Abstract

Stochastic models on the the basis of the Hidden Markov Model (HMM) are established as a standard in automatic speech recognition (ASR). However, an essential drawback of the HMM consists in the so-called conditional independence assumption, i.e. in not directly modelling the statistical dependencies between speech features which are extracted from subsequent frames of the speech signal. This drawback is also related to a preceding model-based speech feature enhancement based on the Gaussian Mixture Model (GMM). The objective of the following work is the integration of the statistical dependencies between subsequent speech features into the statistical speech recognition approach. It is possible to consider inter-frame correlations for speech feature enhancement by modelling the dynamics of the cepstral speech features with a Switching Linear Dynamic Model (SLDM). In this work different possibilities for posterior estimation with SLDMs are investigated. Further a new state space model for the cepstral noise process is introduced. Expectation-Maximization (EM) algorithms are derived for the parameter estimation in this framework. The statistical dependencies between the speech features in the acoustic back-end model are approximated on the level of the HMM mixture weights. An efficient search strategy is developed for the resulting segmental HMM. Further the exchange of information between the speech feature enhancement stage and the speech recognition stage is investigated. A so-called uncertainty decoding is applied for speech recognition in order to exploit the uncertainty of the extracted speech features. Besides, a multi-stage recognition is carried out to consider information from a prior recognition stage for speech feature enhancement. Therefore the speech distribution is influenced with information about the HMM states. It is possible to account for the uncertainty of the recognition results by calculating the posterior of the HMM states instead of the single-best HMM sequence. For this purpose, a wordgraph-based recognizer is developed where the most probable words are determined for each speech frame with a forward-backward algorithm on word level. The posteriors of the HMM states are estimated on the constricted wordgraph with a forward-backward algorithm on state level.