BINAURAL SIGNAL PROCESSING FOR ENHANCED SPEECH RECOGNITION ROBUSTNESS IN COMPLEX LISTENING ENVIRONMENTS

Hendrik Meutzner[1]*, Anton Schlesinger*[2]*, Steffen Zeiler*[1]*, Dorothea Kolossa*[1]

[1]Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[2] GAMPT mbH, Hallesche Strasse 99F, 06217 Merseburg, Germany

hendrik.meutzner@rub.de, anton.schlesinger@gampt.de, steffen.zeiler@rub.de, dorothea.kolossa@rub.de

1. ABSTRACT

This paper addresses the problem of automatic speech recognition (ASR) in the presence of room reverberation, speaker movements and highly non-stationary background noise on the basis of binaural microphone recordings. Investigations are conducted for Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge, posing a small-vocabulary task that requires the recognition of a short keyword control sequence. In order to cope with the severely noisy recordings, we extend our beamforming approach with observation uncertainties from the first CHiME challenge by adding a second, parallel feature extraction based on a binaural time-frequency mask. The output signals of both front-ends, the beamformer and the binaural speech enhancement, are fed to separately trained recognition models. Finally, a late fusion by recognizer output voting error reduction (ROVER) is applied to combine the separate recognition outputs into a jointly optimal transcription. Based on this multi-stage approach, a relative keyword error rate reduction of more than 55 % is achieved compared to the best baseline result of the 2nd CHiME Challenge.

Index Terms— automatic speech recognition, binaural speech processing, cepstral smoothing, time-frequency mask

2. INTRODUCTION

A major challenge to truly ubiquitous use of voice control is still posed by the varying, reverberant noise condition prevalent in everyday living environments [1, 2].

Superdirective beamformers, blind source separation and the multi-channel Wiener filter are three popular spatial sampling schemes that allow for an effective speech enhancement when many microphone channels are available [3, 4]. Considering the binaural recordings of the challenge, however, basic beamforming techniques will only allow for moderate SNR gains. Source separation methods do allow for great gains in signal quality, but the best-performing strategies are typically complex and well-attuned to the task, see, e.g., [5, 6].

The following paper suggests low-complexity alternatives, which are still in tune with the binaural nature of the recordings. It extends our approach from the one used in the 2011 PASCAL CHiME Speech Separation and Recognition Challenge [7], where delay-and-sum beamforming was coupled by observation uncertainty techniques, with a binaural front-end. The utilized front-end is based on time-frequency masking, utilizing the inter-aural phase and level differences. Both techniques, the delay-and-sum beamformer and the binaural front-end, are used in parallel, and recognition outputs are fused by a subsequent late-integration approach, merging the recognition outputs and their associated confidences, to produce one unified recognition hypothesis.

The following description will mainly focus on the newly implemented binaural front-end, which is described in detail in Section 3. Section 4 gives some brief details on the CHiME database and the training method used to obtain the front-end parameters for this dataset. After describing the relevant features of the recognition system in Section 5, we present results on Track 1 of the CHiME dataset and draw conclusions in Sections 6 and 7, respectively.

3. BINAURAL SPEECH ENHANCEMENT

A special version of the multi-channel Wiener filter, which mimics some of the human capability for auditory scene analysis [8], offers a low-complexity and efficient means for enhancing the binaural mixture available in the 2nd CHiME Challenge. Its effectiveness results from the exploitation of the inter-aural transfer function. Based on inter-aural phase and level differences — IPD and ILD, respectively, — almost every direction of sound incidence has a unique and frequency-dependent identifier. From a continuous analysis of the IPD and ILD in a suitable transform domain, such as the short-time Fourier transform (STFT), an amplitude weighting function is generated by comparing the binaural parameters of the noisy mixture signal with *reference* binaural parameters that were obtained in a preceding training phase, which is discussed detailed in Sec. 3.1.

Exploiting the statistics of binaural parameters has shown to result in considerable gains of noise suppression, even in difficult acoustic environments [9, 10, 11]. Generally, there are two methods of including the statistics of the binaural signal. One is histogram-based, the other estimates the parame-

ters of distributions using a Gaussian Mixture Model (GMM). Although the GMM is an efficient approach to handle the abundant data of changing acoustic scenes, histogram-based statistical filtering often allows for higher SNR gains and was therefore chosen for the task at hand [12].

3.1. Derivation of the Binaural Processor

Let $s^{\ell}(n)$ be the band-limited noisy mixture at a sampling frequency $f_{\mathrm{s}} = 16$ kHz at the left input of the binaural speech processor. Using a Hann window h_n, the signal is partitioned into overlapping frames with a frame shift Δp. Subsequently, the STFT of the signal is calculated with an FFT

$$S^{\ell}_{d,m} = \sum_{n=0}^{N_{\mathrm{D}}-1} s^{\ell}_{m\Delta p+n} h_n e^{-j2\pi d \frac{n}{N_{\mathrm{D}}}}, \tag{3.1}$$

where d, m and N_{D} are the frequency index, the frame index and the FFT length, respectively. The noisy mixture at the right input is calculated in the same way, which results in $S^{r}_{d,m}$. Thereafter, the power spectral densities (PSD) are estimated. This is done through a modulus and recursive first order filtering method, known as Welch's averaging method [13]. For the signals at the left and right ear, the Welch method is computed as

$$\begin{bmatrix} \Phi^{\ell}_{d,m} \\ \Phi^{r}_{d,m} \end{bmatrix} = \alpha \begin{bmatrix} \Phi^{\ell}_{d,m-1} \\ \Phi^{r}_{d,m-1} \end{bmatrix} + (1-\alpha) \begin{bmatrix} |S^{\ell}_{d,m}|^2 \\ |S^{r}_{d,m}|^2 \end{bmatrix}, \tag{3.2}$$

where the smoothing factor α is given by

$$\alpha = \exp(-\Delta p/(\tau f_{\mathrm{s}})), \tag{3.3}$$

with τ being the time constant. Furthermore, the cross power spectral density is calculated as

$$\Phi^{\ell r}_{d,m} = \alpha \Phi^{\ell r}_{d,m-1} + (1-\alpha) S^{\ell}_{d,m} \bar{S^{r}}_{d,m}, \tag{3.4}$$

in order to infer binaural temporal differences. Here, $\bar{S^{r}}$ is the complex conjugate of S^{r}. Subsequently, the IPD is computed by

$$\Delta\varphi_{d,m} = \angle \left(\Phi^{\ell r}_{d,m} \right), \tag{3.5}$$

where the symbol $\angle$ denotes the angle in radians and the ILD is found via

$$\Delta L_{d,m} = 10 \log_{10} \frac{\Phi^{\ell}_{d,m}}{\Phi^{r}_{d,m}}. \tag{3.6}$$

For generating bivariate distributions of both directional fine-structure parameters, the binaural feature vector is defined as

$$\Delta_{d,m} = \left[\Delta\varphi_{d,m} \; \Delta L_{d,m} \right]. \tag{3.7}$$

Noise suppression in the binaural speech processor is based on the posterior estimate of the target given a binaural feature vector at each time-frequency bin

$$P_d(\phi_{\mathrm{t}}|\Delta_{d,m}) = \frac{P_d(\Delta_{d,m}|\phi_{\mathrm{t}}) P_d(\phi_{\mathrm{t}})}{\sum_{\phi} P_d(\Delta_{d,m}|\phi) P_d(\phi)}, \tag{3.8}$$

with ϕ being the source azimuth and $\phi_{\mathrm{t}} \in \mathbb{T}$, which denotes a set of target directions of the direct sound and room reflections. Harding et al. showed that this equation can be approximated by the division of two histograms

$$P_d(\phi_{\mathrm{t}}|\Delta_{d,m}) \approx \begin{cases} \frac{H^{\mathrm{t}}_d(\Delta_{d,m})}{H^{\mathrm{a}}_d(\Delta_{d,m})}, & \text{if} \quad H^{\mathrm{a}}_d(\Delta_{d,m}) > \zeta \\ 0, & \text{else} \end{cases}, \tag{3.9}$$

where H^{t}_d and H^{a}_d are the histograms of the labeled target signal and of the noisy mixture, respectively [9]. ζ is a threshold to prevent faulty estimations from insufficient statistical data and numerical noise. Consequently, after the division of these distributions, the filter gain can be read from a look-up table by using $\Delta_{d,m}$.

A soft mask is obtained as a basis for weighting the STFT representation of the noisy mixture signal. For the next steps it is convenient to express the soft mask as

$$\mathcal{M}^{\mathrm{e}}_{d,m} = \max \left(P_d(\phi_{\mathrm{t}}|\Delta_{d,m}), A \right), \tag{3.10}$$

where A is a flooring parameter that allows for balancing the trade-off between noise suppression and signal distortion.

3.2. Post-Processing of Weighting Mask

It is intuitive that the mask of Eq. (3.10) tends to result in a certain degree of non-stationary signal artifacts that are commonly referred to as musical noise. Previous studies have shown that a temporal smoothing of spectral masks in the cepstral domain reduces the effect of musical noise [14, 15] by which the perceived signal quality can be improved significantly. We found that a moderate smoothing of the mask results in improved ASR performance for the specific task of the challenge. The cepstral transform of the mask reads

$$\mathcal{M}^{\mathrm{e,c}}_{q,m} = \frac{1}{N_{\mathrm{Q}}} \sum_{n=0}^{N_{\mathrm{Q}}-1} \ln \left[\mathcal{M}^{\mathrm{e}}_{n,m} \right] \mathrm{e}^{j2\pi q \frac{n}{N_{\mathrm{Q}}}}, \tag{3.11}$$

where q denotes the cepstral index and N_{Q} is the total number of cepstral coefficients. Then, a first-order recursive smoothing is applied frame-wise to the cepstral representation of the mask

$$\widetilde{\mathcal{M}}^{\mathrm{e,c}}_{q,m} = \beta_q \, \widetilde{\mathcal{M}}^{\mathrm{e,c}}_{q,m-1} + (1-\beta_q) \, \mathcal{M}^{\mathrm{e,c}}_{q,m}, \tag{3.12}$$

where β_q is a quefrency dependent smoothing constant that is separately adjusted for different regions in the cepstrum. The smoothing constants β_q should be chosen such that those regions that are crucial for speech intelligibility are not distorted by the temporal smoothing [15, 11]. The smoothing constants have been determined empirically on the development set so as to optimize the ASR performance. For $f_{\mathrm{s}} = 16$ kHz and $N_{\mathrm{Q}} = 512$, their respective values are given by

$$\beta_q = \begin{cases} 0, & q \in [0,7] \cup [505,511] \\ 0.5, & q \in [8,15] \cup [497,504] \\ 0.9, & q \in [16,496] \end{cases}. \tag{3.13}$$

Here, the smoothing constants have been set to zero for the lower cepstral coefficients in order to maintain the spectral envelope of $\mathcal{M}^{\mathrm{e}}_{n,m}$ [14]. The smoothed mask of Eq. (3.12) is transformed back to the frequency domain by applying the inverse transform of Eq. (3.11)

$$\widetilde{\mathcal{M}}^{\mathrm{e}}_{d,m} = \exp\left(\sum_{q=0}^{N_{\mathrm{Q}}-1} \widetilde{\mathcal{M}}^{\mathrm{e,c}}_{q,m} \mathrm{e}^{-j2\pi d\frac{q}{N_{\mathrm{Q}}}}\right). \qquad (3.14)$$

3.3. Application of Mask and Resynthesis

The smoothed version of the weighting mask of Eq. (3.14) is multiplied with the STFT representation of the noisy input signal

$$\begin{bmatrix} \check{S}^{\ell}_{d,m} \\ \check{S}^{r}_{d,m} \end{bmatrix} = \widetilde{\mathcal{M}}^{\mathrm{e}}_{d,m} \begin{bmatrix} |S^{\ell}_{d,m}|e^{j\angle\left(S^{\ell}_{d,m}\right)} \\ |S^{r}_{d,m}|e^{j\angle\left(S^{r}_{d,m}\right)} \end{bmatrix}, \qquad (3.15)$$

so the original phase is left unchanged. Here, $\check{S}^{\ell}_{d,m}$ and $\check{S}^{r}_{d,m}$ denote the STFT representations of the noise-suppressed output signal for the left and the right channel, respectively. As a final processing step, the waveform of the output signal is reconstructed through an inverse STFT, which is then used for the subsequent ASR.

4. EXPERIMENTAL SETUP

4.1. Description of the Database

Our evaluation is based on Track 1 of the 2nd CHiME Challenge and a full description of the challenge set-up is given in [16]. As in the first CHiME Challenge [7], the clean speech signals originate from the Grid corpus [17], which consists of 34 different speakers. The clean speech signals have been filtered with a set of binaural room impulse responses (BRIRs), computed from the recordings of an artificial head, in order to simulate room reverberation and speaker movements. In a last step, the filtered speech signals have been mixed with highly non-stationary background noise, which was recorded in a family living room. The mixing procedure has been designed to yield six different SNR conditions between -6 dB and 9 dB without rescaling the signal amplitudes. More details about the mixing process can be found in [7]. The challenge provides two datasets that can be used for training and tuning, i.e., a training set, consisting of 500 signals from each of the 34 speakers, and a development set, consisting of a total of 600 signals. Both sets include the aforementioned mixture signals and, separately, both the corresponding reverberated and noisy signals. A third dataset, the test set, may only be used for a final evaluation of the system.

4.2. Training the Binaural Processor

For calculating the a-posteriori probability of target presence for each time frame and frequency, as defined in Eq. (3.9), feature histograms for the target speech as well as for the noisy mixture need to be generated. As a means to train the classifier in a supervised fashion, Harding et al. suggested the following ideal binary mask definition for labeling the data

$$\mathcal{M}^{\mathrm{b}}_{d,m} = \begin{cases} 1, & \text{if} \quad 10\log_{10}\frac{\Phi^{s}_{d,m}}{\Phi^{n}_{d,m}} > \xi \\ 0, & \text{else} \end{cases}, \qquad (4.16)$$

where Φ^s and Φ^n are the PSD of the speech signal and the noise signal, respectively, and ξ is a local SNR threshold that categorizes speech and noise [9]. One-channel signals of speech and noise are generated by summing both ear signals, prior to the calculation of Φ^s and Φ^n.

Accordingly, the ideal binary mask is applied to isolate binaural features that correspond to dominant portions of the target signal and these directional parameters (training features) are binned into the target histograms H^{t}_d. Histograms of the noisy mixture, H^{a}_d, on the other hand, are directly binned from the mix of binaural training features. Bivariate histograms were sampled with a grid of 100 $\times$ 100 bins. The parameter ranges were $\pm\pi$ and ± 40 dB for $\Delta\varphi$ and ΔL, respectively.

In order to derive values for Eq. (4.16), the reverberated speech signals of the training set have been mixed with randomly chosen segments of the long-term noise recordings that were provided in addition to the premixed signals. Here, the raw reverberated speech signals have been used to produce new mixture signals at the same SNR levels as they were provided by the challenge. The mixing has been realized by rescaling the amplitudes of the respective noise segments to the desired level, where the SNR has been defined as in [18]. We have generated speaker-dependent histograms for all available speakers, each composed from 500 utterances. Then, a final histogram look-up table has been obtained by averaging these speaker-dependent distributions. The same binaural speech processor, tuned on this set, has been used throughout all the experiments. Algorithmic parameters, used in the training of the binaural speech processor are summarized in Table 1.

Table 1. Training parameters of the binaural processor.

N_{D}	N_{Q}	Δp	τ	A	ζ	ξ
512	512	128	8 ms	0.02	5	-8 dB

5. KEYWORD RECOGNITION

5.1. Speech Recognition System

The JASPER System, introduced in [19], has been used for the majority of the experiments. It has been successfully evaluated on this task in the context of the first CHiME challenge, where its properties were already described in detail [20].

Regarding recognition there are no large differences between JASPER and HTK which both operate in a token passing framework. The main differences are the topology, where

silence models are added at the beginning and end of each sentence and where three states are used per phoneme, the initialization, which occurs per speaker, rather than adapting from a speaker-independent initial model, and the implementation of Baum-Welch training. Output density functions are typically represented by Gaussian mixture models with full covariance matrices, and a mixture-split follows the direction of the first eigenvector with an adaptable split distance. Repeated iterations of EM training are subsequently used to find optimal parameters for the density functions. After two iterations of mixture splitting and parameter optimization, the three-component full-covariance model is projected onto a diagonal model and optimized in a final round of EM iterations. This training strategy has proven advantageous compared to the standard HTK training approach in [20], and was again helpful on the current data set.

5.2. Mixed training

To reduce the mismatch between models and noisy data, a mixed training set was created by augmenting the provided training set. New samples were obtained as a weighted sum of samples from the noise-only database and the reverberated data set for various SNRs. This has given a sevenfold increase in the available amount of training data for mixed training, as compared to the provided isolated training set.

5.3. Feature Extraction

Throughout recognition, the prevalent MFCC features are used, arriving at 39-dimensional feature vectors $\mathbf{x}_m$ composed of the 13 static MFCCs and their delta and acceleration values, calculated as in [20].

Differences exist only in the various pre-processing methods that are used to estimate the features for the JASPER backend from the binaural signal.

The first system, abbreviated by BP for "Binaural Processing" in the following, uses the described binaural processor. It computes a two-channel time-domain signal from the binaural processing front-end, sums the two channels, and calculates 39-dimensional MFCC features from this time-domain representation. A slightly modified system, referred to as BP+LDA, additionally projects the 39-dimensional MFCCs onto a lower-dimensional subspace by means of a linear discriminant analysis (LDA) as further explained in Section 5.4.

The third system is used without modification from [21]. Referred to as "Beamforming + Uncertainty Propagation", or in short "BF+UP", in the following, it uses a delay-and-sum beamformer instead of the binaural processor to extract features and their associated uncertainties in the STFT domain. Subsequently, it transforms these features into the MFCC domain by means of uncertainty propagation. Finally only the 39-dimensional MFCCs are passed to the recognizer backend, because the propagated uncertainties were found not to be very informative for this particular dataset.

5.4. Linear discriminant analysis

The full-covariance models at the intermediate training stages also allow the use of a LDA as a final feature extraction stage. When using LDA, after having trained a full-covariance single-component hidden Markov model (HMM), we find the maximally discriminative projection matrix for the data, the so-termed *LDA matrix* $\mathbf{W}$, by a generalized eigenvector decomposition. As a result, the transformed data

$$\mathbf{x}'_m = \mathbf{W}\mathbf{x}_m \tag{5.17}$$

possesses the maximal ratio between inter- and intra-class covariance. Here, as we perform LDA on the basis of single-mixture full-covariance models, the term *class* is equivalent to one HMM state, so that we actually maximize discrimination between the HMM states of the transformed data model. In the following experiments, this projection was onto 37-dimensional feature vectors $\mathbf{x}'_m$, the same dimensionality that been used in the initial CHiME challenge. While this LDA was not successful in directly enhancing the recognition performance here – unlike in the first CHiME challenge [21] – it did provide improvements when used as one more alternative system in the context of late integration, as described in the next section.

5.5. Late integration of recognizer outputs

During the experiments, numerous results, in form of word graphs with associated scores, have been produced by the recognition systems for different features. To combine these multiple speech recognition outputs into a single one, we employ Recognizer Output Voting Error Reduction (ROVER) [22] in the final step. The fusion enables us to achieve a lower error rate than any of the individual systems alone.

6. RESULTS

Results for the development and test set data are shown in Table 2. Whereas the first block gives results for the standard HTK configuration, including the official do-nothing baseline for the system trained on the isolated set, the other parts show the JASPER results, obtained for noisy and mixed training. The final two blocks, once for the development and once for the test set, contain the results for the ROVER combination of different JASPER outputs.

The results indicate that an exclusive use of the binaural processor yields an average accuracy increase of 8.1 % for the provided reference HTK system. Especially for negative SNRs, the benefit of signal preprocessing becomes clearly visible where the accuracy can be improved by more than 12 %. Employing the JASPER system for ASR shows an improvement of the results as compared to HTK-based recognition. A significant performance gain is achieved for mixed JASPER training, where a late fusion of different system configurations by the ROVER scheme yields a further improvement of approximately 2 % for both evaluation sets.

Table 2. Keyword recognition accuracy in percent evaluated on the development and test set for varying SNRs, and their respective averages. Best results are marked in bold. Systems are trained either on the isolated data set ("isolated") or on the augmented data set ("mixed") for mixed training. The results for the first two systems, BASELINE and the reference HTK system trained with the binaural processor (BP), are shown for comparison. The remaining results are all for different JASPER systems. One system (BF+UP) uses the delay-and-sum beamformer front-end together with uncertainty propagation. There are two systems that use the binaural processor either alone (BP) or in conjunction with a linear discriminant analysis (BP+LDA). The last group of results belongs to a combination of the aforementioned systems (ROVER). Superscript numbers indicate which systems are used for the recognizer output voting error reduction in the late integration scheme.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	average	training set	method
development set	49.67	57.92	67.83	73.67	80.75	82.67	68.75	isolated	BASELINE
	63.67	70.25	75.92	81.67	84.00	85.42	76.82	isolated	HTK BP
	68.33	73.67	79.42	82.25	85.42	87.42	79.58	isolated	JASPER BP [3)]
	64.75	69.50	77.58	83.00	85.67	86.00	77.75	isolated	JASPER BF+UP
	72.08	78.75	82.92	88.42	91.67	92.17	84.33	mixed	JASPER BP [1)]
	68.67	77.33	83.17	88.50	91.67	92.08	83.57	mixed	JASPER BF+UP [2)]
	71.25	77.67	83.00	88.33	90.50	92.25	83.83	mixed	JASPER BP+LDA [4)]
	72.17	79.33	84.42	89.08	92.75	92.50	85.04	mixed	ROVER [1,2]
	74.33	79.83	85.25	88.58	92.58	92.41	85.50	both	ROVER [1,2,3]
	75.00	**80.67**	**85.83**	**90.00**	**92.76**	**93.50**	**86.29**	both	ROVER [1,2,3,4]
test set	71.58	75.33	81.58	86.83	88.00	88.17	81.92	isolated	JASPER BP
	66.08	72.58	80.83	82.58	86.42	86.50	79.16	isolated	JASPER BF+UP
	75.33	78.17	85.08	88.83	91.25	92.83	85.24	mixed	JASPER BP
	71.67	76.42	85.17	88.17	91.42	92.92	84.30	mixed	JASPER BF+UP
	72.17	77.83	84.83	89.33	91.83	92.33	84.72	mixed	JASPER BP+LDA
	74.83	79.17	**87.58**	89.08	92.50	**94.00**	86.19	mixed	ROVER [1,2]
	77.08	**80.08**	86.33	90.08	92.50	93.42	86.58	both	ROVER [1,2,3]
	76.58	**80.08**	87.25	**90.42**	**93.17**	93.75	**86.87**	both	ROVER [1,2,3,4]

7. CONCLUSIONS

The use of a binaural front-end has proven advantageous when faced with the task of reducing the influence of highly reverberant and non-stationary noise, as they occur in typical living environments. On the CHiME challenge data, such a binaural processing stage provides clear improvements relative to the uncertainty-of-observation-based beamforming strategy introduced previously, especially when only the provided training set of isolated, noisy utterances is used.

A further clear improvement of performance is possible when additional, artificial training data is used, generated by adding noise from the considered environments to the clean speech recordings. In this – the *mixed training* – scenario, the binaural front-end and the uncertainty-of-observation-based beamformer are closer in performance, pointing towards the conclusion that mixed training reduces the influence of the applied preprocessing.

The best overall performance for this task can again – as with the first CHiME challenge – be achieved when the different front ends are combined by a final stage of late fusion, using the ROVER approach. On the considered dataset, fusing results from four slightly different versions, three of them based on the binaural and one on the beamforming front-end, leads to the optimal overall performance. In this way, the average keyword error rate can be reduced from originally 31.3 % without preprocessing or mixed training to 13.7 % with the discussed combination of approaches.

8. REFERENCES

[1] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, part 1," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 75–80, 2009.

[2] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, part 2," *Signal Processing Magazine, IEEE*, vol. 26, no. 4, pp. 78–85, 2009.

[3] R. Zelinski, "A microphone array with adaptive post-

filtering for noise reduction in reverberant rooms," in *Proc. Acoustics, Speech, and Signal Processing, International Conference on*, Apr. 1988, vol. 5, pp. 2578 – 2581.

[4] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, no. 1, pp. 27 – 34, 1982.

[5] F. Nesta and M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *Proc. CHiME Workshop on Machine Listening in Multisource Environments*, 2011, pp. 18–23.

[6] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. International Workshop on Machine Listening in Multisource Environments*, 2011, pp. 12–17.

[7] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, Oct. 2012.

[8] W. Gaik and W. Lindemann, "Ein digitales Richtungsfilter, basierend auf der Auswertung interauraler Parameter von Kunstkopfsignalen," in *Fortschr. Akusik–DAGA*, Oldenburg, Germany, 1986, pp. 721–724.

[9] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 58–67, 2005.

[10] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Communication*, vol. 53, no. 5, pp. 606–621, 2011.

[11] A. Schlesinger, *Binaural Model-Based Speech Intelligibility Enhancement and Assessment in Hearing Aids*, Ph.D. thesis, Delft University of Technology, The Netherlands, 2012.

[12] A. Schlesinger and C. Luther, *The technology of binaural listening*, chapter Optimization of binaural algorithms for maximum predicted speech intelligibility, Springer, Berlin–Heidelberg–New York NY, to appear 2013.

[13] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, no. 2, pp. 70–73, 1967.

[14] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 45–48.

[15] T. F. Gerkmann, *Statistical Analysis of Cepstral Coefficients and Applications in Speech Enhancement*, Ph.D. thesis, Ruhr-Universität Bochum, Germany, Dec. 2010.

[16] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," Sept. 2010.

[19] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data," in *Proc. AVSP*, 2009.

[20] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments," *Computer Speech and Language*, 2012.

[21] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "CHiME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *Proc. International Workshop on Machine Listening in Multisource Environments*, 2011.

[22] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 1997, pp. 347 –354.