Improving automatic speech recognition in spatially-aware hearing aids

Hendrik Kayser[1,3], *Constantin Spille*[1,3], *Daniel Marquardt*[2,3], *Bernd T. Meyer*[1,3]

[1]Medizinische Physik, Universität Oldenburg, Germany
[2]Signal Processing, Universität Oldenburg, Germany
[3]Cluster of Excellence "Hearing4all", Universität Oldenburg, Germany

{hendrik.kayser, constantin.spille, daniel.marquardt, bernd.meyer}@uni-oldenburg.de

Abstract

In the context of ambient assisted living, automatic speech recognition (ASR) has the potential to provide textual support for hearing aid users in challenging acoustic conditions. In this paper we therefore investigate possibilities to improve ASR based on binaural hearing aid signals in complex acoustic scenes. Particularly, information about the spatial configuration of sound sources is exploited and estimated using a recently developed method that employs probabilistic information about the location of a target speaker (and a simultaneous localized masker) for robust real-time localization. Two different strategies are investigated: straightforward better-ear listening and a multi-channel beamforming system aiming at enhancement of a target speech source with additional suppression of localized masking sound. The latter method is also complemented by better-ear listening. Both approaches are evaluated in different acoustic scenarios containing moving target and interfering speakers or noise sources. Compared to using non-preprocessed signals, we obtain average relative reductions in word error rate of 28.4% in the presence of a localized interfering noise, 19.2% in the case of a concurrent talker and 23.7% in presence of a concurrent talker in spatially diffuse noise. A post-analysis assesses the relation of localization performance and beamforming for improved speech recognition in complex acoustic scenes.

Index Terms: speech recognition, direction of arrival estimation, hearing aids

1. Introduction

Technologies to support users of assistive devices such as hearing aids gain more and more importance in our aging society [1]. This study analyzes the potential of automatic speech recognition (ASR) operating on the multi-channel signals that are used in today's hearing aids, with application scenarios such as textual support for hearing-impaired users in mind.

Psychoacoustic studies report two strategies that hearing-impaired listeners can profit from in complex acoustic scenes: In [2] it was shown that users of bilateral cochlea implants gain significant benefit in speech intelligibility when estimates of the direction of a moving target is used to steer a beamformer. Secondly, human listeners are able to efficiently exploit the inherent spatial filtering characteristics of the head by simply relying on better-ear listening for better speech intelligibility if target and interferer are spatially separated [3, 4].

Both techniques require a robust estimation of direction of arrival (DOA). In this paper, we explore if our recently developed method for accurate DOA estimation that employs probabilistic information about target and masker can be exploited in binaural machine listening. This real-time DOA estimation method [5] is tailored to the binaural sensor setup by statistical, data-driven learning. The advantage of this method is its real-time capability, robustness in noise and in reverberant conditions without the requirement of further assumptions. This is in contrast to several related studies dealing with DOA estimation, which often rely on additional information such as estimates of target signal properties [6, 7, 8], noise statistics [9, 10], room reverberation [11], sensor position and room geometry [12, 13] from multi-conditional training of the machine localization framework [14].

Aiming at improved ASR in hearing aids by exploiting DOA information, we employ better-ear listening and beamforming for signal enhancement. Both methods are evaluated by linking an ASR back-end to the output of our preprocessing scheme. Since the signals are measured directly at the ears with a behind-the-ear hearing aid setup, it is straightforward to apply the better-ear strategy by choosing the favorable lateral channel on the basis of the DOA estimate. In an attempt to further enhance the signal-to-interference ratio (SIR), multi-channel beamforming is utilized in two ways: amplification of a target speech signal and suppression of a localized interfering sound source. In [15, 16], a similar setup was used that exploited binaural preprocessing combined with bio-inspired feature extraction, which has been shown to significantly enhance ASR performance. However, this method relied on a dynamic model that required the integration of localization features over a complete utterance; thus, it did not offer on-line processing capabilities as required for close-to-realtime transcriptions of speech. For ASR in assistive devices, low-latency processing is an important system feature, which potentially can be achieved with the approach presented here.

The remainder of this paper is structured as follows: In Section 2 the methods for sound source localization and tracking, spatial filtering, and ASR are introduced, as well as the data and the experimental setup. In Section 3, the results of the experiments carried out are presented and discussed, for which the signals without any spatial processing serve as benchmark. The conclusions from our work are presented in Section 4.

2. Methods

In the following, a description of the methods used in our processing chain is given. First, the DOA estimation for target and interfering signals using binaural hearing aid output is described. The estimates are passed to a tracking algorithm, whose output is used to choose the better ear and to steer a beamformer towards the target to be preserved and an interferer to be suppressed. The acoustic output of the beamformer is processed by the ASR back-end.

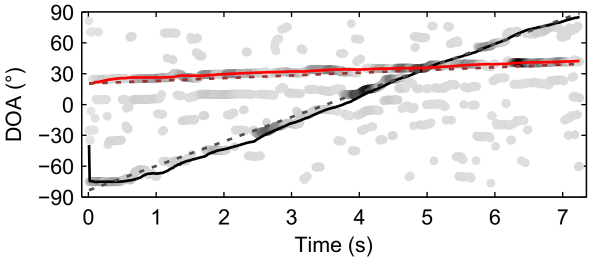

Figure 1: *DOA tracking results in a scenario with two talkers at equal sound level. Gray dots denote DOA estimates with the probability denoted in gray level, solid lines denote the tracks of the target source (red) and the interferer (black) and dashed lines correspond to the ground truth DOA over time.*

2.1. Direction of arrival estimation and source tracking

The direction of arrival of a target speech source and of an interfering sound source, which can be either speech or noise, is estimated using a classification-based approach. The method employed here is described in detail in [5]. It consists of a set of discriminative support vector machine (SVM) classifiers, each one trained to distinguish between presence and absence of a sound source for a given direction, followed by a generalized linear model (GLM) that delivers the probability of sound source incidence from each direction. Short-term generalized cross-correlation functions [17] with phase transform (GCC-PHAT functions) serve as input features. A set of direction-dependent SVM-GLM models was learned that covers the frontal hemisphere of a binaural hearing aid setup with 5° angular resolution. Using noisy, anechoic data in training, purely head-related information is learned and no *a-priori* knowledge of the room characteristics is available to the DOA estimator. DOA probabilities are computed in segments of 10 ms length and subsequently pooled over 200 ms using a sliding average. From the resulting probability functions, the three most likely directions are used for the tracking of the sound sources.

The tracking was carried out with a particle filter algorithm using Kalman filters [18]. The particle filter is able to track multiple sources and to assign each measurement to one of the localized moving sources or to clutter. When a target and a localized interferer have to be tracked, two sets of particles need to be initialized. For the target source, the *a-priori* knowledge of its starting position was used to resolve the target-interferer-ambiguity. Note that this information is not required for accurate DOA, but merely for specifying the correct target. For the second source, particles were initialized equidistantly over the whole hemisphere with 5° resolution, which amounts to 37 particles. Each particle weight was initialized according to the probability of sound source incidence observed during the first 200 ms of the signal, which yields a fast convergence of DOA tracks. Furthermore, DOA probabilities are used in the sequential importance resampling needed during the tracking.

Fig. 1 shows an example of the outputs obtained in the DOA estimation and tracking procedure.

2.2. Spatial filtering approaches

The spatial filtering methods described in the two following sections allow three possible processing strategies: 1) better-ear listening that selects the signal from the favorable side of the head according to the target signal's position; 2) a steered binaural multi-channel beamformer that processes all microphone signals of the hearing aids and is adjusted according to DOA estimates of target and interferer; 3) a combination of 1) and 2), i.e., the better-ear signal from the binaural beamformer output is selected. The difference between 2) and 3) is that in 2) always the same monaural channel is used in contrast to 3), where the channel is selected systematically and adaptively.

2.2.1. Better-ear listening

Better-ear listening exploits the acoustic transmission characteristics of the human head and is implemented in a straightforward way: The signal from the ear that is in the same hemisphere as the target sound source is passed to the ASR back-end. In the transition from the right to the left hemisphere, around a DOA α of 0°, the output signal s_{out} is a linear combination of the signal captured by the front left hearing aid microphone s_l and the front right s_r, respectively:

$$s_{out} = \begin{cases} s_l, & \alpha < -5^\circ \\ s_r, & \alpha > 5^\circ \\ \left(1 - \frac{\alpha+5^\circ}{10^\circ}\right) \cdot s_l + \frac{\alpha+5^\circ}{10^\circ} \cdot s_r, & |\alpha| \leq 5^\circ \end{cases} . \quad (1)$$

2.2.2. Binaural linearly constrained beamformer

In [20] an extension of the LCMV beamformer [19], namely the binaural LCMV beamformer (BLCMV) has been proposed, which aims to minimize the overall noise output power of the left and right hearing aids subject to the constraints of preserving the desired speech source and suppressing the directional interfering source by a prespecified amount determined by the interference rejection parameters η_l and η_r. The optimization criteria for the left and the right hearing aid are

$$\min_{\mathbf{W}_{\{l,r\}}} \ \mathbf{W}^H_{\{l,r\}} \mathbf{R}_v \mathbf{W}_{\{l,r\}} \ \text{ subject to } \mathbf{C}^H \mathbf{W}_{\{l,r\}} = \mathbf{b}_{\{l,r\}} \quad (2)$$

with

$$\mathbf{C} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix}, \quad \mathbf{b}_{\{l,r\}} = \begin{bmatrix} A^*_{\{l,r\}} \\ \eta^*_{\{l,r\}} B^*_{\{l,r\}} \end{bmatrix}, \quad \mathbf{R}_v = P_v \mathbf{\Gamma}.$$

$\mathbf{R}_v$ is the correlation matrix of a diffuse noise field, P_v the power spectral density of the noise component and $\mathbf{\Gamma}$ the spatial coherence matrix. $\mathbf{A}$ and $\mathbf{B}$ are the acoustic transfer functions of the speech component and the directional interference component, respectively. The filter vectors $\mathbf{W}_l$ and $\mathbf{W}_r$ minimizing the expressions given in (2) are equal to [20]

$$\mathbf{W}_{\{l,r\}} = \mathbf{\Gamma}^{-1} \mathbf{C} \left[\mathbf{C}^H \mathbf{\Gamma}^{-1} \mathbf{C} \right]^{-1} \mathbf{b}_{\{l,r\}}. \quad (3)$$

In this paper, we use anechoic head-related transfer functions (equal to those that have been used for the training of the DOA estimator) to represent $\mathbf{A}$ and $\mathbf{B}$ and to calculate the spatial covariance matrix $\mathbf{\Gamma}$. By these means, the HRTFs needed to adjust the beamforming system are selected according to the DOA estimates for the target and the interference. For the interference rejection, we set $\eta_l = \eta_r = 0.2$. Alternatively, expressions for an optimal choice of η_l and η_r are presented in [21], aiming at the maximization of the speech-to-interference+noise ratio (SINR). As this optimization requires an estimate of the correlation matrix of the interference component as well as for the noise component, we adapted the optimization criterion by maximizing the SNR instead of the SINR. Hence, the optimization only depends on $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{\Gamma}$.

In summary, two signal-independent variants of the BLCMV beamformer are used in our experiments: *BLCMV$_{fix}$* with constant $\eta_l = \eta_r = 0.2$ and *BLCMV$_{opt}$* with an adaptive SNR-optimized choice $\eta_l = \eta_r = \eta_{opt}$, whereby η_{opt} is the optimal value relating to the better ear.

2.3. ASR setup

ASR features are calculated by converting the time signals to Mel-Frequency Cepstral Coefficients (MFCCs) [22] with additional cepstral mean and variance normalization. By adding delta and double-delta features, 39-dimensional feature vectors are obtained in 10 ms steps. Feature vectors are spliced with a total temporal context of 11 frames (± 5 frames), subsequently a linear discriminant analysis and maximum likelihood linear transform is performed with a final feature dimension of 40 features per frame. These features are fed into an Hidden Markov Model (HMM) classifier. The HMM classifier has been set up as word models with each word of the vocabulary corresponding to a single HMM. The HMM used sixteen states per word model and five to six Gaussians per mixture and was implemented using the Kaldi speech recognition toolkit [23].

2.4. Speech database and spatial scenes

Speech data was taken from the TIDIGITS speech corpus designed for speaker-independent recognition of connected digit sequences [24], which is divided into a training set (55 male and 57 female speakers) and a test set (56 M, 57 F). Because the utterances are rather short, sequences of the same speaker have been concatenated to create longer sequences suitable for speaker tracking. On average, 23 sequences per speaker with an average duration of 5.9 s were generated. The speech signals were reverberated using binaural room impulse responses (BRIR, [25]) measured with a 6-channel behind-the-ear hearing aid setup in 1 m distance from the sound source in an office room with a reverberation time $T_{60} = 500$ ms. Sound source movement was simulated by partial convolution using 64 ms Hann windows with 50% overlap. This spatialization results in speech sources that move linearly on a half circle in the frontal hemisphere between a pair of starting and end DOA randomly drawn from a -90° to $+90^\circ$ azimuth interval. The moving-speaker signals were mixed with randomly localized stationary speech-shaped noise at SIRs ranging from -5 dB to 20 dB in 5 dB steps. Additionally a set without noise was used, resulting in a total number of 18032 signals (29.5 h) for training.

Three different acoustic scenarios were simulated on the basis of the test speech: The first includes one moving speaker with a fixed localized stationary noise source (*A*) (comparable to the training condition), the second consists of two moving speakers without any additional noise (*B*) and the third scenario is composed of two moving speakers superimposed by spatially diffuse noise (*C*). The same SIRs (-5 to 20 plus *clean*) as for the training set were used, with the exception of *Scenario B* where we did not include the -5 dB condition as we found a masking talker at this SIR to result in ASR chance performance in earlier experiments. In *Scenario C*, interfering talker and noise were mixed at 0 dB before adding the resulting masker to the target.

Initial and final positions were again randomly drawn for all localized sound sources. Parameters of movement were chosen such that the target speaker crossed the localized masker in 50% of all realizations.

2.5. Experiments

For a fair comparison of the spatial processing approaches a separate ASR model was trained for each one of them: *baseline*, i.e., no processing was conducted and a single channel of the left hearing aid was used, *better-ear*, *BLCMV$_{fix}$* and *BLCMV$_{opt}$*. In the training ground truth DOA information was used. ASR was conducted in each of the scenarios described above.

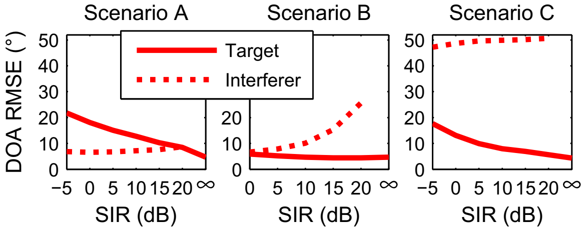

Figure 2: *RMSE of DOA estimates of target and interferer depending on the signal-to-interferer ratio.*

3. Results and discussion

3.1. Localization performance

Fig. 2 shows the localization accuracy expressed in average root mean square error (RMSE) of the estimated angle in reference to the ground truth DOA information. The error was averaged over all realizations for each scenario and SIR condition. In *Scenario A*, the localized interfering noise is localized more accurately than the target speech source even at high positive SIRs. This can be explained by the signal characteristics: In contrast to the modulated speech signal, the noise signal is stationary, i.e., modulation minima of the speech signal can be exploited to localize the noise source also at low levels. At 20 dB SIR, DOA errors for target and interferer are practically identical.

In the *Scenario B*, target and interferer speech signals have similar characteristics. Since the target speech has equal or higher energy compared to the interfering source, it can be localized more accurately. In the 0 dB-SIR condition, the DOAs of both sound sources are estimated with equal accuracy. Accordingly, no systematic advantage of the initially used starting position (cf. Section 2.1) of the target is found.

The third scenario is the most challenging regarding DOA estimation: The target source is masked by a 0-dB mixture of a concurrent speaker and diffuse noise with the same spectral characteristics as in *Scenario A*; hence, the target localization is consistent with *A*. The interfering speech source is to a large extent masked by the noise and additionally by the target, such that DOA estimation is close to chance performance.

3.2. ASR results

ASR results are measured in word error rate (WER). For all preprocessing strategies, we found better-ear listening to lower average WERs in all scenes: The better-ear benefit amounts to 22.1% with better-ear listening as described in Sec. 2.2.1 when expressed in relative improvement in WER of better-ear listening over using always the left ear and averaged over all SIRs $< \infty$ and all scenarios. With beamforming, a relative WER reduction of 4.9% is achieved in case of the BLCMV$_{\text{fix}}$ setup and 8.9% with BLCMV$_{\text{opt}}$. Hence, all results reported in the following include better-ear processing, with exception of the baseline. Comparisons between the setups are drawn in terms of relative improvement (i.e., reduction) in WER over the baseline results. Here, average values computed over improvements in all conditions with an interferer present (SIR $< \infty$) are given.

Absolute WERs obtained from ASR are summarized in Table 1. Regarding the best performing approach, characteristic differences between the scenarios are found: In *Scenario A*, straightforward better-ear listening yields the best ASR results. The averaged relative reduction in WER amounts to 28.4% . With perfect DOA information, the average improvement is 10% higher. In this scenario, the overall ASR performance is highest due to the absence of masking speech.

Table 1: *WERs for each scenario and SIR condition using the different spatial filtering approaches. In the lowermost line average WERs achieved with ground truth DOA information are given.*

	Scenario A			***Scenario B***			***Scenario C***			
SIR (dB)	baseline	better ear	BLCMV$_{\text{fix}}$	baseline	better ear	BLCMV$_{\text{fix}}$	baseline	better ear	BLCMV$_{\text{fix}}$	BLCMV$_{\text{opt}}$
-5	50.03	**46.58**	51.45	–	–	–	69.83	**60.32**	64.48	65.83
0	20.01	**16.08**	25.45	69.96	62.82	**58.09**	52.46	**40.28**	45.73	46.48
5	5.91	**4.44**	10.93	56.89	48.27	**45.66**	36.28	**26.22**	29.45	29.21
10	2.26	**1.42**	4.82	45.43	38.53	**35.99**	24.68	18.23	19.22	**17.97**
15	1.28	**0.78**	2.72	35.93	31.76	**28.87**	17.93	13.24	14.15	**11.91**
20	1.13	**0.65**	2.08	28.84	25.74	**23.33**	12.75	9.51	10.71	**8.17**
∞	1.56	**1.34**	3.14	1.73	**1.56**	3.36	1.69	**1.54**	3.40	4.73
Average	11.74	**10.18**	14.37	39.80	34.78	**32.55**	30.80	**24.19**	26.73	26.33
Average gt	–	**8.76**	12.35	–	34.60	**31.31**	–	23.54	**22.54**	24.50

As could be expected, the concurrent speaker in *Scenario B* results in generally high WERs. In this scenario, the suppression of the interferer becomes particularly important, and consequently the BLCMV$_{\text{fix}}$ beamformer yields the best performance with 19.2% relative WER reduction. Compared to perfect DOA information, this enhancement is only 3% points below the possible maximum. In comparison, straightforward better-ear listening achieves 12.6% relative improvement with estimated DOAs and 13.0% with ground truth DOA.

Scenario C represents a combination of a localized interfering talker and diffuse noise causing a strong degradation of localization performance, especially for the interferer (cf. Fig. 2). This affects ASR performance such that straightforward better-ear listening yields the best results at low SIRs of -5 dB, 0 dB and 5 dB. In all conditions, this methods achieves better results compared to BLCMV$_{\text{fix}}$ and 23.7% relative enhancement in total, which is only 1.4% points below ground-truth-DOA enhancement. Due to the poor localization capabilities regarding the interference, the interferer suppression cannot take its full effect in this scenario with the relatively small fixed value of η_l and η_r. By using the SNR-optimized variant, BLCMV$_{\text{opt}}$, a noticeable reduction of WERs is achieved, in particular at higher SIRs above 10 dB. Note that the use of η_{opt} does make sense only in this scenario, as in there is no diffuse noise present in *Scenarios A* and *B*. If perfect DOA information was available, BLCMV$_{\text{fix}}$ would achieve 31.3% relative reduction in WER. Using DOA estimates, only 16.42% were obtained. By comparison, BLCMV$_{\text{opt}}$ approximates to maximal performance much closer with 22.2% relative reduction in WER compared to 28.6% given perfect DOA information.

ASR results in conditions without interferer show that in such cases beamforming has a detrimental effect, presumably due to distortion of the speech signals and due to increased variability in the training data of the ASR models. If better-ear listening is used in clean conditions, a relative WER reduction of 23.2% averaged over all scenarios is achieved, which can be caused by reduced variability due to a reduced impact of the head-filtering discussed above.

4. Conclusions

In this work, a robust DOA estimation technique with real-time processing capabilities was used to obtain DOA estimates from signals captured with a binaural hearing aid setup, with the advantage of relying on very little *a priori* information, namely only sensor-related characteristics of the setup, such that no specific knowledge about the acoustic scene is required.

Our aim was to improve ASR by exploiting DOA estimation, hence we explored how to optimally apply spatial information in acoustic scenes. Better-ear listening and beamforming (and a combination of those) were tested, which were shown to increase speech intelligibility in human listeners in earlier studies [26, 2]. However, it was unclear if and to what extent a machine listener could profit from such a system configuration. Three acoustic scenarios of different complexity were investigated to cover realistic application scenarios, each including a moving talker masked by different kinds of interfering sources. For each scenario, distinct DOA errors were observed, which strongly depend on the modulation characteristics of the sources. While localization of the target source is sufficiently accurate for better-ear listening (relying solely on target DOA information), interferer localization is increasingly degraded from stationary localized noise (*A*) over a competing speaker (*B*) to competing speaker and diffuse noise (*C*).

When coupled with the probabilistic DOA estimator, the simple method of choosing the better ear signal gave surprisingly good results: For *Scenario A*, choosing the best lateral channel surpasses the performance of the far more complex beamforming approach operating on six channels. Beamforming profits from high-resolution DOA, but also produces signal artifacts with inaccurate estimations that increase ASR error rates as in *Scenario A*. On the other hand, our experiments have shown that accurate DOA estimation is particularly important for suppressing an interferer in scenarios where concurrent speech causes strong masking of the target source, and here beamforming should be employed: in *Scenario B* containing a concurrent speaker without additional noise, the advantage of interferer suppression outperforms better-ear listening. For the most complex *Scenario C*, we see a mixed picture in which better-ear listening should be preferred at low SIRs and in clean conditions, while a beamformer with adaptive optimization of interference rejection should be used at medium SIRs (10 dB to 20 dB).

In future work, our probabilistic DOA algorithm and the beamforming method could be combined with other approaches in machine listening that deliver estimates of additional parameters about the acoustic scene and signals present therein (e.g., room-reverberation, transfer characteristics and noise statistics) to enable an adaption to the acoustic conditions. The results obtained by solely linking probabilistic DOA information with ASR is already encouraging with relative improvements of 19% to 28% compared to unprocessed signals.

5. Acknowledgments

This work was funded by the DFG (Research Unit FOR 1732 “Individualized Hearing Acoustics”, Cluster of Excellence 1077/1‘ “Hearing4All” and the SFB/TRR 31 “The Active Auditory System”). The authors thank Jörg-Hendrik Bach for proofreading this manuscript.

6. References

[1] D. Heger, and I. Holube, “Wie viele Menschen sind schwerhörig (How many people are hearing-impaired)?,” *Zeitschrift fr Audiologie,* 49, pp. 61–70, 2010.

[2] R. M. Baumgïtel, S. Rennebeck, K. Adiloglu, V. Hohmann, B. Kollmeier and M. Dietz, “Evaluation of a steering beamformer algorithm in spatial listening conditions for bilaterally implanted CI-users,” in *18. Jahrestagung deutsche Gesellschaft für Audiologie , March 04–07, Bochum, Germany, Proceedings*, 2015, in press.

[3] D. S. Brungart, and N. Iyer, “Better-ear glimpsing efficiency with symmetrically-placed interfering talkers,” *The Journal of the Acoustical Society of America*, vol. 132, pp. 2545–2556, 2012.

[4] R. Y. Litovsky, A. Parkinson, ad J. Arcaroli, “Spatial Hearing and Speech Intelligibility in Bilateral Cochlear Implant Users,” *Ear and Hearing*, vol. 30, no. 4, pp. 419–431, 2009.

[5] H. Kayser and J. Anemüller, “A discriminative learning approach to probabilistic acoustic source localization,” in *IWAENC 2014 – International Workshop on Acoustic Echo and Noise Control, September 08–11, Antibes, France, Proceedings*, 2014, pp. 100–104.

[6] M.S. Brandstein, “A pitch-based approach to time-delay estimation of reverberant speech,” *WASPAA 1997- Workshop on Applications of Signal Processing to Audio and Acoustics, Proceedings*, 1997, pp. 1–4.

[7] W. Zhang, and B.D. Rao, “A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1913–1928, 2010.

[8] F. Nesta, and M. Omologo, “Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources,” in *WASPAA 2012 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Proceedings*, 2012, pp. 213–216.

[9] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[10] D. Rabinkin, R. Renomeron, J. French, and J. Flanagan, “Estimation of wavefront arrival delay using the cross-power spectrum phase technique,” in *132nd Meeting of the Acoustical Society of America*. 1996.

[11] J. Woodruff, and D. Wang, “Binaural Localization of Multiple Sources in Reverberant and Noisy Environments,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1913–1928, 2012.

[12] A. Deleforge, F. Forbes, and R. Horaud, “Variational EM for binaural sound-source separation and localization,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[13] B. Laufer, R. Talmon, and S. Gannot, “Relative transfer function modeling for supervised source localization,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[14] T. May, S. van de Par, and A. Kohlrausch “A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[15] Spille, C., Dietz, M., Hohmann, V., and Meyer, B. T. (2013). “Using binaural processing for automatic speech recognition in multi-talker scenes,” in *ICASSP 2013 - International Conference on Acoustics, Speech, and Signal Processing , Proceedings*, 2013, pp. 7805–7809.

[16] C. Spille, B. T. Meyer, M. Dietz, and V. Hohmann, Chapter “Binaural scene analysis with multi-dimensional statistical filters,” in *The Technology of Binaural Listening* (Ed. Blauert, J.), Springer, Berlin, 2013.

[17] 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.

[18] S. Särkkä, A. Vehtari, and J. Lampinen, “Rao-Blackwellized particle filter for multiple target tracking,” *Information Fusion*, vol. 8, no. 1, pp. 215, 2007.

[19] E. A. P. Habets, J. Benesty, and P. A. Naylor, “A Speech Distortion and Interference Rejection Constraint Beamformer,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no 3, pp. 854–867, 2012.

[20] E. Hadad, S. Gannot and S. Doclo, “Binaural linearly constrained minimum variance beamformer for hearing aid applications,” in *IWAENC 2012 – International Workshop on Acoustic Echo and Noise Control, September 04–06, Aachen, Germany, Proceedings*, 2014, pp. 1–4.

[21] D. Marquardt, E. Hadad, S. Gannot and S. Doclo, “Optimal binaural LCMV beamformers for combined noise reduction and binaural cue preservation,” in *IWAENC 2014 – International Workshop on Acoustic Echo and Noise Control, September 08–11, Antibes, France, Proceedings*, 2014, pp. 289–293.

[22] S. Davis, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[23] 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,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Proceedings*, 2011, pp. 1–4.

[24] R. Leonard, “A database for speaker-independent digit recognition,” in *ICASSP 1984 – International Workshop on Acoustic Echo and Noise Control, September 08–11, San Diego, California, USA, Proceedings*, 1984.

[25] 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*, vol. 2009, Article ID 298605, pp. 1–11, 2009.

[26] B. A. Edmonds, anf J. F. Culling, “The spatial unmasking of speech: evidence for better-ear listening,” *The Journal of the Acoustical Society of America*, vol. 120(3), pp. 1539–1545, 2006.