Binaural Scene Analysis and Automatic Speech Recognition

Constantin Spille, Mathias Dietz, Volker Hohmann and Bernd T. Meyer
Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
constantin.spille@uni-oldenburg.de, mathias.dietz@uni-oldenburg.de, volker.hohmann@uni-oldenburg.de, bernd.meyer@uni-oldenburg.de

Introduction

The human auditory system is known to be 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, 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, the concept of using multi-source recordings for signal enhancement has been investigated in a number of studies: The approach of an ideal binary mask has been adopted for speaker segregation, e.g., in combination with binaural cues [2], and automatic speech recognition (ASR) [3, 4]. These studies try to find reliable time-frequency regions in which one speaker is dominant and use only these reliable cues instead of all information which seems to have a detrimental effect on the overall performance of the system. More technical approaches use microphone arrays to perform speaker segregation (e.g., [5]). For speech recognition these systems are often combined with beamforming algorithms [6].

While multi-microphone arrays have no physiological basis and binaural cues are often obtained using cross-correlation methods [2], the present paper uses an physiologically based binaural model [7] extracting interaural phase differences (IPD) and interaural level differences (ILD) to achieve robust direction of arrival (DOA) estimation of multiple speakers.

In a two-speaker scenario, we use these DOA estimations to steer a beamformer to enhance the signal of the desired sound source, which mimics the cognitive process of paying attention to one speaker and improves ASR performance significantly [8].

Experimental Setup

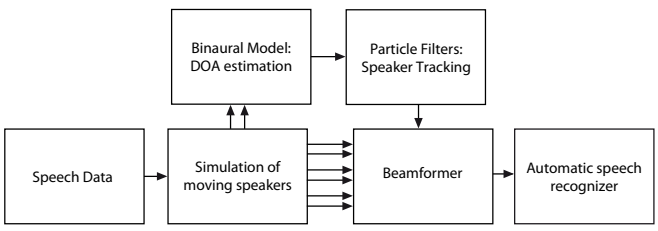

Figure 1: Block diagram of the experimental setup.

Figure 1 shows a block diagram of the whole processing chain from the speech data to the ASR System. Moving speakers are generated by convolving speech data with recorded 6-channel head-related transfer functions (HRIR) (3 channels from each of two behind-the-ear (BTE) hearing aids). The signals of the front microphones are fed into the binaural model which 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 speakers. Its output is used to steer a beamformer, enhancing the 6-channel speech signal that is to be transcribed by an ASR system. In the following sections each of these processing steps is shortly described. For more details see [8] and [9].

Speech Data

The speech data used for the experiments consists of sentences produced by ten speakers – four male, six female. The syntactical structure and the vocabulary were adapted from the Oldenburg Sentence Test, OLSA,, OLSA,[10] where each sentence contains five words with ten alternatives for each word and a syntax that follows the pattern <name><verb><number><adjective><object>, which results in a vocabulary size of 50 words.

Binaural Model

For direction of arrival estimation, we use the interaural phase difference (IPD) model proposed by Dietz et al. [7]. In the following only the conceptually relevant aspects are briefly reviewed. Multi channel signals are analyzed in 23 auditory filters in the range of 200 Hz to 5.0 kHz. Considering the human limit to binaurally exploit fine-structure information above $\sim$1.4 kHz, the fine-structure filter is only implemented in the 12 lowest auditory filters below 1.4 kHz. IPDs are calculated in these fine-structure filters and IPD-to-azimuth mapping is performed with a previously learned mapping function. In this model, the IPD fluctuations are directly accessible and are specified in the form of the interaural vector strength (IVS). The IVS was used to derive a filter mask which consists of a binary weighting of the interaural parameters based on a threshold value $\mathrm{IVS}_0 = 0.98$ [7, 9].

Particle Filter

In this study, a particle filter algorithm provided by Särkkä et al. [11] is used for speaker tracking. The main idea of the algorithm is to split up the problem into two parts ("Rao-Blackwellization"). First, the posterior distribution of the data association is calculated using a Sequential Importance Resampling (SIR) particle filtering algorithm. Second, the single targets are tracked

by an extended Kalman filter that depends on the data associations. Rao-Blackwellization exploits the fact that it is often possible to calculate the filtering equations in closed form. This leads to estimators with less variance compared to the method using particle filtering alone [12]. For more details of the algorithms see [13] and [11] and for details of the actual application the reader is referred to [9].

Steerable beamformer and signal enhancement

The beamformer employed here is a super-directive beamformer based on the minimum variance distortionless response principle [14] that used the six BTE microphone inputs jointly. Let W be the matrix containing the frequency domain filter coefficients of the beamformer, d_1 and d_2 the vectors containing the transfer functions to the microphones of speakers one and two, respectively, and Φ_{VV} the noise power-spectral density (PSD) matrix. Then, the following minimization problem has to be solved

$$\min_{W} W^H \Phi_{VV} W \quad \text{with } W^H d_1 = 1 \text{ and } W^H d_2 = 0 \quad . \tag{1}$$

The solution to this is the minimum variance distortionless response beamformer [15]. The transfer functions in vectors d_1 and d_2 result from the impulse responses which are chosen based on the angle estimation of the tracking algorithm. The coherence matrix which is required to solve Eq. 1 is also estimated using the impulse responses used for generating the signals. For more details see [8, 9].

ASR system

For ASR, the pre-processed signals are first converted to ASR standard features, i.e., Mel-Frequency Cepstral Coefficients (MFCCs) [16]. By adding a delta and double-delta features, 39-dimensional feature vectors were obtained per 10 ms step. The feature vectors are used to train and test the Hidden Markov model (HMM) classifier implemented using the Hidden Markov Toolkit (HTK) [17]. For more details about the HMM see [9].

ASR training was carried out with three different conditions, i.e. clean, multi and matched SNR condition. The training set contained a total of 71 sentences that were used as-is for clean training and in the multi condition training these 71 sentences were additionally mixed five times with a stationary speech shaped noise at SNRs ranging from -5 dB to 20 dB in 5 dB steps, resulting in a total training set of 2201 sentences. The matched SNR training only consisted of the 71 sentences mixed 5 times at a specific SNR, resulting in a total of 355 sentences.

For testing, signals with two moving speakers with identical SNRs as used for training were processed by the complete chain depicted in Fig. 1 (one being the target source, and the other one the suppressed source), and the recognition rate for the words uttered by the target speaker was obtained. The target speaker's data was not contained in the training data, resulting in a speaker-independent ASR system. To increase the number of test items, each speaker was selected as the target speaker once and the training/testing procedure was carried out ten times. The test set contained a total of 781 two-speaker tracks for each SNR, so, the total number of test sentences was 4686 [8].

Results

When using the complete processing chain that included the DOA estimation, tracking, beamforming, and ASR, a word recognition rate (WRR) of 72.7 % was obtained for clean condition training and testing. When the ASR system cannot operate on beamformed signals, but is limited to speech that was converted to mono signals (by selecting one of the 8 channels from the behind-the-ear or in-ear recordings), the average WRR was 29.4 % when testing on clean signals.

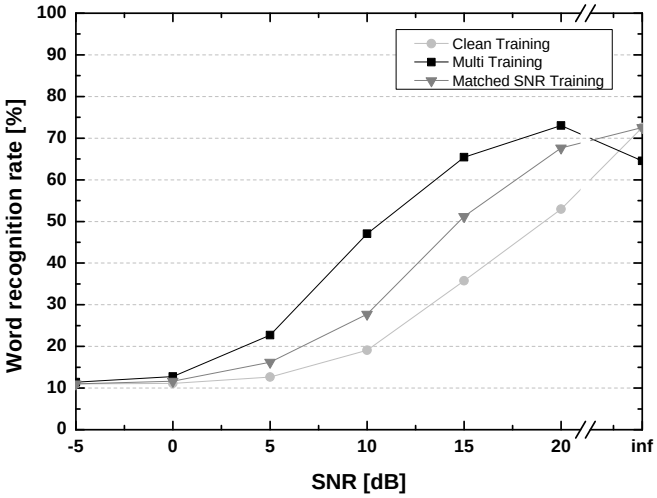

Figure 2: Word recognition rates at all SNRs for clean, multi and matched SNR training.

Figure 2 shows the recognition results for all training conditions. Multi condition training resulted in the highest word recognition rate in all noisy situations. Only in clean testing, multi condition training achieved a lower recognition rate compared to clean training, due to the little amount of clean sentences in the training set compared to the much larger amount of noisy sentences.

In addition, the WRR also depends on the average tracking error. The bottom panel of Fig. 3 shows the dependency of WRR on the average tracking error for 0 dB, 10 dB and 20 dB SNR in the multi-condition training. The WRR is highly dependent on the average tracking error at higher SNRs with higher tracking errors resulting in significantly lower WRRs. This dependency is not observable for 0 dB data, i.e., in a two-speaker scenario with low SNR, the beamforming approach is limited by the presence of the diffuse noise. For a more extensive analysis see [8].

Summary and Outlook

This study provided an overview of computational auditory scene analysis based on binaural information and its application to a speech recognition task. It was

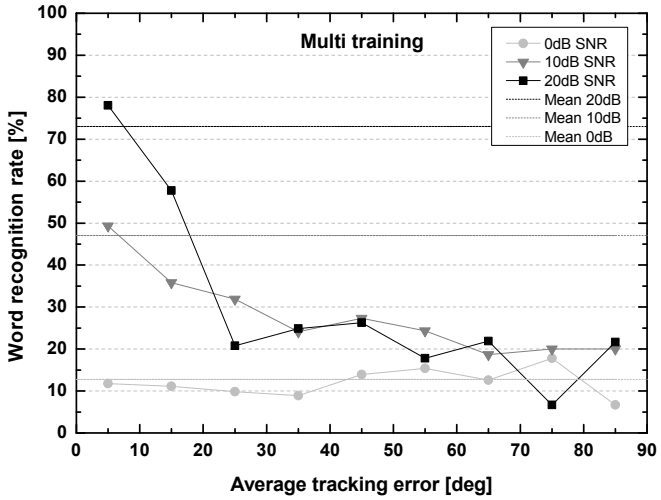

Figure 3: Word recognition rate vs. average tracking error for different signal to noise ratios for multi condition training. Dotted lines show the total word recognition rate for the specific condition.

also shown that the binaural model enables efficient tracking and greatly increases the performance of an automatic speech recognition system in situations with one interfering speaker. The word recognition rate (WRR) was increased from 30.8 % to 72.7 %, which shows the potential of integrating models of binaural hearing into speech processing systems [8].

Further studies in more realistic conditions with reverberation showed that using the system proposed here and incorporating more information about the acoustic scene, i.e. the target-to-noise ratio, increases the average ASR performance. In particular, a relative improvement of 9.7 % of word error rates was achieved on average.

References

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

[2] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[3] N. Ma, J. Barker, H. Christensen, and P. Green, "Combining Speech Fragment Decoding and Adaptive Noise Floor Modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 818–827, Mar. 2012.

[4] T. May, S. Van De Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE T. Audio. Speech.*, vol. 20, pp. 108–121, 2012.

[5] G. Lathoud, I. A. Mccowan, and D. C. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *in Proceedings of Eurospeech 2003, September 2003. IDIAP–RR 03-xx.*

[6] D. Kolossa, F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, and R. Martin, "CHiME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," *Int. Workshop on Machine Listening in Multisource Environments*, vol. 1, pp. 6–11, 2011.

[7] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, May 2011.

[8] C. Spille, M. Dietz, V. Hohmann, and B. T. Meyer, "Using binarual processing for automatic speech recognition in multi-talker scenes," in *Proc. ICASSP 2013*, 2013.

[9] C. Spille, B. T. Meyer, M. Dietz, and V. Hohmann, "Binaural scene analysis with multi-dimensional statistical filters," in *The technology of binaural listening*, J. Blauert, Ed. Berlin–Heidelberg–New York NY: Springer, 2013, ch. 6.

[10] K. C. Wagener and T. Brand, "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters." *International Journal of Audiology*, vol. 44, no. 3, pp. 144–156, 2005.

[11] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-Blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2–15, Jan. 2007.

[12] G. Casella and C. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[13] J. Hartikainen and S. Särkkä, "RBMCDAbox-Matlab Toolbox of Rao-Blackwellized Data Association Particle Filters," *documentation of RBMCDA Toolbox for Matlab V*, 2008.

[14] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[15] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 2.

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal*, vol. 61, 1980.

[17] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Engineering Department*, vol. 3, 2002.