# USING BINARUAL PROCESSING FOR AUTOMATIC SPEECH RECOGNITION IN MULTI-TALKER SCENES

*Constantin Spille, Mathias Dietz, Volker Hohmann, Bernd T. Meyer*

Medical Physics, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

## ABSTRACT

The segregation of concurrent speakers and other sound sources is an important aspect of the human auditory system but is missing in most current systems for automatic speech recognition (ASR), resulting in a large gap between human and machine performance. The present study uses a physiologically-motivated model of binaural hearing to estimate the position of moving speakers in a noisy environment by combining methods from Computational Auditory Scene Analysis (CASA) and ASR. The binaural model is paired with a particle filter and a beamformer to enhance spoken sentences that are transcribed by the ASR system. Results based on an evaluation in clean, anechoic two-speaker condition shows the word recognition rates to be increased from 30.8% to 72.6%, demonstrating the potential of the CASA-based approach. In different noisy environments, improvements were also observed for SNRs of 5 dB and above, which was attributed to the average tracking errors that were consistent over a wide range of SNRs.

***Index Terms***— Automatic speech recognition, particle filter, beamformer, computational auditory scene analyses

## 1. 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 ([3], [4]). These studies try to find reliable time-frequency (T-F) regions in which one speaker is dominant and use only these reliable information instead of all information which seems to have a detrimental effect on the overall performance of the system. In [5], binaural tracking of multiple sources using Hidden Markov models and Kalman filters is discussed, but its application to ASR is not assessed. More technical approaches use microphone arrays to perform speaker segregation (e.g. [6]). For speech recognition these systems are often combined with beamforming algorithms [7].

While these 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 [8] 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. The paper is structured as follows: Section 2 describes the experimental setup and goes into each processing step in more detail. ASR results are presented in Section 3, which are compared to the performance of an ASR system working on unprocessed signals. Finally, we summarize and conclude our study in Section 4.

## 2. EXPERIMENTAL SETUP

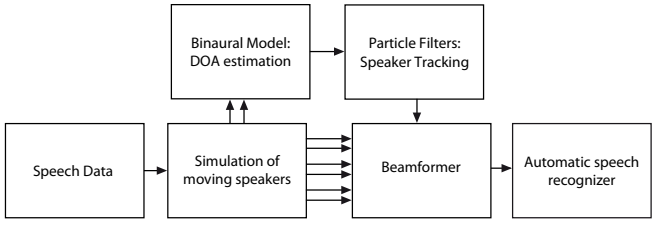


**Fig. 1**. Block diagram of the experimental setup. See text for details.

Fig. 1 shows a block diagram of the whole processing chain from the speech data to the ASR System. Mov-

Supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31).

ing speakers are generated by convolving speech data with recorded 8-channel head-related transfer functions (HRIR) (2 in-ear channels and 3 channels from each of two behind-the-ear (BTE) hearing aids). The in-ear signals 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 described in more detail.

### 2.1. Speech Data

The speech data used for the experiments consists of sentences produced by 10 speakers (4 male, 6 female). The syntactical structure and the vocabulary were adapted from the Oldenburg Sentence Test (olsa) [9], i.e., each sentence contains five words with 10 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. The original recordings with a sampling rate of 44.1 kHz were downsampled to 16 kHz and concatenated (using three sentences from the same speaker) to obtain utterances with a duration of 5 to 10 s, suitable for speaker tracking.

The HRIRs used in this study are a subset of the database described in [10]: Anechoic free-field HRIRs from the frontal horizontal half-plane measured at a distance of 3 meters between microphones and loudspeaker were selected. The HRIRs from the database were measured with a 5° resolution for the azimuth angles, which was interpolated to obtain a 0.5° resolution.

### 2.2. Binaural Model

For direction of arrival estimation, we use the IPD model proposed by Dietz et al.[8]. 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. A problem for fine-structure interaural phase differences in filters above 700 Hz is that their corresponding interaural tie differences do no longer cover the whole range of possible interaural delays, resulting in an ambiguity of direction. Inspired by psychoacoustic findings such as time-intensity trading (e.g., [11]) the sign of the ILD is employed here to extend the unambiguous range of IPDs from $[-\pi, \pi]$ to $[-2\pi, 2\pi]$. Accordingly, the frequency range for unambiguous fine-structure IPD-to-azimuth mapping is extended from $\sim 700$ Hz to 1400 Hz. IPD-to-azimuth mapping itself 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 $\text{IVS}_0 = 0.98$.

By processing each of these high-coherence segments as a single event called "glimpse", a sparse representation of the binaural features is generated from the median value of the azimuth estimation of this segment. If the IVS constantly exceeds $\text{IVS}_0$ for more than 20 ms, a new glimpse is assigned from the same segment.

### 2.3. Particle Filter

The main challenge in the tracking of multiple targets is the mapping from observations (in this case, DOA glimpses) to a specific target, which is a prerequisite for the actual tracking. In this study, an algorithm provided by Särkkä et al.[12] is applied to solve this problem. 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 [13]. For more details of the algorithms see [14] and [12].

The particle filter was initialized with a set of 20 particles using a known starting position of the first speaker (i.e., the location variable of the first target was set to the position for all particles). The location variable of the second target was altered for each particle in equidistant steps throughout the whole azimuth range. Initial velocities were set randomly between $\pm$ 2 m/s for each target in each particle. If no glimpse is observed at time step $t$, the update step of the Kalman filter was skipped for this time step and the prediction was made based on the internal particle states. The range of the predicted angles was limited to the interval [-90°,90°] by setting all predictions outside that range to -90° or 90°, respectively.

### 2.4. Steerable beamformer for source selection

In the proposed application, a position estimate for both the target and concurrent speaker are required to control the beamformer parameters to either enhance the speech of a certain speaker or strongly suppress a concurrent speaker, thereby increasing the overall signal-to-noise ratio and subsequently lower the word error rates of an automatic speech recognizer. The beamformer employed here is a super-directive beamformer based on the minimum variance distortionless response principle [15] 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 $\Phi_{VV}$ the noise power-spectral density (PSD) matrix. Then, the following minimization problem has to be solved

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

The solution to this is the minimum variance distortionless response beamformer [see 16, chap. 2]. 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. Note that relying on the true impulse responses implies the use of *a-priori* knowledgde not available in a real-world application, for which the impulse responses need to be estimated. The beamforming by itself therefore represents an upper bound, and will be extended to be used with estimated impulse responses in future work. However, since the IPD model, the tracking algorithm and the ASR system do not use such *a-priori* knowledge (reflecting realistic conditions), and robust methods for estimation of impulse responses exist, the results should still be transferable to real-world applications.

### 2.5. ASR system

For ASR, the pre-processed signals are first converted to ASR standard features, i.e., Mel-Frequency Cepstral Coefficients (MFCCs) [17]. 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, which has been set up as word model with each word of the vocabulary corresponding to a single HMM. A grammar reflecting the fixed syntax of OLSA sentences is used to ensure a transcription with a valid OLSA sentence structure. The HMM used ten states per word model and six Gaussians per mixture and was implemented using the Hidden Markov Toolkit (HTK) [18].

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 ans in the multi condition training these 71 sentences were additionally mixed with a stationary speech shaped noise at SNRs ranging from -5 dB to 20 dB in 5 dB steps. This procedure was carried out five times using random parts of the noise, 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.

## 3. 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. Although the WRRs in the multi condition training were higher in all other test conditions, the WRRs dropped down to 64.5% in clean testing (see Table 1). This is due to the little amount of clean sentences (71 sentences) in the training material compared to the 2130 sentences with additional noise. The different amount of training material is also the reason why the multi condition training gave better results than the matched SNR training in nearly all conditions. 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. The variations of WRRs between channels were relatively small, ranging from 28.1% to 30.8%. When the best channel for each sentence was selected, i.e., the channel that resulted in the highest WRR for that specific sentence to simulate the best performance when limited to one channel, the average WRR was increased to 38.8%.

| SNR [db] | Average tracking error [deg] | Word recognition rate [%] | | |
|---|---|---|---|---|
| | | Clean | Multi | Matched |
| -5 | 15.19 | 11.00 | 11.42 | 11.03 |
| 0 | 8.17 | 11.11 | 12.73 | 11.65 |
| 5 | 5.84 | 12.65 | 22.72 | 16.24 |
| 10 | 5.44 | 19.09 | 47.04 | 27.73 |
| 15 | 5.88 | 35.75 | 65.41 | 51.20 |
| 20 | 5.43 | 52.93 | 73.05 | 67.64 |
| inf | 5.00 | 72.65 | 64.53 | 72.65 |

**Table 1**. Average tracking error and word recognition rates for all different SNR conditions. See text for details.

The word recognition rate also depends strongly on the localization accuracy which was quantified by calculating the average tracking error, which is the root median squared error between the smoothed tracking estimates and the real azimuth angles of the speakers. Table 1 shows the average tracking error in dependency of the SNR and the corresponding word

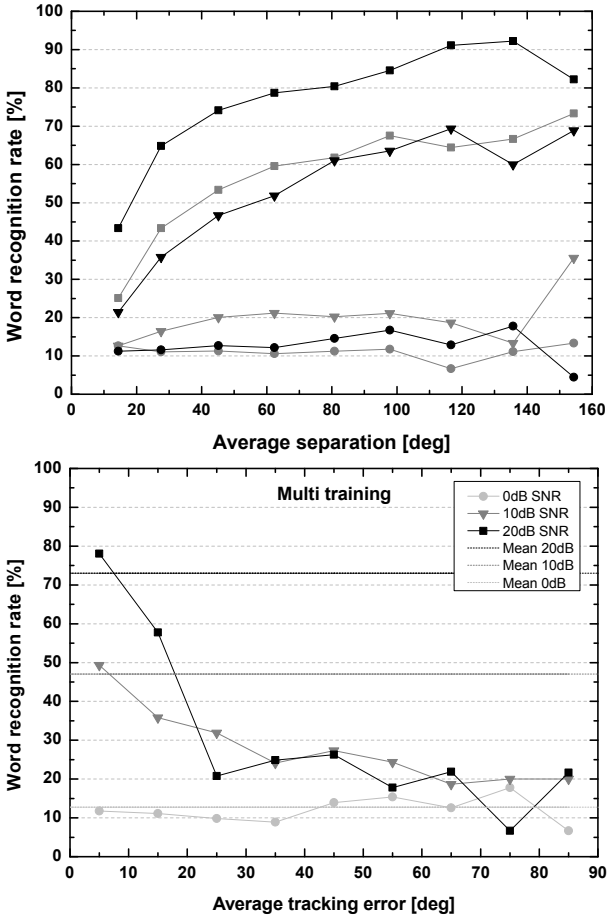


**Fig. 2**. **Top:** Word recognition rate vs. average separation at different SNRs for clean condition training (grey symbols) and multi condition training (black symbols). Circles, triangles and squares represent 0 dB, 10 dB and 20 dB SNR respectively. **Bottom:** Word recognition rate vs. average tracking error for different signal to noise ratios and clean and multi condition training. Dotted lines show the total word recognition rate for the specific condition (see also Table 1).

recognition rates of all training conditions. The average separation of all two-speaker tracks was almost identical in all clean or noisy conditions (ranging from 52.34° to 52.56°). Hence, the different tracking errors can be attributed to the corruption of noise. In particular, the DOA estimation with its coherence mask suffers from the addition of diffuse noise. Fig. 3 presents an exemplary tracking result of two speakers in clean condition; the figure shows that the particle filter is able to accurately track both speakers even when they cross.

The top panel of Fig. 2 shows the dependency of WRR on the average separation. It is obvious that spatially separated speakers interfere much less than spatially close speakers in high-SNR conditions. At 0 dB SNR the WRR does neither depend on the separation of speakers nor on the kind of training material. In addition, the WRR also depends on the average tracking error. The bottom panel of Fig. 2 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.

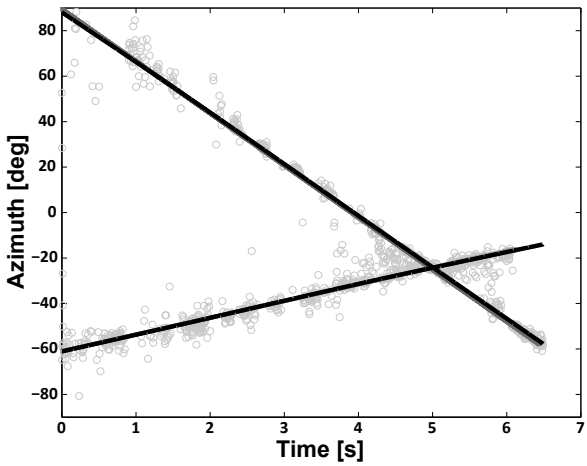


**Fig. 3**. Tracking results of a two-speaker scenario in clean condition. Light grey circles represent the glimpses produced by the binaural model, dark grey lines represent the real azimuth angles of the speakers and the solid black lines show the smoothed estimates obtained by tracking.

## 4. SUMMARY AND CONCLUSION

This study provided an overview of computational auditory scene analysis based on binaural information and its application to a speech recognition task. It was 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. It remains to be seen if this performance gain in anechoic conditions can be validated in real-world scenarios, i.e., in acoustic conditions with strong reverberation, several localized noise sources embedded in a 3D-environment compared to the 2D simulation presented here, or with a changing number of speakers. Follow-up studies in more realistic environments are planned where on the one hand more robust ASR features and on the other hand more information about the acoustic scene is used to improve ASR performance.

# 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] N. Roman and D. Wang, "Binaural Tracking of Multiple Moving Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

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

[7] 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*, pp. 6–11, 2011.

[8] 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.

[9] 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.

[10] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," *EURASIP Journal on Advances in Signal Processing*, , no. 1, pp. 1–11, 2009.

[11] A.-G. Lang and A. Buchner, "Relative influence of interaural time and intensity differences on lateralization is modulated by attention to one or the other cue: 500-Hz sine tones," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2536–2542, 2009.

[12] 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.

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

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

[15] 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.

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

[17] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 357–366, 1980.

[18] 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.