Combining binaural and cortical features for robust speech recognition

Constantin Spille and Bernd T. Meyer
Cluster of Excellence Hearing4all and Medizinische Physik, University of Oldenburg, 26129 Oldenburg, Germany
e-mail: (constantin.spille , bernd.meyer)@uni-oldenburg.de

Abstract

The segregation of concurrent speakers and other sound sources is an important ability of the human auditory system but is missing in most current systems for automatic speech recognition (ASR). This study combines processing related to peripheral and cortical stages of the auditory pathway: A physiologically-motivated binaural model estimates the positions of moving speakers to enhance the desired speech signal. Secondly, signals are converted to spectro-temporal Gabor features that resemble cortical speech representations. Binaural processing improved recognition results in all acoustic conditions under consideration compared to single channel processing. In noisy situations Gabor features perform best, while in clean situations normalized mel-frequency cepstral coefficients should be preferred. A simple decision rule based on the estimated target-to-noise ratio is proposed to select the best processing chain for the particular acoustic scene, which results in a relative improvement of 30.2 % of error rates on average compared to the baseline.

Introduction

The human auditory system is able to easily analyze and decompose complex acoustic scenes into its constituent acoustic sources. This requires the integration of a multitude of acoustic cues, a phenomenon that is often referred to as cocktail-party processing. Auditory scene analysis (ASA), especially the segregation and comprehension of concurrent speakers, is one of the key features in cocktail-party processing [1]. While most of today's ASR systems do not incorporate features estimated from the acoustic scene (such as prior information about the room or the position of speakers), this paper investigates auditory-inspired methods to enhance and optimally represent speech signals recorded by hearing aid microphones. ASR is performed with this system operating in complex acoustic scenes, resembling an application scenario in which transcripts of spoken language could be provided to hearing-impaired listeners. Auditory processing is integrated at two stages in the system, mimicking strategies corresponding to peripheral or central auditory processing: (A) A binaural model [2] extracts interaural phase differences (IPD) and interaural level differences (ILD) to achieve robust direction of arrival (DOA) estimation of multiple speakers. In scenarios with one or two active speakers, we use these DOA estimations to steer a beamformer to enhance the signal of the desired sound source, which has been shown to improve ASR performance significantly [9]. (B) Complex „cortical" ASR features serve as input to the classification system. In this study we use a Gabor filter bank to extract spectro-temporal Gabor features for ASR (Schädler et al. [7], [6]). These features have been shown to increase the robustness of ASR with respect to several additive noise types.

Methods

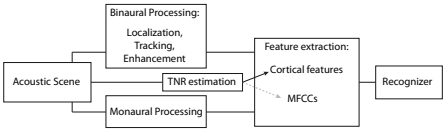

Abbildung 1: Block diagram of the experimental setup. TNR stands for target-to-noise ratio (see section „Scene-specific feature selection").

Figure 1 shows a block diagram of the whole processing chain. Acoustic scenes are simulated by convolving signals with recorded 6-channel head-related impulse respones (HRIR) (3 channels from each of two behind-the-ear (BTE) hearing aids). In the binaural processing step, the signals of the front microphones are fed into the binaural model that is employed to estimate the direction of arrival of spatially distributed speakers. A particle filter is then used to keep track of the positions of the moving sources. Its output is used to steer a beamformer, enhancing the 6-channel speech signal that is to be transcribed by an HMM-GMM ASR system employed using HTK. For monaural processing, each channel is processed independently - see [9] and [8] for more details, especially regarding the ASR system parameters.

Results

In this section we present the results obtained by the proposed binaural system. An application scenario is to use such a system in hearing aids, providing a transcript of spoken language in complex acoustic scenes to hearing-impaired listeners. Signals recorded with six hearing aid microphones are used as input. Single-channel (or monaural) processing serves as baseline, for which the six channels are separately processed and the corresponding word error rates are averaged.

Binaural vs. monaural processing

We first compares the word error rates (WERs) obtained with binaural and monaural processing using cortical Gabor features with cepstral mean and variance norma-

lization (CMVN). In presence of a diffuse noise (1S,DN) at -5 and 0 dB TNR, the binaural processing scheme is slightly worse than monaural processing, but at all other TNRs, binaural processing is outperforms monaural processing. For the localized noise scenario, binaural processing improves ASR performance at low TNRs, i.e. -5 to 5 dB, whereas at higher TNRs monaural processing has a slight advantage. In situations with a concurrent speaker (2S), the binaural processing performs better than the monaural processing at *all* TNRs. However, when averaging over all situations, binaural processing strongly increases ASR performance from 25.4 % to 18.9 % WER.

Cortical Gabor features vs. MFCCs

The properties of features inspired from processing in higher stages of the auditory pathway are analyzed by comparing Gabor features with baseline features. MFCCs with cepstral mean and variance normalization (CMVN) serve as baseline. Since the binaural processing resulted in best results in most conditions, we present results for both feature types combined with the binaural system. Gabor features outperform the baseline features in the presence of noise. However, in the two-speaker scenario or in clean conditions MFCCs perform better than Gabor features. The average WER over all situations is almost identical for Gabor and MFCC features, namely 18.9 % compared to 19.0 %, respectively.

Scene-specific feature selection

The comparison of Gabor and MFCC features showed that each feature type appears to be optimal for specific acoustic conditions. In an oracle experiment in which the best system for the given condition was selected (based on prior information that is not available in a real-world scenario), the WER was reduced to 17.8 %. Due to the complementarity of spectro-temporal Gabor features and MFCCs, it would be desirable to perform feature selection without such a priori information in order to approach the performance of the oracle system. In this study, a simple decision rule based on the estimated TNR is proposed to perform the selection: In each subband the noise level is calculated as a weighted average of spectral magnitude values of the past 50 ms which are below an adaptive threshold[4]. To obtain an estimate of the clean speech signal, the well-established noise reduction algorithm by Ephraim & Malah estimating the optimal minimum mean square error (MMSE) log short-time spectral amplitude (log-STSA) is used [3] (see [5] for details on the algorithm). The energy of the estimated clean speech signal is then compared to the energy of the estimated noise which results in the estimated TNR. Since the system operates on six BTE hearing aid microphones, monaural processing of these microphones serves as a baseline in this study and is carried out by processing each of the six channels independently and averaging the word error rates of the ASR system (see Fig. 2). The combined system improves ASR performance in all situations, the only exception being (1S,LN) for clean condition. In total, the combined system achieved a word error rate of 17.7 % which corresponds to a relative improvement of 30.2 % compared to the baseline.

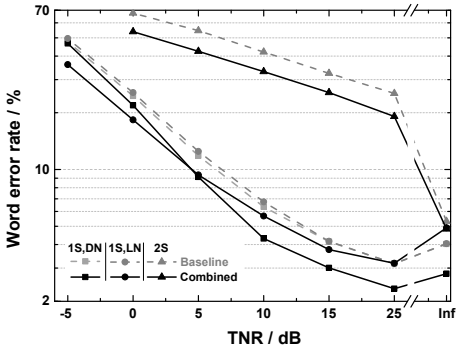

Abbildung 2: Word error rates for the combined system and the baseline system for all situations.

Literatur

[1] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[2] M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, May 2011.

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, Apr. 1985.

[4] H.-G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. ICASSP 1995*, 1995.

[5] T. May, S. Van De Par, and A. Kohlrausch. A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation. *IEEE Transactions On Audio Speech And Language Processing*, 20(7):1–15, 2012.

[6] B. T. Meyer, C. Spille, B. Kollmeier, and N. Morgan. Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition. In *Proc. of Interspeech*, 2012.

[7] M. R. Schädler, B. T. Meyer, and B. Kollmeier. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *Journal of the Acoustical Society of America*, 131(5):4134–4151, May 2012.

[8] C. Spille, M. Dietz, V. Hohmann, and B. Meyer. Using Binaural Processing For Automatic Speech Recognition In Multi-Talker Scenes. In *ICASSP*, pages 7805–7809, 2013.

[9] C. Spille, B. T. Meyer, M. Dietz, and V. Hohmann. Binaural scene analysis with multi-dimensional statistical filters. In J. Blauert, editor, *The technology of binaural listening*, chapter 6. Springer, 2013.