INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Probabilistic spatial filter estimation for signal enhancement in multi-channel automatic speech recognition

Hendrik Kayser[1], *Niko Moritz*[2], *Jörn Anemüller*[1]

[1]Computational Audition, Medizinische Physik and Cluster of Excellence Hearing4all
Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany
[2]Fraunhofer IDMT, Project Group for Hearing, Speech, and Audio Technology
D-26129 Oldenburg, Germany

hendrik.kayser@uni-oldenburg.de, niko.moritz@idmt.fraunhofer.de, joern.anemueller@uni-oldenburg.de

Abstract

Speech recognition in multi-channel environments requires target speaker localization, multi-channel signal enhancement and robust speech recognition. We here propose a system that addresses these problems: Localization is performed with a recently introduced probabilistic localization method that is based on support-vector machine learning of GCC-PHAT weights and that estimates a spatial source probability map. The main contribution of the present work is the introduction of a probabilistic approach to (re-)estimation of location-specific steering vectors based on weighting of observed inter-channel phase differences with the spatial source probability map derived in the localization step. Subsequent speech recognition is carried out with a DNN-HMM system using amplitude modulation filter bank (AMFB) acoustic features which are robust to spectral distortions introduced during spatial filtering.

The system has been evaluated on the CHIME-3 multi-channel ASR dataset. Recognition was carried out with and without probabilistic steering vector re-estimation and with MVDR and delay-and-sum beamforming, respectively. Results indicate that the system attains on real-world evaluation data a relative improvement of 31.98% over the baseline and of 21.44% over a modified baseline. We note that this improvement is achieved without exploiting oracle knowledge about speech/non-speech intervals for noise covariance estimation (which is, however, assumed for baseline processing).

Index Terms: robust distant speech recognition, multi-channel signal enhancement, CHiME-3 challenge

1. Introduction

Spatial signal enhancement forms an important component in the construction of robust distant speech recognition systems that employ multi-channel input from possibly noisy target source signals. Source localization methods for multi-channel data have been proposed in the literature in order to infer source-specific spatial information [1]. These commonly rely on the cross-correlation or empirical covariance matrix [2] of microphone signals. Probabilistic approaches have been shown to enhance robustness against noise and reverberation as uncertainty of the estimates is taken into account, e.g., by interpretation of steered response power (SRP, [3]) functions as source probability [4], application of machine learning techniques [5] and probabilistic incorporation of acoustic room characteristics [6]. Subsequent multi-channel spatial filtering based on estimated source positions requires precise knowledge of inter-sensor transfer function differences, commonly in the form of a steering-vector, in order to significantly improve target-to-interference ratio. While a source localization model may already imply an approximation of this, acoustic variation encountered in realistic scenarios may render such knowledge useless for speech enhancement, unless it is possible to estimate the acoustic parameters with sufficient accuracy for each speech utterance. In real-world scenarios, such information has to be estimated with the possible risk of decreased automatic speech recognition (ASR) performance in case of erroneous estimates. The 3rd CHiME speech recognition challenge [7] provides a platform for the development and evaluation of ASR systems on multi-channel acoustic signals. The CHiME-3 audio corpus provides simulated data as well as real-world recordings allowing for a comparison of ASR systems under both conditions.

The system that we propose here, see Fig. 1 for an overview, consists of a signal enhancement front-end that uses spatial information about the target source which is obtained from inter-channel phase differences weighted with spatial source probability. Spatial probability is estimated using a discriminative classification approach to source localization [8], which has been shown to be robust against noise and mismatch between room conditions in test and training data [9]. This spatial information is used as steering vector in an adaptive delay-and-sum (DS) beamformer or in combination with noise statistics in a minimum-variance-distortionless-response (MVDR) beamformer. The processed data are input to a state-of-the-art ASR system that employs amplitude modulation filter bank (AMFB, [10]) features used together with a hybrid deep neural network (DNN), a hidden Markov model (HMM) ASR back-end and subsequent language model (LM) rescoring based on a recurrent neural network (RNN). Results of AMFB features are compared to Mel-frequency cepstral coefficients (MFCCs) with frame splicing. The effects of the different signal enhancement strategies are observed independent of the specific ASR system, i.e, benefit from the proposed signal enhancement approach is still observed in the best-performing ASR system. We compared our approach to spatial filter estimation with the challenge's baseline enhancement where spatial filters are estimated under the assumption of free-field sound propagation. Furthermore, beamforming was conducted with and without use of estimated noise covariance. The results show that on real-world data, the data-driven probabilistic estimates of spatial filters are most successful for ASR. Another noticeable result is that using noise statistics is advantageous only on the simulated data, where precise information is available. On real data, a detri-

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-1340

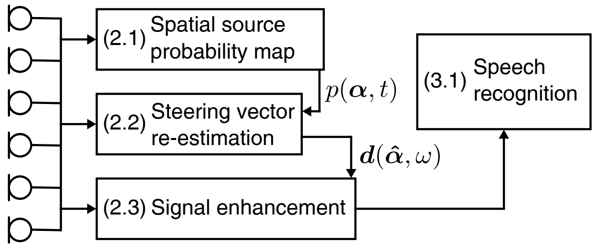
(2.1) Spatial source probability map
(2.2) Steering vector re-estimation
(2.3) Signal enhancement
(3.1) Speech recognition
$p(\boldsymbol{\alpha}, t)$
$\boldsymbol{d}(\hat{\boldsymbol{\alpha}}, \omega)$

Figure 1: *Processing diagram of the multi-channel ASR system with probabilistic re-estimation of steering vectors $\boldsymbol{d}(\hat{\boldsymbol{\alpha}}, \omega)$ based on the spatial source probability map $p(\boldsymbol{\alpha}, t)$ at the maximum a posteriori source position $\hat{\boldsymbol{\alpha}}$. Numbers in brackets indicate sections with details of the respective processing steps.*

ment is observed compared to the simpler signal enhancement approach that uses only target-related spatial information, although knowledge about supposed non-speech intervals is available.

2. Methods

Let the position of a target source be denoted by the generalized location vector $\boldsymbol{\alpha}$ in some arbitrary coordinate system. We here assume that the corresponding physical parameters pertaining to receiver channels $m = 1 \dots M$ are subsumed in frequency-dependent phase variables $\varphi_m(\boldsymbol{\alpha}, \omega)$ that form the steering vector

$$\boldsymbol{d}(\boldsymbol{\alpha}, \omega) = [d_1, \dots, d_M]^T, \quad d_m = e^{i\varphi_m(\boldsymbol{\alpha},\omega)}, \qquad (1)$$

Under the free-field (FF) model, $\varphi_m(\boldsymbol{\alpha}, \omega)$ depends linearly on frequency ω and time-delay $\tau_m(\boldsymbol{\alpha})$ between source and receiver m,

$$\boldsymbol{d}^{\text{FF}}(\boldsymbol{\alpha}, \omega) = [d_1^{\text{FF}}, \dots, d_M^{\text{FF}}]^T, \quad d_m^{\text{FF}} = e^{-2\pi i \omega \tau_m(\boldsymbol{\alpha})}, \qquad (2)$$

which is a good approximation to direct-path sound source propagation from a source to a nearby microphone array, but neglects effects of sound propagation in a real environment and transfer characteristics of the microphones array hardware.

2.1. Probabilistic source localization

The first step for spatial signal enhancement is the estimation of the target source's position relative to the microphone array. The method employed here is a discriminative classification approach to probabilistic sound source localization [8]. It delivers the probability of the sound incidence for a defined set of source locations using short-term generalized cross-correlation functions [11] with phase transform (GCC-PHAT) as input features. The classification part consists of a set of linear discriminative support-vector machines (SVM), trained to distinguish between presence and absence of a sound source for a given position. Each SVM is followed by a generalized linear model (GLM) classifier, that converts SVM decision values into the estimated spatial source probability map $p(\boldsymbol{\alpha}, t)$, providing source presence probability for each position $\boldsymbol{\alpha}$ at time t. In the training procedure, a set of direction-dependent SVM-GLM models is learned on a data set that includes all positions of interest.

2.2. Probabilistic re-estimation of spatial source parameters

While source localization implies knowledge of a spatial source model, this model is in general not sufficiently precise for spatial filtering due variability that is always present under realistic conditions. We, thus, present a novel approach at estimating spatial filters from the multi-channel input signals without explicitly using a model of sound propagation, but still exploiting the source probability map obtained in Section 2.1. The estimated probability map $p(\boldsymbol{\alpha}, t)$ is used as weighting of observed phase differences between all M channels of the sensor setup based on the Cross-power Spectrum Phase (CSP, [12, 13]),

$$\Phi_{nm}(\omega, t) = \frac{x_n^*(\omega, t) \cdot x_m(\omega, t)}{|x_n^*(\omega, t)| \cdot |x_m(\omega, t)|}, \qquad (3)$$

with $x_n(\omega, t)$, $x_m(\omega, t)$ being the short-term Fourier transform (STFT) of the nth, mth input channel, respectively, at time t and frequency ω.

The estimated spatial phase image $\hat{\Phi}_{n,m}$ between channels n and m given source position $\boldsymbol{\alpha}$ is obtained for each position $\boldsymbol{\alpha}$ as the spatial source probability-weighted CSP:

$$\hat{\Phi}_{nm}(\omega|\boldsymbol{\alpha}) = \frac{E\left[\Phi_{nm}(\omega)|\boldsymbol{\alpha}\right]}{|E\left[\Phi_{nm}(\omega)|\boldsymbol{\alpha}\right]|}, \qquad (4)$$

$$E\left[\Phi_{nm}(\omega)|\boldsymbol{\alpha}\right] \approx \sum_t p(\boldsymbol{\alpha}, t)\, \Phi_{nm}(\omega, t), \qquad (5)$$

where n is an arbitrary but fixed reference channel. The resulting probabilistic re-estimation (PR) steering vector is obtained as

$$\boldsymbol{d}^{\text{PR}}(\boldsymbol{\alpha}, \omega) = [d_1^{\text{PR}}, \dots, d_M^{\text{PR}}]^T, \quad d_m^{\text{PR}} = \hat{\Phi}_{nm}(\omega|\boldsymbol{\alpha}). \qquad (6)$$

2.3. Spatial signal enhancement

Spatial signal enhancement is conducted in the frequency domain by multiplying the multi-channel STFT $\boldsymbol{x}(\omega, t)$ of the input signal with a spatial filter vector $\boldsymbol{w}(\boldsymbol{\alpha}, \omega)$ yielding the output $y(\omega, t)$:

$$y(\omega, t) = \boldsymbol{w}^H(\boldsymbol{\alpha}, \omega)\boldsymbol{x}(\omega, t). \qquad (7)$$

Two alternative approaches to spatial filtering are employed here: Delay-and-sum (DS) beamforming uses a filter-vector that is identical to the steering vector,

$$\boldsymbol{w}^{\text{DS}}(\boldsymbol{\alpha}, \omega) = \boldsymbol{d}(\boldsymbol{\alpha}, \omega). \qquad (8)$$

Minimum-variance-distortionless-response (MVDR, [14]) beamforming incorporates the noise covariance matrix $\boldsymbol{R}(\omega)$ into computation of the filter vector according to

$$\boldsymbol{w}^{\text{MVDR}}(\boldsymbol{\alpha}, \omega) = \frac{\left[\boldsymbol{R}(\omega) + r\boldsymbol{I}_M\right]^{-1} \boldsymbol{d}(\boldsymbol{\alpha}, \omega)}{\boldsymbol{d}^H(\boldsymbol{\alpha}, \omega)\left[\boldsymbol{R}(\omega) + r\boldsymbol{I}_M\right]^{-1} \boldsymbol{d}(\boldsymbol{\alpha}, \omega)}, \qquad (9)$$

with a regularization constant r and $\boldsymbol{I}_M$ the identity matrix of size M.

In this study, spatial filters were steered towards the most probable source position $\hat{\boldsymbol{\alpha}}$ obtained by maximum a posteriori estimation

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\text{argmax}} E\left[p(\boldsymbol{\alpha})\right] \approx \underset{\boldsymbol{\alpha}}{\text{argmax}} \sum_t p(\boldsymbol{\alpha}, t) \cdot p_c(\boldsymbol{\alpha}), \qquad (10)$$

where $p_c(\boldsymbol{\alpha})$ is a position-dependent prior that, e.g., assigns higher probability to source positions that are close to the center of the microphone array.

Considering both beamforming approaches and both methods to estimate $\boldsymbol{d}(\boldsymbol{\alpha}, \omega)$, four enhancement systems are obtained as summarized in Tab. 1.

Table 1: *Summary of signal enhancement systems.*

Name	Beamformer	Steering vector
FF-DS	DS (Eq. 8)	free-field (Eq. 2)
PR-DS	DS (Eq. 8)	prob. re-est. (Eq. 6)
FF-MVDR	MVDR (Eq. 9)	free-field (Eq. 2)
PR-MVDR	MVDR (Eq. 9)	prob. re-est. (Eq. 6)

3. Experiments and results

3.1. ASR framework

Most parts of the ASR back-end employed were used for a contribution to the CHiME3 challenge [15]. The ASR back-end consists of a 7-layer hybrid DNN with 2047 sigmoid activation units per hidden-layer. The DNN is pre-trained using stacked restricted Boltzmann machines [16] prior to mini-batch stochastic gradient descent training [17]. The DNN is further discriminatively trained based on the state-level minimum Bayes risk criterion [17]. The AMFB feature extraction analyses temporal dynamics of speech by decomposing critical spectral energies into band-limited amplitude modulation frequency components [10, 18]. MFCC as well as AMFB features are speaker adapted using feature-space maximum likelihood linear regression (fMLLR). As a language model the standard WSJ0 tri-gram with entropy pruning is applied [7] prior to rescoring results using a RNN-based LM [19].

3.2. Baseline signal enhancement system

The baseline enhancement system is provided with the CHiME-3 software package. It consists of an MVDR beamformer that operates on the 6-channel microphone signal captured with the recording hardware used for generating the CHiME3 data set. The target source position is estimated from a non-linear SRP-PHAT pseudo spectrum [20] whose peaks are tracked by the Viterbi algorithm. For the tracking, the SRP-PHAT spectrum is weigthed with the same $p_c(\boldsymbol{\alpha})$ as used in (10) and the transition probabilities between speaker positions are inversely related to the distance between the positions. In the originally provided version, referred to as *base*, the noise covariance matrix for the beamformer is estimated in a time window of 400 ms to 800 ms preceding the utterance. Here, additionally a modified version, *base mod.*, is used in which the estimation is conduced on the time interval of the same length, but after the utterance.

3.3. Speech data

The CHiME-3 data sets contain six-channel speech recordings from different noisy environments, namely public bus transport (BUS), a cafe (CAF), a street junction (STR), and a pedestrian area (PED). Sentences from the Wall Street Journal corpus (WSJ0, [21]) were read by 8 different speakers. In addition, recordings were simulated for each noise scenario by mixing clean WSJ0 recordings with the noise backgrounds using estimated impulse responses of the recording setup. Both of these sets real recordings (*real*) and simulated recordings (*simu*) were divided into a development data set (DEV) and an evaluation set (EVAL) each containing 410 utterances of real recordings and 310 simulated for each noise scenario. The training data set contains 7138 simulated utterances from WSJ0 and 1600 real recordings referred to as the multi-noisy training set. Note that the ASR system used here is trained without any signal enhancement applied to the data, such that the identical acoustic models are used in each experiment. Training was conducted using the Kaldi toolkit [22] provided in the context of the CHiME-3 challenge. All results presented are obtained following rules defined by the CHiME-3 challenge and only the official training and test data sets were used.

Table 2: *Word error rates obtained with the AMFB-RNNLM ASR system in combination with different signal enhancement (SE) systems on the evaluation data set. Average WERs (Avg.) are shown for the simulated data, detailed results for the different noise scenarios, average and mean relative improvement (Rel.) are shown for the real-world recordings.*

	simu	*real*					
SE	Avg.	BUS	STR	CAF	PED	Avg.	Rel.
base	**4.67**	17.72	16.51	11.45	9.97	13.91	—
base mod.	4.68	15.50	13.52	9.88	8.93	11.96	13.70
FF-MVDR	5.52	15.14	14.06	10.03	9.28	12.13	12.18
PR-MVDR	5.33	12.53	10.48	**8.22**	7.94	9.79	28.60
FF-DS	10.65	12.66	10.14	10.18	8.09	10.27	24.27
PR-DS	10.63	**10.99**	**8.44**	9.40	**7.66**	**9.12**	**31.98**

3.4. Training of the localization system

As the probabilistic localization approach requires to learn models for estimation of the spatial source probability map $p(\boldsymbol{\alpha}, t)$ a training data set was compiled from the CHiME-3 data. The clean recorded WSJ0 utterances were mixed with 6-channel noise recordings from all scenarios. The positions of the sources relative to the sensor array were simulated based on the known dimensions of the microphone array. Free-field sound propagation was assumed (which, at least, does not agree with the characteristics of the second array channel which is installed on the back of the tablet used for the recordings). Sound sources were simulated to occur in four grids of 90 cm $\times$ 90 cm with 5 cm resolution in distances of 15 cm, 25 cm, 35 cm and 45 cm to the sensor array plane corresponding to the search space of the baseline localization method. In total 4 sets of $19^2 = 361$ models each were learned - one for each grid separately. For the estimation of $\hat{\boldsymbol{\alpha}}$, a probability map was estimated for each grid independently and the grid with the highest time-integrated probability was chosen afterwards.

3.5. Results

A summary of recognition results is shown in Figure 2. Word error rates (WER) for the simulated and the real data from the development and the evaluation data sets are shown averaged over all noise scenarios. Results are displayed for the MFCC-based ASR system (1st group of six bars in each panel), the AMFB features (2nd group) and the AMBF system with RNNLM rescoring (3rd group).

3.5.1. ASR systems

Regarding the performance of the different ASR systems, AMBF features outperform the MFCC system.

The relative WER improvement, averaged across all noise scenarios and signal enhancement approaches, of AMFB features compared to the MFCC setup amounts to 6.37% (*simu*) and 8.95% (*real*), respectively, on the development test data and to 14.53% (*simu*) and 12.80% (*real*) using the evaluation test data, respectively. The improved ASR performance of AMFB features compared to the MFCC plus frame splicing approach is the result of a better generalization effect by using the AMFB [18]. A further improvement is achieved by the RNNLM rescor-

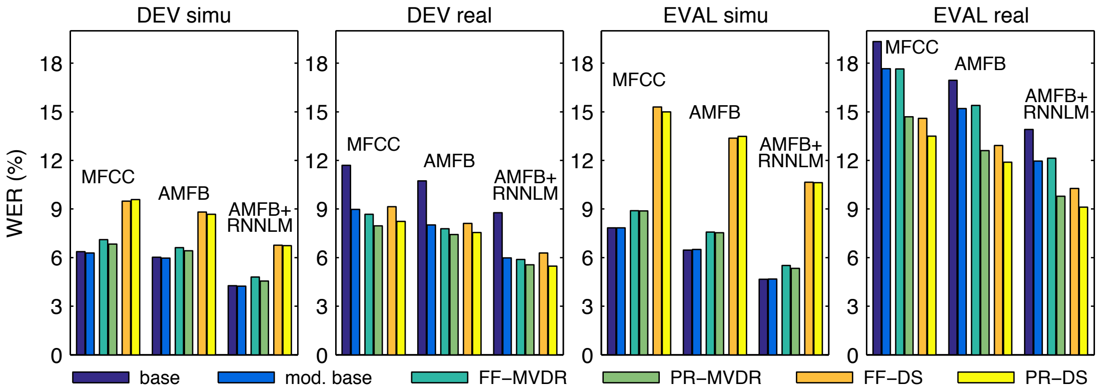

Figure 2: *Word error rates (WER) achieved on the simulated (panels "DEV simu" and "EVAL simu") and real (panels "DEV real" and "EVAL real") CHiME-3 development ("DEV") and evaluation ("EVAL") data sets. Results are shown for three DNN-based ASR systems ("MFCC", "AMFB", "AMFB+RNNLM" groups of six results in each panel) and the six different signal enhacement methods (see legend, Table 1 and text for explanation).*

ing, which consistently yields an average relative improvement on all data sets of 27.05% and 23.90% on DEV *simu* and *real*, 25.42% and 21.13% on EVAL.

3.5.2. Signal enhancement approaches

The average WERs achieved with the different signal enhancement (SE) methods shown in Figure 2 can be grouped into groups of two by the underlying localization and beamforming methods. The first two bars in each group of six are the baseline SE system (base) and the modified baseline with noise covariance estimation after the utterance (base mod.). For the simulated data results of both variants of the baseline are almost equal, average relative differences are less than 1.2% for all ASR systems and 0.87% on average over all ASR systems. On the *real* data sets, significant differences occur: Estimation of noise covariance after the utterance achieves a relative improvement of MFCC: 24.08%, AMFB: 25.98% and AMFB+RNNLM: 32.04% on the DEV data and 8.33%, 10.13% and 13.70%, respectively on the EVAL set.

Similar observations are made regarding the method for estimating the steering vectors in both the other groups of two, MVDR and DS. Relative differences between FF-MVDR and PR-MVDR are small for simulated data, average values over ASR systems are DEV: 3.96% and EVAL: 1.23%, while larger differences are found for the *real* DEV set 5.90% and the largest for *real* EVAL: 17.45%. In the DS group the average relative differences for *simu* are DEV: 0.49% and EVAL: 0.58% and for *real* DEV: 9.61% and EVAl: 8.30%.

In summary, on both simulated data sets the baseline methods yields the best ASR performance. This is potentially due to the (slight) adaptation of the steering filters according to the source tracking included in the baseline SE system during a single utterance, while in the proposed SE filters are kept fixed at the most probable source position for the duration of the utterance. The DS approaches, that do not exploit noise statistics, are far behind MVDR-based systems. For the real-world recordings the picture is clearly different. The time interval in which the noise covariance is estimated has a noticeable effect on the ASR performance. This is supposedly due to unlabeled samples of the target speakers voice that accidentally occur before the actual utterance starts, e.g., by aborted erroneous recordings. These are more likely to precede an utterance than occur afterwards. The modified baseline approach, FF-MVDR and FF-DS are on a similar performance level on the DEV *real* data and a slight advantage is taken from the PR-based spatial filters. On the EVAL *real* data set using spatial filters obtained with the proposed probabilistic phase re-estimation yields an noticeable enhancement of ASR performance in combination with noise information (PR-MVDR vs. FF-MVDR) of 18.80% for the best-performing ASR system AMFB+RNNLF. Without noise information (PR-DS vs. FF-DS) the relative improvement amounts to 10.73%. Thereby PR-DS achieves the best overall performance on the real-world evaluation data. Detailed absolute results in WERs for this data set are shown in Tab. 2 as well as average WERs for the simulated evaluation data.

With the CAF noise scenario as the only exception, the PR-DS approach, using only target-related spatial information for signal enhancement, achieves the best results. In the CAF scenario, PR-MVDR clearly outperforms the FF-based approaches.

4. Conclusions

(1) Probabilistic estimation of source-related spatial filter characteristics without the assumption of free-field sound propagation consistently yields higher ASR performance on real data recordings.

(2) On the other hand, experiments with simulated data show that reliable ("oracle") information about noise statistics may yield significant benefits over using target information only. Hence, robust estimation of noise statistics remains important as its lack may severely degrade ASR performance and may explain why the signal enhancement method proposed here is most successful on real data.

(3) AMFB acoustic features prove to be robust under the spatial filtering approaches investigated here and provide improved generalization capabilities from development to evaluation test data set. Incorporation of prior information about human speech processing through modulation frequency decomposition [18] is hypothesized to be one factor contributing to the improvement over entirely data-driven temporal processing approaches.

5. Acknowledgments

H.K. and J.A. acknowledge funding by the German Research Foundation (DFG) through projects FOR 1732 "Individualized Hearing Acoustics" and SFB/TRR 31 "The Active Auditory System". N.M. acknowledges funding by the Federal Ministry of Education and Research (BMBF) and the European Commission (project KNOTS).

6. References

[1] C. Blandin, A. Ozerov, and E. Vincent, “Multi-source TDOA estimation in reverberant audio using angular spectra and clustering,” *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, aug 2012.

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

[3] J. Dibiase, “A High-Accurate, Low-Latency Technique for Talker Localization in Reverberation Environments Using Microphone Array,” Ph.D. dissertation, Brown University, 2000.

[4] Y. Oualil, M. Magimai, F. Faubel, and D. Klakow, “Joint Detection and Localization of Multiple Speakers Using a Probabilistic Interpretation of the Steered Response Power,” in *Statistical and Perceptual Audition Workshop*, 2012.

[5] 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.

[6] H. Kayser, V. Hohmann, S. D. Ewert, B. Kollmeier, and J. Anemüller, “Robust auditory localization using probabilistic inference and coherence-based weighting of interaural cues.” *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2635–2648, 2015.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ Speech Separation and Recognition Challenge: Dataset, task and baselines,” in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015, pp. 504–511.

[8] 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*, 2014, pp. 100–104.

[9] H. Kayser, C. Spille, D. Marquardt, and B. T. Meyer, “Improving Automatic Speech Recognition in Spatially-Aware Hearing Aids,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 175–179, 2015.

[10] N. Moritz, J. Anemüller, and B. Kollmeier, “An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 11, pp. 1926–1937, 2015.

[11] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[12] M. Omologo and P. Svaizer, “Acoustic event localization using a crosspower-spectrum phase based technique,” *Proceedings of ICASSP ’94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. ii, no. 2, pp. II/273–II/276, 1994.

[13] ——, “Use of the crosspower-spectrum phase in acoustic event location,” *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 288–292, 1997.

[14] X. Mestre and M. A. Lagunas, “On diagonal loading for minimum variance beamformers,” in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2003*, 2003, pp. 459–462.

[15] N. Moritz, S. Gerlach, K. Adiloglu, J. Anemüller, B. Kollmeier, and S. Goetze, “A chime-3 challenge system: Long-term acoustic features for noise robust automatic speech recognition,” in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015, pp. 468–474.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[17] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. 1, pp. 2345–2349, 2013.

[18] N. Moritz, B. Kollmeier, and J. Anemüller, “Integration of optimized modulation filter sets into deep neural networks for automatic speech recognition,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, (submitted) 2016.

[19] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, “Recurrent Neural Network based Language Model,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. September, 2015, pp. 1045–1048.

[20] B. Loesch and B. Yang, “Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions,” *Latent Variable Analysis and Signal Separation*, pp. 41–48, 2010.

[21] D. B. Paul and J. M. Baker, “The design for the wall street journal-based csr corpus,” in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT ’91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362.

[22] 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*. IEEE Signal Processing Society, 2011.