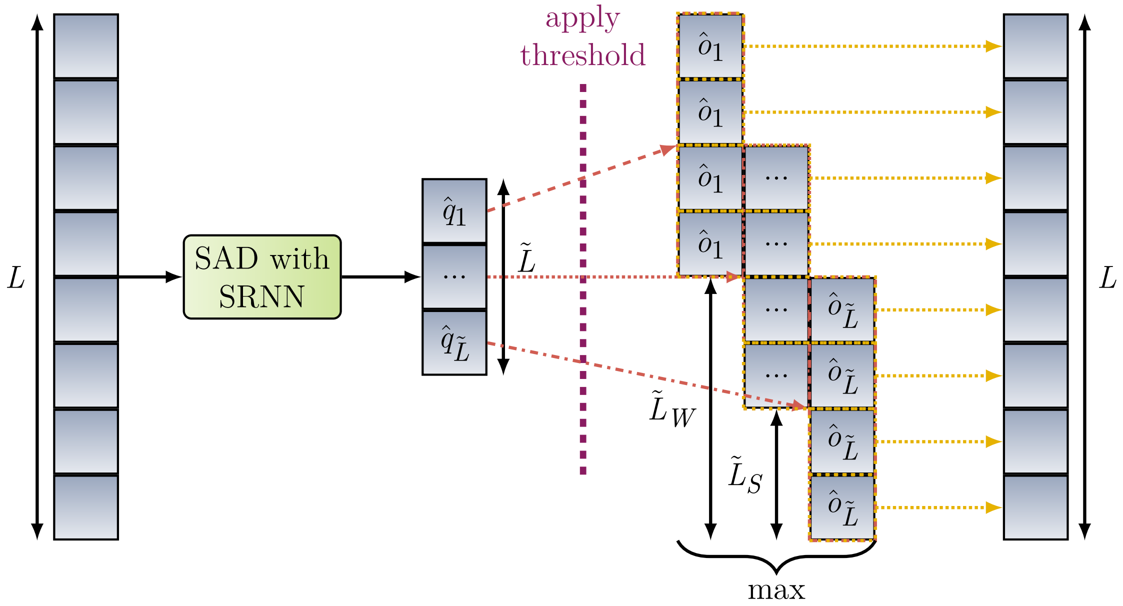


Figure 5.5: Block diagram of the post-processing after a SAD model with a SRNN temporal layer. A threshold is applied to the network output $\hat{q}_i$ from the result an estimate per frame is calculated using the maximum operation. Each blue block represents a feature vector.

$$\hat{o}_i = \begin{cases} 1 \text{ if } \hat{q}_i > \text{th}^{\text{nn,fixed}} \\ 0 \text{ else} \end{cases} . \tag{5.2}$$

A frame-level decision is obtained by declaring a frame active if at least one segment containing that frame indicates speech activity. This is similar to the overlap-add, except that here the maximum operation is calculated instead of the sum. Thereby, possible fluctuation between speech activity and silence are reduced at the cost of overestimating the activity. A block diagram of the post-processing steps during evaluation of a SAD system with a SRNN temporal layer is depicted in Figure 5.5.

The resulting smoothing can be adjusted via the segment length $\tilde{L}_W$ and shift $\tilde{L}_S$. Reducing the length $\tilde{L}_W$ decreases the number of frames affected by a segment with erroneous detected activity but also reduces the temporal context of the GRU. Increasing the shift $\tilde{L}_S$ leads to a reduced overlap between segments and thereby to a lower number of segments $\tilde{L}$. This entails a lower computational complexity since the number of chunks processed in parallel by the GRU is decreased. However, the number of neighboring segments contributing to the activity estimation for each frame is reduced as well.

5.3 Cost function

All networks are trained using the binary cross entropy (BCE) as cost function, since activity detection is a binary classification task. The loss is calculated as follows:

$$\mathcal{L}^{\mathrm{BCE}}(\hat{\mathbf{q}}^{\mathrm{SAD}}, \mathbf{q}^{\mathrm{SAD}}) = -\frac{1}{\tilde{L}} \sum_{\ell=1}^{\tilde{L}} q_\ell \cdot \log \hat{q}_\ell + (1 - q_\ell) \cdot \log(1 - \hat{q}_\ell) \tag{5.3}$$

where $\hat{\mathbf{q}}^{\mathrm{SAD}} = [\hat{q}_1, ..., \hat{q}_{\tilde{L}}]^\mathsf{T}$ and $\mathbf{q}^{\mathrm{SAD}} = [q_1, ..., q_{\tilde{L}}]^\mathsf{T}$ are the network output and the target vector, respectively. $\tilde{L}$ represents the number of time frames in both the output and target vector. For the SRNN $\tilde{L}$ is smaller than the number of STFT window frames L due to the segmentation steps with $\tilde{L} = \left\lfloor \left(L + \tilde{L}_S - 1\right) / \tilde{L}_S \right\rfloor$. Maximum pooling over the time dimension with a pool size $\tilde{L}$ and stride $\tilde{L}_S$ is applied to the target sequences to adjust the targets to the smaller time resolution of the network output. For all networks with a temporal layer other than the SRNN holds $\tilde{L} = L$. The following section introduces the statistical baseline system as described in [OC5].

5.4 Baseline

Statistical SAD has been a subject of research for many years [166] and has shown to achieve impressive results, even on challenging data [167]. For this work a baseline similar to [OC4] was chosen, which uses a strong denoising front-end with subsequent adaptive thresholding and a HMM-Gaussian mixture model (GMM) smoothing algorithm.

The block diagram of the statistical SAD system is depicted in Figure 5.6 and consists of two stages. First, a denoising stage is applied, which consists of multiple runs of the minimum statistics (MS) algorithm with subsequent application of a Wiener filter [92] and a high pass to suppress noise in frequencies above 2.7 kHz. Afterwards, a 1st-order linear predictive

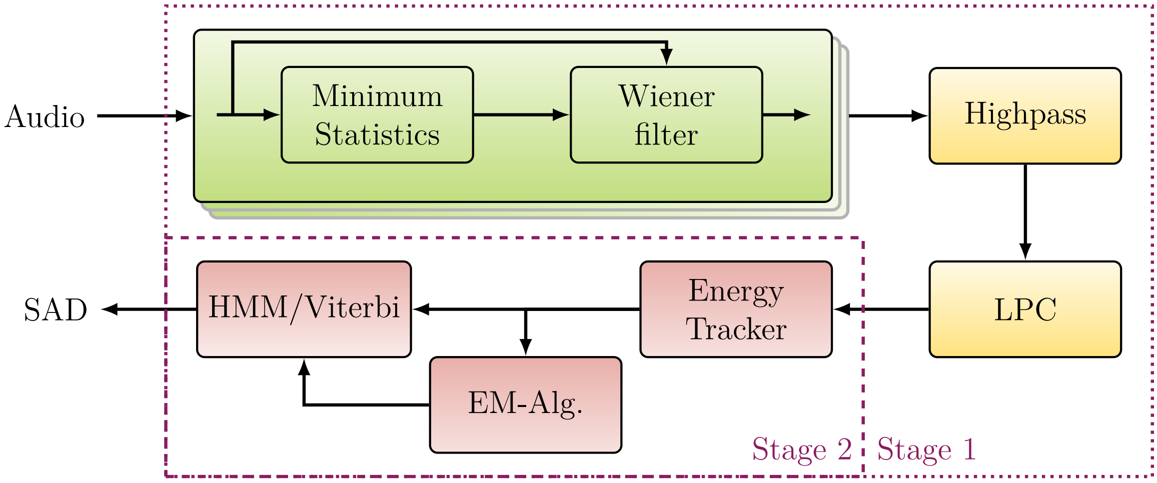


Figure 5.6: Block diagram of the statistical SAD system.

coding (LPC) filter reduces the signal to its well predictable parts like the highly correlated speech signal. As a second step, an adaptive threshold is calculated using sub-band energy information. The resulting threshold is combined with a subsequent smoothing using a HMM trained with the expectation maximization (EM)-algorithm and decoded with the Viterbi algorithm [168]. For further information on the statistical system please refer to [OC5] or [OC4].

This statistical system has no trainable parameters, only a hyperparameter optimization may be necessary to adapt the system to a new environment. It is designed for high frequency transmission data without a carrier frequency offset (CFO) ($\mathit{\Delta}^{\mathrm{f}} = 0\,\mathrm{Hz}$) and will be used as baseline for the NN-based SAD system described above.

In the next section, the discussed SAD systems are evaluated and the impact of additional simulated training data as described in Section 4.9 for SAD on signals with a CFO are investigated.

5.5 Evaluation

Afterwards, the hyperparameters ($\tilde{L}_S$ and $\tilde{L_W}$) of the SRNN layer are fine-tuned and the performance of the resulting network is compared to the temporal layers described in Section 5.1 and the statistical system. These two experiments are also performed for both the development set of the Fearless Steps Challenge [80] and the evaluation set of our own real HF recordings. Note, that the hyperparamters for the statistical baseline are optimized individually for each database. For experiments on the Fearless Steps database, the configuration described in [OC4] is chosen, and for our HF recordings, the system is parameterized according to [OC5].

The best network architecture is further evaluated on both the evaluation set with a CFO greater than zero and the Russian data described in Section 4.6. These experiments are only performed on our recordings because the Fearless Steps database does not provide different language data or signals with a CFO. To improve the results, additional data is simulated as described in Section 4.9 and mixed with the real training data. Finally, the best system is evaluated for its performance on the recorded in-phase and quadrature (IQ) signals instead of the baseband signal for negative and positive CFOs.

Both the NN-based SAD systems and the statistical baseline use the STFT as feature extraction. For the NN a window length of $L_W = 25\,\mathrm{ms}$ with a shift between the windows of $L_S = 10\,\mathrm{ms}$ and a DFT size of 256 is chosen, as described in Section 5.1. The statistical SAD system uses a window length of $L_W = 64\,\mathrm{ms}$ with the shift $L_S = 32\,\mathrm{ms}$ and a DFT size of 1024.

For both databases the input CNN of the NN uses the following hyperparamters (see Figure 5.2): $C_1 = 16$, $C_2 = 32$, $C_3 = 64$, $P_1 = 4$ and $P_2 = P_3 = 8$. The temporal GRU for the experiments on the database of the Fearless Steps challenge contains only one layer with 256 units instead of two as in the experiments with our HF recordings. The additional layer is added to account for the lower signal to noise ratio (SNR), i.e., the more difficult task, in our HF recordings. During training all input audio sequences are split into 4 s long segments to improve convergence. For optimization during training the ADAM algorithm [169] with a learning rate of 0.001 is chosen and each model is trained until the BCE does not improve for two epochs.

The systems are compared using receiver operating characteristic (ROC) curves which show the true positive rate (TPR) over the false positive rate (FPR) for different thresholds. Additionally, the decision cost function (DCF), recall (REC) and precision (PRC) value for the threshold with the lowest DCF are calculated with

$$\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad (5.4) \qquad\qquad \mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad (5.5)$$

$$\mathrm{DCF} = 0.75 \cdot \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} + 0.25 \cdot \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}, \qquad (5.6)$$

where TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative frame-wise predictions. For the TP, FP, TN and FN calculation a collar of 1 s length around speech active segments are not scored to allow the systems to slightly overestimate speech activity without loss in accuracy as suggested in the NIST openSAD evaluation [170]. Underestimation of the activity with these collars still results in the same error rates as if no collars had been used. These collars improve the real world significance of the calculated metrics because most applications require the SAD to correctly detect every second of speech, while declaring preceding silence frames as speech is considered less severe.

Additionally, the real time factor (RTF) which is defined as the ratio between processing time and signal length is calculated for some systems. The RTF is measured on an Intel®Xeon®CPU E3-1240 v6 with 3.70 GHz and 8 GB RAM to compare the computational complexity.

5.5.1 SAD on optimally demodulated speech

This section gives an overview of the results achieved on speech demodulated without an error in the demodulation frequency ($\varDelta^{\mathrm{f}} = 0\,\mathrm{Hz}$).

First, the statistical SAD baseline is compared with the NNs with a CNN or GRU temporal layer. Here, the NN-based SAD estimation is smoothed with a median filter after thresholding. The median filter was chosen over smoothing with an HMM since our experiments in [OC4] suggest that both perform similarly despite the lower computational complexity of the median filter.

Figure 5.7 depicts the ROC curves for the three systems together with a line representing the equal error rate (EER) (FNR = FPR) on the Fearless Steps database and the evaluation set of our recordings without a CFO.

The statistical SAD slightly outperforms both NN-based systems in terms of EER for our HF recordings, while the opposite is true on the Fearless Steps database. For the statistical SAD the performance is similar for both databases with an EER around 5 %. However, the NN-based systems achieve a low EER close to 2 % on the Fearless Steps database and a

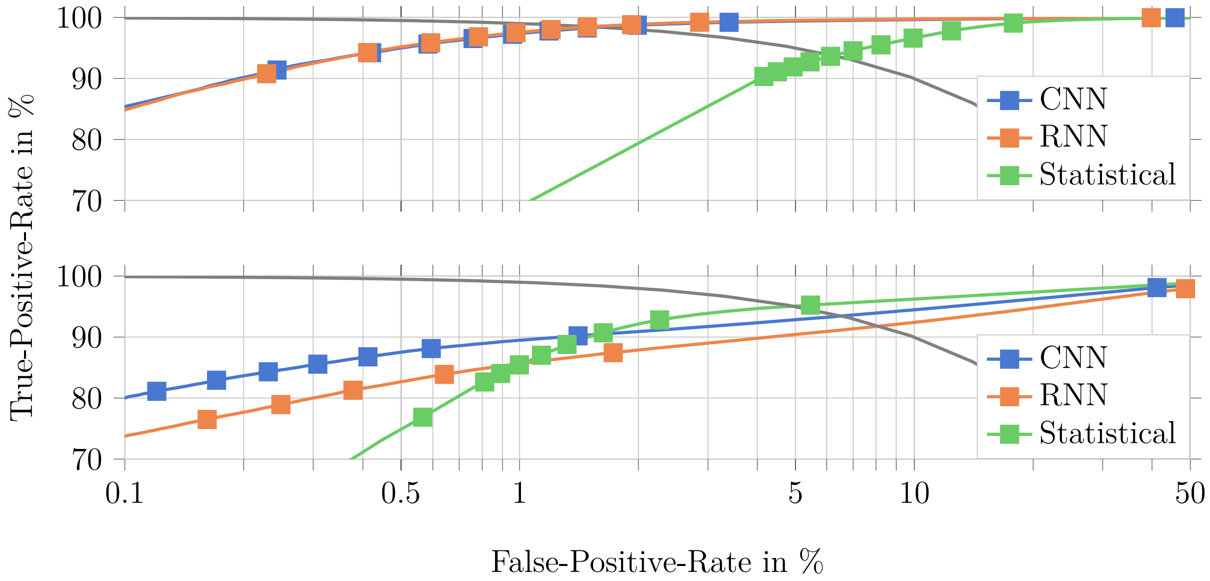


Figure 5.7: ROC curve for different SAD systems on the development set of the Fearless Steps challenge (first row) and the evaluation set of our real HF data (second row). Each square represents a different threshold. The gray line symbolizes the EER.

higher EER around 7 % on our HF recordings. These results indicate a larger mismatch between the training and test data for our HF recordings compared to the Fearless Steps challenge because the statistical system, without a training phase, performs similar on both sets, while the NN-based models perform clearly worse on our recorded data. Two reasons for the difference in the achieved EER for the two databases might be the concurrent speakers discussed in Section 4.8 and the low SNR, that can mask some activity segments in our recorded data.

To improve upon those results, the SRNN architecture is chosen as temporal layer. First, however, a parameter tuning is performed to find the best combination of shift $\tilde{L}_S$ and segment length $\tilde{L}_W$ for the SRNN for both databases. The DCF results for different $\tilde{L}_S$ and $\tilde{L}_W$ for the development set of the Fearless Steps database and our recordings are displayed in Figure 5.8. Here, only the results for the threshold with the lowest DCF value are shown, and the value range of $\tilde{L}_S$ and $\tilde{L}_W$ is limited to those with the best results for each database.

For both databases, the NN with a SRNN and an appropriate configuration outperform the system with the lowest EER from the initial experiment. Still, which configuration leads to the best results, depicted by a blue asterisk, differs between the databases. For the Fearless Steps database lowering $\tilde{L}_S$ and $\tilde{L}_W$ improved the DCF value with $\tilde{L}_W = 5\,\text{ms}$ and $\tilde{L}_S = 1\,\text{ms}$ as the best configuration. On our recordings the systems with $\tilde{L}_W = 1000\,\text{ms}$ outperform systems with larger or shorter window lengths. The difference to the results on the Fearless Steps database may be explained by the lower SNR on our recordings, where a larger context can be used to capture speech statistics even for signals with high noise energy. Additionally, the large window increases the possibility to exploit short intervals with high speech energy in multiple segments.

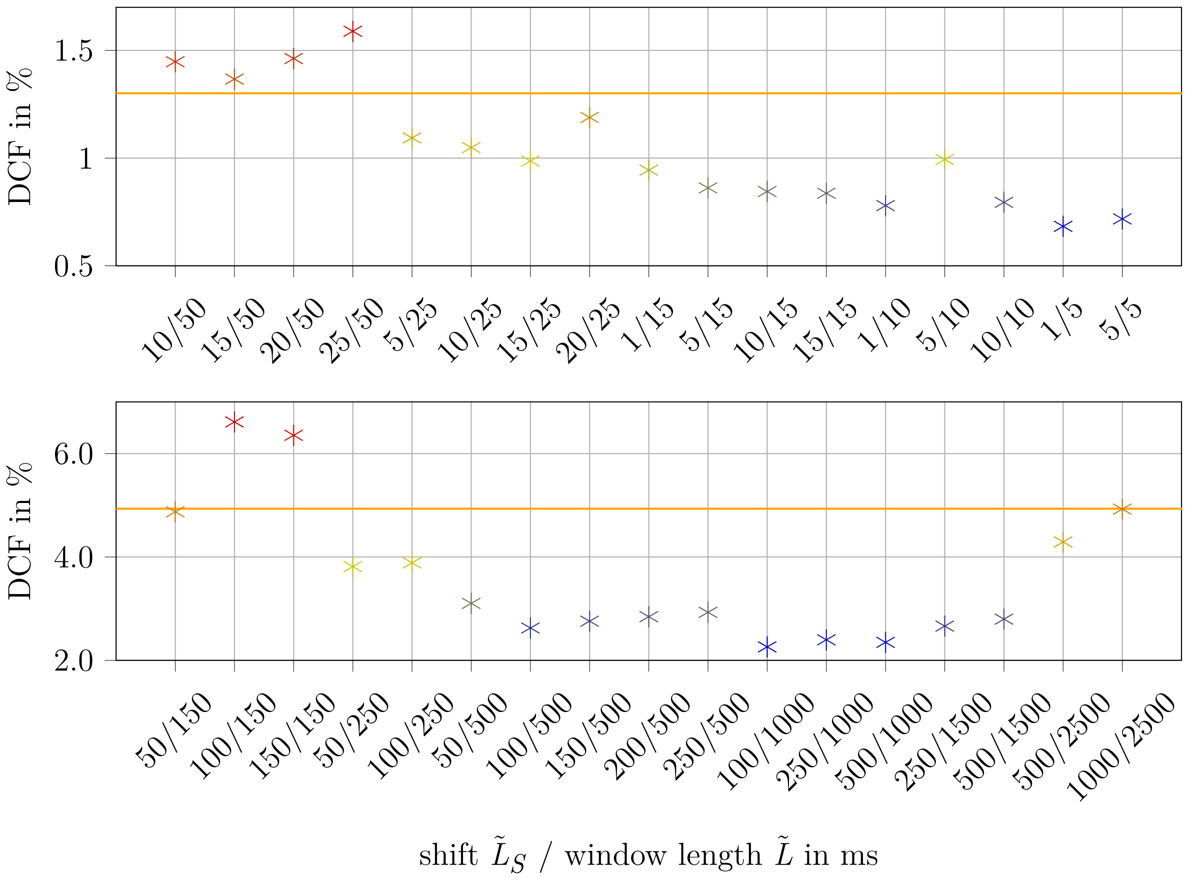


Figure 5.8: DCF results for the SRNN temporal layer on the development set of the Fearless Steps database (first row) and our recorded HF data (second row) with different shifts $\tilde{L}_S$ and segment length $\tilde{L}_W$ for the respective optimal threshold. The orange line symbolizes the DCF value of the best result from the prior experiment (CNN for Fearless, statistical SAD for our data).

For our recordings the DCF values for different shifts $\tilde{L}_S$ with a constant $\tilde{L}_W$ of one second are similar. Therefore, the shift $\tilde{L}_S = 0.5\,\mathrm{s}$ is chosen which reduces the DCF by only 0.077 % compared to the best system with $\tilde{L}_S = 0.1\,\mathrm{s}$ but improves the RTF from 0.112 to 0.032. The increased processing speed is due to the lower overlap of 50 %, which results in a lower number of segments $\tilde{L}$ to be processed by the SRNN.

During the following experiment, the chosen SRNN temporal layer is compared to both the statistical baseline and the other temporal layers described in Section 5.1 on both databases. Here, the SRNN based system clearly outperforms all other systems on our database with higher TPR values for all FPRs and an EER of 2.8 %. For the Fearless Steps database the depiction of the ROC curve does not show a large gain for the SRNN. To highlight the strong detection results of the SRNN, Table 5.1 displays the metrics from Equations (5.4) to (5.6) for each system as well as the RTF for both databases. All results are calculated using the threshold with the lowest DCF value on the development set for each model.

The SRNN outperforms all other systems for both databases in all presented metrics, while

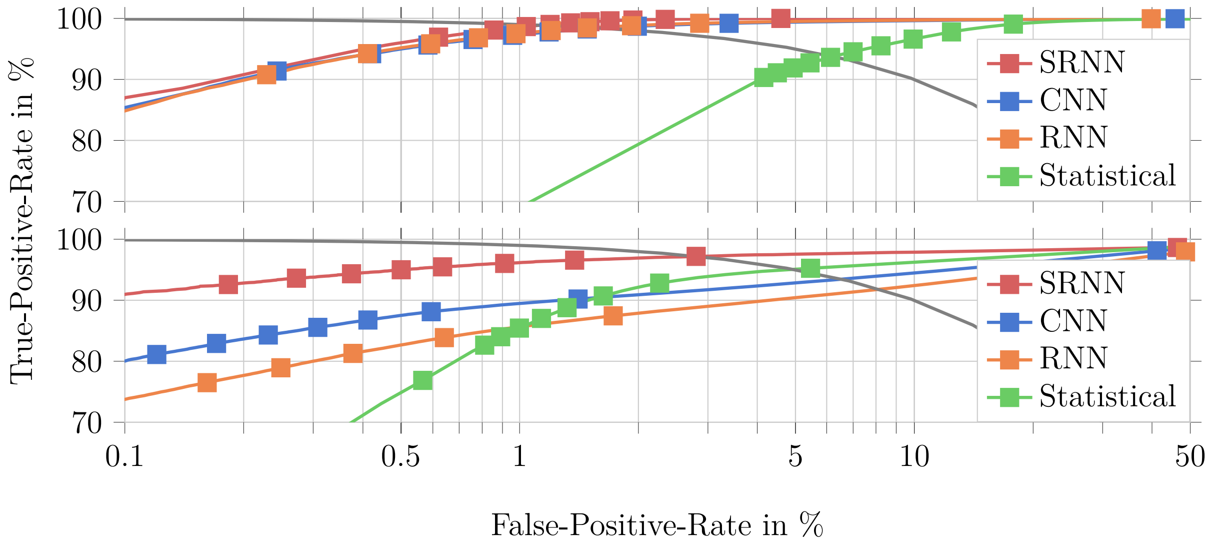


Figure 5.9: ROC curve for all presented SAD systems on the development set of the Fearless Steps challenge (first row) and the evaluation set of our real HF data (second row). Each square represents a different threshold. The gray line symbolizes the EER.

still allowing real-time processing. In particular, the DCF value is improved compared to the system with the next lowest DCF. For this metric the SRNN leads to a relative reduction of 47.7 % on the Fearless Steps database and 43.6 % on our recordings. Although the RTF is higher than for all other systems, it is still very low.

One should note, that the improvements with the SRNN as temporal layer are due to the more sophisticated architecture and not just the high number of parameters. The RNN layers used for the RNN and the SRNN temporal layer both have 1,677,312 trainable parameters. Although, this is more than the 26,250 parameters of the CNN this still cannot

Table 5.1: Results for all presented systems for different metrics on the development set for the Fearless Steps database and the evaluation set for our real HF radio database. The arrow after each metric name indicate whether higher (↑) or lower (↓) values are an improvement for this metric. The best result for each metric and database is shown in bold font.

Database	System	PRC/% ↑	REC/% ↑	F1/% ↑	DCF/% ↓	RTF ↓
Fearless	Statistical	80.91	97.97	88.63	4.76	**0.0047**
	CNN	95.17	99.12	97.10	1.30	0.0181
	RNN	94.98	99.01	96.95	1.41	0.0276
	SRNN	**96.19**	**99.77**	**97.95**	**0.68**	0.0320
Our HF recordings	Statistical	79.06	94.74	86.19	4.93	**0.0047**
	CNN	71.17	93.31	80.75	6.25	0.0181
	RNN	62.87	92.04	74.71	7.11	0.0376
	SRNN	**85.60**	**97.14**	**91.00**	**2.35**	0.0445

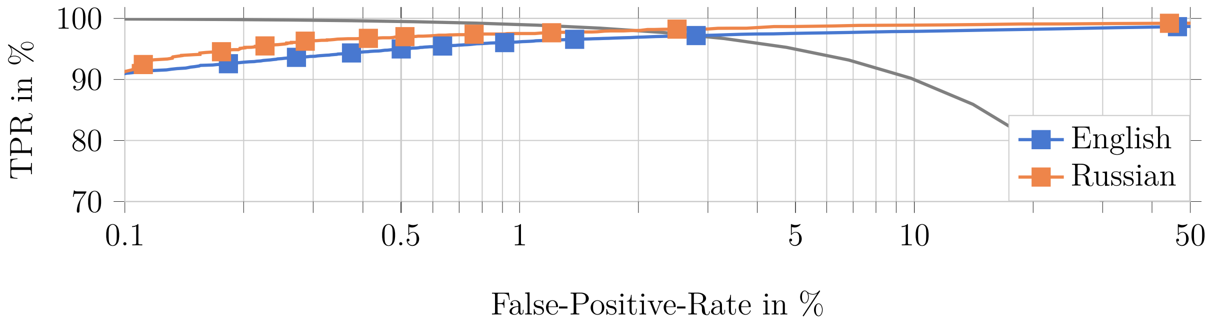


Figure 5.10: ROC curve for the SRNN SAD systems on the English and Russian evaluation set. Each square represents a different threshold. The gray line symbolizes the EER.

explain the superior results because otherwise the RNN temporal layer would outperform the CNN.

To assess whether the presented system is robust to a language mismatch between training and evaluation data, the model with the SRNN layer chosen above is applied to the Russian evaluation set without CFO. In Figure 5.10 the ROC curve of the SRNN on the Russian evaluation set is compared to the results on the English evaluation set.

Although no Russian data was used during training, the SAD system shows strong detection performance on the Russian evaluation set. The results on the Russian data are even slightly better than the results on the English data, which might be explained by a smaller number of examples with concurrent speakers. As the system performs very well on this previously unseen language, we can only assume a certain language independence.

5.5.2 SAD on recordings with concurrent speakers

In the previous experiments the models are compared to ground truth labels that have been obtained automatically that means that concurrent speakers are not annotated as activity as discussed in Section 4.8. Next, the influence of concurrent speakers on the presented system is examined by evaluating the activity detector with manual annotations as ground truth, where concurrent speakers are marked as activity. The resulting ROC curve is compared to the previous results on the automatic annotation in Figure 5.11.

If the activity detected by neural network is compared with the manual annotations the TPR is much lower than for a comparison with the automatic annotations. This is to be expected, since the network is trained on the automatic annotations as ground truth, where the concurrent speakers recorded are not annotated as activity. Since the concurrent speakers are recorded mainly with a CFO above 1000 Hz, the network learns to focus on the lower frequencies and dismiss activity in the upper frequencies.

The network is able to learn this difference between a speech signal in the higher and lower frequencies despite the potential shift equivariance of the input CNN with subsequent pooling. Shift equivariance to translation means that a CNN has a limited receptive field, so that during training the CNN learns to recognize certain patterns in respect to the neighboring bins, independent of the specific position of the neighborhood in the input signal [38, p. 329]. The chosen CNN configuration leads to a large receptive field over the frequency dimension of $((1 \cdot 8 + 2 + 2) \cdot 8 + 2 + 2) \cdot 4 + 2 + 2 = 404$, which surpasses the size of the frequency dimension. Therefore, the CNN is able to exploit padding at the spectrum edges to encode positional information in the output values. Initial experiments have shown that the given network architecture is also able to tag concurrent speaker as active, if the annotation of the training targets is adjusted accordingly.

Although it is beneficial for the network to ignore activity in the higher frequencies in case of concurrent speakers, this is no longer true for recordings with a CFO. The next section is examining the impact of CFO in the evaluation data on the detection results. Furthermore, the influence of additional simulated data during training on the evaluation results in the case of a CFO is evaluated.

5.5.3 SAD on frequency shifted speech

As described in Section 3.2, the mismatch between the modulation and demodulation frequency and the resulting frequency shift has a major impact on speech quality. In [85] the resulting speech is described as ”chipmunk-like”, which perfectly describes the challenge of recognizing these recordings as speech, even though the network was trained only on signals with perfect demodulation ($\mathit{\Delta}^{\mathrm{f}} = 0\,\mathrm{Hz}$).

To evaluate the impact of a CFO in the evaluation set, the best performing model on the evaluation set with $\mathit{\Delta}^{\mathrm{f}} = 0$ is applied to the English HF recordings with $\mathit{\Delta}^{\mathrm{f}} \geq 0\,\mathrm{Hz}$ described in Section 4.6. The resulting ROC curves are displayed in Figure 5.12.

The activity detection becomes less accurate for increasing CFOs, which is expected, as the prior experiments to concurrent speakers have shown that the network learns to ignore activity in the higher frequencies if no activity is detected in the lower frequencies. For a CFO up to 300 Hz the network still detects most of the activity without a high FPR, but the results for signals with a CFO of 500 Hz and 1000 Hz are not acceptable with EERs above 10 %.

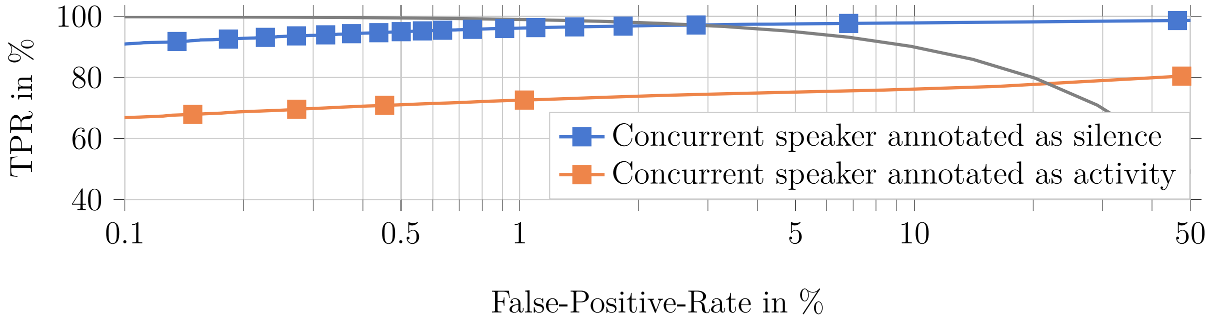


Figure 5.11: ROC curve for the SRNN SAD systems on the English evaluation set for the original annotation of the evaluation data, where concurrent speakers are not marked as activity, and the manual one, where they are annotated as activity. Each square represents a different threshold. The gray line symbolizes the EER.

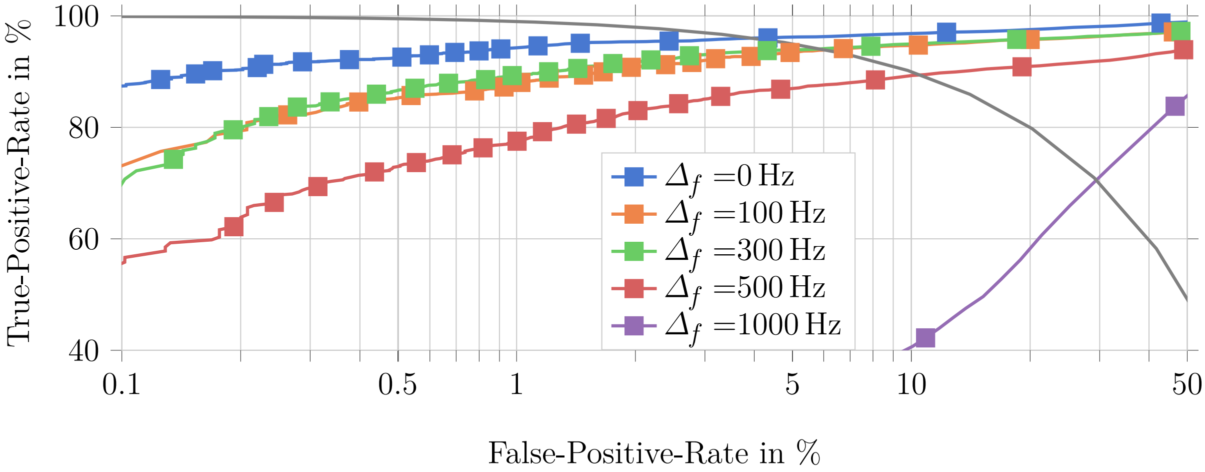


Figure 5.12: ROC curve for the SRNN SAD systems on the English evaluation set. Each square represents a different threshold. The gray line symbolizes the EER.

To improve the detection results, the mismatch between training and evaluation is reduced by augmenting the training data with additional simulated signals with a CFO as described in Section 4.9. For the simulation an additive noise with a random SNR between −10 DB and 5 DB is chosen. Since the network should not tag concurrent speakers in the higher frequencies as active, the additional training data is limited to CFOs between 0 Hz and 1500 Hz. Thereby, the network is still rewarded for ignoring activity in higher frequencies above 1500 Hz, which is considered a distortion in this work.

In Table 5.2 the results with different ratios between real and simulated training data are shown. For all models with simulated data in the training set, the threshold is calculated on a purely simulated development set. In contrast, a threshold is determined on the development set with $\Delta^{\mathrm{f}} = 0\,\mathrm{Hz}$ for the model solely trained on real data, as discussed for the last experiments.

For the model trained only on the real training set without a CFO the metric values shown in Table 5.2 are calculated from the same TP, FP, TN and FN values that lead to the ROC curves displayed in Figure 5.12. Training only on simulated data already provides improved results for most metrics and $\Delta^{\mathrm{f}} \geq 0$. However, for the examples with $\Delta^{\mathrm{f}} = 0\,\mathrm{Hz}$ the results are worse than the results with the original model trained only on real recordings without a CFO. This is especially clear for the EER, which is increased from 4.63 % for the training without a CFO to 6.46 % for the purely simulated training data with $\Delta^{\mathrm{f}} \geq 0$. Experiments on the evaluation set without a CFO for the model trained only on simulated signals with a CFO confirm this deficit.

Therefore, the effect of a mixture of real and simulated training data is evaluated. While the model trained on data equally distributed between real and simulated data show improved results for all examples with $\Delta^{\mathrm{f}} \leq 500\,\mathrm{Hz}$, there is a significant drop in performance for $\Delta^{\mathrm{f}} = 1000\,\mathrm{Hz}$ compared to the model trained only on simulated signals that exhibit a CFO. To increase the accuracy for higher errors in the demodulation frequency the share of simulated data in the training set is raised to 90 % simulated and 10 % real data. For more information about the reasoning behind this specific ratio please refer to Appendix A.2.

Table 5.2: Results for models with different ratios of real and simulated data in the training set on the evaluation set with $|\mathit{\Delta}^{\mathrm{f}}| \geq 0\,\mathrm{Hz}$ of the real HF radio database. The arrow after each metric name indicate whether higher (↑) or lower (↓) values are an improvement for this metric. All metrics are given in % and the best result per metric and $\mathit{\Delta}^{\mathrm{f}}$ is shown in bold font.

$\mathit{\Delta}^{\mathrm{f}}$ / Hz	Share of the training data		EER ↓	PRC ↑	REC ↑	F1 ↑	DCF ↓
	Real ($\mathit{\Delta}^{\mathrm{f}} = 0$)	Simu ($\mathit{\Delta}^{\mathrm{f}} \geq 0$)					
0	1	0	4.63	82.69	95.14	88.48	4.60
	0	1	6.46	67.96	95.83	79.52	5.29
	1/2	1/2	3.71	94.37	94.75	94.56	4.21
	1/10	9/10	4.88	80.98	94.75	87.33	5.00
100	1	0	7.34	67.97	92.83	78.48	7.43
	0	1	8.36	52.62	97.59	68.38	5.92
	1/2	1/2	6.36	84.55	91.74	88.00	6.98
	1/10	9/10	6.57	69.52	93.97	79.92	6.45
300	1	0	8.05	58.41	92.52	71.61	8.69
	0	1	3.01	78.23	97.53	86.82	3.12
	1/2	1/2	5.73	91.46	90.79	91.13	7.30
	1/10	9/10	4.35	88.91	94.42	91.58	4.74
500	1	0	14.87	45.56	85.98	59.55	15.23
	0	1	3.93	71.48	97.40	82.45	3.73
	1/2	1/2	7.77	92.99	87.09	89.94	9.98
	1/10	9/10	4.37	87.63	94.52	90.94	4.72
1000	1	0	41.42	42.22	25.42	31.74	57.62
	0	1	3.93	61.99	98.36	76.05	4.16
	1/2	1/2	14.78	86.50	48.96	62.53	38.65
	1/10	9/10	4.98	84.84	93.08	88.77	6.00

The higher share of simulated data improves the EER from 14.78 % to 4.98 % compared to the fifty-fifty ratio on the data with $\mathit{\Delta}^{\mathrm{f}} = 1000\,\mathrm{Hz}$. This is slightly worse than the 3.93 % of the model trained only on simulated data. However, the model with 90 % simulated training data outperforms the network trained only on simulated data on the signals with a low CFO of 0 Hz or 100 Hz by over 1.5 % EER.

To visualize the gain from the simulated data the EER is displayed in Figure 5.13 for the models with different training data over the five CFOs. The distinct difference in performance between the different demodulation frequency offsets for the original model trained only on real recordings without a CFO are significantly lowered for the new training data with simulated CFO. Furthermore, this change only leads to a negligible loss for $\Delta^{\mathrm{f}} = 0\,\mathrm{Hz}$.

For experiments with the models trained on simulated data on the Russian recordings refer to Appendix A.3. Here, the model trained on 90 % simulated signals also outperforms the original model and achieves similar results for all CFOs. Therefore, all future experiments will be performed with the 1:9 model if not stated otherwise.

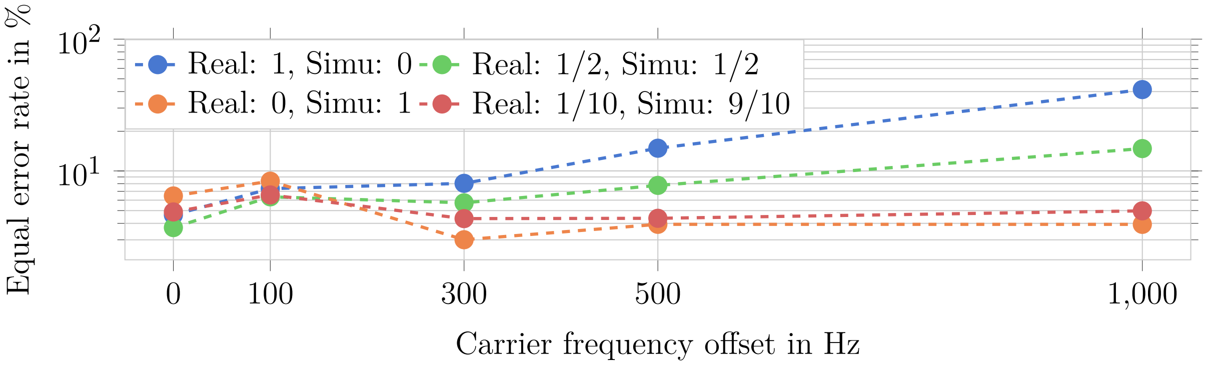


Figure 5.13: EER curve for the SRNN model trained on a combination of real data without a CFO and simulate signals with a CFO over the CFO in the evaluation data.

In the experiments in this section only positive CFO are considered since negative CFOs lead to a loss of the pitch and its first harmonics, so that the speech signal cannot be identified easily. On the other hand, if the activity detection is performed on the IQ-components of the baseband signal instead of the baseband signal itself, it is possible to detect speech even for negative CFOs. Therefore, the next section is going to examine the performance of the presented SAD system on the recorded IQ evaluation set.

5.5.4 SAD on IQ signals

As discussed in Section 4.7 there are benefits from using the complex baseband signal $y_n^+ = i_n + \mathrm{j}q_n$, which consists of the IQs-components $(i_n,\ q_n)$, instead of the real-valued demodulated signals. Therefore, the network with training data consisting of 90 % simulated signal is evaluated on the IQ evaluation set.

Since the networks are not shift-invariant, as discussed above, and are trained on signals with a Nyquist frequency of $F_{\max} = 4\,\mathrm{kHz}$, they cannot be applied directly to the complex baseband signal with a frequency range of 8 kHz. Therefore, the the complex baseband signal is divided into three overlapping segments, each consisting of a 4 kHz band and together covering the entire 8 kHz spectrum. An example for this segmentation of the frequency spectrum is shown in Figure 5.14. Note, that the experiments on the IQ signals are performed on lower sideband (LSB) transmissions by mirroring the spectrum on the time axis. Thereby, an equivalent upper sideband (USB) transmission is created, which can be processed by the network trained on real-valued demodulated signals.

To evaluate the performance for different CFOs, the IQ-components of the transmission are shifted to the baseband with a random CFO $\Delta^{\mathrm{f}} \in [0, 100, 300, 500, 1000]$ and the SAD is performed on both the complex-valued baseband signal and the real-valued demodulated signal. To illustrate the benefits of the complex-valued baseband signal another experiment is performed where both negative and positive CFOs are considered. Therefore, the set of possible CFOs is extended by [−100, −300, −500, −1000]. The results are presented in Figure 5.15

The EER for the estimator on the real-valued demodulated signals is higher than in previous experiments, which indicates that the recordings in the IQ evaluation set are more challenging than the ones in previous experiments. However, since the estimator on the real-valued demodulated signals has shown to perform well on other data sets with positive CFOs it is considered a strong baseline.

For positive CFOs both the network on the real-valued demodulated signals and the complex-valued baseband signal perform similarly, so that the activity estimation on the IQ-signal can be considered successful. If the signals also include negative CFOs the network processing the real-valued demodulated signals has a lower TPR, while the results on the complex-valued baseband are similar to the ones in the experiments without negative CFOs. These results are to be expected because the real-valued demodulated signals looses a high percentage of speech energy when the CFO is negative, while the complex-valued baseband still includes the entire signal spectrum as discussed in Section 4.7.

In the discussed experiments, we have only considered a small excerpt from each of the recorded 64 kHz IQ-signals. Even so, the results can be transferred from the 8 kHz excerpt to the entire 64 kHz signal. Since the SAD system is only processing 4 kHz segments of the considered frequency range, the entire spectrum can be segmented and each segment can be processed independently by the SAD system. Therefore, the presented approach can be used to search the entire 64 kHz spectrum for activity by a single-sideband transmission.

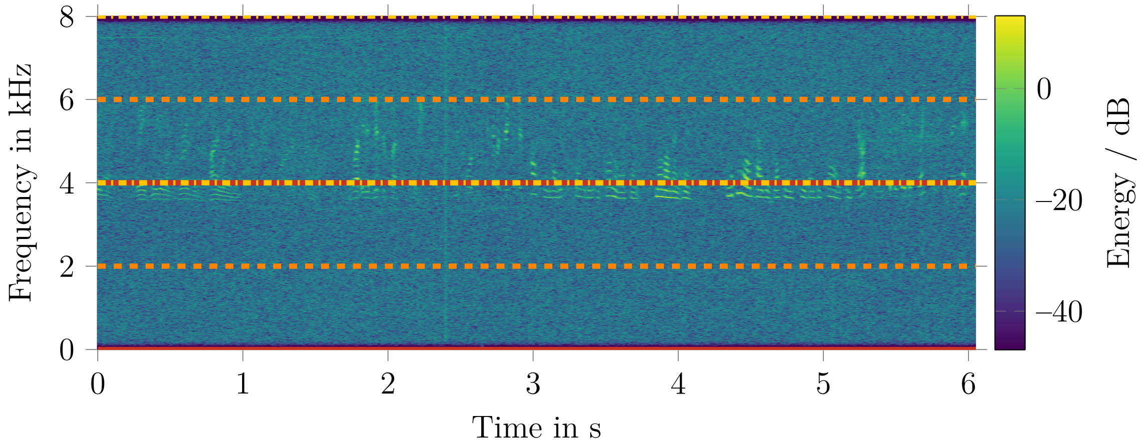


Figure 5.14: Example for the segmentation of the complex baseband signal for SAD with NNs trained on real-valued recordings, where the segments are 0 Hz-4000 Hz (red line), 2000 Hz-6000 Hz (orange dashed line) and 4000 Hz-8000 Hz (yellow dashed and dotted line)

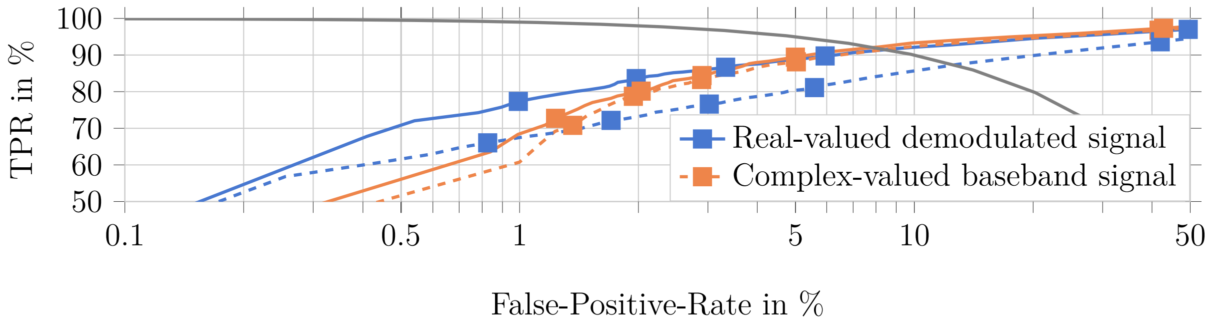


Figure 5.15: ROC curve for the SRNN SAD systems on the IQ evaluation set with simulated, positive CFOs (solid) or simulated positive and negative CFOs (dashed). The gray line symbolizes the EER.

5.6 Summary

In this chapter a SAD system with a novel temporal layer called SRNN layer is presented. It is shown that the neural network is able to detect speech activity for both English and Russian recordings, despite a training set consisting only of English recordings. However, errors in the demodulation frequency $\mathit{\Delta}^{\mathrm{f}}$ still lead to high errors in the detection rate, which can be mitigated by training on additional simulated signals that exhibit a CFO. This significantly improves the results on real recordings with positive CFOs ($\mathit{\Delta}^{\mathrm{f}} > 0$), without sacrificing accuracy on signals with $\mathit{\Delta}^{\mathrm{f}} = 0$. Therefore, all following experiments are run with the following parameters if not stated otherwise:

- SRNN with $\tilde{L}_W = 1000\,\mathrm{ms}$ and $\tilde{L}_S = 500\,\mathrm{ms}$ as temporal layer
- Training data that consists of 10 % real and 90 % simulated signals

Additionally, this best performing SAD system is evaluated on the complex-valued instead of the real-valued recordings. It is shown that the system can perform similarly on the complex signal and even improve the detection for negative CFOs.

The presented SAD system is the first building block in the processing pipeline introduced in Chapter 1. The next chapter will deal with the CFO estimation and correction, assuming perfect activity information provided by the SAD system.

6 Carrier Frequency Offset Estimation and Correction

The speech activity detection (SAD) algorithm described in the last chapter is shown to detect speech despite the possible frequency shifts in the recorded signals due to an error in the demodulation frequency, i.e., a carrier frequency offset (CFO). However, most other tasks become significantly more difficult in case of a frequency shift $\mathit{\Delta}^{\mathrm{f}}$. Already a small CFO can lead to a loss in intelligibility of the recorded speech signal as discussed in Section 3.2. Therefore, this chapter deals with estimating demodulation frequency errors and reverting the subsequent frequency shift.

First, three systems for CFO estimation are introduced, a pitch-tracking based algorithm called RAKE and two neural network (NN)-based approaches. The estimated CFO $\hat{\mathit{\Delta}}^{\mathrm{f}}$ is then used to revert the frequency shift by following steps similar to the simulation framework discussed in Section 4.9. Afterwards, the estimation algorithms are compared in terms of their accuracy and the impact of estimations errors on the speech signal's quality and intelligibility. Finally, the estimators are compared with respect to their performance on signals with a negative CFO and the application of the estimators to the complex-baseband signal instead of real-valued recording used in the other experiments is proposed to improve the results.

6.1 RAKE estimator

The following estimator exploits the structure of the short time Fourier transform (STFT) representation of voiced speech signals, which can be considered to show an approximately harmonic pattern [57, p. 182]. These patterns allow the algorithm to recognize a shifted speech signal by its fundamental frequency, i.e., its pitch, and the harmonic repetitions. An intuitive explanation for the estimator is that filters representing different pitches f^{p} and their harmonics $(2 \cdot f^{\mathrm{p}}, 3 \cdot f^{\mathrm{p}}, ...)$ are moved through the frequency spectrum of the input signal for different shifts $\mathit{\Delta}^{\mathrm{f}}$, and the correlation with the input signal is calculated for each pitch and shift combination $(\mathit{\Delta}^{\mathrm{f}} + f^{\mathrm{p}}, \mathit{\Delta}^{\mathrm{f}} + 2 \cdot f^{\mathrm{p}}, \mathit{\Delta}^{\mathrm{f}} + 3 \cdot f^{\mathrm{p}}, ...)$. The pitch and shift combination with the highest correlation with the input signal is chosen as the estimate for pitch and CFO. The pitch- and shift-varying filter is comparable to dragging a rake through the spectrum, so that the algorithm is called RAKE.

We presented the RAKE algorithm in [OC15], where a filterbank $H_f(f^{\mathrm{p}})$ with varying center-frequency indices is used to locate the pitch index f^{p} and its harmonics in the noisy speech. In Figure 6.1 examples for the filters $H_f(f^{\mathrm{p}})$ are shown. The filters are designed

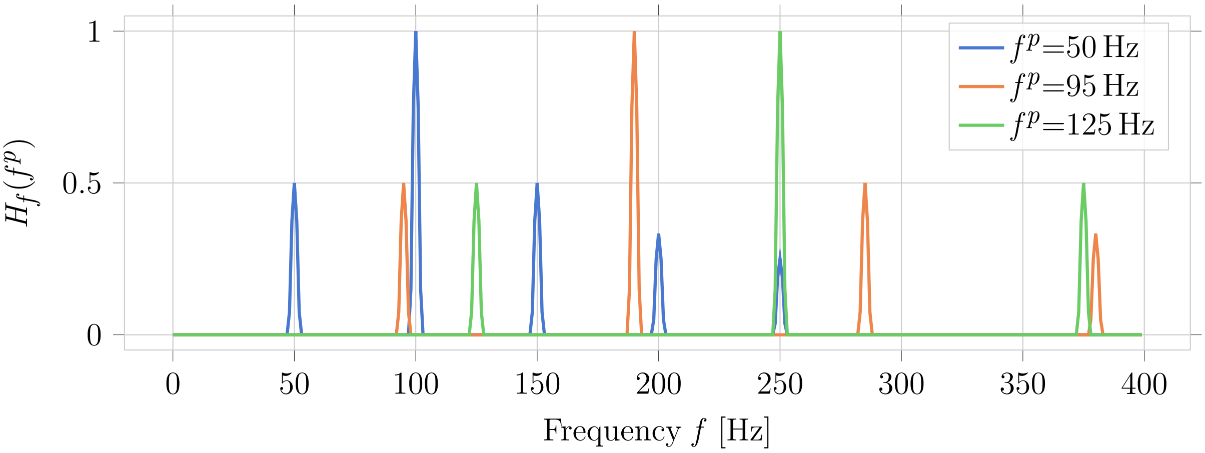


Figure 6.1: Examples for the filters $H_f(f^{\mathrm{p}})$ with $O = 4$ and $W = 2$.

such that they are zero for all frequency indices except the pitch, its harmonics and their vicinity:

$$H_f(f^{\mathrm{p}}) = \sum_{o=1}^{O} \sum_{\nu=-W}^{W} \omega(o, \nu) \cdot \gamma\left(f - o \cdot f^{\mathrm{p}} - \nu\right). \tag{6.1}$$

Here, O stands for the number of considered harmonics and W defines how many of the frequencies surrounding a harmonic are considered. γ is the Kronecker delta and f represents the frequency bin index. The weights are defined as

$$w(o, \nu) = \begin{cases} 0.5 \cdot \wedge(\nu) & \text{if } o = 1 \\ 1/(o-1) \cdot \wedge(\nu) & \text{for } 1 < o <= O \\ 0 & \text{else} \end{cases}, \tag{6.2}$$

with $\wedge(\nu)$ as a filter in triangular shape. The weight for the pitch ($o = 1$) is lower than the weight of the first harmonic ($o = 2$), since in several publication such as [171] it is reported that in many recordings the pitch is only weakly observable, while the first harmonic is clearly visible. A similar pattern is witnessed in multiple of the real high frequency (HF) recordings discussed in **??**.

As the pitch of human voices is between 80 Hz and 500 Hz [57, p. 65], the considered values for f^{p} are also restricted to this interval with $\Omega^{f^{\mathrm{p}}}$ as the set of possible pitch frequencies. To accommodate rapid changes in the pitch over time [172] the pitch frequency is considered time-varying for the following discussion: $f^{\mathrm{p}} = f^{\mathrm{p}}_{\ell}$ with ℓ being the frame index. As an intermediate result the correlation between the filters $H_f(f^{\mathrm{p}}_{\ell})$ and the logarithmic power spectral density (PSD) of the STFT coefficients of the observed signal $Y_{\ell,f}$ are calculated

$$\Gamma_{\ell,f}\left(f^{\mathrm{p}}_{\ell}\right) = \log\left(\left|Y_{\ell,f}\right|^2\right) * H_f(f^{\mathrm{p}}) = \sum_{o=0}^{O} \sum_{\nu=-W}^{W} \omega(o, \nu) \cdot \log\left(\left|Y_{\ell,f-o\cdot f^{\mathrm{p}}_{\ell}-\nu}\right|^2\right), \tag{6.3}$$

where $*$ symbolizes the discrete convolution operation.

For each frequency bin there is a sequence of pitches $\mathbf{f}^{\mathrm{p}} = [f_0^{\mathrm{p}}, f_1^{\mathrm{p}}, ..., f_{L-1}^{\mathrm{p}}]^{\mathsf{T}}$ that maximizes the summation of $\mathit{\Gamma}_{\ell,f}\left(f_\ell^{\mathrm{p}}\right)$ over time

$$\hat{\mathbf{f}}^{\mathrm{p}} = \underset{\mathbf{f}^{\mathrm{p}} \in \mathit{\Omega}^{f^{\mathrm{p}}}}{\operatorname{argmax}} \sum_{\ell=0}^{L-1} \mathit{\Gamma}_{\ell,f}\left(f_\ell^{\mathrm{p}}\right) = \underset{\mathbf{f}^{\mathrm{p}} \in \mathit{\Omega}^{f^{\mathrm{p}}}}{\operatorname{argmax}} \mathit{\Gamma}_f\left(\mathbf{f}^{\mathrm{p}}\right), \tag{6.4}$$

where L represents the number of frames in an input signal. Then, the frequency bin with the largest $\mathit{\Gamma}_f\left(\hat{\mathbf{f}}^{\mathrm{p}}\right)$ is chosen as an estimate for the CFO $\mathit{\Delta}^{\mathrm{f}}$

$$\hat{\mathit{\Delta}}^{\mathrm{f}} = \underset{f \in \mathit{\Omega}^{\mathit{\Delta}^{\mathrm{f}}}}{\operatorname{argmax}} \tilde{\mathit{\Gamma}}_f\left(\hat{\mathbf{f}}^{\mathrm{p}}\right) \tag{6.5}$$

with $\mathit{\Omega}^{\mathit{\Delta}^{\mathrm{f}}}$ as a set of possible carrier frequency differences. Note, that the accuracy of the estimation is limited by the frequency resolution of the chosen STFT size.

In the original work the entire frequency spectrum is searched for the considered pitch and harmonic combinations. To reduce the computational complexity one can restrict $\mathit{\Omega}^{\mathit{\Delta}^{\mathrm{f}}}$ to a smaller subset if prior knowledge about the possible frequency differences is given with $\mathit{\Delta}^{\mathrm{f}} \in [\mathit{\Delta}^{\mathrm{f}}_{\mathrm{min}}, \mathit{\Delta}^{\mathrm{f}}_{\mathrm{max}}]$. Additionally, the limited range of $\mathit{\Delta}^{\mathrm{f}}$ prevents the algorithm from estimating the CFO of concurrent speakers instead of the target CFO.

In Figure 6.2 the pitch tracking result of the RAKE approach is shown for a real example utterance with $\mathit{\Delta}^{\mathrm{f}} = 500\,\mathrm{Hz}$. The red lines, which are a smoothed estimate for the pitch and its harmonics, clearly fit the speech signal. For this input signal the pitch f_ℓ^{p} has been estimated assuming a CFO of $\mathit{\Delta}^{\mathrm{f}} = 500\,\mathrm{Hz}$. The pitch search for any other $\mathit{\Delta}^{\mathrm{f}}$ results in a lower $\tilde{\mathit{\Gamma}}_f\left(\hat{\mathbf{f}}_f^{\mathrm{p}}\right)$ because either less harmonics or a wrong pitch are found. Therefore, $\hat{\mathit{\Delta}}^{\mathrm{f}} = 500\,\mathrm{Hz}$ is chosen as the estimate of the RAKE algorithm for the CFO for this example.

The discussed estimator has a high computational complexity, which can be reduced by replacing the correlation in Equation (6.3) with a multiplication in the cepstral domain. The complete algorithm is depicted as a block diagram in Figure 6.3. For a more extensive description of the algorithm, see [OC15].

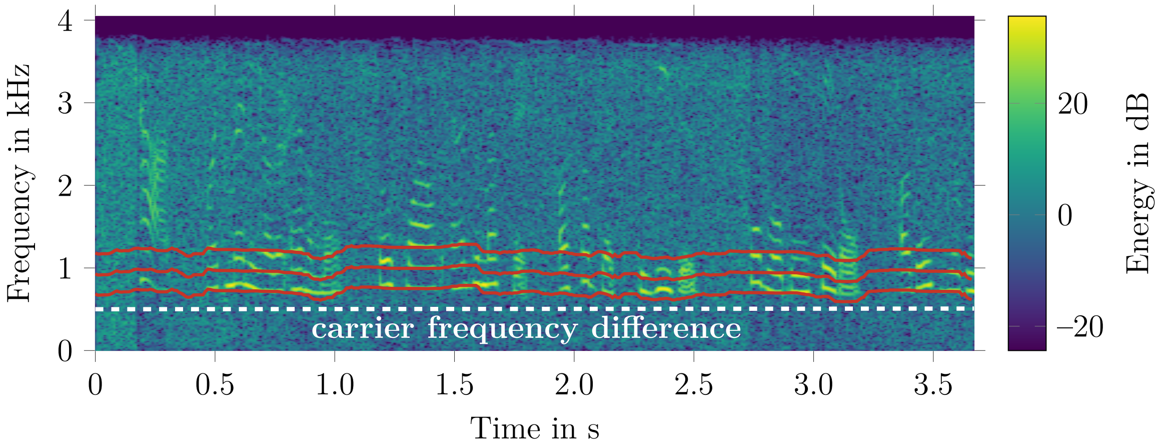


Figure 6.2: Example for the pitch tracking (red) in case of a CFO $\mathit{\Delta}^{\mathrm{f}} = 500\,\mathrm{Hz}$.

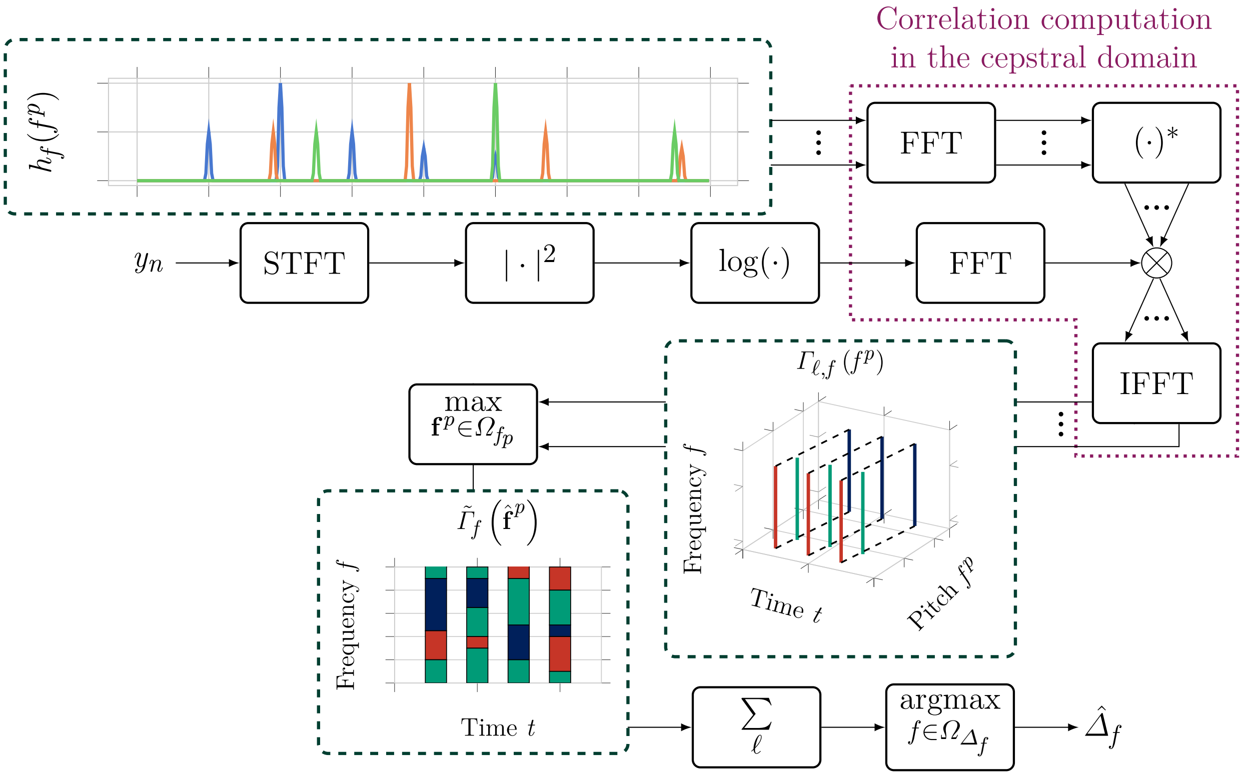


Figure 6.3: Block diagram of the RAKE carrier frequency difference estimation. Multiple arrows indicate that more than one pitch is considered. The three dots indicate that different possible pitch values and therefore different filter functions are considered.

6.2 NN-based estimators

In this section two NN-based estimators are presented. One uses information from the entire frequency spectrum of the input signal to estimate the position of speech activity in this spectrum. This allows the network to ignore activity in higher frequencies to prevent errors due to concurrent speakers, which are also present in some of the training data as discussed in Section 4.9. During the following discussion this network architecture is called full-band classifier. We published a first evaluation of this network architecture in [OC6].

As another network, a frequency-shift invariant convolutional neural network (CNN) architecture is chosen, which estimates for each frequency bin whether it includes the pitch of a speech signal. This network can be applied to signals with different frequency spectrum sizes, which makes it adaptable to the in-phase and quadrature (IQ) recordings. Due to the shift-invariant architecture, the network estimates the CFO for different excerpts from the frequency spectrum independently from each other. Therefore, it is called sub-band classifier during the following discussion. For both architectures the same loss function is used, which is described in Section 6.2.3

6.2.1 Full-band classifier

The full-band classifier is designed as a two-stage approach, where the first stage calculates an attention mask that is multiplied with the observed signal. Thereby, highlighting frequencies with high energy so that the second stage is less affected by additive noise. The second stage processes the masked observation to calculate a final CFO estimate.

As input for the mask estimation layer the magnitude of the input STFT signal $y_{\ell,f}$ is chosen. This input is represented by a vector $\mathbf{y}_\ell = [y_{\ell,0}, ..., y_{\ell,F-1}]^\mathsf{T}$, which is divided into $N_{\tilde{F}}$ sub-band vectors $\tilde{\mathbf{y}}_{\ell,i}$ with size $\tilde{F} = \lfloor F/N_{\tilde{F}} \rfloor$. Here, i represents the sub-band index and F the number of frequency bins. The sub-bands do not overlap and, possible remaining frequencies $F - N_{\tilde{F}} \cdot \tilde{F}$ near the Nyquist frequency are set to zero in the final mask. Each sub-band i is independently processed by a sub-band layer (SBL) to calculate a sub-band activity masking vector $\tilde{\mathbf{m}}_{\ell,i}$ of size $\tilde{F}$, where all SBLs share the same parameters. Estimating a mask per sub-band allows the network to highlight speech and suppress noise independent of the frequency band that contained the observed noise during training [OC6]

The masking vector $\mathbf{m}_\ell$ of size F is constructed from the sub-band mask vectors with $\mathbf{m}_\ell = [\tilde{\mathbf{m}}_{\ell,0}^T, ..., \tilde{\mathbf{m}}_{\ell,N_{\tilde{F}-1}}^T, \mathbf{0}^T]^T$, where $\mathbf{0}$ is a vector of zeros with the length $F - N_{\tilde{F}} \cdot \tilde{F}$ that represents the possible remaining frequencies after the segmentation of the input vector.

In [OC6] the vector $\mathbf{m}_\ell$ is multiplied to the input vector $\mathbf{y}_\ell$ resulting in the $\mathbf{o}_\ell^{\text{masked}} = \mathbf{m}_\ell \circ \mathbf{y}_\ell$ with $\circ$ as the Hadamard product. The masked vector is then further processed by the second stage of the network called full-band classification layer. However, the second stage can also work directly on the output vector of the first stage $\mathbf{o}_\ell^{\text{direct}} = \mathbf{m}_\ell$ to allow the first network stage to encode the important information per sub-band in the SBL output. Regardless of the input vector design, the second stage of the network processes the entire frequency range to predict the CFO value. The output vector $\hat{\mathbf{q}}_\ell$ consists of $\mathit{\Delta}_{\text{max}}^{\text{f}} - \mathit{\Delta}_{\text{min}}^{\text{f}} + 1$ values, where $\mathit{\Delta}_{\text{min}}^{\text{f}}$ and $\mathit{\Delta}_{\text{max}}^{\text{f}}$ represent the lowest and highest considered CFO values, respectively.

For the SBL two architectures as discussed in [OC6] are compared, which are a multi-layer recurrent neural network (RNN) or a 1D-CNN block consisting of three CNN layers with batch normalization [163]. The output activation for both architectures is either a Sigmoid function if the input to the second network stage is the masked input vector $\mathbf{o}_\ell^{\text{masked}}$ or a rectified linear unit (ReLU) function for $\mathbf{o}_\ell^{\text{direct}}$ as the input to the full-band classification layer.

As the full-band classification layer a 1D-CNN block with a subsequent fully-connected (FC) layer is chosen, which is the same architecture we proposed in [OC6]. The 1D-Block consists of three CNN layers each including a batch normalization and a ReLU activation function. For the FC block two linear layers are used, the first with a ReLU and the second with a Softmax activation function. The full network is displayed in Figure 6.4.

As discussed above, the presented architecture allows the network to combine information from the whole frequency spectrum. This offers the possibility to ignore speech activity in higher frequencies which can be attributed to concurrent speakers and only evaluate speech

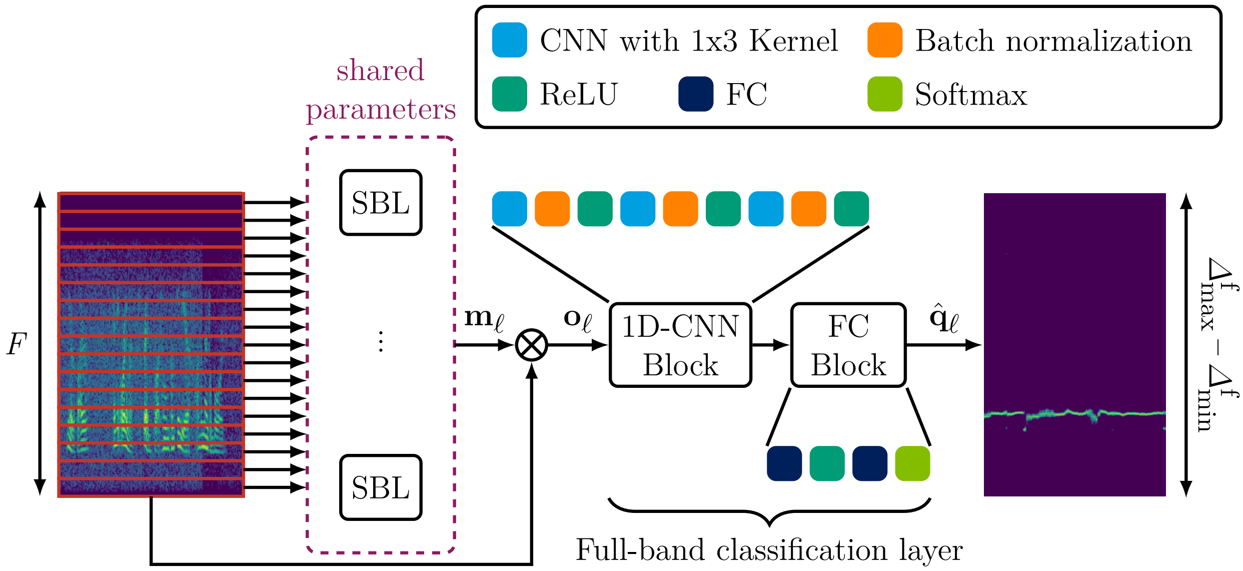


Figure 6.4: Block diagram of the NN-based full-band classifier for CFO estimation, first presented in [OC6] with $N_{\tilde{F}} = 128$.

signals in the frequency range seen during training. However, this also leads to a network that can only be applied to a predefined frequency range. Therefore, it cannot easily be adapted to process signals with frequency range larger than the one seen during training, e.g., the complex-valued baseband signal. The next section introduces a sub-band classifier network, which can be applied to signals with different frequency ranges without retraining or other adaptation schemes.

6.2.2 Sub-band classifier

Another considered network is inspired by the success of the WaveNet architecture [46] in speech signal generation. Here, a combination of CNN layers with increasing dilation allows the layers to exploit an exponentially growing context without reducing the resolution [173]. In CNNs dilation represents a distance between two bins that are simultaneously processed by one kernel. For example, a dilation of 3 with a kernel of size 2 means that the n-th output is calculated by calculating the convolution between the CNN kernel and a vector consisting of the n-th and $n + 3$-th input values.

For the CFO estimation the properties of a dilated convolution are exploited to increase the context over the frequency dimension without reducing the resolution of the final estimate. Therefore, a 2D-CNN architecture is designed, which consists of multiple CNN layers with an exponentially increasing dilation over the frequency dimension for each subsequent layer. Maximum pooling over the time dimension is used to increase the temporal receptive field. Each CNN layer has a larger number of channels compared to the previous layer, to aggregate the information from the growing frequency context. For each frequency bin, the output of the CNN-block is processed by applying a FC layer to reduce the channel dimension to one. Here, all FC layers share their parameters to allow a shift-independent training of these output layers. The network output is passed through a Sigmoid activation to limit the results

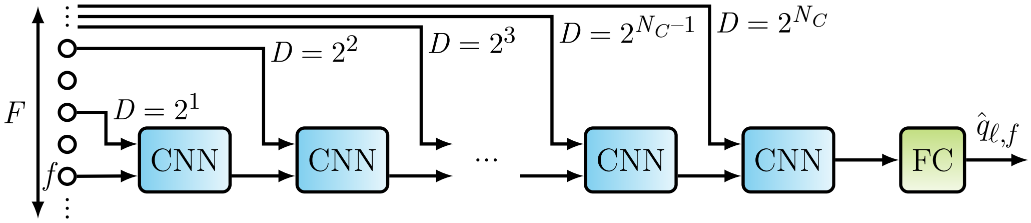


Figure 6.5: Illustration of the receptive field along the frequency dimension for the sub-band classifier with CNN kernels of size 2 along the frequency dimension. The illustration is designed to clarify the receptive field for one example time frame ℓ and frequency bin f. D is the dilation factor.

to values between 0 and 1. This allows the output to be interpreted as a confidence for each frequency bin whether it contains the pitch of the transmitted speech.

In the case of exponential to the base of two increasing dilations over the frequency dimension, the CNN has a receptive field of 2^{N_C} with N_C as the number of CNN layers. Figure 6.5 illustrates the receptive field for one frequency bin f.

Both the sub-band and full-band classifier share the same loss function, which is described in the following section.

6.2.3 Loss function

In general, the CFO estimation could be considered a typical regression task. However, preliminary results have shown that a regression loss function like the mean squared error (MSE) leads to poor results. Inspired by the observation presented in [132] that some regression tasks benefit from being redefined as a classification task, a binary cross entropy (BCE) loss, as used in Equation (5.3) for the SAD model, is considered for the CFO estimation.

For the full-band classifier a discrete number of possible CFOs in the range from $\varDelta^{\mathrm{f}}_{\mathrm{min}}$ to $\varDelta^{\mathrm{f}}_{\mathrm{max}}$ is chosen, while the sub-band estimator estimates a value for each frequency bin. The target vector $\mathbf{q} = [q_{\varDelta^{\mathrm{f}}_{\mathrm{min}}}, ..., q_{\varDelta^{\mathrm{f}}_{\mathrm{max}}}]^{\mathsf{T}}$ is compared to the time average $\hat{\mathbf{q}} = [\hat{q}_{\varDelta^{\mathrm{f}}_{\mathrm{min}}}, ..., \hat{q}_{\varDelta^{\mathrm{f}}_{\mathrm{max}}}]^{\mathsf{T}}$ of the network output $\hat{\mathbf{q}}_{\ell}$. $\mathbf{q}$ is designed as an one-hot vector $\mathbf{I}^{\varDelta^{\mathrm{f}}}$ with length $\varDelta^{\mathrm{f}}_{\mathrm{max}} - \varDelta^{\mathrm{f}}_{\mathrm{min}} + 1$, which is only one for the index representing the true CFO $\varDelta^{\mathrm{f}}$ and zero else. Therefore, an error of 1 Hz leads to the same loss as an error of 1000 Hz. This behavior is undesirable because small errors cannot be perceived by human listeners and consequently the uniform consideration of errors is not appropriate.

Similar solutions have been reported in literature, e.g., the ranking regression loss in [174]. However, most of these loss functions are continuous over the error, while in CFO estimation errors up to 5 Hz [86] do not impact the listening experience and should therefore be minimally considered. All other errors should be weighted higher, as they affect speech intelligibility.

To take these consideration into account, an auxiliary target vector

$$\tilde{\mathbf{q}}^{\mathrm{CFO}} = [\tilde{q}_{\mathit{\Delta}^{\mathrm{f}}_{\mathrm{min}}}, .., \tilde{q}_i, ..., \tilde{q}_{\mathit{\Delta}^{\mathrm{f}}_{\mathrm{max}}}]^{\mathsf{T}}$$

is introduced with

$$\tilde{q}_i = \begin{cases} \left(1 - \frac{|i - \mathit{\Delta}^{\mathrm{f}} \cdot \frac{F}{F_{\mathrm{max}}}|}{L_{\mathrm{I}}}\right)^2 & \text{if } |i - \mathit{\Delta}^{\mathrm{f}} \cdot \frac{F}{F_{\mathrm{max}}}| < L_{\mathrm{I}} \\ 0 & \text{else} \end{cases}, \tag{6.6}$$

where L_{I} refers to the length of an error tolerance interval, for example the number of frequency bins equivalent to 5 Hz and F_{max} to the Nyquist frequency. The weighted target vector $\tilde{\mathbf{q}}$ defines a reduced penalty for smaller deviations from the target CFO.

For the final loss a weighted sum over the both BCE losses is calculated

$$\mathcal{L}^{\mathrm{CFO}}\left(\hat{\mathbf{q}}^{\mathrm{CFO}}, \mathbf{q}^{\mathrm{CFO}}\right) = (1 - \alpha) \cdot \mathcal{L}^{\mathrm{BCE}}\left(\hat{\mathbf{q}}^{\mathrm{CFO}}, \mathbf{I}^{\mathit{\Delta}^{\mathrm{f}}}\right) + \alpha \cdot \mathcal{L}^{\mathrm{BCE}}\left(\hat{\mathbf{q}}^{\mathrm{CFO}}, \tilde{\mathbf{q}}^{\mathrm{CFO}}\right) \tag{6.7}$$

with α as the loss weight.

During evaluation, the time dimension of the network output $\hat{\mathbf{q}}_\ell$ is removed by computing the sum over it. This makes the estimation more robust to large silence segments in the input signals. Afterwards, the median of the five possible CFOs with the highest corresponding output value is chosen as the estimate to reduce the influence of possible outliers. Afterwards, the CFO estimation can be used to reverse the CFO. A possible approach for this is discussed in the next section.

6.3 CFO correction

After the CFO is estimated, the calculated offset $\hat{\mathit{\Delta}}^{\mathrm{f}}$ can be corrected by following the steps similar to the CFO simulation in Section 4.9. In short, the input signal is interpolated and the frequency spectrum is shifted to a higher frequency which corresponds to an adjusted carrier frequency $\tilde{f}_0 = f_0 - \hat{\mathit{\Delta}}^{\mathrm{f}}$. Afterwards, the repetition introduced by shifting the signal to a higher frequency is removed by limiting the bandwidth using a band-pass filter with the transmission bandwidth 2.7 kHz. The remaining single-sideband (SSB) signal is shifted to the zero frequency using the simulated carrier frequency f_0 and the signal is low-pass filtered with the bandwidth as cut-off frequency. At last the signal is downsampled to the original sample rate.

If $\mathit{\Delta}^{\mathrm{f}}$ is zero, a perfect reconstruction of the band-limited original signal is possible, by applying a low-pass filter to limit the signal to the 2.7 kHz transmission bandwidth. However, in case of a CFO ($|\mathit{\Delta}^{\mathrm{f}}| \neq 0$) the spectrum is shifted compared to the original spectrum and some signal information may be lost due to the filtering during demodulation. Since more than 80 % of the energy of speech signals is located in only 10 % of the lower frequencies [87], most of the information is retained for a positive CFO. In case of a shift towards the negative frequencies because of a negative CFO the signal cannot be easily reconstructed for larger differences between the demodulation and modulation frequency since most of the speech energy may be lost. To account for a negative CFOs, the shift estimation and correction can be performed on the IQ components instead of the real-valued recordings, as discussed in Section 4.7 and further elaborated in Section 6.4.5. This discussion assumes a upper sideband (USB) transmission but the same considerations are true for lower sideband (LSB) modulation only that a negative CFO leads to a shift towards the higher frequencies and a positive shift leads to a shift towards the lower frequencies.

In the following section the CFO estimators discussed above are evaluated with respect to their estimation accuracy and the influence of their estimation errors on the signal quality and intelligibility after the CFO correction. These investigations are performed on the real-valued recordings for positive CFOs and then extended to negative CFOs by using the IQ components.

6.4 Evaluation

This section includes a comparison of the RAKE and NN-based CFO estimators. First the parameter of the NN-based estimators are fine-tuned. The resulting best network is compared to the RAKE estimator.

Most of the experiments are performed on the evaluation set of the database of real recordings described in Section 4.6. Only the parameter tuning for the NN-based CFO estimators is performed on simulated data to prevent an overfitting to the real HF data. The following section gives more insight in the considered evaluation data sets.

6.4.1 Database

Two databases are used during evaluation. One consists of simulated HF data as described in Section 4.9, where the clean audio signal is taken from the LibriSpeech data set. This data is used for training and fine tuning of the NN parameters. During evaluation of the CFO estimators, the CFO $\mathit{\Delta}^{\mathrm{f}}$ is a random value between 0 Hz and 1500 Hz and the signal to noise ratio (SNR) is randomly drawn from the interval –5 dB to 5 dB.

The other database is a HF database of real HF transmissions described in Section 4.6, which is only used during evaluation. Results are presented for both the Russian and English recordings with a CFO. In Table 6.1 all three evaluation sets are compared. The real recordings contain both audio with and without a CFO.

The next section gives an overview of different performance measures used to evaluate the presented CFO estimators.

6.4.2 Performance measures

For the real HF data, two ways to calculate the performance measures are considered. Either the scores are calculated on a concatenation of all activity segments in a recording, or, in order to assess the influence of shorter utterances, the performance is evaluated individually

Table 6.1: Comparison of the evaluation sets for CFO estimation and correction.

Name	Simulated	CFO	Duration in h	Speech activity in %	SDR in dB	
					min	max
HF Simu	✓	✓	10.81	83.06	–5.0	5.0
Evaluation	✗	✗	37.12	12.53	–21.4	5.8
Evaluation Shift	✗	✓	23.53	14.16	–24.2	5.8
Russian	✗	✗	6.50	5.84	–18.0	7.9
Russian Shift	✗	✓	25.54	5.53	–21.3	10.5

for each of the five excerpts from a LibriSpeech utterance per recording. For more information about the design of the transmitted clean signals and the LibriSpeech excerpts refer to Section 4.2. If not stated otherwise, the experiments are performed on the activity segments of an entire recording. The CFO correction has two aspects that can be evaluated. First, the accuracy of the CFO estimation and secondly, the impact of the correction on speech enhancement.

Carrier frequency offset estimation

Two evaluation measures are considered to assess the quality of the CFO estimation. Intuitively, the estimation error $e^{\varDelta^{\mathrm{f}}} = \varDelta^{\mathrm{f}} - \hat{\varDelta}^{\mathrm{f}}$ or the MSE would be chosen as a score. However, for an improved visualization either the error class affiliation (ECA) or the cumulative distribution function (CDF) of the estimation error is calculated.

For the ECA, the errors $e^{\varDelta^{\mathrm{f}}}$ are grouped in four classes representing different level of signal quality after the CFO correction. The first class includes all errors smaller than 5 Hz, which lead to a remaining CFO that is not audible to most human ears [85]. As a second class the errors between 5 and 10 Hz are grouped, which produce a CFO that only slightly impacts the sound quality [86]. For the third error class the errors between 10 and 50 Hz are chosen, which lead to signals that are still intelligible despite the strongly diminished sound quality. The last error class includes all estimation errors greater than 50 Hz. Even after correction, the remaining frequency shift of more than 50 Hz results in signals that are either very difficult to understand or even completely incomprehensible.

Carrier frequency offset correction

After estimating the CFO it is corrected to enhance the recorded signal. The speech enhancement success is evaluated with both perceptual evaluation of speech quality (PESQ) [175], a speech quality measure, and the short-time objective intelligibility (STOI) score [159].

PESQ represents a prediction of the subjective mean opinion score [57, p. 85] with a value range between 1.0 and 4.5 [175] where higher values represent a better speech quality. It consists of a time alignment, equalization of gain variation and other normalizations, which allow for reliable quality measurement even for the HF audio with amplitude variations due to the automatic gain control (AGC).

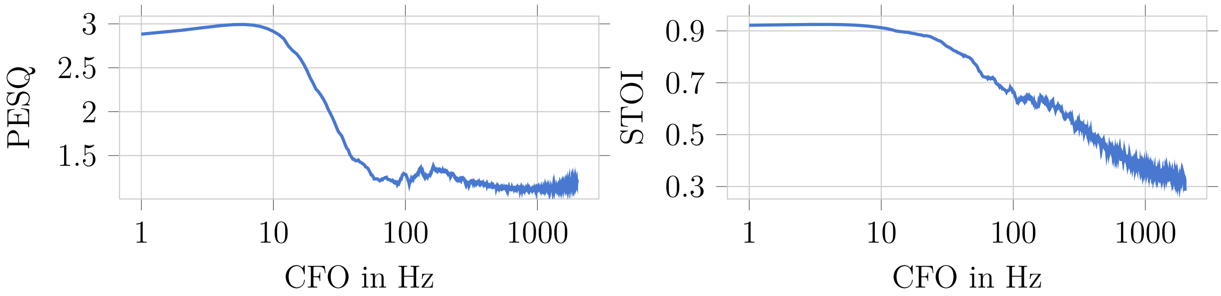


Figure 6.6: PESQ and STOI values for 10000 signals with a high SNR between 25 and 30 dB and a simulated CFO (First published in [OC6]).

STOI measures the intelligibility by comparing the transformed clean and distorted signals. Here, the value range is limited to a range between 0.0 and 1.0 with higher values representing better intelligibility. Similar to PESQ the score is calculated for short segments, where the observed signal is normalized to the amplitude level of the target signal. The normalization reduces the influence of the changes in the amplitude introduced by the AGC on the performance measure.

For both scores, the reference speech signals are low-pass filtered to account for the limited transmission bandwidth of 2.7 kHz. This reduction of the frequency range reduces the difference between the target signal and the recorded signal. Since both performance measures are not designed to evaluate the impact of a CFO on a speech signal, it is first examined whether there is a correlation between a reduction in the score and an increasing CFO. We performed such an investigation in [OC6], where the PESQ and STOI values are calculated for a set of 10000 signals with varying CFOs. The CFOs are simulated as described in Section 4.9 with a SNR uniformly drawn at random between 25 and 30 dB. From the graphs in Section 6.4.2 a clear correlation between the PESQ score and the CFO between 10 Hz and 100 Hz can be seen. This correlation does not extend to values outside of this interval. Nevertheless, the PESQ score is still meaningful for signals with a CFO, since it still distinguishes between high speech quality for signals with a CFO below 10 Hz and very low quality for CFOs above 100 Hz. The STOI score displays a clear correlation with CFOs larger than 10 Hz. Therefore, both performance measures can be considered appropriate choices to examine the influence of the CFO correction on the signal quality.

The evaluation of the CFO estimators starts with a parameter tuning for the NN-based CFO estimators on the simulated HF database.

6.4.3 NN parameter tuning

All NN-based CFO estimators are trained on the training set of the LibriSpeech database [155]. The signals are downsampled to 8 kHz and distorted with the simulation setup explained in Section 4.9. A similar configuration of the simulation is used as described for the SAD training in Section 5.5.3, with a SNR in the range between −5 dB and 10 dB. The CFOs for the full-band classifier are chosen in the range $\Delta^{\mathrm{f}} \in [-100\,\mathrm{Hz}, 1500\,\mathrm{Hz}]$ and for the sub-band classifier between 0 Hz and 3000 Hz. The different CFOs are chosen to account for the different tasks of classifiers. For the full-band classifier the limited range of CFOs seen during training allows the network to ignore speech activity in higher frequencies and thereby deal with concurrent speakers with a CFO above 1500 Hz. The sub-band classifier cannot easily learn to ignore specific frequency ranges because of its shift-invariant configuration. Therefore, it is trained on a larger range of possible CFOs. The possible CFOs are only limited by the information loss for very high CFOs due to the Nyquist frequency of 4 kHz.

For both architectures, the magnitude of the STFT coefficients of the input signal with a window size of 25 ms and a shift of 10 ms is chosen as a feature. The NN-based CFO estimators are dependent on multiple hyper-parameters that are optimized for the given task in the following subsections.

Full-band classifier

The full-band classifier consists of two stages a sub-band layer (SBL) and a full-band classification layer, as shown in Figure 6.4. A single sub-band in the SBL consists of $\tilde{F}$ frequency bins, where $\tilde{F}$ is fixed to a value representing 250 Hz. For the SBL, either a CNN or a RNN with configurations as displayed in Table 6.2 is chosen. Both SBL variants are designed to consists of roughly 300 k parameters.

For the full-band classification layer a fixed combination of CNN and FC layers is chosen with 4.7 M parameters and the configuration as displayed in Table 6.3 . Each CNN consists of the convolutional layer with or without a subsequent batch normalization [163] and an activation function.

Table 6.2: RNN and CNN configuration used in the SBL of the full-band classifier.

Sub-band layer	Index	Parameters	Batch-Norm	Activation
CNN	1	3x1 Kernel / channels=256	✓	ReLU
	2	3x1 Kernel / channels=256	✓	ReLU
	3	3x1 Kernel / channels=$\frac{F \cdot N_F}{F_{\text{max}}}$	✗	–
RNN	1	#units: 256	✗	–
	2	#units: 256	✗	–
	3	#units: $\frac{F \cdot N_F}{F_{\text{max}}}$	✗	–

Table 6.3: Architecture for the classifier layer of the NN-based CFO estimator.

Index	Layer type	Parameters	Batch-Norm	Activation
1	CNN	3x1 Kernel / channels=512	✓	ReLU
2	CNN	3x1 Kernel / channels=256	✓	ReLU
3	CNN	3x1 Kernel / channels=128	✓	ReLU
4	FC	#units: 512	✗	ReLU
5	FC	#units: $\mathit{\Delta}^{\text{f}}_{\text{max}} - \mathit{\Delta}^{\text{f}}_{\text{min}} + 1$	✗	Softmax

Table 6.4: Comparison of the NN-based CFO estimators with different $\mathbf{o}_\ell$ on the simulated HF evaluation data regarding their error class affiliation. All values are given in percentage of estimates in a given error class. For all results $F = 2049$ is fixed.

Sub-band layer	Masking	<5 Hz	5-10 Hz	10-50 Hz	>50 Hz
CNN	✓	58.97	39.72	1.02	0.29
	✗	47.52	44.42	8.06	0.00
RNN	✓	55.68	40.10	4.22	0.00
	✗	58.11	41.60	0.19	0.10

The first experiment evaluates whether masking improves the results over a direct calculation of $\mathbf{o}_\ell$ for both the CNN and RNN-SBL without the auxiliary loss, i.e., with a fixed loss weight $\alpha = 0$. For the feature size F a fixed value of 2049 is chosen as a compromise between high frequency resolution and computational complexity. The results are displayed in Table 6.4.

For the CNN-SBL the masking with $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{masked}}$ improves the results over the direct approach with $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ by reducing the fraction of errors above 10 Hz by almost 7 %. However, this is not true for the RNN-SBL, where $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ leads to about 4 % lower fraction of errors above 10 Hz, while the number of errors with $e^{\Delta^{\text{f}}} > 50\,\text{Hz}$ increases slightly.

To illustrate the difference between the networks with $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{masked}}$ and $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$, example signals for both types of $\mathbf{o}_\ell$ are depicted in Figure 6.7. For the CNN-SBL the vector $\mathbf{o}_\ell^{\text{masked}}$ includes only the pitch and the high energy harmonics of the speech signal. The rest of the signal spectrum is greatly reduced by the attention mask. Thereby, the network can follow a similar approach as RAKE to first detect the harmonics and then calculate the CFO via the frequency shift of the pitch. The direct approach also slightly highlights the speech active parts, however it also adds a smoothing over the time dimension.

For the RNN-SBL, the attention mask does not only highlight speech activity but also some of the background noise. Although $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ leads to the best results for the RNN excerpt layer, no discernible speech patterns remain. This signal representation might be beneficial for the NN but hinders an interpretation by the human eye. This is similar to the behavior of the time-domain audio separation network (TasNet) discussed in [OC3] and Section 7.5.

One explanation for the differences in $\mathbf{o}_\ell$ for the two possible SBL is the different temporal context. For the RNN layer $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ can encode the information for all previous frame in the current frame, while the CNN layer only aggregates information from eight neighboring frames. This is clearly seen in the example for $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$, where the CNN only uses local information so that a relation to the original signal can still be observed, which is not true for the estimation with a RNN-SBL.

Following the results in Table 6.4, all future experiments with the full-band classifier use $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{masked}}$ for the CNN and $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ for the RNN-SBL.

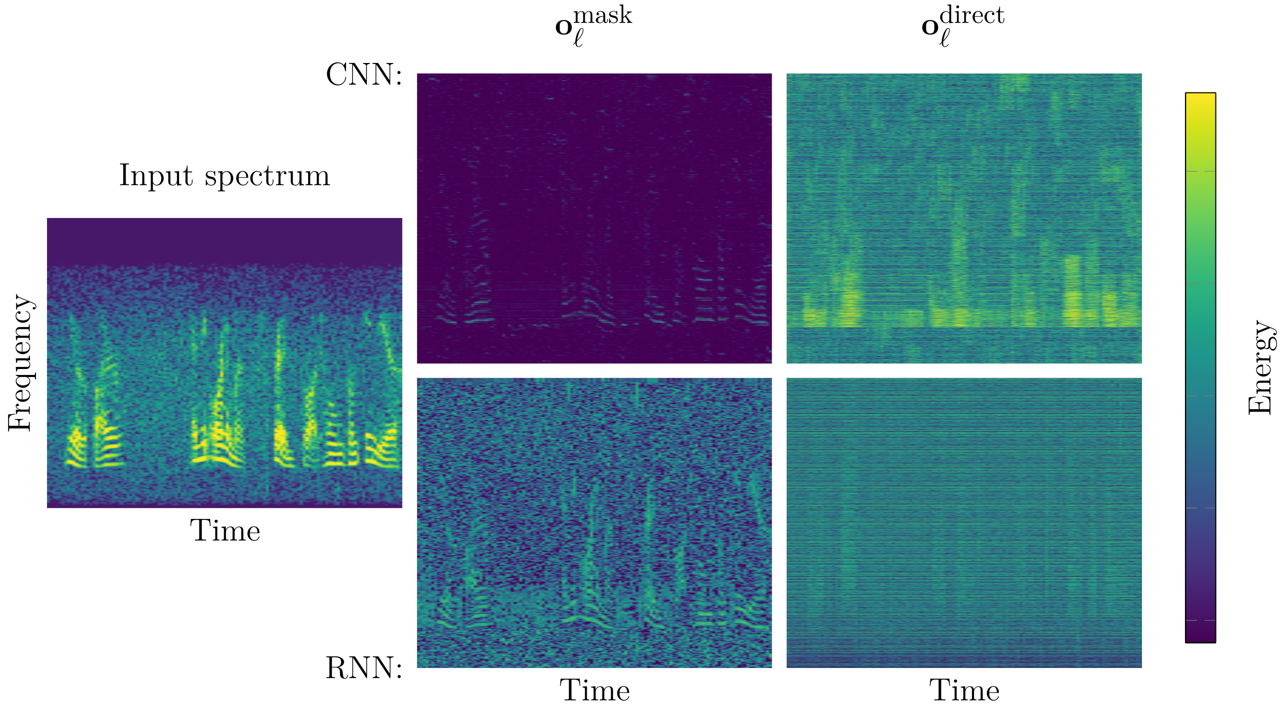


Figure 6.7: Depiction of $\mathbf{o}_\ell^{\text{masked}}$ and $\mathbf{o}_\ell^{\text{direct}}$ for the CNN (first row) and RNN (second row) excerpt layers.

Sub-band classifier

For the sub-band classifier eight CNN layers with an increasing number of channels are chosen. The first layer uses 16 channels and for each second CNN the number of channels is doubled, which leads to a last layer with 128 channels. While the dilation over the frequency dimension is increasing exponentially, no dilation is used along the time dimension. The temporal context is increased by applying maximum pooling with a window size and stride of two after every second CNN. In combination with a kernel size of three, this leads to a temporal receptive field of 150 frames or 1.5 s. For a larger discussion of the receptive of a CNN refer to [38] or Section 5.1.

Loss weights

Both the sub-band and full-band classifier networks use the loss function described in Section 6.2.3 with an auxiliary loss weighted by α. The next experiment focuses on dependency of the networks on this loss weight. For this experiment the size of the STFT input vector is fixed to $F = 2049$ and for the full-band classifier with the CNN-SBL the SBL output is used as an attention mask and for the network with the RNN-SBL it is directly sent to the full-band classification layer as discussed above. To reduce the number of variables the length L_I of the error tolerance interval of the auxiliary loss is set to a fixed value of

Table 6.5: Comparison of NN-based CFO estimators for different loss weight α on the simulated HF evaluation data regarding their error class affiliation. All values are given in %. For all results F is set to 2049 and for the full-band classifier holds $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{masked}}$ for the CNN-SBL and $\mathbf{o}_\ell = \mathbf{o}_\ell^{\text{direct}}$ for the RNN-SBL. The gray rows mark the value of α with the best results for each network.

Classifier Type	Sub-band layer	α	<5 Hz	5-10 Hz	10-50 Hz	>50 Hz
Full-band	CNN	0.0	58.97	39.72	1.02	0.29
		0.2	57.92	40.71	1.37	0.00
		0.5	56.00	43.08	0.83	0.10
		1.0	55.96	41.89	2.05	0.10
	RNN	0.0	55.68	40.10	4.22	0.00
		0.2	54.46	42.60	2.85	0.10
		0.5	57.18	39.85	2.97	0.00
		1.0	55.68	39.14	5.08	0.10
Sub-band	n.a.	0.0	71.59	15.48	7.35	5.58
		0.2	73.30	14.43	4.21	8.06
		0.5	81.24	9.09	4.21	5.45
		1.0	76.26	6.41	7.14	10.18

3 frequency bins, which corresponds to approximately 5 Hz. The results are displayed in Table 6.5.

For all networks an influence of the loss weight α can be observed. While both considered full-band classifier networks are only slightly impacted by a change in α, the results for the sub-band classifier are more heavily affected by a change of α. Overall, the estimation with the full-band classifier results in less errors greater than 50 Hz compared to the sub-band classifier. This may be partly due to concurrent speakers recorded with the additive noise (Section 4.9), which can interfere with the sub-band classifier but not with the full-band one as discussed above.

For the full-band classifier with a CNN-SBL $\alpha = 0.2$ and for the RNN-SBL $\alpha = 0.5$ are chosen, because these configuration do not lead to errors greater than 50 Hz and for the RNN $\alpha = 0.5$ also leads to the smallest number of errors greater than 10 Hz. One could argue, that for the network with a CNN-SBL $\alpha = 0.5$ or $\alpha = 0.0$ should be chosen because they lead to an improvement of 0.3-0.5 % in the error range between 10 and 50 Hz. However, they also slightly increases the number of errors greater 50 Hz, which is a much more severe error. For the sub-band classifier the best result is achieved with $\alpha = 0.5$. After choosing the hyperparameters for the neural networks, the best models can be compared to the RAKE algorithm.

6.4.4 CFO estimation

The previous experiments are designed to tune the NN-based estimators on the simulated data. In this experiment, the tuned estimators are compared to the statistical RAKE system described in Section 6.1. In contrast to [OC15] the set $\Omega^{\Delta^{\mathrm{f}}}$ of possible CFOs for the RAKE algorithm is upper bounded with $\Delta^{\mathrm{f}}_{\mathrm{max}} = 1500\,\mathrm{Hz}$ since the full-band classifiers are only trained on signals with a CFO below 1500 Hz.

The systems are compared for different STFT features sizes F since it affects both the speed as well as the performance of the CFO estimation. One reason for the impact on the performance is the frequency resolution. A smaller feature size F results in a lower resolution so that the algorithm has to interpolate between frequency bins to get an accurate estimate. For the 8 kHz audio signals in the evaluation set with the Nyquist frequency $F_{\mathrm{max}} = 4\,\mathrm{kHz}$ the STFT with feature size 4097 is chosen which leads to each frequency bin representing roughly 1 Hz. All experiments with the NNs use a fixed window length of 25 ms and a shift of 10 ms for the STFT, while the RAKE algorithm operates with a window length of 64 ms and a shift of 20 ms to be comparable to the experiments in [OC15].

In all following experiments the CFO estimation is only performed on activity segments that are detected by an ideal prior SAD system. Therefore, the target activity is calculated from the alignment information as discussed in **??**. The results for the different estimators are shown in Table 6.6.

For all feature sizes F the NN-based systems outperform the statistical RAKE in the sense, that they produce less errors greater than 10 Hz. This improvement is especially large for $F = 1025$ which indicates that the neural network is able to deal with the low frequency

Table 6.6: Comparison of the NN-based and the statistical CFO estimation on the real HF English evaluation data for more than 10 s of speech activity regarding their error class affiliation. All values are given in %.

System	Feature size F	<5 Hz	5-10 Hz	10-50 Hz	>50 Hz	RTF
Statistical RAKE	1025	20.49	39.53	37.81	2.16	–
Full-band (CNN)		95.13	2.66	1.08	1.13	0.058
Full-band (RNN)		94.18	3.02	1.98	0.81	0.058
Sub-band		47.87	45.68	1.08	5.37	0.121
Statistical RAKE	2049	96.44	1.40	0.00	2.16	–
Full-band (CNN)		95.67	4.06	0.00	0.27	0.031
Full-band (RNN)		99.46	0.18	0.09	0.27	0.039
Sub-band		85.52	13.31	0.00	1.17	0.364
Statistical RAKE	4097	96.66	1.17	0.00	2.16	–
Full-band (CNN)		96.89	1.35	0.36	1.40	0.138
Full-band (RNN)		95.26	2.66	0.63	1.44	0.150
Sub-band		81.32	15.65	0.36	2.66	0.762

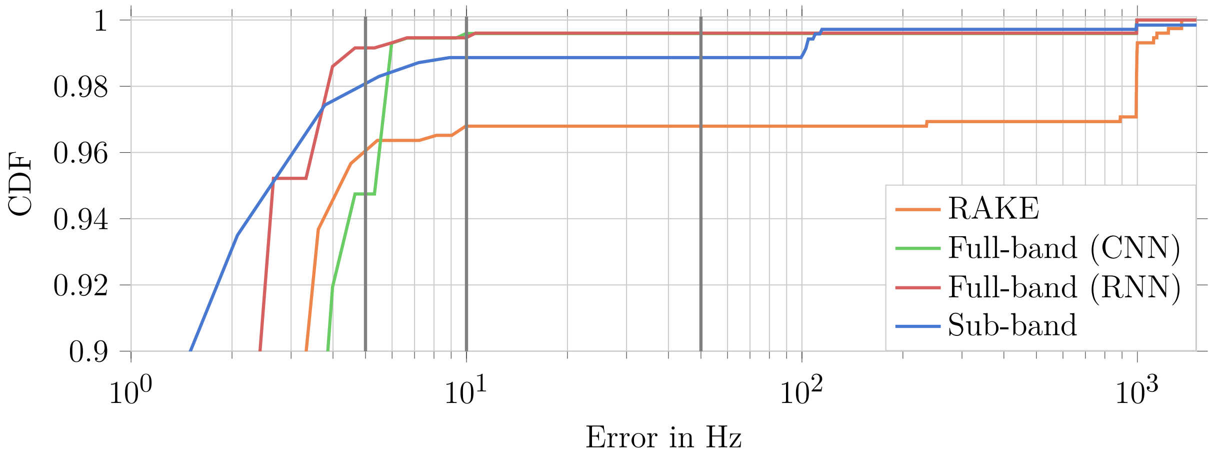


Figure 6.8: CDF of the CFO estimation error for the NN-based systems compared with the RAKE algorithm for segments with more than 10 s of speech activity. The number of errors used as boundaries for the ECA are marked with gray lines.

resolution and still output a highly accurate estimation. Additionally, the real time factor (RTF) of the models is significantly smaller than 1 for all feature sizes F. Therefore, both are able to process signals in real time, although the sub-band classifier is slower than both the full-band classifier. The RTF numbers for the RAKE algorithm are not presented, since it is implemented in standard Python without an optimization of the runtime. The networks on the other hand are implemented in Pytorch [176], which internally optimizes the runtime of the networks, so that a RTF comparison would favor the NNs. All future experiments are performed using $F = 2049$ since a lower feature size increases the error rate especially for the statistical RAKE algorithm and a higher feature size increases the RTF without any significant gain. To offer a visualization for the performance of the four estimators with the feature size $F = 2049$, the CDF of the error is depicted in Figure 6.8.

A detailed analysis shows that the errors above 10 Hz are mostly due to concurrent speakers. While a manual review found only one example with an error above 10 Hz for the full-band classifiers that cannot be explained by a concurrent speaker (refer to Figure 6.9b for the signal spectrum) in the HF transmission channel, the RAKE algorithm generates 28.6 % of the errors above 10 Hz in recordings without a concurrent speaker (refer to Figure 6.9a for an example spectrum). Another 17 % of the estimation errors generated by the RAKE algorithm are originating from an interfering narrow-band signal that is active on periodically repeating frequencies, which is misinterpreted as speech harmonics. Such an error type is also seen in the additive noise during network training as discussed in Section 4.9. An example is displayed in Figure 4.11 in **??**. Therefore, the NN-based systems are not distracted by these interfering narrow-band signals since similar distortions occur during training.

Astonishingly, for the sub-band classifier only one of the errors above 10 Hz is caused by concurrent speakers. All other errors are caused by the remainder of the suppressed sideband signal. Therefore, these errors only occur in signals with a very high SNR (refer to Figure 6.9c for an example spectrum). An example for a recording with a remainder of the suppressed sideband signal is displayed in Figure 6.9.

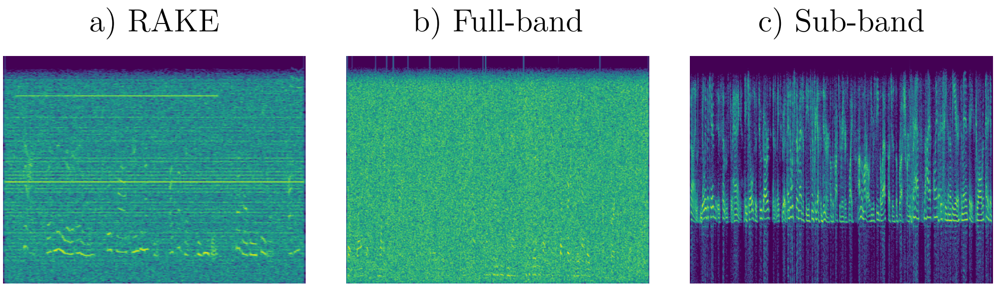


Figure 6.9: Spectrum of a signal with an estimation error above 10 Hz for each system.

One explanation for the strong results on signals with concurrent speakers is the large activity duration of more than 10 s distributed over up to 2 minutes. During this time the target speaker is always active, while a concurrent speaker has only short activity intervals. Therefore, the sum over the time dimension during evaluation leads to a larger accumulation of activity bins for the transmitted signal than for the concurrent speaker.

To further compare the four estimators the ECA are displayed in Figure 6.10 for different speech activity durations similar to our experiments published in [OC15]. The results for examples with more than 10 s of speech activity are calculated on the concatenated activity of entire recordings and the results for shorter activity segments are calculated individually on each of the LibriSpeech utterance in a recording. In contrast to [OC15] only segments with activity longer than one second are considered because all systems do not perform well for shorter segments. In particular, the sub-band classifier suffers for shorter segments since it is designed with a temporal context of 1.5 s.

Considering errors below 10 Hz as acceptable, all estimators perform a nearly flawless CFO estimation for the 158 recordings with more than 20 s of activity. For an activity duration greater than 2 s but smaller than 20 s the full-band classifiers outperform both the sub-band

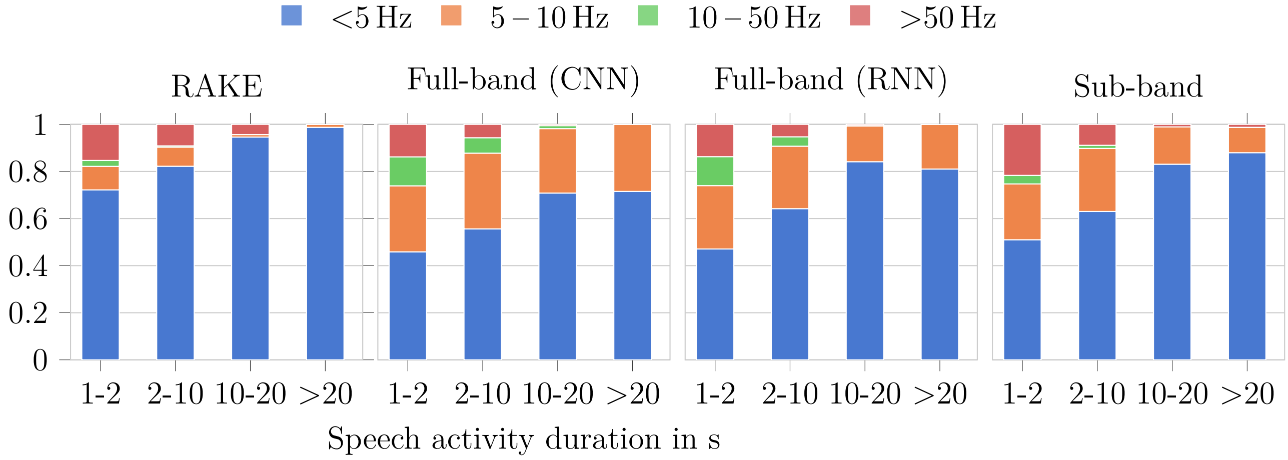


Figure 6.10: Depiction of error class affiliation for the CNN, RNN and RAKE CFO estimators for different speech activity duration.

classifier and the RAKE algorithm. However, for shorter activity segments between 1 and 2 seconds all three NN-based systems perform worse than the statistical RAKE approach. In particular, the sub-band classifier experiences a drop in performance for these shorter segments, which makes sense given its temporal context of around 1.5 s.

For all prior experiments, the full-band classifier outperforms the sub-band classifier. To assess the value of a sub-band estimation the models are applied to the IQ recordings with positive and negative CFOs.

6.4.5 CFO estimation on IQ signals

In this section, it is shown that there are benefits of the shift-invariant sub-band classifier over the full-band classifier. All estimators are designed for positive CFOs since a reconstruction of negative CFOs from the real-valued recording is not possible as discussed in Section 6.3. However, using the complex-valued baseband signal allows a reconstruction of the transmitted signal even for negative CFOs.

As discussed in Section 4.7 the frequency range of the complex-valued baseband signal is twice as large as the spectrum of the real-valued recording seen during training. This difference is due to the symmetry of the frequency spectrum of a real-valued signal. Therefore, an estimation of the CFO on the complex-valued baseband signal requires the estimator to be independent of the size of the input spectrum. While the full-band classifier cannot be applied to a larger frequency range than the one for which it was trained, the sub-band classifier has no such restriction.

This difference is due to the full-band classification layer in the full-band classifier, which uses the entire input spectrum for the CFO estimation. For the sub-band classifier on the other hand an estimation is performed for each frequency bin individually and the estimate is only informed by a limited context, so that this network can be applied to spectra of varying size. To adapt the full-band classifier to this larger input spectrum the network would have to be retrained on spectra of complex-valued baseband signals.

To emphasize the advantage of working on the IQ components, the CDFs for all estimators are displayed in Figure 6.11 for the IQ recordings described in Section 4.7 with simulated CFOs ($\mathit{\Delta}^{\mathrm{f}} \in [0, 100, 300, 500, 1000]$). For the evaluation of the influence of negative CFOs on the error rate, the set of possible CFOs is extended by $[-100, -300, -500, -1000]$. Note, that only the sub-band classifier is directly applied to the complex-valued baseband signal calculated from the IQ components. The full-band classifiers and the RAKE algorithm still process the real-valued recordings.

As expected, the sub-band classifier significantly outperforms the full-band classifiers if both positive and negative CFOs are considered. The full-band classifier could be applied to the larger spectrum of the complex-valued baseband signal by applying it twice. Once for the negative and once for the positive frequencies. However, this still does not solve the problem for small negative CFOs, which would be detected both in the positive and negative half of the IQ spectrum. Although the RAKE algorithm could also be applied directly to the IQ data, this is not pursued further in this work due to the significantly worse

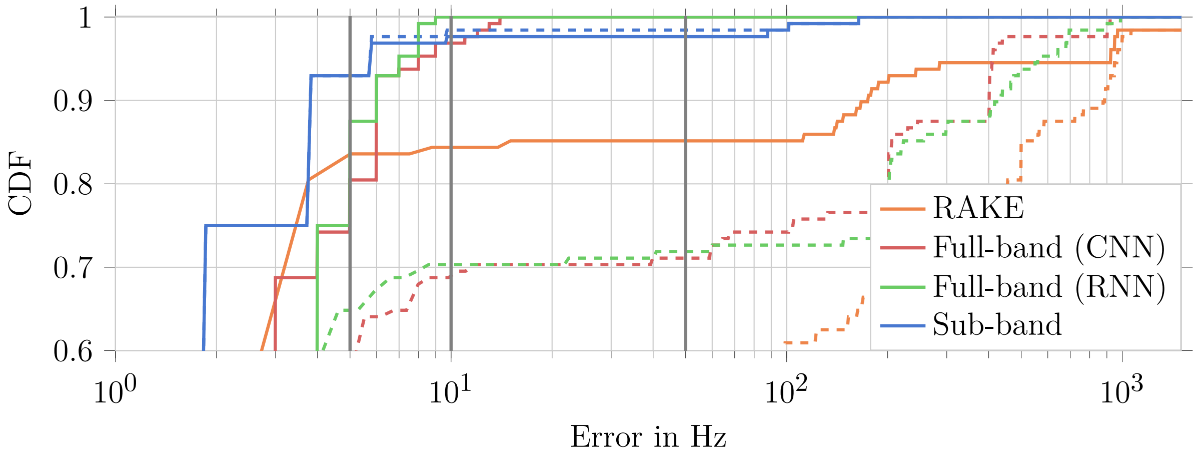


Figure 6.11: CDF of the CFO estimation error for the NN-based systems compared with the RAKE algorithm on the IQ evaluation set with simulated positive CFOs (solid) or simulated positive and negative CFOs (dashed). The errors used as boundaries for the ECA are marked with gray lines.

performance compared to the NN-based estimators for positive CFOs on the "evaluation IQ" data set.

After establishing, the strong performance of the CFO estimators, the influence of the CFO correction on the speech signal quality and intelligibility is evaluated in the next section.

6.4.6 CFO correction

As discussed in Section 3.2 a high CFO significantly reduces the speech quality and intelligibility. For the English HF evaluation set without a CFO the average PESQ value is 1.55 and the STOI value is 57.1 %. The low values indicate the distortion of the recorded signal due to the transmission independent of a CFO. However, these values are still high compared to the PESQ and STOI values for the evaluation set with CFOs greater than zero, which are 1.20 and 42.0 %. In the next experiments, the estimated CFOs are used to correct the frequency shift and thereby enhance the signals. The absolute improvement compared to the recorded signal are displayed in Table 6.7.

The CFO correction leads to similar results independent of the CFO estimator. For all estimators the performance measures are close to the ones for the correction with the true CFO (Oracle), which is not surprising considering the accuracy of the CFO estimation. For the recordings without a CFO, limitation of the signal frequencies to the transmission bandwidth using a low-pass filter leads to a small gain in the PESQ and STOI score.

One can conclude that the few errors in the CFO estimation do not have a high impact on the speech signal quality after enhancement. This is surprising for the larger estimation errors, which lead to a highly distorted signal. One explanation for the small difference between the oracle results and the performance of the estimation systems is that signal segments with

Table 6.7: Evaluation of the signal enhancement due to CFO correction for different CFO estimators with $F = 2049$ on the real HF English evaluation data with ($\mathit{\Delta}^{\mathrm{f}} \geq 0$) and without ($\mathit{\Delta}^{\mathrm{f}} = 0$) a CFO.

CFO estimator	$\mathit{\Delta}^{\mathrm{f}} = 0$		$\mathit{\Delta}^{\mathrm{f}} > 0$	
	PESQ	STOI %	PESQ	STOI / %
None	1.55	57.1	1.20	42.0
Oracle	1.55	58.8	1.71	62.9
Statistical RAKE	1.56	58.8	1.71	61.9
Full-band (CNN)	1.57	59.3	1.70	62.6
Full-band (RNN)	1.57	59.3	1.71	62.7
Sub-band	1.56	58.9	1.69	62.7

high CFO prediction error rate are already highly distorted, so the additional distortion due to the erroneous CFO correction is concealed by the existing distortion. To emphasize this conjecture, Figure 6.12 displays the distribution of PESQ values for the recorded signal and the signal after CFO correction with oracle information. While the CFO correction leads to a gain for some examples, most occurrence are still centered around a low PESQ value of 1.5. For these examples, the distortion introduced during transmission mask the gain from the CFO correction.

All previous experiments are performed on English audio. To evaluate whether the systems are robust to a language change, the CFO correction is performed on the Russian HF evaluation data. The results are presented in Table 6.8.

The results on the Russian data are very similar to the results on the English recordings indicating that the systems are robust to a different language. Surprisingly, all systems lead to slightly higher score than the oracle CFO correction. This indicates that some of the CFO labels for the Russian signals might be noisy.

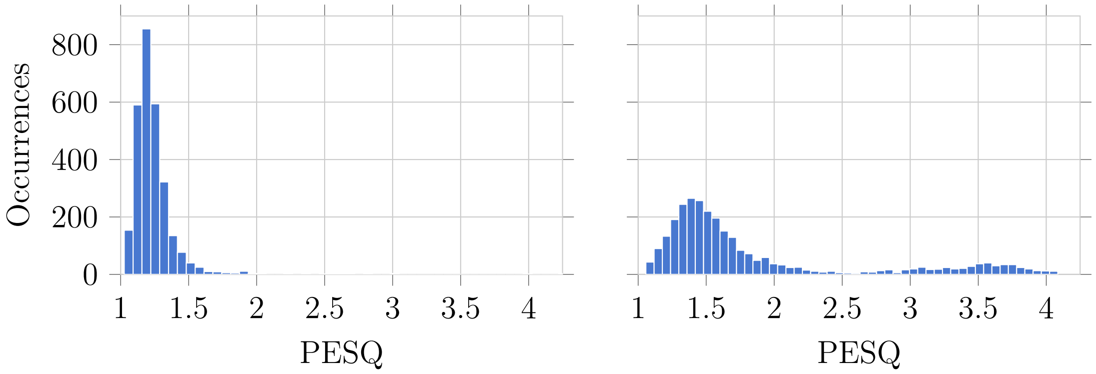


Figure 6.12: Histogram of the PESQ value of the recordings before (left) and after (right) CFO correction with oracle information for recordings with a CFO $\mathit{\Delta}^{\mathrm{f}} > 0$.

Table 6.8: Evaluation of the signal enhancement due to CFO correction for different CFO estimators with $F = 2049$ on the real HF Russian evaluation data with ($\mathit{\Delta}^{\mathrm{f}} \geq 0$) and without a CFO ($\mathit{\Delta}^{\mathrm{f}} = 0$).

CFO estimator	$\mathit{\Delta}^{\mathrm{f}} > 0$		$\mathit{\Delta}^{\mathrm{f}} = 0$	
	PESQ	STOI / %	PESQ	STOI / %
None	1.42	59.3	1.20	40.2
Oracle	1.53	59.3	1.60	54.4
Statistical RAKE	1.56	60.4	1.63	55.1
Full-band (CNN)	1.56	60.5	1.62	55.1
Full-band (RNN)	1.56	60.5	1.63	55.2
Sub-band	1.55	59.3	1.63	55.3

6.5 Summary

In this chapter a statistical and two NN-based approaches to estimate the CFO of HF recordings are introduced and compared. The two NNs architectures achieve a performance similar to the statistical RAKE approach and even outperform it on the the “evaluation IQ” data set.

Furthermore, the benefits of the CFO estimation on IQ signals with the sub-band classifier for negative CFOs are shown. Finally, the large impact of a CFO correction on the speech quality and intelligibility is shown, where the CFO estimates for all three systems lead to results close to those obtained if the true CFO value is known. Since the results of all estimators are similar for the positive CFOs in the “evaluation shift” data set both NN-based systems and the statistical approach are considered in future experiments.

After evaluating both the CFO estimation and correction, the following experiments deal with blind source separation (BSS) and noise reduction (NR) assuming absence (or perfect correction) of a CFO.

7 Speech Enhancement

The last chapters introduced systems for speech activity detection (SAD) and carrier frequency offset (CFO) correction, which are assumed as a preprocessing step, so that all following systems can be designed for segments with speech activity and without a CFO. However, the speech signal is still distorted by the high frequency (HF) transmission channel and additive noise.

Therefore, the observation can be modeled as

$$y_n = b_n * \tilde{x}_n + d_n = x_n + d_n, \tag{7.1}$$

with y_n denoting the observation, b_n as the transmission channel, d_n as the noise and $\tilde{x}_n$ as the clean signal prior to the distortion due to the transmission. The final definition of the observed signal is a simplification of the original signal model introduced in **??** because the degradation caused by the HF channel is not considered.

Since many of the current neural network (NN)-based noise reduction (NR) approaches have their origin in blind source separation (BSS) [108], [150], this chapter starts by introducing a modular NN-based BSS system. This system is a combination of the estimator suggested in [115] and the time-domain audio separation network (TasNet) introduced in [116]. Note, that the described system can also be applied to the NR task.

For the following description noise is treated as a source so that the training scheme for NR and BSS with two sources are indistinguishable and the signal model is further simplified to

$$y_n = x_{0,n} + x_{1,n} = \sum_{k=0}^{K-1} x_{k,n}, \tag{7.2}$$

with K as the number of sources ($K = 2$). For noise reduction $x_{1,n}$ represents the noise ($x_{1,n} = d_n$) and $x_{0,n}$ the clean signal. Note that the noise reduction network is trained to reconstruct the noise signals as well as the source signal, as suggested in [108]. During evaluation only the reconstructed speech signal is considered. To make the equations valid for more scenarios, K is kept undefined although this work only considers scenarios with $K = 2$.

This section starts with an overview of an universal BSS system and then gives insights into each of its components. Finally, the presented system is evaluated on a speaker separation and the NR task on HF signals in Section 7.5.

7.1 System overview

In this work we consider the received signal y_n to be a sum of two sources $x_{0,n}$ and $x_{1,n}$. For the NR task one of these sources is the channel noise. A common BSS system is depicted in Figure 7.1, it consists of three blocks. The first block is the encoder, which transforms the time domain signal into a feature domain suitable for enhancement. This block is followed by a mask estimator, which predicts an enhancement mask $\hat{M}_{k,\ell,f}$ for each source k, with ℓ as the frame index and f as the feature index. The estimated masks are multiplied with the transformed signal $Y_{\ell,f}$ to separate the sources, leading to the estimate $\hat{X}_{k,\ell,f}$. The mask multiplication is followed by a decoder, which transforms the enhanced signal back to the time domain, thereby calculating the final estimate $\hat{x}_{k,n}$.

Both the encoder and decoder can either be learned transformations as in [116] or a transformation with a fixed parameterization. There are different possibilities for a fixed transformation [OC3], [177], [178]. However, in this work only the short time Fourier transform (STFT) and inverse STFT are considered as suggested in [137] and [OC3]. A loss function can either be calculated on the enhanced signal in the latent or time domain, using $X_{k,\ell,f}$ or $x_{k,n}$ as targets.

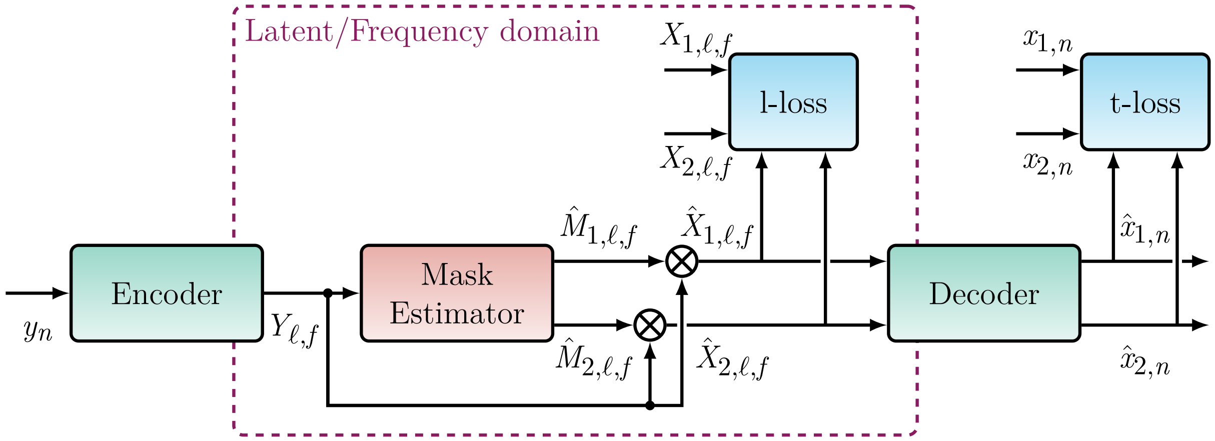


Figure 7.1: Block diagram of a NN-based BSS and NR system.

7.2 Encoder and decoder

As a first step of the transformation, the input signal y_n is segmented into L frames. Each frame has the length L_W with $L_W - L_S$ samples overlapping with the previous frame, where L_S is the advance between frames. After segmentation, the signal is written as $y_{i,\ell} := y(i + \ell \cdot L_S)$ with i as the sample index within the frame ℓ. Then, the transformation is applied to the segmented signal.

For source separation the STFT is a common transformation [179] with

$$Y_{\ell,f} = \sum_{i=0}^{L_W-1} y_{i,\ell} \cdot w_i \cdot \exp\left(-\mathrm{j}2\pi \cdot i \cdot f / F_{\mathrm{DFT}}\right), \tag{7.3}$$

where w_i represents a window function and F_{DFT} the size of the discrete Fourier transformation (DFT). In this equation, it is assumed that L_W is identical to F_{DFT}. If $L_W < F_{\mathrm{DFT}}$ the segments are zero-padded.

Since the input signal is real-valued, the STFT leads to a symmetric output if F_{DFT} is an even number. To take advantage of the redundancy due to the symmetry only frequency bins f in the range from 0 to $F-1$ are considered during separation with $F = F_{\mathrm{DFT}}/2+1$. The dropped frequency bins are restored during decoding with the inverse STFT.

For the learned transformation a one-dimensional convolutional neural network (CNN) layer is used for both the encoder

$$Y_{\ell,f} = g\left(\sum_{i=0}^{L_W-1} y_{i,\ell} \cdot u_{i,f}\right), \tag{7.4}$$

and the decoder

$$\hat{x}_{k,n} = \sum_{\ell=0}^{L-1} \ell\mathrm{rect}\left(\frac{n-\ell L_S - \frac{L_W}{2}}{L_W}\right) \sum_{f=0}^{F-1} \tilde{u}_{i=n-\ell L_S,f} \cdot \hat{X}_{k,\ell,f} \tag{7.5}$$

where $u_{i,f}$ and $\tilde{u}_{i,f}$ are the learnable, one-dimensional CNN kernels and $\hat{x}_{k,n}$ is the time domain estimate of the k-th source signal. $g(\cdot)$ is the activation function of the encoder, which commonly is either the identity function or a rectified linear unit (ReLU) [135]. The time domain estimate is calculated from the feature domain estimate by computing the weighted sum over the feature dimension for each frame and deriving the per-sample result using the overlap-add method.

Both the STFT and inverse STFT can be written as a one-dimensional convolution with a fixed instead of a learned kernel. In this case the complex STFT operation is split into a real and imaginary part leading to two separate kernels

$$u_{i,f}^{\mathrm{real}} = w_i \cdot \cos\left(2\pi \cdot i \cdot \frac{f}{F_{\mathrm{DFT}}}\right), \tag{7.6}$$

$$u_{i,f}^{\mathrm{imag}} = -w_i \cdot \sin\left(2\pi \cdot i \cdot \frac{f}{F_{\mathrm{DFT}}}\right), \tag{7.7}$$

The complex signal resulting from the two convolutions $Y_{\ell,f} = \sum_i y_{i,\ell} \cdot u_{i,f}^{\mathrm{real}} + \mathrm{j} \sum_i y_{i,\ell} \cdot u_{i,f}^{\mathrm{imag}}$ is identical to the STFT result in Equation (7.3).

7.3 Mask estimator

The mask estimator operates on the transformed signal to calculate activity information for both sources per time frequency bin. For the STFT as encoder, it is common practice to present the magnitude of the complex spectrum to the estimator [115]. However, due to this data representation the phase information of the complex signal is unknown to the mask estimator. To maintain this information the real and imaginary part of the STFT transformed signal can be concatenated so that the estimator input has $F_{\mathrm{DFT}} + 2$ real valued features and can calculate an activity mask for both real and imaginary part of the signal [180].[1]

For many publications the mask estimation output activation either is a Softmax or a Sigmoid function to limit the mask to the interval $[0, 1]$ [101]. In [132] it is suggested that a "mask" with values greater than one can be used to amplify the target signal components and thereby achieve an improved separation performance. Therefore, in the following discussion the mask is not confined to the interval $[0, 1]$ but can take on any real, positive value ($\hat{M}_{k,\ell,f} \in \mathcal{R}^+$).

There are many different architectures for the mask estimator ranging from recurrent layers [116] to a transformer structure [134]. In this work, only the convolutional architecture described in [135] and the recurrent one published in [44] are compared, since they are two of the most popular mask estimators for source separation.

7.3.1 Convolutional mask estimator network

The convolutional mask estimator introduced in [135] is a temporal convolutional network [181]. It consists of an input normalization with subsequent projection of the input signal with size F to the network feature size B_C with a one-dimensional CNN layer with a 1x1 kernel. The projected signal is further processed by R_C CNN blocks which consist of D_C one-dimensional convolution blocks (Conv1D-blocks) with increasing dilation. In Figure 7.2 an example for a Conv1D-block is depicted. It consists of an input normalization, a one-dimensional CNN with 1x1 kernel as input projection, a parametric rectified linear unit (PReLU) activation and the central one-dimensional CNN with a 1x3 kernel and the block specific dilation. As the last two steps of the Conv1D-block the signal is processed by a PReLU activation and a one-dimensional CNN with 1x1 kernel as output projection to return the feature size to the network feature size B_C. After the R_C CNN blocks the signal is projected to the size $k \cdot F$ so that the network outputs an estimate for each source $k = 1, .., K$. Finally, an activation function is applied to calculate a time frequency mask. In [135] a Sigmoid or Softmax activation is suggested. However, as discussed in [132] a less restrictive activation may lead to improved results. Therefore, a ReLU activation function is used in this work.

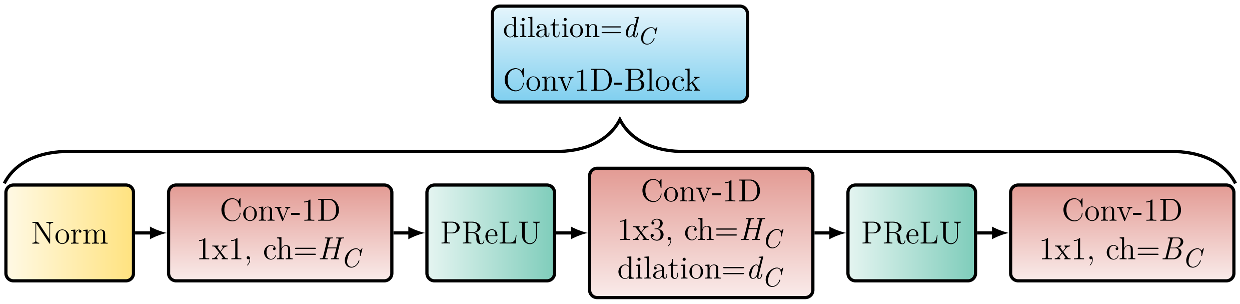


Figure 7.2: Block diagram of a Conv1d-block of the convolutional mask estimator

[1] Due to the symmetry of the DFT only F_{DFT} features of the transformed signal are unique. For simplicity $F_{\mathrm{DFT}} + 2$ values are calculated.

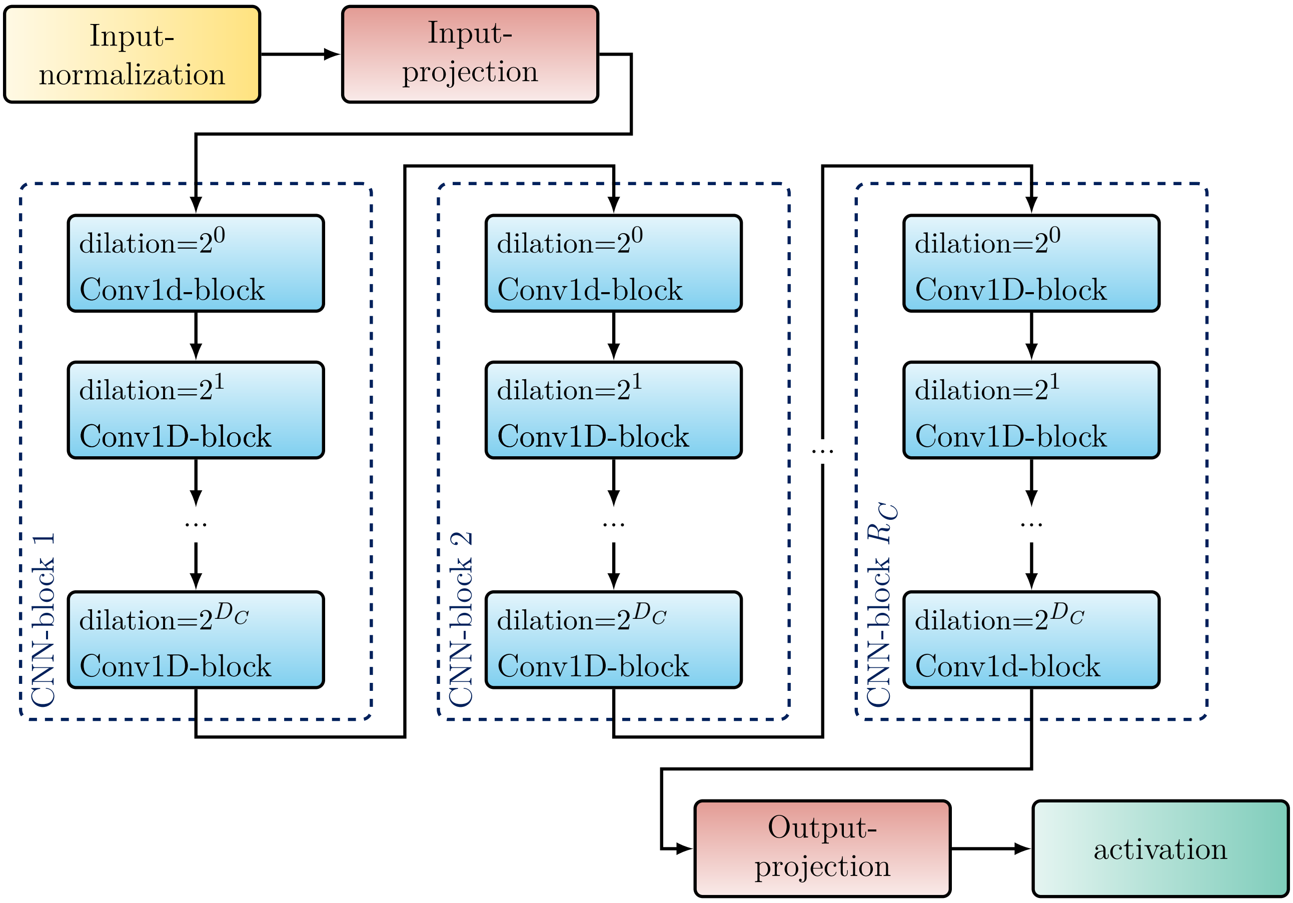


Figure 7.3: Block diagram of the convolutional mask estimator.

The combination of multiple CNN layers with different dilation factors leads to a large temporal context of roughly $\frac{2^{D_C} \cdot R_C \cdot L_S}{F_{\max}}$, which allows the network to aggregate source information for the mask estimation even in case of short silence segments. The whole convolutional estimator is depicted in Figure 7.3.

7.3.2 Recurrent mask estimator network

The second architecture considered here is a dual-path recurrent neural network (DPRNN) network as introduced in [44]. This architecture consists of R_R DPRNN-blocks, where each block includes two recurrent neural networks (RNNs) layers each with H_R nodes. As preprocessing for the DPRNN-blocks the L frames of the encoded input signal are split into $\tilde{L}$ segments of length $\tilde{L}_W$ with a shift $\tilde{L}_S$, which results in an overlap $\tilde{L}_W - \tilde{L}_S$ between the segments. This is similar to the segmentation in Section 5.2 and allows to define the temporal context presented to the first RNN layer. Next, the input signal is projected to the expected feature space with a feature size B_R. Before the signal is processed by the second RNN layer it is again projected to the network feature size B_R and normalized using layer normalization [182]. Additionally, the segmented input signal and the normalization output are connected with a residual connection.

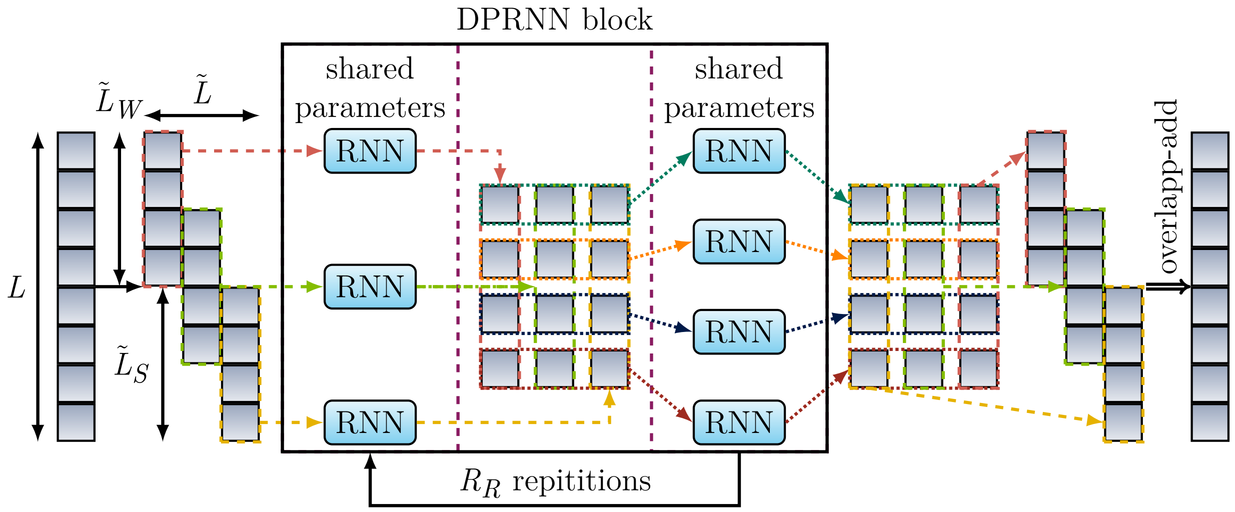


Figure 7.4: Block diagram of the DPRNN layer, where each blue block represents a feature vector for one frame and the colors indicate which frames are processed together in the following step.

The second RNN processes inter segment information by stacking the $\tilde{\ell}$-th value of all segments to get an input vector for a recurrent layer with $\tilde{\ell} \in [0, \tilde{L}-1]$. Thereby, allowing the network to process local information in the first RNN and using this information in a global context in the second RNN. The second RNN layer includes an output projection, layer normalization and a recurrent connection exactly as the first RNN. The $\tilde{L}$ frames of the segmented output of the RNN-blocks are added together using the overlapp-add method to calculate the activity mask for each of the L frames. Figure 7.4 depicts a block diagram of the described recurrent mask estimator, where the different colors and line types show that the first RNNs are applied to the each segment individually, while the second RNNs process values from each segment to gather inter-segment information. Note, that the output of a DPRNN blocks is sent to the next block, and only the output of the last layer is further processed to get a frame-wise representation again.

In this work, the focus lies on the original version of DPRNN despite the recently published improvements to the DPRNN, which were shown to lead to stronger separation results [183] or to require less computational power [184]. This decision can be justified by the assumption that the insides gained during this work can be transferred to all versions of DPRNN.

7.4 Cost function

The loss for both the BSS and the NR network can be either calculated in the feature or in the time domain. In Figure 7.1 this is represented by the l- and t-loss blocks. For the BSS task the permutation problem is solved using the utterance-level permutation invariant training (u-PIT) method [115]. As discussed previously, the permutation problem does not occur for the NR task because the noise and speaker statistics are distinct, allowing the network to learn an unambiguous assignment between estimated signals and targets.

One common loss function for noise reduction calculates the deviation of the estimated mask $\hat{M}_{k,\ell,f}$ from an oracle mask $M_{k,\ell,f}$ using the binary cross entropy (BCE)

$$\mathcal{L}^{\text{M–BSE}} = \frac{1}{L \cdot F \cdot K} \sum_{k=0}^{K-1} \sum_{\ell=0}^{L-1} \sum_{f=0}^{F-1} \Big[-M_{k,\ell,f} \log\left(\hat{M}_{k,\ell,f}\right)$$

$$- \left(1 - M_{k,\ell,f} \log\left(1 - \hat{M}_{k,\ell,f}\right)\right)\Big] \,. \tag{7.8}$$

An oracle mask is calculated from the observation $Y_{\ell,f}$ and the paired clean speech signal $X_{k,\ell,f}$. There are various possible oracle masks as discussed in [101]. In this work the ideal ratio mask (IRM) and the ideal complex mask (ICM) are considered as target masks. Both are discussed in more details in Section 7.5.3.

An alternative loss function is the mean squared error (MSE) between the enhanced signal $\hat{X}$ and the source signal X. For a frequency domain system with the magnitude spectrum as input there are two common MSE-based loss functions. The first ist a simple MSE loss

$$\mathcal{L}^{\text{MSE}} = \frac{1}{L \cdot F \cdot K} \sum_{k=0}^{K-1} \sum_{\ell=0}^{L-1} \sum_{f=0}^{F-1} \left| \left|\hat{X}_{k,\ell,f}\right| - \left|X_{k,\ell,f}\right| \right|^2 , \tag{7.9}$$

which takes the phase information into account if the input is a concatenation of the real and imaginary part of the observation instead of the magnitude. Otherwise, no phase information is considered in the loss calculation.

Therefore, in [115] a phase sensitive loss $\mathcal{L}^{\text{P–MSE}}$ is introduced, which adds phase information to the simple MSE loss

$$\mathcal{L}^{\text{P–MSE}} = \frac{1}{L \cdot F \cdot K} \sum_{k=0}^{K-1} \sum_{\ell=0}^{L-1} \sum_{f=0}^{F-1} \Big| \left|\hat{X}_{k,\ell,f}\right|$$

$$- \left|X_{k,\ell,f}\right| \cdot \cos\left(\varDelta\theta_{k,\ell,f}\right) \Big|^2 , \tag{7.10}$$

with $\varDelta\theta_{k,\ell,f} = \theta_{\ell,f} - \theta_{k,\ell,f}$ as the difference between the observed phase $\theta_{\ell,f}$ and the phase information for each speaker $\theta_{k,\ell,f}$. Using the cosine to convey the phase information can be shown to be the best real-valued approximation of the true complex-valued signal [185]. For a more in depth discussion of frequency domain loss functions refer to [186].

For both BSS and NR the scale invariant signal to distortion ratio (SI-SDR) is a common time domain loss function [116], [150] with

$$\mathcal{L}^{\text{SI–SDR}} = -10 \frac{1}{K} \sum_{k=0}^{K-1} \log_{10} \frac{\sum_{n=0}^{N-1} \left|\alpha \cdot x_{k,n}^2\right|}{\sum_{n=0}^{N-1} \left|\alpha \cdot x_{k,n} - \hat{x}_{k,n}\right|^2} \quad \text{and} \quad \alpha = \frac{\sum_{n=0}^{N-1} \hat{x}_{k,n} \cdot x_{k,n}}{\sum_{n=0}^{N-1} \left|x_{k,n}\right|^2} , \tag{7.11}$$

where α is used as a scaling factor. In [OC3] we showed that with only a few reformulations a logarithmic MSE in the time domain can be derived from the SI-SDR. Therefore, a new scaling factor $\beta = \frac{1}{\alpha}$ is introduced and all terms without a dependency on learnable parameters are removed.

$$\begin{aligned}\mathcal{L}^{\text{SI–SDR}} &= -10\frac{1}{K}\sum_{k=0}^{K-1}\log_{10}\frac{\sum\limits_{n=0}^{N-1}\left|x_{k,n}\right|^2}{\sum\limits_{n=0}^{N-1}\left|x_{k,n}-\beta\cdot\hat{x}_{k,n}\right|^2} \\ &\propto 10\frac{1}{K}\sum_{k=0}^{K-1}\log_{10}\sum_{n=0}^{N-1}\left|x_{k,n}-\beta\cdot\hat{x}_{k,n}\right|^2. \end{aligned} \tag{7.12}$$

Setting β to 1 leads to the logarithmic MSE in the time domain

$$\mathcal{L}^{\text{T–LMSE}} = 10\frac{1}{K}\sum_{k=0}^{K-1}\log_{10}\sum_{n=0}^{N-1}\left|x_{k,n}-\hat{x}_{k,n}\right|^2. \tag{7.13}$$

When assuming that one of the sources is inactive during a whole utterance, then α approaches infinity following Equation (7.11) which leads to an information devoid gradient for $\mathcal{L}^{\text{SI–SDR}}$ as a loss function, while the gradient for $\mathcal{L}^{\text{T–LMSE}}$ is not affected as long as some of the predicted values are greater than zero. Therefore, $\mathcal{L}^{\text{T–LMSE}}$ is both simpler than $\mathcal{L}^{\text{SI–SDR}}$ and provides a more informative gradient in case of inactivity for one of the sources. However, the scale invariance is lost for $\mathcal{L}^{\text{T–LMSE}}$ as a loss function compared to $\mathcal{L}^{\text{SI–SDR}}$. For both $\mathcal{L}^{\text{T–LMSE}}$ and $\mathcal{L}^{\text{SI–SDR}}$ the gradients are strongly influenced by examples from the training set with already good predictions ($\hat{x}_{k,n}$ close to $x_{k,n}$) because the loss approaches minus infinity. To mitigate the influence of these stronger predictions a clipping or thresholding is suggested in [4] and [187]. These upper bounds to the time domain loss functions are only considered for noise reduction to keep the comparability to [OC3] for the BSS experiments.

The system for BSS and NR introduced in this chapter is evaluated in the next section.

7.5 Evaluation

In this section the presented system is evaluated for both the BSS and NR task. Here, the experiments are conducted to evaluate whether frequency domain separation and noise reduction networks can still achieve competitive results compared to time domain systems. Therefore, first the evaluation data sets and measures are discussed.

Afterwards, a top and a baseline system are presented for the different data sets. Then, a step-by-step analysis of the system components is performed for both the BSS and NR data. From this evaluation a design recommendation for both tasks specific to the considered data is derived.

7.5.1 Database

Two distinct databases are used in the evaluation. The first is the WSJ0-2mix database [114], which is used to evaluate the described system on the BSS task. It consists of mixtures of two utterances from the Wall Street Journal (WSJ) database [188] spoken by different speakers. The signals are mixed at a random signal to distortion ratio (SDR) between –2.5 dB and 2.5 dB. In contrast to many publications the length of the utterance is chosen to be the length of the longest utterance as suggested in [OC3] to ensure that for both speakers the whole utterance is included in the observed signal. This configuration leads to about 0.5 dB lower SDR results compared to the common approach to choose the length of the shortest utterance [OC3]. The database consists of three data sets with 30 h of training, 10 h of validation and 5 h of testing data.

Similar to the CFO systems, all NR models are only trained on HF signals of English speakers simulated as described in Section 4.9. For the evaluation on the NR task a subset of simulated data (HF Simu) and the real recordings discussed in Section 4.6 are considered. The evaluation sets without a CFO described in Section 4.6 are used to evaluate the NR performance of the discussed system on real recordings. Most of the evaluations on real HF signals are performed on the recordings of English speakers (HF English) and only in Section 7.5.6 is the Russian evaluation set (HF Russian) used to examine the impact of an unknown language on the NR performance. In Table 7.1 an overview over the considered data sets is displayed. For all NR signals a perfect CFO correction is applied prior to the NR network.

The next section gives an overview of the evaluation measures used during the following experiments.

Table 7.1: Comparison of the speech enhancement evaluation sets.

Name	Simulated	Speech Overlap	Noisy	SDR	
				min	max
WSJ0-2mix	✓	✓	✗	–2.5	2.5
HF Simu	✓	✗	✓	–5.0	5.0
Evaluation	✗	✗	✓	–21.4	5.8
Russian	✗	✗	✓	–21.3	10.5

7.5.2 Evaluation Measure

Both the NR and BSS performance can be assessed using the short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) evaluation measures discussed in Section 6.4.2. Additionally, three common evaluation measures for enhancement that are influenced by the amplitude variation due to the automatic gain control (AGC) are used to evaluate the performance on simulated HF signals.

The first two additional measures are MIR-SDR [189], [190] and SI-SDR [191], which are common BSS evaluation measures that compare the power of the target x_n and estimated signal $\hat{x}_n$. Therefore, both measures cannot be used for the real HF data, which is processed by an AGC. Although the SI-SDR calculates a scaling factor to compensate for amplitude differences, this factor is only estimated globally for an utterance so that the changing factor of the AGC cannot be counteracted. Additionally, the SI-SDR is susceptible to small temporal delays between the estimated and target signal, which do not reduce the quality or intelligibility as we have shown in [118]. This leads to very low SI-SDR values for the real HF signals in case of small errors in the synchronization of the transmitted and recorded signal described in **??**. For brevity, MIR-SDR will be called SDR for the rest of this work.

As a third evaluation measure, the word error rate (WER) is calculated to show the impact of the enhancement systems on a downstream automatic speech recognition (ASR) system. Two acoustic models are trained, one to evaluate the BSS results and another for the NR output. Both models use a factorized time-delayed neural network (TDNN-F) [192] with a configuration based on the WSJ recipe from the Kaldi toolbox [156], [157]. The training procedure follows the one we describe in [118], where a GMM-HMM is trained on clean audio to extract alignments. These alignments are used to train the acoustic model with 8 TDNN-F layers. During evaluation the default tri-gram Kaldi language model for either the WSJ or LibriSpeech database is used. No language model rescoring is used for this work. The WER is calculated by accumulating the insertion, deletion and substitutions over all utterances and dividing the sum by the number of words in the data set. Therefore, the value of the WER is a positive number and lower values represent an improvement in the recognition.

Two ASR systems are trained, one for the WSJ0-2mix and one for the HF data. For the WSJ0-2mix database both the acoustic model and the HMM-GMM for alignment extraction are trained on the WSJ data. To train an acoustic model for the HF Simu data set, the alignments are extracted from the clean LibriSpeech audio and then used to train the acoustic model on the simulated training data.

For the evaluation on the real HF recordings the NR is performed on each of the five activity segments in a recording individually. The presented results are an average over the values calculated per utterance.

To offer a comparison for the enhancement performance of the presented NN for speech enhancement during evaluation both a baseline and a topline system are discussed in the next section.

7.5.3 Top- and baseline

In this section a baseline and a topline for BSS and NR are introduced and evaluated on the task specific, simulated evaluation sets. Additionally, two baselines for the noise reduction task on the real HF signals are discussed.

Simulated signals

For the simulated databases WSJ0-2mix and HF Simu a multiplication of the frequency spectrum of the input signal with an oracle mask is chosen as the topline, with $\hat{X}_{k,\ell,f} = M_{k,\ell,f} \cdot Y_{\ell,f}$. The following experiments compare the IRM, Wiener-like mask (WLM) and ICM oracle masks for both tasks.

The masks are calculated in the frequency domain with $L_W = 64\,\mathrm{ms}$ and $L_S = 16\,\mathrm{ms}$. For each considered mask the equation and average result on the simulated evaluation sets for both tasks are displayed in Table 7.2 for the PESQ and SDR evaluation measures. To compare additional evaluation measures please refer to Appendix A.5. Additionally, the results for the observation signal y_n without any enhancement are presented as a baseline.

For both BSS and NR the WLM shows stronger distortion reduction than the IRM which can be explained by the more sparse representation due to the quadratic terms.

For these simulated signals, the enhancement with the ICM leads to a nearly perfect reconstruction, which is expected as the observation is a linear combination of the sources. To emphasize the differences between the masks examples for the IRM and ICM are displayed in Figure 7.5. The WLM is not displayed since it has a behavior similar to the IRM. For the ICM mask the absolute value is displayed. The low values in the higher frequencies for the NR masks are due to the bandwidth restriction to 2.7 kHz of the simulated and real HF signals as discussed in **??**.

For a long time the described real valued masks were considered an upper bound on performance for mask estimation with neural networks [115], [193]. However, with the introduction of a time-domain loss these oracle masks are surpassed regarding the SDR evaluation measure [138]. Note, that only the real valued masks are outperformed by the new architectures. The ICM still is an upper limit for linear distortions.

To give a sense of the difficulty of noise reduction on the real recordings the next section introduces two baseline NR systems and their results on the real HF signals.

Table 7.2: Comparison of the oracle masks for BSS on the test data of the WSJ0-2mix database and NR on the evaluation set of the HF-Simu database

Name	Equation	Source Separation		Noise Reduction	
		PESQ	SDR / dB	PESQ	SDR / dB
No masking	1	1.86	0.1	1.44	0.2
Ideal ratio mask (IRM)	$\frac{\lvert x_{k,\ell,f}\rvert}{\lvert Y_{\ell,f}\rvert}$	3.67	12.4	3.51	13.3
Wiener-like mask (WLM)	$\frac{\lvert x_{k,\ell,f}\rvert^2}{\lvert Y_{\ell,f}\rvert^2}$	3.60	13.2	3.25	14.1
Ideal complex mask (ICM)	$\frac{x_{k,\ell,f}}{y_{\ell,f}}$	4.55	155.0	4.55	74.9

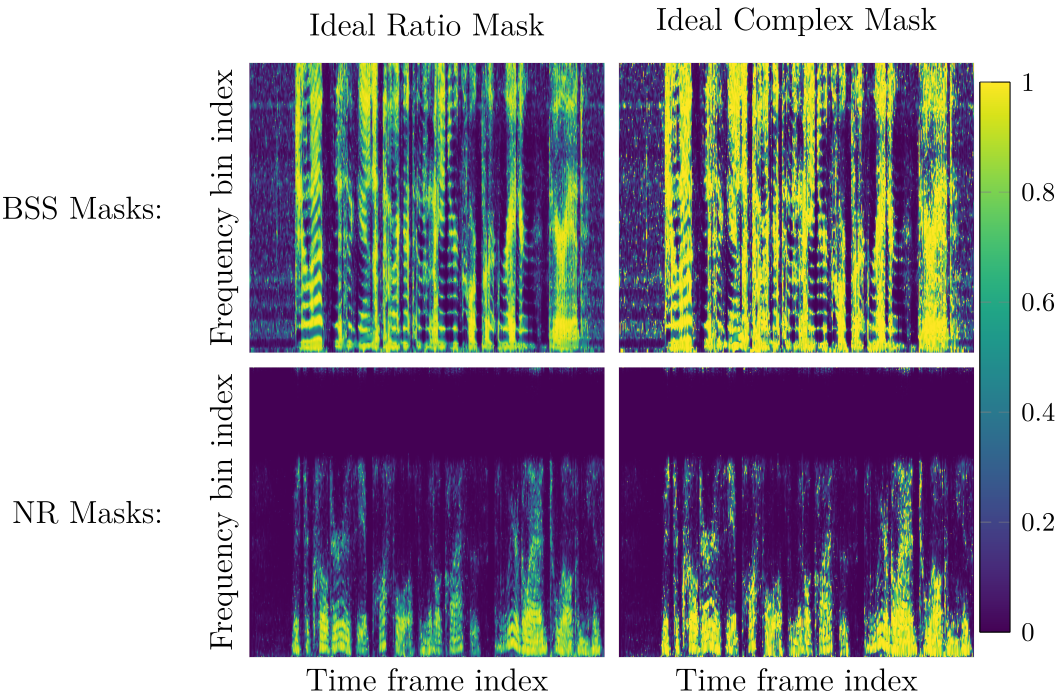


Figure 7.5: Examples for oracle masks for BSS (first row) and NR (second row) on the respective simulated evaluation sets

Real high frequency signals

For the real HF recordings only the observation and the paired clean speech but no noise data is recorded so that the masks discussed above cannot be calculated. Therefore, a simple quantile mask $M^S_{\ell,f}$ is derived from the clean signals, which is set to one for the 10 % of the time-frequency points with the highest energy and zero for all remaining time-frequency points. The noise mask $M^D_{\ell,f}$ is defined as the opposite to the speech mask with $M^D_{\ell,f} = 1 - M^S_{\ell,f}$ Both masks are multiplied with the input signal to get an estimate of the respective clean signal, which is then used to calculate a Wiener filter [92]:

$$W_{\ell,f} = \max\left(\frac{\left| M^S_{\ell,f} \cdot Y_{\ell,f} \right|^2}{\left| M^D_{\ell,f} \cdot Y_{\ell,f} \right|^2}, G_{\min} \right), \tag{7.14}$$

with $G_{\min}$ as a lower bound of the Wiener filter.

Another baseline system for the real HF data is spectral subtraction in the α-domain as suggested in [194] and for example applied in [OC7]. Performing the subtraction not on the power spectrum ($\alpha = 2$), but with a lower exponent α reduces the kortosis ratio [194], which is an indicator for musical noise in the processed signal [195]. Therefore, the estimated clean speech signal $\hat{S}_{\ell,f}$ can be described as

$$\hat{S}_{\ell,f} = \left(\left| Y_{\ell,f} \right|^{\alpha} - \overline{\left| D_f \right|^{\alpha}} \right)^{\frac{1}{\alpha}}, \tag{7.15}$$

with $\overline{\left|D_f\right|^\alpha}$ as the averaged noise spectrum in the α domain. To get an estimate of the noise statistics the silence prior to the observed signal is accumulated and the average operation is performed over the alpha domain signal

$$\overline{\left|D_f\right|^\alpha} = \frac{1}{L_D} \sum_{\ell=0}^{L_D-1} \left|D_{\ell,f}\right|^\alpha . \tag{7.16}$$

The estimated noise and speech power spectrum are then used to calculate a Wiener filter. For experiments on the simulated data the time-averaged noise power $\overline{\left|D_f\right|^\alpha}$ can be replaced by time-varying values $\left|D_{\ell,f}\right|^\alpha$ since parallel noise data is available.

In the following experiments the NR performance of the oracle mask and the statistical NR explained above are examined. Both baseline algorithms are performed in the frequency domain with $L_W = 64$ ms and $L_S = 16$ ms. As discussed in Section 7.5.2 neither the SDR nor SI-SDR evaluation measure is calculated on the real data because they are distorted by the AGC in the Kiwi-SDR receiver. In Table 7.3 the results of the oracle and statistical NR in comparison to the evaluation measure calculated directly on the recorded signal are displayed.

While both the statistical and the oracle results lead to improved PESQ and STOI values on the real data compared to the unprocessed signal, both evaluation measures are significantly lower than the ones calculated for the simulated data. There are multiple reasons for the lower results. First, there is the non-linear distortion of the original signal during the transmission. This is especially harmful to the considered systems since both enhancement algorithms are designed for linear distortions. Secondly, the real data includes some utterances with such low signal to noise ratio (SNR) that the speech signals are incomprehensible. For the statistical algorithm an additional impediment is the missing parallel noise information.

For the simulated data the switch from time-varying to an time-averaged noise signal leads to a loss of about 20 % in STOI. Even with the averaged noise information the statistical approach still achieves results close to the oracle mask with Wiener filtering. For STOI the difference amounts to around 3 % and for PESQ to 0.12.

Table 7.3: Comparison of the baseline results on the real HF data.

		Simulated		Real	
Method	Wiener filter	PESQ	STOI %	PESQ	STOI %
No enhancement	✗	1.44	70.1	1.69	58.7
Oracle Mask	✗	1.79	87.2	1.59	65.1
	✓	2.10	86.5	1.99	64.3
Statistical $\left\|D_{\ell,f}\right\|^\alpha$	✓	2.73	88.5	–	–
Statistical $\overline{\left\|D_f\right\|^\alpha}$	✓	1.47	69.5	1.87	61.5

These results on both the simulated and real data are presented to offer a value range to which the presented BSS and NR systems discussed in the next sections can be compared. First, the BSS systems are evaluated.

7.5.4 Source separation

In this section the modular source separation network is evaluated and different configurations are compared to show the potential of frequency domain separation despite the current trend towards latent domain networks. For these experiments both the convolutional (Section 7.3.1) and recurrent (Section 7.3.2) mask estimator are considered. For both mask estimators a configuration similar to the original publication is chosen with $R_C = 4$, $D_C = 8$, $B_C = 256$ and $H_C = 256$ for the convolutional mask estimator and $R_R = 6$, $\tilde{L}_S = 100$, $\tilde{L}_S = 50$, $B_R = 64$ and $H_R = 128$ for the recurrent one. We published a subset of the following experiments in [OC3].

Similar to [OC3] the Adam method [169] is used for optimization with a step size of 0.001 The step size is divided by 2 each time the SDR on the validation set is not reduced for 5 consecutive training epochs. Additionally, each time the step size is reduced, the learned parameters of the network are reset to the last state with the highest SDR. As in [OC3] the convolutional mask estimator is designed with a ReLU activation function and a global layer normalization [135]. In contrast to [OC3], the window size L_W and shift L_S are reduced from 4 ms to 2 ms and from 2 ms to 1 ms, respectively. This reduction increases the comparability to [44] and [135].

From frequency to time domain source separation

As in [OC3] the STFT-based mask estimator with a phase sensitive frequency loss $\mathcal{L}^{\mathrm{P-MSE}}$ is used as baseline. Starting from this baseline the encoder and decoder as well as the loss function is changed step by step until the resulting combination is a TasNet architecture. The results for the different steps are displayed in Table 7.4, where the domain name “Magnitude” represents the frequency magnitude spectrum and “Real+Imag” the concatenation of the real and imaginary part of the frequency spectrum.

For the network configuration with the MSE frequency domain loss $\mathcal{L}^{\mathrm{MSE}}$ a worse performance is observed than for the phase-sensitive baseline loss $\mathcal{L}^{\mathrm{P-MSE}}$, which is expected since the network does not learn to consider the phase for the mask estimation. However, the additional phase information gained by estimating a mask for both the real and imaginary part of the observation leads to the worst results of all experiments. Here, the additional phase information does not improve the network training. While the performance can be improved by calculating the SI-SDR loss $\mathcal{L}^{\mathrm{SI-SDR}}$, it is still slightly worse than the baseline results. Reducing the window size L_W and shift L_S to 2 ms and 1 ms during encoding and decoding leads to a relative improvement of more than 100 % for the SDR, SI-SDR and WER evaluation measures compared to $L_W = 64$ ms and $L_S = 16$ ms. This gain indicates the importance of small window sizes for enhancement with a time-domain loss function on non-reverberated data. This conclusion is supported by the experiments presented in [177] and [178]. The final step from the baseline to the TasNet separator is to replace the STFT encoder and inverse

Table 7.4: A step-by-step comparison of frequency and time domain source separation on the test set of the WSJ0-2mix database with the convolutional mask estimator. The line in the table separates configurations with a frequency and time domain loss function.

Domain	$L_\mathrm{W}/L_\mathrm{S}$ ms/ms	Loss-Fn	PESQ	STOI %	SI-SDR dB	SDR dB	WER %
Magnitude	64/16	$\mathcal{L}^\mathrm{P\text{–}MSE}$	2.29	86.3	7.9	8.3	43.36
Magnitude	64/16	$\mathcal{L}^\mathrm{MSE}$	2.36	88.0	7.7	8.0	45.04
Real+Imag	64/16	$\mathcal{L}^\mathrm{MSE}$	2.12	84.6	2.9	5.4	50.28
Real+Imag	64/16	$\mathcal{L}^\mathrm{SI\text{–}SDR}$	2.29	86.6	9.2	9.7	44.47
Real+Imag	2/1	$\mathcal{L}^\mathrm{SI\text{–}SDR}$	2.82	92.8	12.9	13.2	26.16
Latent	2/1	$\mathcal{L}^\mathrm{SI\text{–}SDR}$	3.21	94.5	15.4	15.8	20.83
Latent	2/1	$\mathcal{L}^\mathrm{T\text{–}LMSE}$	3.26	94.6	15.4	15.8	20.80

STFT decoder with a learned encoder and decoder. Due to the separation in the latent space, the SDR is improved by additional 2 dB. Switching from the SI-SDR loss $\mathcal{L}^\mathrm{SI\text{–}SDR}$ to $\mathcal{L}^\mathrm{T\text{–}LMSE}$ leads to only small improvements.

In summary, the main advantage of the TasNet architecture compared to the baseline system is a combination of the time domain loss and a small window size and shift in the encoder and decoder. In other words, frequency domain source separation offers the possibility to achieve competitive results if the configuration is optimized. Because the difference between the $\mathcal{L}^\mathrm{T\text{–}LMSE}$ and $\mathcal{L}^\mathrm{SI\text{–}SDR}$ loss function is quite small and the $\mathcal{L}^\mathrm{SI\text{–}SDR}$ loss offers better comparability to other publications, all following experiments with time domain loss functions will be performed using the $\mathcal{L}^\mathrm{SI\text{–}SDR}$ loss.

Component comparison for source separation

To further investigate the influence of the latent domain on the separation results different combinations of the learned and STFT-based en- and decoder are compared. To ensure that these observations are not specific to the convolutional mask estimator the same experiments are performed using the recurrent mask estimator. The results are displayed in Table 7.5.

While the architecture with both learned encoder and decoder achieves the best results, the gap in SDR between the frequency and latent domain is reduced to 0.9 dB from 2.6 dB by using an learned inverse transform. A learned encoder with the fixed inverse STFT decoder gives a similar improvement of 1.5 dB in SDR to 14.7 dB. For the recurrent mask estimator the differences between the models are comparable to the ones witnessed for the convolutional mask estimator. Only for the learned encoder with an inverse STFT the results deviate, where the recurrent mask estimator achieves a SDR of 15.60 dB which is close to the result of the convolutional mask estimator with learned encoder and decoder. The WER of this model is even lower than the result of the systems with learned encoder and decoder, indicating the benefits of the regularization of the latent space due to the fixed decoder. Most of the

Table 7.5: Comparison of different encoder and decoder combination for both the recurrent and convolutional mask estimator using $L_{\mathrm{W}} = 2\,\mathrm{ms}$ and $L_{\mathrm{S}} = 1\,\mathrm{ms}$ on the test set of the WSJ0-2mix database.

Mask estimator	Encoder	Decoder	PESQ	STOI %	SI-SDR dB	SDR dB	WER %
Convolutional	learned	learned	3.21	94.5	15.4	15.8	20.83
	STFT	learned	3.09	94.4	14.6	14.9	22.57
	learned	inverse STFT	3.08	92.9	14.3	14.7	23.48
	STFT	inverse STFT	2.82	92.8	12.9	13.2	26.16
Recurrent	learned	learned	3.27	95.1	15.7	16.1	20.82
	STFT	learned	2.98	94.4	14.1	14.5	22.13
	learned	inverse STFT	3.18	95.0	15.3	15.6	20.64
	STFT	inverse STFT	2.99	93.0	13.7	14.1	24.45

evaluated configurations with a recurrent model achieve improved results compared to the corresponding model with a convolutional mask estimator.

The slightly worse results of the fixed compared to the learned encoder/decoder are somewhat compensated by the higher interpretability of the transformed signals for the STFT compared to the latent encoder. Furthermore, the STFT encoder allows a combination of the NN-based separation with well known statistical enhancement and extraction algorithms in the spectral domain like mixture models [124], beamforming [29] or Wiener filtering [92]. Overall, the frequency domain separation is still a valid alternative to the latent space network and achieves strong separation results.

Both the separation with the recurrent and the convolutional mask estimator achieve similiar results with a window size L_W and shift L_S of 2 ms and 1 ms, respectively. These results fit the observation published in [44] that the DPRNN significantly outperforms the convolutional mask estimator for smaller L_W, with $L_W = 0.25\,\mathrm{ms}$ giving the best results. These small window sizes are not considered in this work since they drastically increase the computational complexity of the model. The 16.1 dB of the enhanced signal are close to the results presented in [44], where the small difference may be attributed to the different data preparation described in Section 7.5.1 and the differences in optimization.

The simulations presented in this work indicate, that both the recurrent and convolutional mask estimator can achieve competitive separation results for both frequency and latent domain separation. Therefore, both will be considered in the following experiments. During the next section, the system is evaluated for noisy HF recordings. The experiments serve to clarify whether the changes from frequency to time domain separation described in Table 7.4 lead to similar improvements in the noise reduction task.

7.5.5 Noise reduction

This section deals with two questions. The first question is whether the gains of time domain source separation can be transferred to the NR task. Second, it is examined whether the insights gained from the experiments on the simulated HF signals can be transferred to the real HF recordings.

From frequency to time domain noise reduction for simulated data

First, the experiments described in Section 7.5.4 are repeated for the NR task on the simulated HF data. The training on the simulated HF signals is similar to the one described above for the BSS task. Only the comparison between the training steps differs from the BSS tasks, because the SDR evaluation measure is calculated for the estimation of the first source signal but not for the second. Since the second source represents a noise signal for the NR task, it is not important for a successful noise reduction and is discarded. Only the enhanced speech signal is used during evaluation. Additionally, a clipping of the time-domain losses as suggested in [4] is used in the following experiments, since preliminary results suggested that this clipping is beneficial for the NR task.

As above, the baseline system is a frequency domain mask estimator with a phase-sensitive loss. The encoder, decoder, window and shift sizes as well as the loss function of this baseline are varied until the network configuration is changed to a time domain noise reduction network. The results for the different stages are displayed in Table 7.6.

Changing the parameters in the NR architecture leads to similar improvements as for BSS. Only the influence of the latent domain is smaller than the one witnessed for BSS. Nevertheless, the NR in the latent domain with a time domain loss achieves the strongest performance with an SDR of 14.8 dB. The SDR and all other evaluation measures are worse than the ones in BSS which can be explained by the missing upper frequencies due to the 2.7 kHz bandwidth restriction. Preliminary experiments without the bandwidth restriction lead to significantly higher evaluation measures close to those seen for BSS.

Table 7.6: A step-by-step comparison of frequency and time domain noise reduction on the evaluation set of the simulated HF database with the convolutional mask estimator. The line in the table separates configurations with a frequency and time domain loss function.

Domain	$L_\mathrm{W}/L_\mathrm{S}$	Loss-Fn	PESQ	STOI	SI-SDR	SDR	WER
	ms/ms			%	dB	dB	%
Magnitude	64/16	$\mathcal{L}^{\mathrm{P-MSE}}$	2.33	86.4	9.8	10.4	27.87
Magnitude	64/16	$\mathcal{L}^{\mathrm{MSE}}$	2.29	85.5	10.6	11.4	26.88
Real+Imag	64/16	$\mathcal{L}^{\mathrm{MSE}}$	2.07	81.1	8.7	11.0	34.89
Real+Imag	64/16	$\mathcal{L}^{\mathrm{SI-SDR}}$	2.43	87.5	12.0	13.2	27.51
Real+Imag	2/1	$\mathcal{L}^{\mathrm{SI-SDR}}$	2.52	88.7	13.3	14.2	21.90
Latent	2/1	$\mathcal{L}^{\mathrm{SI-SDR}}$	2.70	90.1	13.8	14.8	21.42

Furthermore, the time-domain NN outperforms the real-valued oracle masks discussed in Section 7.5.3 similar to the model for the BSS task. Note, that many of the NR models lead to an improved WER compared to the error rate of the unprocessed signals with a WER of 29.00 %. This gain is noteworthy because it contradicts the assumption that in general a ASR system is more inhibited by distortions introduced by a NR system than by the additive noise in the observed signal [108]. One explanation for these improvements is the high non-stationarity of the noise, which degrades the recognition results even though the network is trained on similar signals. A stronger ASR architecture might no longer profit from this single channel enhancement as suggested in [196].

Mask estimator comparison for noise reduction on simulated data

To verify the smaller influence of the latent domain, the next experiments compare the NR results in the latent and time domain for both mask estimators. The results are displayed in Table 7.7. For both the convolutional and recurrent mask estimator the choice of the feature domain only slightly influences the noise reduction results. While the network with STFT encoder and inverse STFT decoder performs worse than the system with the learned encoder and decoder, the difference is not as large as in Table 7.5. The intermediate steps with learned encoder and fixed decoder, as well as fixed encoder and learned decoder, are not considered here, since they do not lead to a significant gain over the network with fixed encoder and decoder.

Note that the recurrent mask estimator is outperformed by the convolutional estimator independent of the chosen feature domain, which is not consistent with the results observed for the BSS task. This difference to the separation can be explained by the low temporal auto-correlation of the noise signal compared to the speech signal of an overlapping speaker. Therefore, the larger temporal context of the recurrent estimator due to the second, inter-segment RNN layer is not as beneficial. Note, that the STFT domain signal is limited to the lower frequencies due to the low pass filtering during demodulation, the learned encoder is not limited and learns a latent domain where the whole feature dimension contains signal information.

Overall, the presented NR experiments on the simulated HF mirror the results for BSS. In both BSS and NR the SDR evaluation measure for the better performing neural networks outperform the real valued oracle masks in Section 7.5.3. This is true for both frequency

Table 7.7: Comparison of the recurrent and convolutional mask estimator using $L_\mathrm{W} = 2\,\mathrm{ms}$ and $L_\mathrm{S} = 1\,\mathrm{ms}$ on the evaluation set of the simulated HF database.

Mask estimator	Encoder	Decoder	PESQ	STOI	SI-SDR	SDR	WER
				%	dB	dB	%
Convolutional	learned	learned	2.70	90.1	13.8	14.8	21.42
	STFT	inverse STFT	2.52	88.7	13.3	14.2	21.90
Recurrent	learned	learned	2.62	88.9	13.3	14.4	23.25
	STFT	inverse STFT	2.50	87.8	12.9	13.9	23.15

domain and latent domain enhancement systems. Especially for BSS the best performing system outperforms the oracle results by more than 2 dB. However, the real valued oracle masks still outperform the systems in regards to speech quality represented by the PESQ value. These results indicate, that there is still information to be gained from the oracle masks that can improve the results by reducing the focus of the network training on SDR improvement.

In the next sections the systems evaluated for NR are examined with regards to their performance on the real HF data. These experiment show whether the non-linear distortion during transmission and the mismatch between training and evaluation data lead to a drop in performance.

From frequency to time domain noise reduction for real data

In the following experiments, the networks trained on the simulated data for the experiments in the last section are applied to the real HF data without retraining. The results are displayed in Table 7.8.

As depicted in Table 7.8 the differences between the results with the different parameters are lower than the ones presented in Table 7.4. Overall, the NR on real recordings leads to worse results for both the PESQ and STOI evaluation measure compared to NR on the simulated signals. This drop in performance can be partially explained by a shorter average utterance length, the lower SNR and the non-linear distortion due to the transmission and the AGC. All systems outperform both the statistical baseline and the oracle mask with regards to PESQ and only the frequency domain system with the MSE loss $\mathcal{L}^{\mathrm{MSE}}$ achieves a worse STOI result than both baselines. Therefore, we can assume that all trained systems are effective for the real HF recordings.

The best result is achieved by a noise reduction in the frequency domain using a time domain loss with a PESQ value of 2.32 and a STOI value of 72.6 %. A similar result is achieved with the phase sensitive MSE loss. Both systems estimate the mask in the frequency domain with

Table 7.8: A step-by-step comparison of frequency and time domain noise reduction on the evaluation set of the real HF database with the convolutional mask estimator. The line in the table separates configurations with a frequency and time domain loss function.

Domain	$L_{\mathrm{W}}/L_{\mathrm{S}}$	Loss-Fn	PESQ	STOI
	ms/ms			%
Magnitude	64/16	$\mathcal{L}^{\mathrm{P–MSE}}$	2.23	71.4
Magnitude	64/16	$\mathcal{L}^{\mathrm{MSE}}$	2.17	69.8
Real+Imag	64/16	$\mathcal{L}^{\mathrm{MSE}}$	2.07	58.9
Real+Imag	64/16	$\mathcal{L}^{\mathrm{SI–SDR}}$	2.32	72.6
Real+Imag	2/1	$\mathcal{L}^{\mathrm{SI–SDR}}$	2.17	69.4
Latent	2/1	$\mathcal{L}^{\mathrm{SI–SDR}}$	2.27	69.3
Latent	64/16	$\mathcal{L}^{\mathrm{SI–SDR}}$	2.41	72.8

$L_W = 64\,\mathrm{ms}$ and $L_S = 16\,\mathrm{ms}$. Therefore, a network in the latent domain with the larger L_W and L_S is trained to examine whether the STFT domain generally improves the results or if the improvements can be attributed to the larger L_W and L_S. This separation in the latent domain further improves the enhancement with a PESQ value of 2.41 and a STOI value of 72.8 %. These results are comparable to the experiments presented in [OC3], where for the more realistic database, a larger shift and window size in the encoder is beneficial for the network. One conclusion is that the enhancement network does not profit from the small window L_W and shift size L_S if there is a larger mismatch between training and evaluation. In other words, the larger window and shift size can increase the robustness of the network at the cost of degraded results for signals that are very similar to the training set.

Mask estimator comparison for noise reduction on real data

To asses the difference between the latent and frequency domain noise reduction for real recordings, both mask estimators are compared in both domains. The results are presented in Table 7.9 for $L_W = 64\,\mathrm{ms}$ and $L_S = 16\,\mathrm{ms}$. The larger sizes are chosen because they consistently produce improved enhancement results. For a comparison with results for a shorter L_W and L_S refer to the appendix (Table A.6).

Again, the convolutional mask estimator slightly outperforms the recurrent mask estimator for both feature domains. Note, that the difference between the results of the frequency and latent domain networks is even smaller than for the experiments on the simulated signals. One explanation for this difference between the experiments on the real and simulated HF signals is that the fixed transformation prevents the network from overfitting the latent domain to the simulated data. This overfitting might still improve the results on the simulated evaluation set, but no longer leads to an improvement for the real HF data with larger mismatch to the training data.

Table 7.9: Comparison of the recurrent and convolutional mask estimator using $L_\mathrm{W} = 64\,\mathrm{ms}$ and $L_\mathrm{S} = 16\,\mathrm{ms}$ on the evaluation set of the real HF database.

Mask estimator	Encoder	Decoder	PESQ	STOI %
Convolutional	learned	learned	2.41	72.8
	STFT	inverse STFT	2.32	72.6
Recurrent	learned	learned	2.30	72.3
	STFT	inverse STFT	2.32	72.1

7.5.6 Noise reduction on an unseen language

In previous experiments all models were trained and tested on English speech for both the simulated and real HF data. In this section different models trained on simulated signals with English speakers are compared for real HF recordings of Russian speech signals. Without any

Table 7.10: Comparison of a NR system with both recurrent and convolutional mask estimator on the evaluation set of the real Russian HF database.

Mask estimator	L_W/L_S	Fixed encoder/decoder		Trained encoder/decoder	
		PESQ	STOI / %	PESQ	STOI / %
Convolutional	4/2	2.14	74.5	2.18	74.6
	64/16	2.23	76.4	2.33	77.1
Recurrent	4/2	2.03	74.1	2.15	74.6
	64/16	2.25	76.7	2.21	76.8

NR an average PESQ value of 1.70 and a STOI value of 59.1 % is calculated. These values are comparable to those calculated for the real English HF recordings in Section 7.5.3. The results are displayed in Table 7.10.

Similar to the results for the English recordings, the models with fixed and learned encoder/decoder achieve a similar performance. This is especially true for the recurrent mask estimator, where the STFT encoder/decoder even slightly outperforms the separation in the learned latent space in both evaluation measures. Overall, the convolutional mask estimator with the larger L_W and L_S using a trained encoder and decoder achieves the best noise reduction with a PESQ value of 2.33 and a STOI value of 77.1 %. For all models the intelligibility after NR is higher for the Russian signals compared to the English recordings. During the experiments in Section 5.5.1 the SAD systems delivered a stronger performance on the Russian than on the English data. This difference is attributed to a lower amount of concurrent speakers in the Russian data, which could also explain the difference in NR performance.

7.6 Summary

In this chapter a model architecture for BSS and NR is described and evaluated. Although current system for BSS and NR rely on masking in a learned latent space, the experiments presented in this chapter show that STFT-based separation can still achieve competitive results and in some metrics even outperform the latent domain networks. Additionally, the evaluation results show the benefits of a larger window size L_W and shift L_S for signals with more realistic distortions. This is true for both BSS and NR. However, other results cannot be transferred from BSS to NR. While the recurrent mask estimator outperforms the convolutional mask estimator for source separation, it leads to worse results in all of the NR experiments.

For future experiments on real recordings only models with larger window length $L_W = 64\,\text{ms}$ and shift $L_S = 16\,\text{ms}$ in the encoder and decoder are considered. These models have shown to consistently outperform the models with small L_W and L_S for experiments on both English and Russian recordings.

In this chapter the NR system is evaluated on signals with perfect information about the CFO and speech activity. The next chapter takes a step towards a more realistic system by combining the three presented algorithms for SAD, CFO correction and NR to a multi-stage approach and evaluates the dependencies between the stages.

8 Multi-Stage Processing

In this chapter the multi-stage system discussed in Chapter 1 is evaluated. It consists of three subsystems:

- the speech activity detection (SAD) networks as presented in Chapter 5,
- the carrier frequency offset (CFO) correction discussed in Chapter 6,
- the noise reduction (NR) networks as introduced in Chapter 7.

In Chapter 7 the NR is trained and evaluated on signals without a CFO, while assuming a prior CFO correction. However, the NR could also be performed on signals with a CFO $\varDelta^{\mathrm{f}} \geq 0$ without correcting the CFO if the model's training data is changed accordingly. Therefore, two different multi-stage systems are considered as depicted in Figure 8.1: One, where the CFO is corrected before the NR network, and a second, where NR is performed on the shifted signals with a subsequent CFO estimation and correction. For both enhancement approaches only the segments with speech activity are concatenated and sent to the enhancement stage.

To reduce the scope of the investigations only a limited number of the subsystems discussed in the previous chapters are considered. These systems for CFO and activity estimation

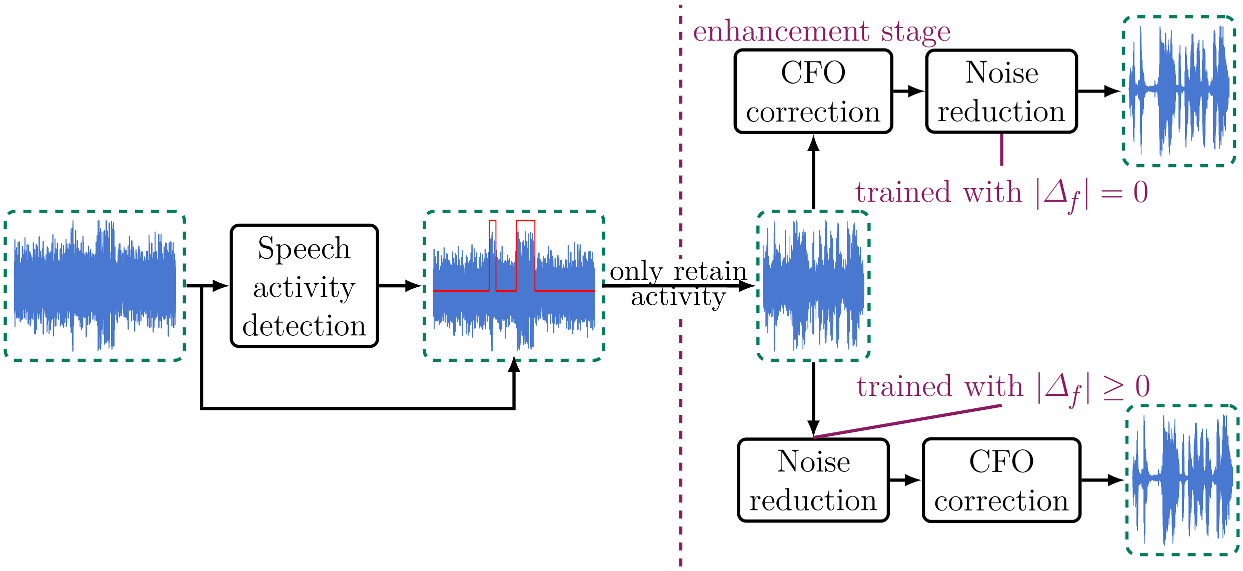


Figure 8.1: Block diagram of both considered multi-stage systems, where the enhancement consists of a CFO correction with subsequent NR (upper path) or the same blocks in the reversed order (lower path).

Table 8.1: Subsystems

Task	System Name	Parameters
SAD	Oracle	Oracle activity information
	Real	SRNN trained only on real data without a CFO
	RealSimu	SRNN trained on real and simulated data
CFO estimation	Oracle	Oracle CFO information
	RAKE	Statistical CFO estimation algorithm
	Full-band (CNN)	Full-band classifier with a CNN sub-band layer
	Full-band(RNN)	Full-band classifier with a RNN sub-band layer
	Sub-band	Shift-invariant sub-band classifier
NR	Convolutional	ConvNet with trained encoder / decoder
	Recurrent	DPRNN with trained encoder / decoder

are presented in Table 8.1 with a short explanation. For both SAD systems the threshold with results closest to the equal error rate (EER) on the simulated high frequency (HF) development set is chosen.

Both the convolutional and recurrent mask estimator with trainable encoder and decoder are used and compared as NR systems. Following the experiments discussed in Section 7.5 all models for NR are designed with the larger window length and shift $L_W = 64\,\mathrm{ms}$ and $L_S = 16\,\mathrm{ms}$ to achieve the best performance on the real recordings.

In the following sections different combination of sub-systems will be evaluated with the evaluation measures discussed in Section 7.5.2 and Section 8.2. All experiments will be performed on the real recordings described in **??**. The next section examines the enhancement stage, which consists of a CFO correction and a NR system.

8.1 Evaluation of the enhancement stage

Prior experiments have considered noise reduction and CFO correction as separate tasks. However, since the second tasks is dependent on the performance of the prior task The next experiments will evaluate how the errors due to an imperfect CFO estimation affects the NR.

Note, that these experiments are performed on each utterance in a recording to be comparable to the experiments discussed in Section 7.5. This is a change to the setup for prior CFO estimation experiments in Section 6.4. Here, the CFO estimator is applied to each utterance without speech activity information. Oracle information is used to define the utterance boundaries. Therefore, the speech activity in the input is below 8 s and interspersed with silence which is different to the CFO experiments in Section 6.4. In Figure 8.2 the cumulative distribution functions (CDFs) of the estimation error on this new setup for the four considered CFO estimators are displayed. All estimators show a worse performance on these shorter signals than on the large activity segments in Section 6.4. This is to be expected, since

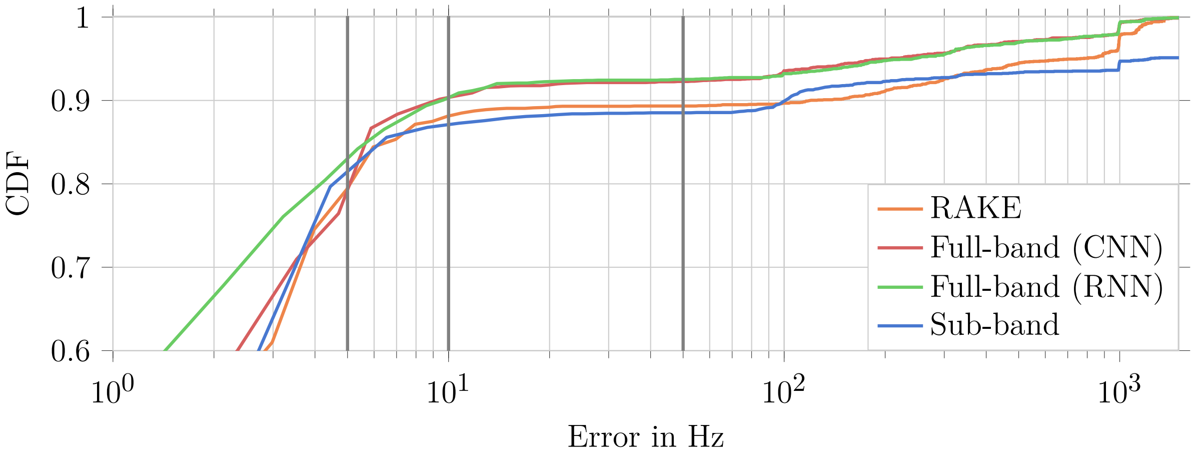


Figure 8.2: CDF of the CFO estimation error for the neural network (NN)-based systems compared with the RAKE algorithm for activity segments shorter than 8 s. The number of errors used as boundaries for the error class affiliation (ECA) are marked with gray lines.

experiments in Section 6.4 already showed the benefits of input segments with lengths above 10 s for the CFO estimation.

After evaluating the CFO estimation on its own, it is combined with the noise reduction system for the following experiments. The system combinations are evaluated with respect to the speech enhancement metrics short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ).

The results for both variations of the enhancement stage in case of an estimated CFO are displayed in Table 8.2. Here, a CFO correction prior to the NR model outperforms the enhancement stage with a noise reduced input to the CFO.

A CFO correction followed by a NR network is expected to outperform a system with a

Table 8.2: Comparison of NR systems with and without a prior CFO correction during training and evaluation on the real HF data with a CFO for both variations of the enhancement stage.

Mask estimator	CFO estimator	CFO → NR PESQ	CFO → NR STOI / %	NR → CFO PESQ	NR → CFO STOI / %
Convolutional	Oracle	2.55	78.4	2.40	70.1
	RAKE	2.52	75.7	2.01	64.0
	Full-band (CNN)	2.54	76.9	2.27	68.4
	Full-band (RNN)	2.54	76.9	2.31	68.9
	Sub-band	2.43	75.8	2.23	67.1
Recurrent	Oracle	2.38	77.8	2.29	68.6
	RAKE	2.31	74.1	1.93	62.4
	Full-band (CNN)	2.37	76.3	2.19	67.0
	Full-band (RNN)	2.37	76.3	2.22	67.4
	Sub-band	2.27	75.0	2.12	65.5

CFO correction after the NR, since the CFO estimators are trained to deal with noise but not with the distortions introduced by the noise reduction system. Therefore, the following experiments will only be performed using the NR network trained on signals without a CFO.

Note, that the enhancement results using an estimated CFO are very close to those with oracle CFO information. Especially, for the full-band classifiers the PESQ values are close to identical to the oracle experiments. This indicates that small errors in the CFO estimation do not impact the noise reduction system. Additionally, for some signals with a high CFO estimation error, it is reasonable to assume that the signal is so distorted that the noise reduction network does not perform well even with the oracle CFO correction. This is in-line with the evaluation of the CFO correction in Section 6.4.6, where some of the high CFO estimation errors are masked by the low signal quality. For a broader evaluation of influence of a remaining CFO on the noise reduction performance please refer to Appendix A.7.

The presented experiments all use oracle activity information to process each utterance individually. However, this information is not available in a realistic scenario. Therefore, evaluation measures for experiments without utterance boundary information are discussed in the next section.

8.2 Performance measure for imperfect activity information

The combination of subsystems is mostly evaluated with the evaluation measures described in Section 7.5.2. However, in case of imperfect activity information the enhanced signal and the transmitted clean signal have a different length. In this case, the evaluation measures can not be applied since they are defined only for an equal length of the reference and estimated signal. Therefore, the sizes are adjusted in two steps. First, all detected activity in a recording is concatenated and sent through the enhancement algorithm. Then, the missing activity (false negative) is added as zeros to the enhanced signal. Now the enhanced signal is larger than the activity in target signal because of possible false positives in the activity estimation. Therefore, for each sample with a false positive activity estimation, zeros are inserted into the target signal to equalize the length between the target and the estimated signals. Adding these zeros to the estimated and target signal penalizes the false positives and false negatives in the enhancement evaluation measures. The resulting target signal is compared to the enhanced signal using the evaluation measures discussed in Section 7.5.2. An illustration of these adjustment to the evaluation scheme is displayed in Figure 8.3.

In Table 8.3 the evaluation measures for the real HF data are displayed for the unprocessed signals. These simulations serve as a baseline and can be compared to the evaluation measures calculated per utterance.

For oracle activity information with oracle CFO correction the evaluation on whole recordings leads to a PESQ value of 1.55 and a STOI value of 58.76 % on the evaluation set of the real HF database without any additional processing. The described results are similar to those of the experiments per utterance in Section 6.4.6. Furthermore, the different results for the

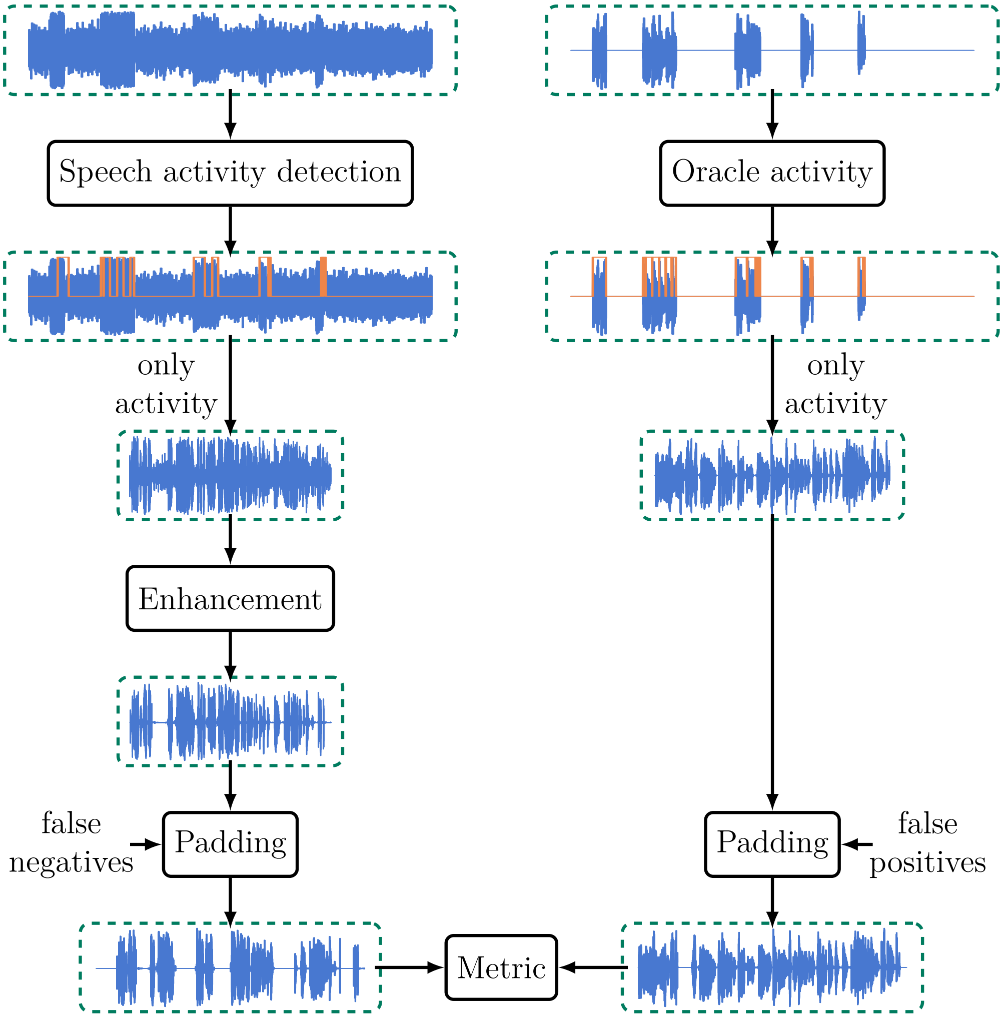


Figure 8.3: Illustration of the preprocessing for the calculation of performance measures in case of imperfect activity information.

“Real” and “RealSimu” SAD system show the influence of CFOs in the training data for evaluation on signals with a CFO.

Note, that the PESQ value increases for imperfect activity information on signals without a CFO, while the STOI decreases as expected. This result can be explained by the zeros introduced by false positives of the activity detection, where silence is compared to noise. These effects have to be considered when interpreting the NR evaluation measures on systems using imperfect activity information.

For recordings without a CFO the CFO correction leads to an improved result although there is no shift to correct. This gain can be explained by the low pass applied during correction, which leads to a reduction of the noise energy in the higher frequencies without speech information, as discussed in Section 6.4.6.

In the following section, the described evaluation measures are used to examine the overall system, consisting of SAD, CFO correction and NR, regarding its influence on the speech quality and intelligibility.

Table 8.3: Comparison of NR evaluation measures for the real HF input data with and without a CFO with oracle CFO correction for imperfect activity information.

SAD	CFO	$\Delta^{\mathrm{f}} = 0$		$\Delta^{\mathrm{f}} \geq 0$	
	correction	PESQ	STOI / %	PESQ	STOI / %
Oracle	✗	1.47	50.0	1.25	40.5
	✓	1.55	58.8	1.71	62.9
Real	✗	1.54	46.2	1.39	32.9
	✓	1.62	57.2	1.67	49.8
RealSimu	✗	1.56	45.9	1.35	38.0
	✓	1.64	56.8	1.71	59.9

8.3 Evaluation of the multi-stage system

Each of the systems in the multi-stage approach displayed in Figure 8.1 has shown to achieve strong results as a single system. In this section, the multi-stage system is applied to the real HF recordings for different system combinations and evaluated in terms of its NR performance. The resulting evaluation measures for both the real HF recordings with and without a CFO are shown in Table 8.4.

While the gap between the performance of the convolutional mask estimator and the recurrent mask estimator is smaller compared to most previous NR experiments, the convolutional architecture still slightly outperforms the recurrent network, especially in the PESQ evaluation measure. This difference can partially be explained by the larger activity segments used. All studies in previous sections combining NR and CFO estimation for real recordings were performed individually for each excerpt of an utterance from the LibriSpeech database, but here all speech activity segments in a recording are concatenated. Therefore, the input duration is no longer limited to an activity below 8 s, but is increased to more than 10 s for all examples. Since the temporal context of the convolutional mask estimator is limited as discussed in Section 7.3.1, it cannot take advantage of this additional information. On the other hand, the recurrent mask estimator is designed to exploit large temporal contexts, allowing it to use the additional information.

For both NN-based SAD systems the results on the recordings without a CFO are comparable to the NR with oracle SAD information. Nevertheless, the slightly stronger SAD performance of the SAD system only trained on real signals leads to small improvements regarding the STOI evaluation measure. As discussed in Section 8.2 errors during SAD leads to a large drop in performance in regards to the STOI evaluation measure, while only slightly impacting the PESQ measure. Therefore, the difference in the STOI evaluation measure can be be attributed to a different SAD performance despite the similar PESQ values.

For the recordings with $\Delta^{\mathrm{f}} \geq 0$ the difference between the multi-stage system with and without oracle SAD information is more prominent, which is expected since it reflects the evaluation of the single SAD systems in Section 5.5. Especially, the results for the network only trained on real recordings without a CFO are worse with a drop of around 16 % in the

Table 8.4: Comparison of different configuration of the multi-stage system on the real HF data with ($\varDelta^{\mathrm{f}} \geq 0$) and without ($\varDelta^{\mathrm{f}} = 0$) a CFO. For all three SAD systems the highest value for each evaluation measures is marked as bold.

SAD	CFO	Mask	$\varDelta_f = 0$		$\varDelta_f \geq 0$	
	estimator	estimator	PESQ	STOI / %	PESQ	STOI / %
Oracle	Oracle	Convolutional	2.27	76.8	2.36	78.6
		Recurrent	2.15	76.2	2.21	78.2
	RAKE	Convolutional	2.28	76.8	**2.38**	74.1
		Recurrent	2.16	76.1	2.20	76.7
	Full-band (CNN)	Convolutional	2.28	77.3	2.34	78.1
		Recurrent	2.17	76.7	2.19	77.7
	Full-band (RNN)	Convolutional	2.28	77.3	2.35	78.2
		Recurrent	2.23	77.5	2.20	77.9
	Sub-band	Convolutional	**2.34**	**77.6**	2.34	**78.7**
		Recurrent	2.21	76.8	2.19	78.1
Real	Oracle	Convolutional	2.29	75.0	2.14	62.3
		Recurrent	2.21	74.5	2.06	62.3
	RAKE	Convolutional	2.28	73.1	2.14	60.5
		Recurrent	2.21	72.1	2.09	60.2
	Full-band (CNN)	Convolutional	2.30	75.4	**2.15**	**61.8**
		Recurrent	2.22	74.9	2.06	61.0
	Full-band (RNN)	Convolutional	**2.36**	**76.2**	**2.15**	61.6
		Recurrent	2.22	74.9	2.07	60.8
	Sub-band	Convolutional	2.35	75.9	2.03	60.5
		Recurrent	2.27	75.3	1.97	60.0
RealSimu	Oracle	Convolutional	2.28	74.4	2.30	75.0
		Recurrent	2.21	74.0	2.21	74.8
	RAKE	Convolutional	2.27	68.4	**2.33**	72.2
		Recurrent	2.27	70.2	2.20	70.1
	Full-band (CNN)	Convolutional	2.39	**73.4**	2.30	72.2
		Recurrent	2.30	72.9	2.20	72.0
	Full-band (RNN)	Convolutional	**2.41**	72.7	**2.33**	**73.2**
		Recurrent	2.33	72.1	2.23	72.9
	Sub-band	Convolutional	**2.41**	72.5	2.25	72.5
		Recurrent	2.32	71.8	2.16	71.9

STOI value for both mask estimators with oracle information regarding the CFO. Using the SAD system trained on real recordings without a CFO and simulated signals with a CFO reduces the loss to 4 % for the STOI evaluation measure compared to the oracle system. Therefore, the following experiments only use the SAD system trained on both real and simulated signals for the recordings since it on average shows the best performance for both data sets.

8.4 Impact of an unseen language

During previous chapters all single systems are examined for their performance on a language not seen during training. In this section the performance of the multi-stage system is evaluated on recordings of utterance in a previously unseen language.

Overall, the difference for both the SAD and the CFO estimation to the oracle information is limited. Therefore, the NR network is the bottleneck for stronger speech enhancement so only the two mask estimator networks are compared, with the best performing CFO estimation and SAD combination as front-end. Here, the experiments are performed using the full-band classifier with a recurrent neural network (RNN) sub-band layer (SBL) for CFO estimation and the SAD model trained on both real and simulated signals.

As expected the NR performance of the multi-stage system on the Russian language data is comparable to the results on English recordings presented in Section 7.5.6. Therefore, the multi-stage system can be considered as robust to a change in the recorded language as the single NR system.

Table 8.5: Comparison of NR performance of the multi-stage system on the real Russian HF data with ($\mathit{\Delta}^{\mathrm{f}} \geq 0$) and without ($\mathit{\Delta}^{\mathrm{f}} = 0$) a CFO. The CFO estimation is always performed with the full-band classifier with a RNN SBL and for SAD the model trained on both real and simulated signals is chosen.

SAD / CFO	Mask	$\mathit{\Delta}_f = 0$		$\mathit{\Delta}_f \geq 0$	
	estimator	PESQ	STOI / %	PESQ	STOI / %
Oracle	Convolutional	2.04	76.6	2.19	73.6
	Recurrent	1.92	76.3	2.05	73.0
Estimated	Convolutional	2.26	77.6	2.24	72.7
	Recurrent	2.26	77.7	2.20	72.4

8.5 Summary

In this chapter the systems presented in the previous chapters are combined to a multi-stage system. Here, two possible combinations of the CFO correction and the NR are considered. However, it is shown that the performance of the CFO estimation is reduced by a prior noise reduction with the considered networks. Therefore, the multi-stage system is composed of a SAD network followed by an enhancement stage consisting of a CFO estimation and correction with subsequent NR.

For this multi-stage system, different configurations are evaluated, where the best performance is achieved using the following configuration:

- a SAD system trained both on real recordings without a CFO and simulated signals with a CFO,

- a full-band classifier with a RNN SBL as CFO estimator,
- a NR network using the convolutional mask estimator with two-step training.

This best performing system configuration is evaluated on its performance on a language not seen during training and it achieves results comparable to those reported for the English recordings.

9 Conclusion and Outlook

9.1 Main contributions

This work offers new insights into three areas of research. First, an architecture for neural network (NN)-based speech activity detection (SAD) is introduced in Section 5.2 that achieved the best published performance on the Fearless Steps challenge data to date. This architecture is shown to outperform comparable architectures on single-sideband (SSB)-modulated speech transmissions in Section 5.5.1 and is adapted to detect speech shifted due to a carrier frequency offset (CFO) in Section 5.5.3. Initial experiments have shown that the trained model is able to detect speech from Russian speakers although it is only trained on the English language, indicating a certain robustness against a prior unseen language. Furthermore, the SAD network is extended in Section 5.5.4 to detect speech activity on the spectrum of complex-valued baseband signals, which allows a strong detection performance even for negative CFOs.

The second developed building block is the CFO correction. Here, a statistical approach called RAKE is described in Section 6.1 and in Section 6.2, two NN architectures for CFO estimation are presented, the so-called full-band and sub-band classifiers. All three CFO estimators are shown to predict the CFO with a low error rate for input signals with a duration of at least 2 s. Additionally, it is shown in Section 6.4.5 that the sub-band classifier allows the prediction of negative CFOs if the complex-valued baseband signal is provided. The impact of the CFO correction on the speech quality and intelligibility is evaluated in a further experiment where the CFO estimation of all three systems lead to similar speech quality improvements as the correction with the true CFO. In a final experiment all three CFO estimation systems are shown to perform well on Russian SSB recordings, although they are not trained or designed for this language.

As the third contribution, a source separation architecture, which processes the observation either in a latent or the frequency domain, is adapted to noise reduction (NR) on recordings of high frequency (HF) transmissions. For this model different combinations of the encoder and the decoder are evaluated with respect to their impact on the model performance for different loss functions. These experiments in Section 7.5.4 illustrate the importance of an appropriate segmentation window size during encoding to achieve strong separation results. Additionally, the positive effects of a time domain loss function for frequency domain separation are emphasized. Further experiments in Section 7.5.5 show that these improvements can be transferred to the NR task. For the SSB transmission data a larger segmentation window size of around 64 ms leads to the best performing system over the common window size of 4 ms. During the last chapter it was shown that all three building blocks can be combined to a multi-stage system, which achieves strong NR results for all considered recordings of SSB-modulated speech transmissions.

9.2 Outlook

Each of the three tasks, SAD, CFO estimation and NR, which lead to an improved listening experience for HF recordings, offers further opportunities for research. For the SAD system a shift-invariant architecture which is trained to recognize both sidebands as speech could allow a direct application to the in-phase and quadrature (IQ) data without a prior segmentation of the frequency dimension. Additionally, this design allows the network to learn to recognize the modulation type, which would further improve the detection.

A similar adaptation could also improve the CFO estimation if the transmitted sideband is unknown to the listener. This could be achieved by adapting the the sub-band classifier to predict the CFO for both sidebands allowing for an increased automation of the demodulation.

As a final remark, the presented NN-based systems have shown strong performance for all three speech processing tasks on the real SSB recordings. Therefore, other common tasks for HF transmission, like speaker and language recognition could also benefit from specifically designed NNs.

A Appendix

A.1 Analysis of the delay correction on the recording

Even after the synchronization of the recordings with the transmitted signal, as discussed in Section 4.4, a small delay may remain between the signals. During most of this work only recordings with a delay correction using generalized cross-correlation with phase transform (GCC-PHAT) for the delay estimation as described in Section 4.5 are considered. To show that this delay correction leads to an improvement perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) values are shown for all data sets for both the corrected and non-corrected data. Both evaluation measures are calculated on the speech active segments of the audio sequences.

The STOI values clearly improve after the delay correction for all data sets. Although the change in PESQ is smaller, it is still an improvement. Note, that the highest possible delay correction is 16 ms. Therefore, most of the gains have to come from an improved synchronization and not from the introduction of zeros to the signal during correction. One can conclude that the delay correction improved the synchronization.

Table A.1: Comparison of the data before and after delay correction during post processing.

Data set	non-corrected		corrected	
	PESQ	STOI	PESQ	STOI
Training	1.51	0.51	1.57	0.63
Development	1.58	0.53	1.63	0.64
Evaluation	1.59	0.49	1.65	0.61
Evaluation with CFO	1.33	0.41	1.52	0.56
Russian	1.62	0.53	1.74	0.65
Russian with CFO	1.33	0.42	1.59	0.57

A.2 Comparison of real to simulated training data ratios

This section offers some additional comparisons of the segment recurrent neural network (SRNN)-based SAD models for real HF signals demodulation with a CFO Δ^{f}.

Table A.2: Results for models with different ratios of real and simulated data in the training set on the evaluation set of the real HF radio database with $|\Delta^{\mathrm{f}}| > 0\,\mathrm{Hz}$. All evaluation measures are given in % and the best result per evaluation measure and Δ^{f} is marked by a gray background.

Δ^{f}	Share		EER	PRC	REC	F1	DCF
	Real	Simu					
0	1/2	1/2	3.71	94.37	94.75	94.56	4.21
	1/4	3/4	4.50	86.06	93.74	89.74	5.43
	1/6	5/6	4.71	83.27	94.74	88.63	4.86
	1/8	7/8	4.52	84.63	94.74	89.40	4.77
	1/10	9/10	4.88	80.98	94.75	87.33	5.00
	1/12	11/12	4.66	84.12	94.30	88.92	5.13
	1/14	13/14	4.97	83.13	93.86	88.17	5.52
100	1/2	1/2	6.36	84.55	91.74	88.00	6.98
	1/4	3/4	6.54	74.67	92.97	82.82	6.75
	1/6	5/6	5.82	75.70	94.21	83.95	5.76
	1/8	7/8	6.68	66.80	94.36	78.22	6.42
	1/10	9/10	6.57	69.52	93.97	79.92	6.45
	1/12	11/12	5.88	75.00	94.17	83.50	5.85
	1/14	13/14	6.69	68.88	93.96	79.49	6.52
300	1/2	1/2	5.73	91.46	90.79	91.13	7.30
	1/4	3/4	5.34	88.15	91.85	89.96	6.69
	1/6	5/6	4.75	92.70	94.08	93.39	4.78
	1/8	7/8	4.86	89.87	93.77	91.78	5.17
	1/10	9/10	4.35	88.91	94.42	91.58	4.74
	1/12	11/12	4.09	87.34	94.89	90.96	4.48
	1/14	13/14	4.18	83.23	95.59	88.98	4.21
500	1/2	1/2	7.77	92.99	87.09	89.94	9.98
	1/4	3/4	6.25	86.11	90.97	88.47	7.44
	1/6	5/6	6.39	88.49	92.01	90.21	6.54
	1/8	7/8	6.23	84.91	92.16	88.39	6.63
	1/10	9/10	4.37	87.63	94.52	90.94	4.72
	1/12	11/12	4.98	82.99	94.05	88.17	5.35
	1/14	13/14	5.28	81.62	94.15	87.44	5.36
1000	1/2	1/2	14.78	86.50	48.96	62.53	38.65
	1/4	3/4	8.99	85.69	79.90	82.70	15.72
	1/6	5/6	7.55	87.49	86.95	87.22	10.39
	1/8	7/8	8.84	77.64	88.20	82.58	10.09
	1/10	9/10	4.98	84.84	93.08	88.77	6.00
	1/12	11/12	5.16	80.69	93.20	86.54	6.24
	1/14	13/14	5.65	80.93	93.32	86.69	6.08

In Table A.2 models trained on varying ratios of real and simulated data are compared. Starting from a 1:9 ratio of real and simulated data the models achieve very similar results. Although some of the models with a lower share of simulated data outperform the model trained on 90 % simulated data for some Δ^{f}, none achieve a similar consistency over all considered CFOs. Therefore, the 1:9 model is used during this work.

A.3 Evaluation on Russian data with a CFO

The SRNN-based SAD models described in Chapter 5 are trained either on real HF recordings without a CFO or simulated signals with a CFO. In this section a model trained with a 1:9 ratio of real and simulated signal is compared with a model only trained on the real recordings for the Russian evaluation set with different CFOs. Both models are only trained on signals with English speakers. The results are displayed in Figure A.1.

Similar to the experiments discussed in Section 5.5.3 the simulated training data improves the SAD results for CFOs above 100 Hz. This indicates some robustness against speech signals of a previously unseen language similar to Section 5.5.1.

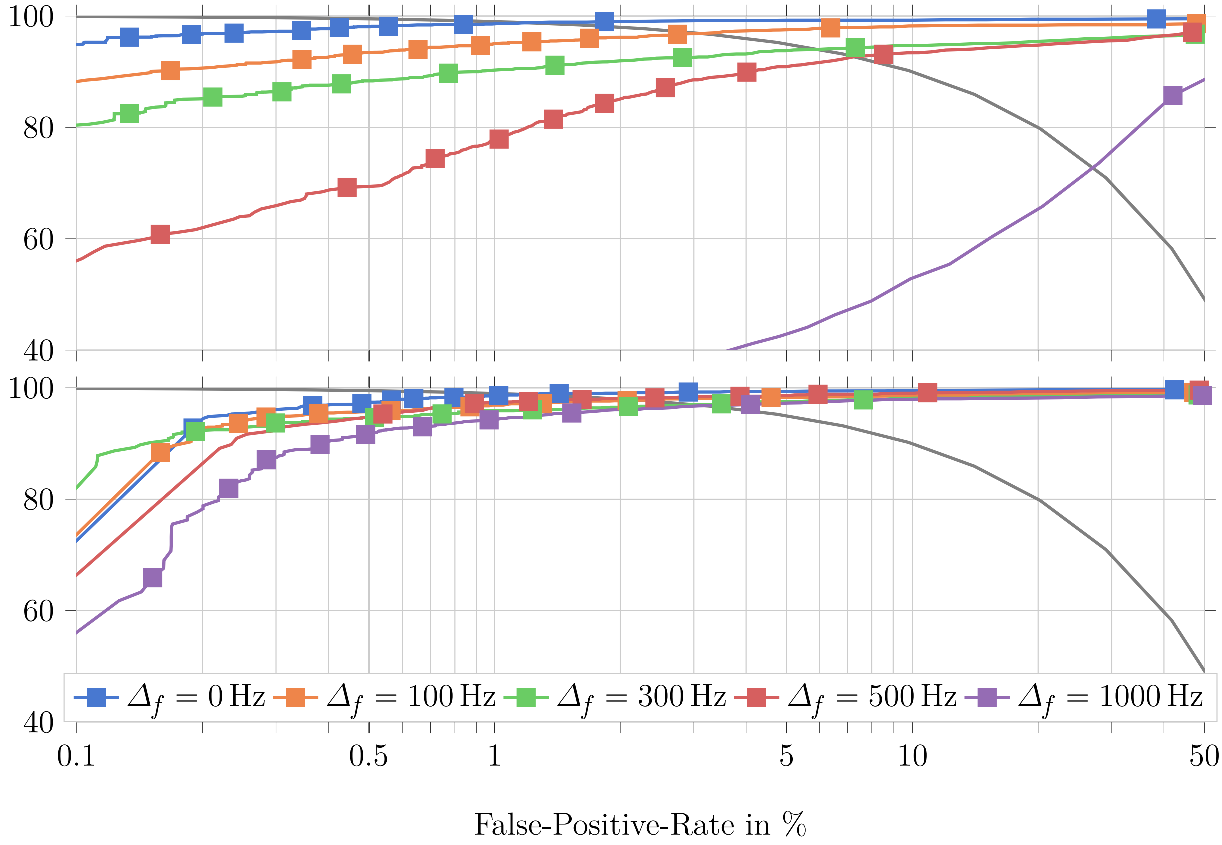


Figure A.1: ROC curve for the SRNN model trained on real English speech recordings with $\Delta^{\mathrm{f}} = 0$ Hz (upper) and the model trained on both simulated and real data (lower) for Russian speech recordings with different CFOs. The gray line symbolizes the EER.

A.4 CFO estimation on overlapping speakers

Table A.3: Comparison of the full-band classifier for CFO estimation on the test set of the WSJ0-2mix database with simulated CFO regarding their error class affiliation. All values are given in %.

# speakers	System	<5 Hz	5-10 Hz	10-50 Hz	>50 Hz
1	NN-based (RNN)	57.18	39.85	2.97	0.00
	NN-based (CNN)	57.92	40.71	1.37	0.00
2	NN-based (RNN)	64.620536	34.93	0.45	0.00
	NN-based (CNN)	64.843750	32.59	1.34	1.23

All evaluations of the CFO estimation in this work are performed on signal with one speaker. However, if a signal with two overlapping speakers is transmitted over a HF channel and demodulated with a CFO it is interesting to know whether the CFO estimation can deal with such a signal. Therefore, the signal with overlapping speaker from the test set of the WSJ0-2mix database described in Section 7.5.1 are shifted as described in Section 4.9 to simulate a CFO. Afterwards, the signals are processed by the NN-based estimator described in Section 6.2. The error class affiliation of the estimations are displayed in Table A.3. For both considered architectures, the CFO estimation on the overlapping speech signal is comparable to the estimation on signals with a single speaker.

A.5 Enhancement with oracle masks

In Section 7.5.3 some topline results for blind source separation (BSS) and NR are presented, but not all evaluation measures are shown. In this section the missing results are presented. The Table A.4 and Table A.5 display PESQ, STOI, signal to distortion ratio (SDR), scale invariant signal to distortion ratio (SI-SDR) and word error rate (WER) values for all considered oracle masks for the two tasks.

For the BSS task the observations discussed in Section 7.5.3 are corroborated by the additional evaluation measures. Interestingly, the WER of the ideal ratio mask (IRM) is more than 2 % lower than the one achieved with the Wiener-like mask (WLM) mask, despite the improved

Table A.4: Comparison of the oracle masks for source separation on the test data of the WSJ0-2mix database

Name	Equation	PESQ	STOI	SI-SDR	SDR	WER
			%	dB	dB	%
Ideal ratio mask (IRM)	$\frac{\lvert X_{k,\ell,f}\rvert}{\lvert Y_{\ell,f}\rvert}$	3.51	0.95	11.7	12.4	10.26
Wiener-like mask (WLM)	$\frac{\lvert x_{k,\ell,f}\rvert^2}{\lvert Y_{\ell,f}\rvert^2}$	3.25	0.94	12.4	13.2	12.84
Ideal complex mask (ICM)	$\frac{X_{k,\ell,f}}{Y_{\ell,f}}$	4.55	1.00	155.1	155.0	9.78

Table A.5: Comparison of the oracle masks for noise reduction and on the simulated HF evaluation data

Name	Equation	PESQ	STOI %	SI-SDR dB	SDR dB	WER %
Ideal ratio mask (IRM)	$\frac{\lvert X_{k,\ell,f}\rvert}{\lvert Y_{\ell,f}\rvert}$	3.67	0.97	12.9	13.3	14.94
Wiener-like mask (WLM)	$\frac{\lvert X_{k,\ell,f}\rvert^2}{\lvert Y_{\ell,f}\rvert^2}$	3.60	0.96	13.8	14.1	16.38
Ideal complex mask (ICM)	$\frac{X_{k,\ell,f}}{Y_{\ell,f}}$	4.55	1.00	74.8	74.9	14.41

SI-SDR and SDR. This indicates that even slight improvements in STOI and PESQ are more correlated to the WER than the two power-based evaluation measures.

For the NR tasks, the results are similar to the BSS task. A small gain in PESQ and STOI for the IRM compared to the WLM leads to a 1.5 % lower WER, despite the lower SDR values for the IRM.

A.6 Small encoder and decoder windows for NR on real HF signals

In Section 7.5.5 it is shown, that the NR on real HF data profits from a higher window size L_W and shift L_S in the encoder and decoder. Therefore, all further experiments are performed using these larger values. To show that the higher $L_\mathrm{W} = 64\,\mathrm{ms}$ and $L_\mathrm{S} = 16\,\mathrm{ms}$ are beneficial on the real recordings independent of the architecture, the experiments in Table 7.9 are repeated with $L_\mathrm{W} = 2\,\mathrm{ms}$ and $L_\mathrm{S} = 1\,\mathrm{ms}$. These experiments are performed for a trainable as well as the STFT-based encoder and decoder. Additionally, the recurrent and convolutional mask estimator described in Section 7.3 are compared. The results are presented in Table A.6.

Table A.6: Comparison of different encoder and decoder combinations using $L_\mathrm{W} = 2\,\mathrm{ms}$ and $L_\mathrm{S} = 1\,\mathrm{ms}$ on the evaluation set of the real HF database.

Mask estimator	Encoder	Decoder	PESQ	STOI %
Convolutional	learned	learned	2.27	69.3
	short time Fourier transform (STFT)	learned	2.22	70.7
	learned	inverse STFT	2.32	71.4
	STFT	inverse STFT	2.17	69.4
Recurrent	learned	learned	2.26	69.9
	STFT	learned	2.28	69.8
	learned	inverse STFT	2.23	70.0
	STFT	inverse STFT	2.18	69.3

The presented results are worse for both mask estimators and all encoder and decoder combinations. Therefore, it can be concluded that the larger window size and shift a in general beneficiary for the real HF recordings.

A.7 Evaluation of noise reduction on signals with a CFO

To evaluate the dependence of a NR system on a prior CFO correction the two described NR architectures are applied to signals with a CFO. Therefore, the simulation scheme for both training and evaluation is changed. So far all models are both trained and evaluated on simulated data without a CFO. For the following experiment the models trained on signals without a CFO are compared to those trained on signals with a CFO. Both models are evaluated on input signals with and without prior CFO correction. For the models trained with a CFO the same offset simulated on the input signals is added to the target signals during training by following the steps in Section 4.9. Adding the CFO to the target signal forces the network to reduce the noise without addressing the CFO.

The comparison of the models with and without CFO correction is performed for both simulated and real HF evaluation data. To quantify the enhancement results the following evaluation measures as discussed in Section 7.5.2 are used: PESQ, STOI, SI-SDR, SDR and WER for the simulated signals and only PESQ and STOI for the real recordings.

All reference signals for the performance measures are adjusted during evaluation by simulating the CFO of the observation and then correcting this CFO with oracle information to ensure that only the NR performance is judged. Otherwise, the evaluation measures might decrease due to missing signal power in higher frequencies in case of CFOs greater than the difference between the Nyquist frequency $F_{\text{max}} = 4\,\text{kHz}$ and the bandwidth, here $4\,\text{kHz} - 2.7\,\text{kHz} = 1.3\,\text{kHz}$. The power in these frequencies is lost due to the transmission and not the enhancement. Therefore, these frequencies should not be considered during evaluation of the NR.

Simulated signals

In a first step to examine the influence of a CFO on the NR, the models trained without a CFO are evaluated on the simulated HF data without CFO correction. Here, the CFO is corrected after the NR is performed to ensure comparability with the results presented in Section 7.5.5. The calculated evaluation measures are displayed in Table A.7 and compared to the results on the evaluation data with perfect CFO correction presented in Section 7.5.5.

As expected all evaluation measures indicate a worse performance for signals with a CFO if the models are trained on simulated audio without a CFO. Training the models on signals with a CFO reduces this difference for all evaluation measures except the WER. The WER suffers because of missing frequencies above the Nyquest frequency F_{max} as described above.

Table A.7: Comparison of NR systems with and without prior CFO correction during training and evaluation on the simulated HF data.

Mask estimator	CFO Train	CFO Eval	PESQ	STOI %	SI-SDR dB	SDR dB	WER %
Convolutional	✗	✗	2.70	90.1	13.8	14.8	21.42
	✗	✓	1.55	53.5	–3.4	0.9	82.30
	✓	✗	2.34	87.1	12.6	13.5	25.41
	✓	✓	2.22	86.1	12.2	13.2	60.30
Recurrent	✗	✗	2.62	88.9	13.3	14.4	23.25
	✗	✓	1.67	67.5	2.6	5.2	74.52
	✓	✗	2.39	86.2	12.2	13.2	25.53
	✓	✓	2.19	85.4	11.8	12.8	58.24

This effect is not as prominent for the other evaluation measures because of the discussed adjustment of the target signals.

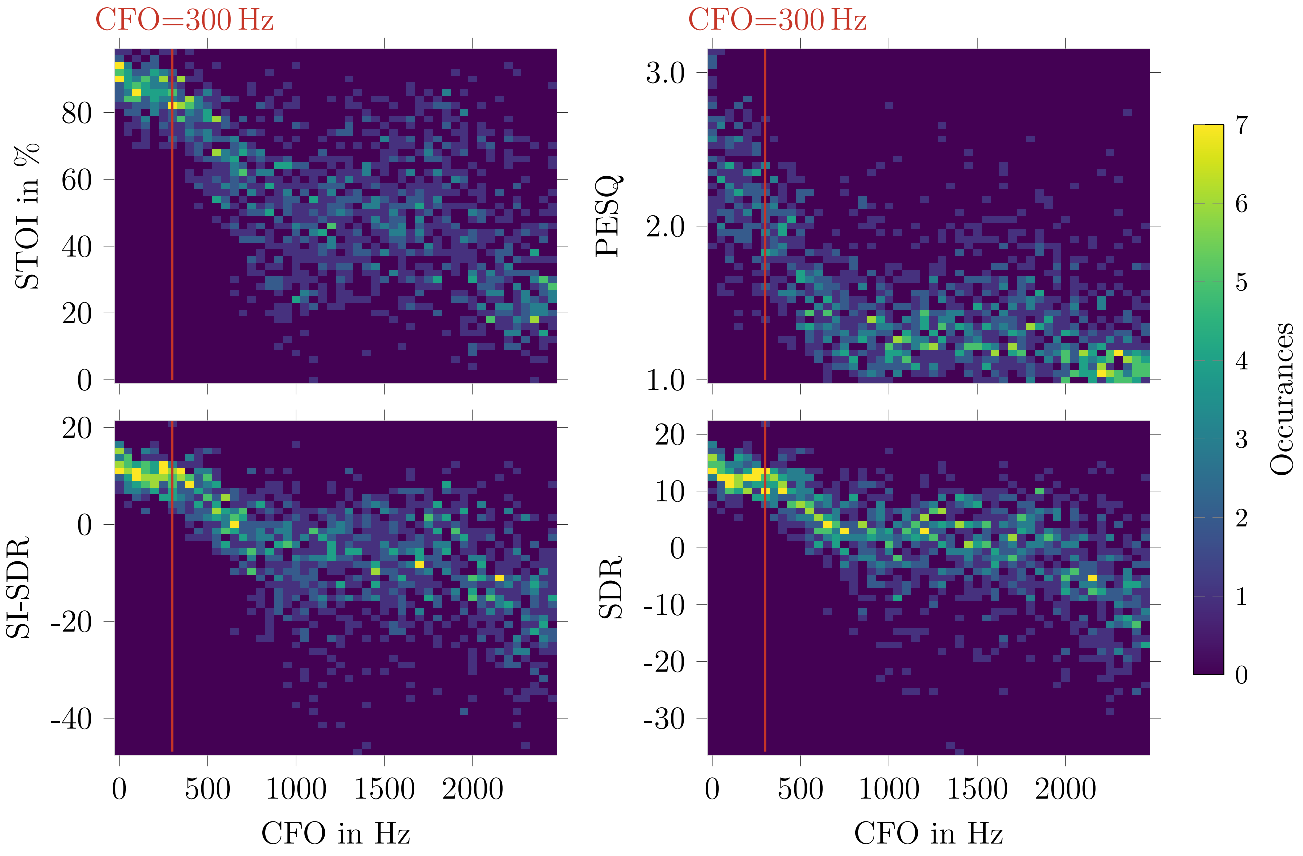


Figure A.2: Heatmaps of PESQ, SDR, SI-SDR and STOI values over the CFO after NR with the convolutional mask estimator trained on signals without CFO for the simulated HF data.

Note, that the recurrent estimator is less affected by the CFO if it is not trained on signals with a CFO, and the difference between the estimators is smaller if the models are trained on signals with a CFO. This indicates that the recurrent mask estimator is more robust to CFOs. However, this observation does not transfer to the real HF signals, as shown below.

To emphasize the negative influence of the CFO on NR a scatter plots of the evaluation measures is plotted over the CFO for the convolutional mask estimator with learned encoder and decoder. The resulting graphs are displayed in Figure A.2. From these heatmaps it appears that the CFO does affect the NR, but it does not lead to a distorted output signal for every offset. While the evaluation measures for signals with an CFO below 300 Hz are in the range expected from a successful NR, higher CFOs mostly lead to highly deteriorated results. Therefore, a CFO does not prevent the NR system from improving the signal, but for some examples causes the NR network to distort the input signal during enhancement. The amount of examples that are distorted by the NR network increases with the CFOs. For models trained on signals with a CFO the NR performance is largely independent of the CFO as shown in Figure A.3. Note, that further experiments in Appendix A.7 and Section 8.1 show that the NR performance in case of a CFO correction is stronger if the network was not trained on signals with a CFO.

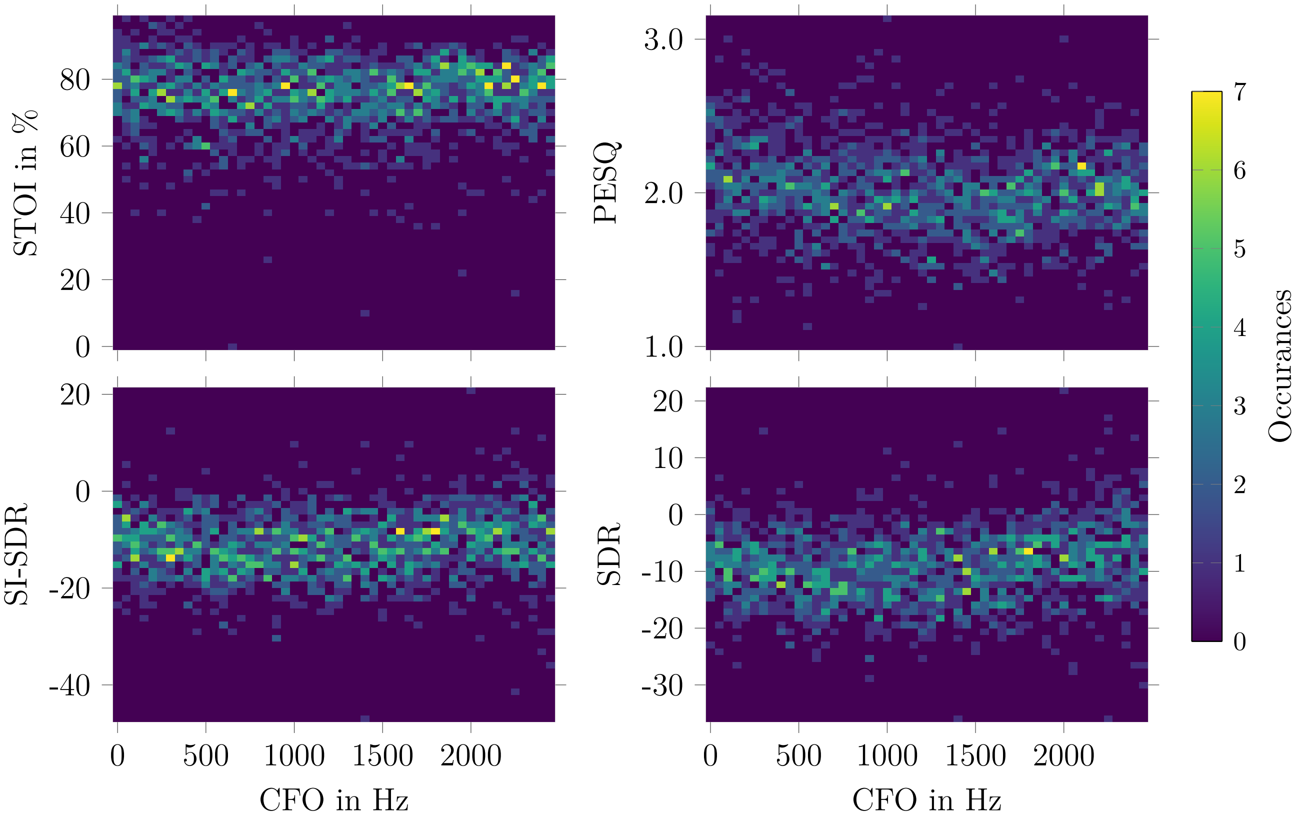


Figure A.3: Heatmaps of PESQ, SDR, SI-SDR and STOI values over the CFO for NR with the convolutional mask estimator trained on signals with a CFO for the simulated HF data.

These experiments have shown that the NR with subsequent CFO correction achieves similar results to NR on signals without a CFO if the training data is adjusted accordingly. The following experiments are examining the NR on real data with and without a CFO.

Real recordings

Here, three kind of evaluation recordings are considered: without a CFO, with a CFO and with a CFO that was corrected before NR is applied. For the signals without a CFO the real HF evaluation set with $\varDelta^{\mathrm{f}} = 0$ is chosen. The remaining two evaluations use the real HF evaluation set with $\varDelta^{\mathrm{f}} \geq 0$, where the CFO is corrected with oracle information for the experiment with a corrected CFO. The resulting evaluation measures are presented in Table A.8

Most of the findings from the experiments on the simulated data can be transferred to the real recordings. Training on simulated signals with a CFO improves the results on recordings with a CFO for both mask estimators. However, if the CFO is corrected using oracle information, the NR trained on signals without a CFO outperforms the network trained on shifted data for both the PESQ and STOI evaluation measures. Surprisingly, the network trained on signals with a CFO achieves a higher STOI value if the CFO is corrected before the network is applied, but a slightly higher PESQ value if the NR is performed first. These differences might be due to the noise energy in the frequencies above 2.7 kHz, which are removed by the CFO correction. Removing these frequencies might deprive the network of contextual information that while slightly improving the signal quality, decreases the intelligibility.

For the real signals, the recurrent mask estimator no longer outperforms the convolutional one for signals with a CFO. Here, the gains of the convolutional network on signals without a CFO can be transferred to the recordings with a CFO both with and without a prior CFO correction.

Table A.8: Comparison of NR systems with and without a prior CFO correction during training on different evaluation data on the real evaluation sets of the HF data. Here, “without CFO” refers to the evaluation set with $\varDelta^{\mathrm{f}} = 0$, “with CFO” represents the evaluation set with $\varDelta^{\mathrm{f}} \geq 0$ without prior CFO correction and “Corrected CFO” refers to the evaluation set with $\varDelta^{\mathrm{f}} \geq 0$, where the CFO is corrected before the NR is applied.

Mask estimator	CFO in	without CFO		with CFO		corrected CFO	
	Train	PESQ	STOI / %	PESQ	STOI / %	PESQ	STOI / %
Convolutional	✗	2.55	78.4	1.84	57.2	2.55	78.4
	✓	2.08	72.0	2.40	70.1	2.21	73.4
Recurrent	✗	2.30	72.3	1.69	52.6	2.38	77.8
	✓	2.14	71.8	2.29	68.6	2.26	73.1

Acronyms

ADC analog-to-digital converter.

AGC automatic gain control.

ASR automatic speech recognition.

BCE binary cross entropy.

BSS blind source separation.

CDF cumulative distribution function.

CFO carrier frequency offset.

CNN convolutional neural network.

Conv1D-block one-dimensional convolution block.

DAC digital-to-analog converter.

DCF decision cost function.

DFT discrete Fourier transformation.

DNN deep neural network.

DPRNN dual-path recurrent neural network.

DSB double sideband.

ECA error class affiliation.

EER equal error rate.

EM expectation maximization.

FC fully-connected.

FPR false positive rate.

GCC-PHAT generalized cross-correlation with phase transform.

GMM Gaussian mixture model.

GRU gated recurrent unit.

HF high frequency.

HMM hidden Markov model.

ICM ideal complex mask.

IF intermediate frequency.

IQ in-phase and quadrature.

IRM ideal ratio mask.

ITU International Telecommunication Union.

LPC linear predictive coding.

LSB lower sideband.

LSTM long short-term memory.

MFCC Mel frequency cepstral coefficients.

MS minimum statistics.

MSE mean squared error.

NN neural network.

NR noise reduction.

PESQ perceptual evaluation of speech quality.

PRC precision.

PReLU parametric rectified linear unit.

PSD power spectral density.

RATS Robust Automatic Transcription of Speech.

REC recall.

ReLU rectified linear unit.

RNN recurrent neural network.

ROC receiver operating characteristic.

RTF real time factor.

SAD speech activity detection.

SBL sub-band layer.

SDR signal to distortion ratio.

SE speech enhancement.

SI-SDR scale invariant signal to distortion ratio.

SID speaker identity detection.

SNR signal to noise ratio.

SPP speech presence probability.

SRNN segment recurrent neural network.

SSB single-sideband.

STFT short time Fourier transform.

STOI short-time objective intelligibility.

TasNet time-domain audio separation network.

TDNN-F factorized time-delayed neural network.

TPR true positive rate.

u-PIT utterance-level permutation invariant training.

USB upper sideband.

WER word error rate.

WLM Wiener-like mask.

WSJ Wall Street Journal.

Bibliography

List of peer-reviewed publications with own contributions (OC)

[OC1] J. Heitkaemper, J. Heymann, and R. Haeb-Umbach, “Smoothing along frequency in online neural network supported acoustic beamforming,” in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[OC2] J. Heitkaemper, T. Fehér, M. Freitag, and R. Haeb-Umbach, “A study on online source extraction in the presence of changing speaker positions,” in *Proceedings of the 7th International Conference on Statistical Language and Speech Processing (SLSP)*, C. Martín-Vide, M. Purver, and S. Pollak, Eds., ser. Lecture Notes in Computer Science, vol. 11816, Springer, 2019, pp. 198–209, ISBN: 978-3-030-31372-2.

[OC3] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, “Demystifying TasNet: A dissecting approach,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6359–6363 (cit. on pp. 11, 18, 20, 21, 66, 77, 82–84, 89, 95).

[OC4] J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, “Statistical and neural network based speech activity detection in non-stationary acoustic environments,” in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2597–2601 (cit. on pp. 9, 11, 13, 14, 36, 38, 41–43).

[OC5] J. Heitkaemper, J. Schmalenstroeer, J. Ullmann, V. Ion, and R. Haeb-Umbach, “A database for research on detection and enhancement of speech transmitted over hf links,” in *Speech Communication; 14th ITG Conference*, 2021 (cit. on pp. 9, 22, 30, 36, 41, 42).

[OC6] J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, “Neural network based carrier frequency offset estimation from speech transmitted over high frequency channels,” in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, 2022 (cit. on pp. 15, 57–59, 64).

[OC7] A. Chinaev, J. Heitkaemper, and R. Haeb-Umbach, “A priori snr estimation using weibull mixture model,” in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5 (cit. on pp. 20, 87).

[OC8] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 35–40 (cit. on p. 17).

[OC9] M. Kitza, W. Michel, C. Boeddeker, J. Heitkaemper, T. Menne, R. Schlüter, H. Ney, J. Schmalenstroeer, L. Drude, J. Heymann, et al., "The RWTH/UPB system combination for the CHiME 2018 workshop," in *Proceedings of the 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, 2018, pp. 53–57.

[OC10] J. Ebbers, J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, "Benchmarking neural network architectures for acoustic sensor networks," in *Speech Communication; 13th ITG-Symposium*, VDE, 2018, pp. 1–5.

[OC11] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1248–1252.

[OC12] J. M. Martín-Doñas, J. Heitkaemper, R. Haeb-Umbach, A. M. Gomez, and A. M. Peinado, "Multi-channel block-online source extraction based on utterance adaptation," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 96–100.

[OC13] C. Boeddeker, T. Cord-Landwehr, J. Heitkaemper, C. Zorila, D. Hayakawa, M. Li, M. Liu, R. Doddipatla, and R. Haeb-Umbach, "Towards a speaker diarization system for the chime 2020 dinner party transcription," in *Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 42–47.

[OC14] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The pytorch-based audio source separation toolkit for researchers," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2637–2641.

[OC15] J. Schmalenstroeer, J. Heitkaemper, J. Ullmann, and R. Haeb-Umbach, "Open range pitch tracking for carrier frequency difference estimation from hf transmitted speech," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1–5 (cit. on pp. 15, 28, 54, 56, 69, 71).

[OC16] T. Gburrek, J. Schmalenstroeer, J. Heitkaemper, and R. Haeb-Umbach, "Informed vs. blind beamforming in ad-hoc acoustic sensor networks for meeting transcription," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

References

[1] Y. Chen, B. Dong, X. Zhang, P. Gao, and S. Li, “A hybrid deep-learning approach for single channel HF-SSB speech enhancement,” *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2165–2169, 2021. DOI: 10.1109/LWC.2021.3095383 (cit. on pp. 1, 20).

[2] H. Xing and J. H. L. Hansen, “Single sideband frequency offset estimation and correction for quality enhancement and speaker recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 124–136, 2017. DOI: 10.1109/TASLP.2016.2623563 (cit. on pp. 1, 3, 9, 14, 15).

[3] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, “Toward human parity in conversational speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017. DOI: 10.1109/TASLP.2017.2756440 (cit. on pp. 1, 11).

[4] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021. DOI: 10.1109/TASLP.2021.3099291 (cit. on pp. 1, 18, 83, 92).

[5] Z. Chen, S. Wang, and Y. Qian, “Multi-modality matters: A performance leap on VoxCeleb,” in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2252–2256. DOI: 10.21437/Interspeech.2020-2229 (cit. on p. 1).

[6] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, “Speech processing for digital home assistants: Combining signal processing with deep-learning techniques,” *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019. DOI: 10.1109/MSP.2019.2918706 (cit. on p. 1).

[7] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, “Service chatbots: A systematic review,” *Expert Systems with Applications*, vol. 184, p. 115 461, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115461 (cit. on p. 1).

[8] T. Seymour and A. Shaheen, “History of wireless communication,” *Review of Business Information Systems (RBIS)*, vol. 15, no. 2, pp. 37–42, 2011. DOI: 10.19030/rbis.v15i2.4202 (cit. on p. 3).

[9] A. F. Molisch, Ed., *Wireless communications*, 2nd ed. Wiley - IEEE, 2011 (cit. on p. 3).

[10] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 3).

[11] T. S. Rappaport, Ed., *Wireless Communications: Principles and Practice, 2nd Edition*, 2nd ed., ser. Communications Engineering and Emerging Technologies Series. Prentice Hall, Dec. 2001, vol. 2, ISBN: 0130422320 (cit. on pp. 3–5, 32).

[12] J. Barry, E. Lee, and D. Messerschmitt, *Digital Communication*, 3rd ed. Springer US, 2012 (cit. on p. 3).

[13] R. Gibson and R. Wells, "The potential of SSB for land mobile radio," in *29th IEEE Vehicular Technology Conference*, vol. 29, 1979, pp. 90–94. DOI: 10.1109/VTC.1979.1622670 (cit. on p. 3).

[14] J. Reed, *Software Radio: A Modern Approach to Radio Engineering*, ser. Prentice Hall Communications E. Prentice Hall, 2002, ISBN: 9780130811585 (cit. on pp. 3, 7, 8).

[15] A. Wyglinski, R. Getz, T. Collins, and D. Pu, *Software-Defined Radio for Engineers*, ser. Artech House mobile communications series. Artech House, 2018, ISBN: 9781630814595 (cit. on pp. 3, 6, 7).

[16] J. R. Carson, "Method and means for signaling with high-frequency waves," 1 449 382, 1915 (cit. on p. 4).

[17] Collins Radio Company, *Fundamentals of Single Side Band*, 2nd, Collins Radio Company, Ed. Cedar Rapids: Collins Radio Co, 1959 (cit. on pp. 4, 6, 7, 9).

[18] United States. Department of the Army and United States. Department of the Air Force, *Fundamentals of Single-sideband Communication*, ser. Air Force TO. 1961 (cit. on pp. 4, 9).

[19] International Telecommunication Union, *Rec.ITU-R BS.640-1 single sideband (SSB) system for HF broadcasting*, Okt 1990 (cit. on pp. 4, 22).

[20] ——, *Recommendation ITU-R M.1173-1: Technical characteristics of single-sideband transmitters used in the maritime mobile service for radiotelephony in the bands between 1 606.5 kHz (1 605 kHz Region 2) and 4 000 kHz and between 4 000 kHz and 27 500 kHz*, Mar. 2012 (cit. on p. 4).

[21] A. B. Carlson and P. B. Crilly, *Communication Systems: An Introduction to Signals and Noise in Electrical Communication*, 5. McGraw-Hill, 2010, ISBN: 978–0–07–338040–7 (cit. on pp. 4–8).

[22] E. Bedrosian, "The analytic signal representation of modulated waveforms," *Proceedings of the IRE*, vol. 50, no. 10, pp. 2071–2076, 1962. DOI: 10.1109/JRPROC.1962.288236 (cit. on p. 4).

[23] F. W. King, *Hilbert Transforms*, ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009, vol. 1. DOI: 10.1017/CBO9780511721458 (cit. on p. 4).

[24] D. K. Weaver, "A third method of generation and detection of single-sideband signals," *Proceedings of the IRE*, vol. 44, no. 12, pp. 1703–1705, 1956. DOI: 10.1109/JRPROC.1956.275061 (cit. on pp. 5, 30).

[25] International Telecommunication Union, *Recommendation ITU-R V.431-8 Nomenclature of the frequency and wavelengh bands used in telecommunications*, Aug. 2015 (cit. on p. 5).

[26] P. de Fornel and H. Sizun, *Radio Wave Propagation for Telecommunication Applications*, ser. Signals and Communication Technology. Springer Berlin Heidelberg, 2006, ISBN: 9783540266686 (cit. on pp. 5, 6).

[27] S. F. Mahmoud and Y. M. M. Antar, "High frequency ground wave propagation," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 11, pp. 5841–5846, 2014. DOI: 10.1109/TAP.2014.2346211 (cit. on p. 5).

[28] C. Johnson and W. Sethares, *Telecommunication Breakdown: Concepts of Communication Transmitted Via Software-defined Radio*, ser. Telecommunication Breakdown: Concepts of Communication Transmitted Via Software-defined Radio Bd. 1. Pearson Education Incorporated, 2004, ISBN: 9780131430471 (cit. on pp. 6, 8).

[29] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2017, ISSN: 2329-9290 (cit. on pp. 6, 16, 91).

[30] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2012)*, 2012, pp. 291–297 (cit. on pp. 9, 14).

[31] J. Hansen, A. Sangwan, A. Joglekar, A. Bulut, L. Kaushik, and C. yu, "Fearless Steps: Apollo-11 corpus advancements for speech technologies from Earth to the Moon," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 2758–2762. DOI: 10.21437/Interspeech.2018-1942 (cit. on pp. 9, 11, 14).

[32] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," *Proceedings of the 13th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Jan. 2012 (cit. on pp. 9, 13, 14).

[33] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3497–3501, 2013 (cit. on pp. 9, 12–14).

[34] Q. Lin, T. Li, and M. Li, "The DKU speech activity detection and speaker identification systems for Fearless Steps Challenge Phase-02," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2607–2611. DOI: 10.21437/Interspeech.2020-1915 (cit. on p. 9).

[35] O. Plchot, M. Diez, M. Soufifar, and L. Burget, "PLLR features in language recognition system for RATS," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 3047–3051. DOI: 10.21437/Interspeech.2014-611 (cit. on p. 9).

[36] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, "Medium-duration modulation cepstral feature for robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1749–1753. DOI: 10.1109/ICASSP.2014.6853898 (cit. on p. 9).

[37] A. Gorin, D. Kulko, S. Grima, and A. Glasman, “This is Houston. Say again, please. The Behavox system for the Apollo-11 Fearless Steps Challenge (Phase II),” in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2612–2616. DOI: 10.21437/Interspeech.2020-2822 (cit. on p. 9).

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2016, ISBN: 9780262337373 (cit. on pp. 11, 37, 48, 67).

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, Oct. 1986 (cit. on pp. 11, 12).

[40] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, ISSN: 0899-7667. DOI: 10.1162/neco.2006.18.7.1527 (cit. on p. 11).

[41] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019. DOI: 10.1109/ACCESS.2019.2896880 (cit. on p. 11).

[42] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “ICASSP 2021 deep noise suppression challenge,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6623–6627. DOI: 10.1109/ICASSP39728.2021.9415105 (cit. on p. 11).

[43] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, “The third DIHARD diarization challenge,” in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 3570–3574. DOI: 10.21437/Interspeech.2021-1208 (cit. on p. 11).

[44] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50. DOI: 10.1109/ICASSP40776.2020.9054266 (cit. on pp. 11, 18, 39, 79, 80, 89, 91).

[45] M. Alam, M. Samad, L. Vidyaratne, A. Glandon, and K. Iftekharuddin, “Survey on deep neural networks in speech and vision systems,” *Neurocomputing*, vol. 417, pp. 302–321, 2020, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.07.053 (cit. on pp. 11, 12).

[46] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” in *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125 (cit. on pp. 11, 59).

[47] J. S. Chung, A. Nagrani, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, “Voxsrc 2019: The first VoxCeleb speaker recognition challenge,” *ISCA Challenges*, 2019 (cit. on p. 11).

[48] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech &Language*, vol. 46, pp. 535–557, 2017, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2016.11.005 (cit. on pp. 11, 20).

[49] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines,” in *Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 1561–1565. DOI: 10.21437/Interspeech.2018-1768 (cit. on pp. 11, 17).

[50] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013 (cit. on p. 11).

[51] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, “First DIHARD challenge evaluation plan,” *Zenodo*, Mar. 2018. DOI: 10.5281/zenodo.1199638 (cit. on p. 11).

[52] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014. DOI: 10.1109/TASLP.2014.2304637 (cit. on p. 11).

[53] 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,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, Dec. 2016 (cit. on p. 11).

[54] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. DOI: 10.1162/neco.1989.1.4.541 (cit. on p. 12).

[55] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, A. Moschitti, B. Pang, and W. Daelemans, Eds., ACL, Okt 2014, pp. 1724–1734. DOI: 10.3115/v1/d14-1179 (cit. on pp. 12, 38).

[56] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 12).

[57] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, ser. Springer Handbooks. Springer, 2008, ISBN: 978-3-540-49127-9 (cit. on pp. 12, 16, 54, 55, 63).

[58] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001, ISSN: 1063-6676 (cit. on p. 12).

[59] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012 (cit. on p. 12).

[60] J. Ramirez, J. M. Górriz, and J. C. Segura, “Voice activity detection. fundamentals and speech recognition system robustness,” in, K. K. Michael Grimm, Ed. BoD – Books on Demand, 2007, ch. 1, pp. 1–22. DOI: 10.5772/4740 (cit. on p. 12).

[61] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999, ISSN: 1558-2361. DOI: 10.1109/97.736233 (cit. on pp. 12, 13, 28).

[62] S. Graf, T. Herbig, M. Buck, and G. Schmidt, “Features for voice activity detection: A comparative analysis,” *EURASIP Journal on Advances in Signal Processing*, Nov. 2015 (cit. on pp. 12, 13).

[63] A. Ziaei, L. Kaushik, A. Sangwan, J. Hansen, and D. Oard, “Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions,” in *Proceedings of the 15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014 (cit. on pp. 12, 13).

[64] J.-H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006 (cit. on p. 12).

[65] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. L. Hansen, A. Janin, B.-s. Lee, Y. Lei, V. Mitra, N. Morgan, S. O. Sadjadi, T. Tsai, N. Scheffer, L. N. Tan, and B. Williams, “All for one: Feature combination for highly channel-degraded speech activity detection,” in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 709–713 (cit. on p. 12).

[66] B. Liu, Z. Wang, S. Guo, H. Yu, Y. Gong, J. Yang, and L. Shi, “An energy-efficient voice activity detector using deep neural networks and approximate computing,” *Microelectronics Journal*, vol. 87, pp. 12–21, 2019, ISSN: 0026-2692. DOI: https://doi.org/10.1016/j.mejo.2019.03.009 (cit. on pp. 12, 13).

[67] T. Hughes and K. Mierle, “Recurrent neural networks for voice activity detection,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382. DOI: 10.1109/ICASSP.2013.6639096 (cit. on p. 13).

[68] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, “Two-dimensional convolutional recurrent neural networks for speech activity detection,” in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2045–2049. DOI: 10.21437/Interspeech.2019-1354 (cit. on pp. 13, 38).

[69] D. Rosenbaum, *Human Motor Control*. Elsevier Science, 2014, ISBN: 9780080571089 (cit. on p. 13).

[70] L. Mateju., P. Cerva., and J. Zdansky., *Study on the Use of Deep Neural Networks for Speech Activity Detection in Broadcast Recordings*. SciTePress, 2016, pp. 45–51, ISBN: 978-989-758-196-0. DOI: 10.5220/0005952700450051 (cit. on p. 13).

[71] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5695–5699. DOI: 10.1109/ICASSP.2016.7472768 (cit. on p. 13).

[72] J. Ebbers, L. Drude, A. Brendel, W. Kellermann, and R. Haeb-Umbach, "Weakly supervised sound activity detection and event classification in acoustic sensor networks," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Guadeloupe, West Indies, 2019 (cit. on p. 13).

[73] S. [Katagiri, *Handbook of neural networks for speech processing*, 2000 (cit. on p. 13).

[74] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016 (cit. on p. 13).

[75] R. Zazo, T. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2016, pp. 3668–3672 (cit. on p. 13).

[76] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Apr. 2013 (cit. on p. 13).

[77] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2014, pp. 890–894 (cit. on p. 13).

[78] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 133–137, 2009. DOI: 10.1109/TCSVT.2008.2009262 (cit. on p. 13).

[79] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 433–439. DOI: 10.21437/Odyssey.2020-62 (cit. on p. 13).

[80] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "Fearless Steps Challenge (FS-2): Supervised learning with massive naturalistic apollo data," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2617–2621. DOI: 10.21437/Interspeech.2020-3054 (cit. on pp. 14, 36, 42).

[81] T. Vuong, Y. Xia, and R. M. Stern, “The application of learnable STRF kernels to the 2021 Fearless Steps Phase-03 SAD challenge,” in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 4364–4368. DOI: 10.21437/Interspeech.2021-651 (cit. on p. 14).

[82] A. Joglekar, S. O. Sadjadi, M. Chandra-Shekar, C. Cieri, and J. H. Hansen, “Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for unseen channel and mission data across NASA Apollo audio,” in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 986–990. DOI: 10.21437/Interspeech.2021-2011 (cit. on p. 14).

[83] D. Cole, S. Sridharan, and M. Moody, “Frequency offset correction for hf radio speech reception,” *IEEE Transactions on Industrial Electronics*, vol. 47, no. 2, pp. 438–443, 2000. DOI: 10.1109/41.836360 (cit. on pp. 14, 15).

[84] P. F. Assmann, S. Dembling, and T. M. Nearey, “Effects of frequency shifts on perceived naturalness and gender information in speech,” in *Proceedings of the 7th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2006, paper 1710–Tue1BuP.10. DOI: 10.21437/Interspeech.2006-297 (cit. on p. 14).

[85] P. Clark, S. H. Mallidi, A. Jansen, and H. Hermansky, “Frequency offset correction in speech without detecting pitch,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7020–7024. DOI: 10.1109/ICASSP.2013.6639023 (cit. on pp. 14, 15, 48, 63).

[86] T. Gülzow, U. Heute, and H. J. Kolb, “SSB-carrier mismatch detection from speech characteristics: Extension beyond the range of uniqueness,” in *Proceedings of the 11th European Signal Processing Conference (EUSIPCO)*, 2002, pp. 1–4 (cit. on pp. 15, 60, 63).

[87] V. Stahl, A. Fischer, and R. Bippus, “Quantile based noise estimation for spectral subtraction and wiener filtering,” in *2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, 1875–1878 vol.3. DOI: 10.1109/ICASSP.2000.862122 (cit. on pp. 15, 61).

[88] S. Ganapathy and J. Pelecanos, “Enhancing frequency shifted speech signals in single side-band communication,” *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1231–1234, 2013. DOI: 10.1109/LSP.2013.2285361 (cit. on p. 15).

[89] J. Benesty, S. Makino, and J. Chen, “Speech enhancement,” in, J. Benesty, S. Makino, and J. Chen, Eds., ser. Signals and communication technology. Berlin [u.a.] : Springer, 2005, ch. 1.1, p. 1, Literaturangaben, ISBN: 3-540-24039-X (cit. on pp. 15, 16).

[90] J. S. Lim, “Speech enhancement,” in, J. S. Lim, Ed., ser. Prentice-Hall signal processing series. Prentice-Hall, 1983, p. 3, Bibliography: p351-363, ISBN: 0-13-829705-3 (cit. on p. 16).

[91] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209 (cit. on p. 16).

[92] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* Wiley, New York, 1949, ISBN: 0262730057 (cit. on pp. 16, 41, 87, 91).

[93] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. DOI: 10.1109/PROC.1979.11540 (cit. on p. 16).

[94] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995. DOI: 10.1109/89.397090 (cit. on p. 16).

[95] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994, ISSN: 0885-2308. DOI: https://doi.org/10.1006/csla.1994.1016 (cit. on p. 16).

[96] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *1987 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, 1987, pp. 177–180. DOI: 10.1109/ICASSP.1987.1169756 (cit. on p. 16).

[97] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980. DOI: 10.1109/TASSP.1980.1163394 (cit. on p. 16).

[98] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994, Higher Order Statistics, ISSN: 0165-1684 (cit. on p. 16).

[99] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010. DOI: 10.1109/TASL.2009.2031510 (cit. on p. 16).

[100] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings of 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581. DOI: 10.1109/GlobalSIP.2014.7032183 (cit. on p. 16).

[101] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712. DOI: 10.1109/ICASSP.2015.7178061 (cit. on pp. 16, 79, 82).

[102] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013. DOI: 10.1109/TASL.2013.2250961 (cit. on p. 16).

[103] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014. DOI: 10.1109/LSP.2013.2291240 (cit. on p. 16).

[104] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3734–3738. DOI: 10.1109/ICASSP.2014.6854299 (cit. on p. 16).

[105] A. Chinaev, J. Heymann, L. Drude, and R. Haeb-Umbach, "Noise-presence-probability-based noise PSD estimation by using DNNs," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5 (cit. on pp. 16, 20).

[106] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. I–I. DOI: 10.1109/ICASSP.2006.1660065 (cit. on p. 16).

[107] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875–6879. DOI: 10.1109/ICASSP.2019.8683634 (cit. on pp. 16, 20).

[108] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7009–7013. DOI: 10.1109/ICASSP40776.2020.9053266 (cit. on pp. 16, 17, 20, 76, 93).

[109] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 2811–2815. DOI: 10.21437/Interspeech.2021-296 (cit. on p. 16).

[110] S. Makino, H. Sawada, and T.-W. Lee, Eds., *Blind Speech Separation*, ser. Signals and Communication Technology. Springer, 2007, ISBN: 978-1-4020-6478-4. DOI: 10.1007/978-1-4020-6479-1. [Online]. Available: https://doi.org/10.1007/978-1-4020-6479-1 (cit. on p. 17).

[111] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953. DOI: 10.1121/1.1907229. eprint: https://doi.org/10.1121/1.1907229 (cit. on p. 17).

[112] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566. DOI: 10.1109/ICASSP.2014.6853860 (cit. on p. 17).

[113] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. DOI: 10.1109/ICASSP.2017.7952154 (cit. on p. 17).

[114] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 31–35 (cit. on pp. 17, 18, 84).

[115] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017. DOI: 10.1109/TASLP.2017.2726762 (cit. on pp. 17, 18, 76, 78, 81, 82, 86).

[116] Y. Luo and N. Mesgarani, “TaSNet: Time-domain audio separation network for real-time, single-channel speech separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700. DOI: 10.1109/ICASSP.2018.8462116 (cit. on pp. 17, 18, 76, 77, 79, 82).

[117] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, “WHAMR!: Noisy and reverberant single-channel speech separation,” *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2020 (cit. on p. 17).

[118] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, “SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition,” *ArXiv e-prints*, 2019 (cit. on pp. 17, 85).

[119] S. Renals, T. Hain, and H. Bourlard, “Recognition and understanding of meetings the AMI and AMIDA projects,” in *2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2007, pp. 238–247 (cit. on p. 17).

[120] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317 (cit. on p. 17).

[121] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018 (cit. on p. 17).

[122] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, “MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition,” in *2019 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 237–244. DOI: 10.1109/ASRU46091.2019.9003986 (cit. on p. 17).

[123] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, “Cracking the cocktail party problem by multi-beam deep attractor network,” in *2017 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017, pp. 437–444. DOI: 10.1109/ASRU.2017.8268969 (cit. on p. 17).

[124] L. Drude and R. Haeb-Umbach, “Integration of neural networks and probabilistic spatial models for acoustic blind source separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, 2019. DOI: 10.1109/JSTSP.2019.2912565 (cit. on pp. 17, 91).

[125] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157. DOI: 10.1109/EUSIPCO.2016.7760429 (cit. on p. 17).

[126] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *2007 IEEE International Symposium on Circuits and Systems*, 2007, pp. 3247–3250. DOI: 10.1109/ISCAS.2007.378164 (cit. on p. 17).

[127] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250. DOI: 10.1109/ICASSP.2017.7952155 (cit. on pp. 17, 18).

[128] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 686–690. DOI: 10.1109/ICASSP.2018.8462507 (cit. on p. 18).

[129] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982. DOI: 10.1109/TASSP.1982.1163920 (cit. on p. 18).

[130] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2010.12.003 (cit. on p. 18).

[131] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016. DOI: 10.1109/TASLP.2015.2512042 (cit. on p. 18).

[132] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 370–382, 2019. DOI: 10.1109/JSTSP.2019.2904183 (cit. on pp. 18, 60, 79).

[133] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 2713–2717. DOI: 10.21437/Interspeech.2018-1773 (cit. on p. 18).

[134] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25. DOI: 10.1109/ICASSP39728.2021.9413901 (cit. on pp. 18, 79).

[135] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, May 2019. DOI: 10.1109/TASLP.2019.2915167 (cit. on pp. 18, 20, 78, 79, 89).

[136] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *MultiMedia Modeling - 26th International Conference, MMM*, ser. Lecture Notes in Computer Science, vol. 11961, Springer, Jan. 2020, pp. 653–665. DOI: 10.1007/978-3-030-37731-1_53 (cit. on p. 18).

[137] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2019, pp. 4574–4578. DOI: 10.21437/Interspeech.2019-3181 (cit. on pp. 18, 20, 77).

[138] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 31–35. DOI: 10.1109/ICASSP40776.2020.9054172 (cit. on pp. 18, 86).

[139] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018. DOI: 10.1109/TASLP.2018.2795749 (cit. on p. 18).

[140] Y. Luo, E. Ceolini, C. Han, S.-C. Liu, and N. Mesgarani, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 260–267. DOI: 10.1109/ASRU46091.2019.9003849 (cit. on p. 20).

[141] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7319–7323. DOI: 10.1109/ICASSP40776.2020.9053092 (cit. on p. 20).

[142] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6389–6393. DOI: 10.1109/ICASSP40776.2020.9053833 (cit. on p. 20).

[143] X. Lu, Y. Tsao, S. Matsud, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 436–440 (cit. on p. 20).

[144] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837 (cit. on p. 20).

[145] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443. DOI: 10.1109/ASRU.2015.7404828 (cit. on p. 20).

[146] J. Heymann, L. Drude, and R. Haeb-Umbach, “Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition,” in *Proceedings of the CHiME 2016 Workshop on Speech Processing in Everyday Environments*, Sep. 2016, pp. 12–17 (cit. on p. 20).

[147] L. Drude, J. Heymann, and R. Haeb-Umbach, “Unsupervised Training of Neural Mask-Based Beamforming,” in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1253–1257. DOI: 10.21437/Interspeech.2019-2549 (cit. on p. 20).

[148] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, “Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 286–290. DOI: 10.1109/ICASSP.2017.7952163 (cit. on p. 20).

[149] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015. DOI: 10.1109/TASLP.2014.2364452 (cit. on p. 20).

[150] M. Kolbaek, Z.-H. Tan, S. H. Jensen, and J. Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020. DOI: 10.1109/TASLP.2020.2968738 (cit. on pp. 20, 76, 82).

[151] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073. DOI: 10.1109/ICASSP.2018.8462417 (cit. on p. 20).

[152] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, “A comparative study of time and frequency domain approaches to deep learning based speech enhancement,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206928 (cit. on pp. 20, 21).

[153] J. Seamons, *Kiwi-SDR*, http://kiwisdr.com, 2021 (accessed July 21, 2021) (cit. on p. 22).

[154] Bundesnetzagentur, *Ordinance concerning the amateur radio act*, Aug. 2019 (cit. on p. 23).

[155] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964 (cit. on pp. 24, 64).

[156] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, . Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Dec. 2011, ISBN: 978-1-4673-0366-8 (cit. on pp. 24, 85).

[157] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” Idiap, Tech. Rep., Jan. 2012 (cit. on pp. 24, 25, 85).

[158] R. Skaug, J. Hjelmstad, and I. of Electrical Engineers, *Spread Spectrum in Communication*, ser. IEE telecommunications series. P. Peregrinus, 1985, ISBN: 9780863410345 (cit. on p. 25).

[159] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, ISSN: 1558-7916 (cit. on pp. 27, 63).

[160] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976. DOI: 10.1109/TASSP.1976.1162830 (cit. on p. 27).

[161] W. E. Sabin and E. O. Schoenike, *Single-sideband Systems and Circuits*, 2., W. E. Sabin and E. O. Schoenike, Eds. McGraw-Hill, 1987, ISBN: 9780070544079 (cit. on p. 28).

[162] A. Veysov, A. Voytsekhovskiy, P. Denisov, and Y. Baburov, *Russian Open Speech To Text (STT/ASR) Dataset*, https://github.com/snakers4/open_stt/, 2021 (accessed July 21, 2021) (cit. on p. 28).

[163] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167 (cit. on pp. 37, 58, 65).

[164] L. Kong, C. Dyer, and N. A. Smith, “Segmental recurrent neural networks,” May 2016 (cit. on p. 38).

[165] L. Pang, Y. Lan, J. Xu, J. Guo, and X. Cheng, “Locally smoothed neural networks,” in *Proceedings of the Ninth Asian Conference on Machine Learning*, M.-L. Zhang and Y.-K. Noh, Eds., ser. Proceedings of Machine Learning Research, vol. 77, Yonsei University, Seoul, Republic of Korea: PMLR, 15–17 Nov 2017, pp. 177–191 (cit. on p. 38).

[166] S. V. Gerven and F. Xie, “A comparative study of speech detection methods,” in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997, pp. 1095–1098 (cit. on p. 41).

[167] B. Sharma, R. Das, and H. Li, “Multi-level adaptive speech activity detector for speech in naturalistic environments,” in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2015–2019. DOI: 10.21437/Interspeech.2019-1928 (cit. on p. 41).

[168] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967. DOI: 10.1109/TIT.1967.1054010 (cit. on p. 42).

[169] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014 (cit. on pp. 43, 89).

[170] NIST U.S. Department of Commerce, *NIST Open Speech-Activity-Detection Evaluation*, 2016 (accessed April 9, 2020). [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation (cit. on p. 43).

[171] S. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *The Journal of the Acoustical Society of America*, vol. 123, pp. 4559–71, Jul. 2008 (cit. on p. 55).

[172] Y. Gong and J.-P. Haton, “Time domain harmonic matching pitch estimation using time-dependent speech modeling,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1386–1400, 1987. DOI: 10.1109/TASSP.1987.1165056 (cit. on p. 55).

[173] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *International Conference on Learning Representations (ICLR)*, May 2016 (cit. on p. 59).

[174] J. Cheng, Z. Wang, and G. Pollastri, “A neural network approach to ordinal regression,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1279–1284. DOI: 10.1109/IJCNN.2008.4633963 (cit. on p. 60).

[175] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 749–752 (cit. on p. 63).

[176] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on p. 70).

[177] D. Ditter and T. Gerkmann, “A multi-phase gammatone filterbank for speech separation via tasnet,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 36–40. DOI: 10.1109/ICASSP40776.2020.9053602 (cit. on pp. 77, 89).

[178] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, “Filterbank design for end-to-end speech separation,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6364–6368. DOI: 10.1109/ICASSP40776.2020.9053038 (cit. on pp. 77, 89).

[179] J. Burred and T. Sikora, “On the use of auditory representations for sparsity-based sound source separation,” in *2005 5th International Conference on Information Communications Signal Processing*, 2005, pp. 1466–1470. DOI: 10.1109/ICICS.2005.1689302 (cit. on p. 77).

[180] A. Pandey and D. Wang, “Exploring deep complex networks for complex spectrogram enhancement,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6885–6889. DOI: 10.1109/ICASSP.2019.8682169 (cit. on p. 79).

[181] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks: A unified approach to action segmentation,” in *European Conference on Computer Vision*, Springer, 2016, pp. 47–54 (cit. on p. 79).

[182] L. J. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *CoRR*, vol. abs/1607.06450, 2016. eprint: 1607.06450 (cit. on p. 80).

[183] J. Chen, Q. Mao, and D. Liu, “Dual-Path Transformer Network: Direct context-aware modeling for end-to-end monaural speech separation,” in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 2642–2646. DOI: 10.21437/Interspeech.2020-2205 (cit. on p. 81).

[184] Y. Luo, C. Han, and N. Mesgarani, “Group communication with context codec for lightweight source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1752–1761, 2021. DOI: 10.1109/TASLP.2021.3078640 (cit. on p. 81).

[185] J. R. Erdogan Hakan and Hershey, S. Watanabe, and J. Le Roux, “Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio,” in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Eds. Springer International Publishing, 2017, pp. 165–186 (cit. on p. 82).

[186] S. Braun and I. Tashev, “A consolidated view of loss functions for supervised deep learning-based speech enhancement,” in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, 2021, pp. 72–76. DOI: 10.1109/TSP52935.2021.9522648 (cit. on p. 82).

[187] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, “Unsupervised sound separation using mixture invariant training,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 3846–3857 (cit. on p. 83).

[188] J. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete LDC93S6A,” *Web Download. Philadelphia: Linguistic Data Consortium*, 1993 (cit. on p. 84).

[189] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. W. Ellis, C. C. Raffel, B. Mcfee, and E. J. Humphrey, “MIR-Eval: A transparent implementation of common MIR metrics,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014 (cit. on p. 84).

[190] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006. DOI: 10.1109/TSA.2005.858005 (cit. on p. 84).

[191] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?" In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630. DOI: 10.1109/ICASSP.2019.8683855 (cit. on p. 84).

[192] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 3743–3747. DOI: 10.21437/Interspeech.2018-1417 (cit. on p. 85).

[193] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiME challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 444–451. DOI: 10.1109/ASRU.2015.7404829 (cit. on p. 86).

[194] T. Inoue, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Rondo, "Theoretical analysis of musical noise in generalized spectral subtraction: Why should not use power/amplitude subtraction?" In *2010 18th European Signal Processing Conference*, 2010, pp. 994–998 (cit. on p. 87).

[195] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *IWAENC2008: the 11th International Workshop on Acoustic Echo and Noise Control,*, Seattle, Washington USA, Sep. 2008 (cit. on p. 87).

[196] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 1571–1575. DOI: 10.21437/Interspeech.2018-1262 (cit. on p. 93).