Front-End Processing for the CHiME-5 Dinner Party Scenario

Christoph Boeddecker, Jens Heitkaemper*, Joerg Schmalenstroeer,*
Lukas Drude, Jahn Heymann, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Paderborn, Germany
{boeddeker, heitkaemper, schmalen, drude, heymann, haeb}@nt.upb.de

Abstract

This contribution presents a speech enhancement system for the CHiME-5 Dinner Party Scenario. The front-end employs multi-channel linear time-variant filtering and achieves its gains without the use of a neural network. We present an adaptation of blind source separation techniques to the CHiME-5 database which we call *Guided Source Separation* (GSS). Using the baseline acoustic and language model, the combination of Weighted Prediction Error based dereverberation, guided source separation, and beamforming reduces the WER by 10.54 % (relative) for the single array track and by 21.12 % (relative) on the multiple array track.

1. Introduction

During the past decade various new speech enhancement techniques, supported by the rise of Neural Networks (NNs), have been developed [1, 2]. Many of these developments have been spurred by recent challenges like REVERB [3], CHiME-3 [4] and CHiME-4 [5], which showcased the benefits of a strong speech enhancement front-end for Automatic Speech Recognition (ASR). In particular, supervised learning approaches for time-frequency mask estimation employing a neural network [1, 6] with subsequent beamforming, achieved excellent results. In recent years similar improvements in ASR were achieved with NN supported mask estimation for source separation [7, 8, 9]. However, these NN-based front-ends rely on clean training targets for efficient training making their application to a real word scenario challenging, where the required supervision information is often missing.

The CHiME-5 challenge [10], introduces a database consisting only of real multi-channel recordings in a dinner party scenario. Since the clean, uncorrupted speech is not available for computing the targets for a mask estimating neural network, one has to resort to unsupervised mask estimation techniques, which, e.g., have been used in the context of Blind Source Separation (BSS) employing spatial mixture models. In this contribution we modify the BSS approach introduced in [11, 12] to make efficient use of the available time and speaker annotations provided with the challenge data. Therefore, this novel separation system is called *Guided Source Separation* (GSS) [13] in the sequel. The GSS outputs time-frequency masks, from these mask spatial covariance matrices are estimated, and from these matrices, in turn, the coefficients of the statistically optimum Minimum Variance Distortionless Response (MVDR) beamformer are computed. Separation performance can be further improved by applying spectral masking to the beamformed signal, however, at the cost of introducing spectral distortions. Additionally, we apply multi-channel dereverberation based on the Weighted Prediction Error (WPE) principle [14, 15, 16] as a preprocessing step for GSS. We will also show how the proposed methods can be extended from single array enhancement to the multi array scenario.

We evaluate the proposed front-end processing techniques by computing Word Error Rates (WERs) using the baseline ASR back-end provided by the challenge organizers, and observe significant WER improvements. Yet better WERs can be achieved by combining the presented front-end with a stronger back-end, as is shown in [17].

In Section 2 a short overview over the CHiME-5 database is provided, and the components of the presented system are described in Section 3. Section 4 presents the evaluation of the proposed system, which is followed by conclusions and an outlook in Section 5.

2. CHiME-5 database

The CHiME-5 database consists of real recordings from 20 separate dinner parties divided into 16 for training, two for development and two parties for evaluation. Each party scenario, from now on called session, consists of the recordings of the conversation of four friends who spend around two hours in a real home. Each session is divided into three parts, which take place in different locations of the home and which last at least half an hour. The first part is the meal preparation in the kitchen, the second part takes place in the dining area and for the last part the friends move to the living room. The sessions are recorded by six Microsoft Kinect devices with four audio channels each and two Kinects per room. Additionally, each speaker wore two in-ear microphones. Those in-ear microphone signals are only provided for the training and development sessions. A session is split into multiple utterances for which time stamps and the target speaker id are provided.

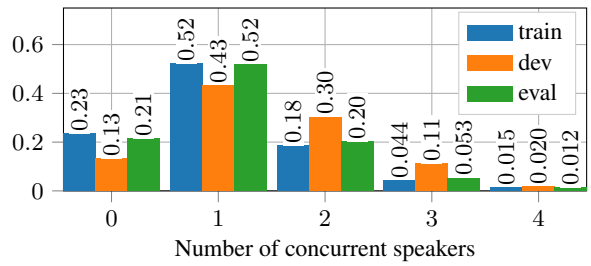

Figure 1: *Histograms of the number of active speakers for the train, development and evaluation dataset.*

Figure 1 shows that during the conversations a high amount of cross talk is present with up to four active speakers at a time. Since this is a likely source of recognition errors, our efforts concentrated on reducing the cross talk to enable the acoustic model to focus on the target speaker to improve recognition re-

*Both authors contributed equally.

sults.

3. System Overview

Let $\mathbf{Y}_{t,f}$ denote the multi-channel signal, consisting of D microphone signals, in the Short time Fourier Transform (STFT) domain, where t and f are the time frame and frequency bin indices, respectively. We model it as

$$\mathbf{Y}_{t,f} = \sum_k \mathbf{X}^{\text{early}}_{t,f,k} + \underbrace{\sum_k \mathbf{X}^{\text{tail}}_{t,f,k}}_{\mathbf{X}^{\text{tail}}_{t,f}} + \mathbf{N}_{t,f}, \tag{1}$$

with $\mathbf{X}^{\text{early}}_{t,f,k}$ and $\mathbf{X}^{\text{tail}}_{t,f,k}$ being the STFT coefficients for the early and late reverberated signals corresponding to the kth speaker. $\mathbf{N}_{t,f}$ represents the STFT coefficients of the noise signal.

Figure 2 depicts the structure of the enhancement system which will be described in the following.

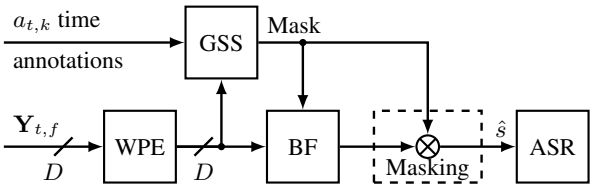

Figure 2: *Overview of speech enhancement system*

3.1. Weighted prediction error dereverberation

The first algorithm applied to the microphone signals is multi-channel WPE [16] for dereverberation, using our implementation [18] available on GitHub[1]. WPE estimates the reverberation tail $\mathbf{X}^{\text{tail}}_{t,f}$ and subtracts it from the observed signal $\mathbf{Y}_{t,f}$.

Note that WPE has been originally derived assuming absence of noise and overlapping speakers. However, in [19] we showed that WPE achieves reasonable WER reduction even in environments with significant amounts of additive noise. In [20] WPE and BSS has been jointly considered leading to a rather complex algorithm which alternates between dererberation and source separation. For complexity reasons we did not consider this tight integration here and applied WPE simply as a front end to all further processing.

3.2. Guided source separation

Assuming sparsity of speech in the STFT domain, BSS can be achieved by estimating, in an alternating fashion, first which source is dominant in each time-frequency (tf) bin, and then the statistics of each source from the tf bins it dominates. To this end we employed complex Angular Central Gaussian Mixture Model (cACGMM) [12] of which our implementation is available on GitHub[2]. The Probability Density Function (PDF) of the cACGMM is:

$$p\left(\tilde{\mathbf{Y}}_{t,f}; \boldsymbol{\theta}_f\right) = \sum_k \pi_{f,k} \mathcal{A}\left(\tilde{\mathbf{Y}}_{t,f}; \mathbf{B}_{f,k}\right), \tag{2}$$

[1] https://github.com/fgnt/nara_wpe
[2] https://github.com/fgnt/pb_bss

where

$$\mathcal{A}\left(\tilde{\mathbf{Y}}; \mathbf{B}\right) = \frac{(D-1)!}{2\pi^D \det(\mathbf{B})} \cdot \frac{1}{\left(\tilde{\mathbf{Y}}^{\mathrm{H}} \mathbf{B}^{-1} \tilde{\mathbf{Y}}\right)^D}, \tag{3}$$

$$\tilde{\mathbf{Y}}_{t,f} = \frac{\mathbf{Y}_{t,f}}{\|\mathbf{Y}_{t,f}\|}. \tag{4}$$

Here, $\tilde{\mathbf{Y}}_{t,f}$, $\boldsymbol{\theta}_f$, $\pi_{f,k}$ and $\mathbf{B}_{f,k}$ are the normalized observation vector, the parameter set, the mixture weights and the matrix parameter of the complex Angular Central Gaussian distribution [21], respectively.

Because the parameters of the model are estimated on each frequency bin independently, the well-known frequency permutation problem arises: the same mixture index k may correspond to different speakers in different frequency bins. Further, the number of mixture components, a.k.a. the number of speakers, has to be known.

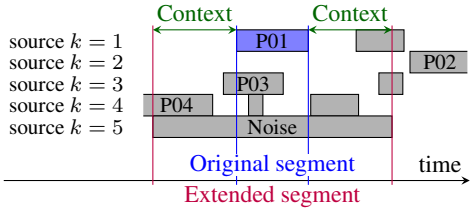

Figure 3: *Time annotation visualization. All utterance segments are aligned on the target array. Relative to a desired utterance is an extended segment selected for the enhancement.*

To overcome these difficulties we exploited the time annotations provided by the challenge organizers (visualized in Figure 3), which indicate when a particular speaker is active. The source activity pattern derived from the annotations guides the estimation of the mixture model parameters and avoids the need to solve the frequency permutation and the global speaker permutation problem. The latter refers to the fact, that the speaker order may also be a permuted version of the provided speaker order from the challenge organizers. Furthermore, it renders the estimation of the number of active sources unnecessary.

From these time annotations we compute a variable $a_{t,k}$, which takes the values one or zero depending on whether speaker k is active or inactive at time frame t. The time-invariant mixture weights $\pi_{f,k}$ are converted to time variant weights $\pi_{t,f,k}$ using the annotations $a_{t,k}$:

$$\pi_{t,f,k} = \frac{\pi_{f,k} a_{t,k}}{\sum_{k'} \pi_{f,k'} a_{t,k'}}. \tag{5}$$

With this modification the Expectation Maximization (EM) algorithm for the cACGMM from [12] has to be slightly modified. The E-step now is:

$$\begin{aligned}\gamma_{t,f,k} &= \frac{\pi_{t,f,k} \frac{1}{\det(\mathbf{B}_{f,k})} \frac{1}{\left(\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}} \mathbf{B}_{f,k}^{-1} \tilde{\mathbf{Y}}_{t,f}\right)^D}}{\sum_{k'} \pi_{t,f,k'} \frac{1}{\det(\mathbf{B}_{f,k'})} \frac{1}{\left(\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}} \mathbf{B}_{f,k'}^{-1} \tilde{\mathbf{Y}}_{t,f}\right)^D}} \\ &= \frac{\pi_{f,k} a_{t,k} \frac{1}{\det(\mathbf{B}_{f,k})} \frac{1}{\left(\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}} \mathbf{B}_{f,k}^{-1} \tilde{\mathbf{Y}}_{t,f}\right)^D}}{\sum_{k'} \pi_{f,k'} a_{t,k'} \frac{1}{\det(\mathbf{B}_{f,k'})} \frac{1}{\left(\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}} \mathbf{B}_{f,k'}^{-1} \tilde{\mathbf{Y}}_{t,f}\right)^D}},\end{aligned} \tag{6}$$

and the M-step is:

$$\pi_{f,k} = \frac{1}{T}\sum_t \gamma_{t,f,k}, \tag{7}$$

$$\mathbf{B}_{f,k} = D\frac{\sum_t \gamma_{t,f,k}\frac{\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}}\tilde{\mathbf{Y}}_{t,f}}{\tilde{\mathbf{Y}}_{t,f}^{\mathrm{H}}\mathbf{B}_{f,k}^{-1}\tilde{\mathbf{Y}}_{t,f}}}{\sum_t \gamma_{t,f,k}} \tag{8}$$

To account for background noise we use an additional noise class, whose activity $a_{t,k}$ is set to be always one. This mixture component can be considered as a garbage class which is supposed to collect distortions. This additional class introduces a permutation problem with the target speaker (Figure 3 original segment), because the activity $a_{t,k}$ for the target speaker and the noise class is equal. To reduce the permutation problem, a context is used (Figure 3 extended segment). With this context it is likely that the activity $a_{t,k}$ for the target speaker contains enough zeros so that the GSS system does no longer permute the target speaker with the noise.

3.3. Beamforming and masking

The estimated masks, i.e., the posteriors $\gamma_{t,f,k}$, are used for beamforming and/or mask-based source extraction. We employed the MVDR beamformer according to [22, 23]:

$$\mathbf{w}_f(\mathbf{r}) = \frac{\mathbf{\Phi}_{NN,f}^{-1}\mathbf{\Phi}_{XX,f}\,\mathbf{r}}{\mathrm{tr}\left\{\mathbf{\Phi}_{NN,f}^{-1}\mathbf{\Phi}_{XX,f}\right\}}, \tag{9}$$

where

$$\mathbf{\Phi}_{\nu\nu,f} = \frac{1}{T}\sum_t \gamma_{t,f,\nu}\mathbf{Y}_{t,f}\mathbf{Y}_{t,f}^{\mathrm{H}}, \quad \nu \in \{X, N\} \tag{10}$$

$$\mathbf{r} = \underset{\mathbf{e}}{\mathrm{argmax}}\left\{\frac{\mathbf{w}_f^{\mathrm{H}}(\mathbf{e})\mathbf{\Phi}_{XX,f}\mathbf{w}_f(\mathbf{e})}{\mathbf{w}_f^{\mathrm{H}}(\mathbf{e})\mathbf{\Phi}_{NN,f}\mathbf{w}_f(\mathbf{e})}\right\}. \tag{11}$$

Here, $\mathbf{e}$ is a unit vector, which selects the reference microphone, where the reference microphone is defined as the one with the largest output SNR, see Equation (11). The target mask $\gamma_{t,f,X}$ is the posterior $\gamma_{t,f,k}$ where the source index k corresponds to the target speaker and the distortion mask $\gamma_{t,f,N}$ is the sum of all remaining posteriors $\gamma_{t,f,k}$. Further the Blind Analytic Normalization (BAN) postfilter [24]

$$\mathbf{w}_f \leftarrow \frac{\sqrt{|\mathbf{w}_f^{\mathrm{H}}\mathbf{\Phi}_{NN,f}\mathbf{\Phi}_{NN,f}\mathbf{w}_f|}}{\mathbf{w}_f^{\mathrm{H}}\mathbf{\Phi}_{NN,f}\mathbf{w}_f}\mathbf{w}_f, \tag{12}$$

is applied to the beamformer output.

To improve the separation performance if the spatial diversity of the speakers is low, additional spectral masking is performed. However, this introduces additional spectral distortions.

3.4. Multi array enhancement

The above described system is developed for single array enhancement. However, we propose to apply it also to multiple arrays by stacking all microphone signals. In general, speech enhancement on multiple arrays without a tight synchronization of their individual sampling clocks can result in poor performance. Further offsets between the signals of different arrays can occur if chunks of samples are lost by the signal capturing device. However, the challenge organizers provided a synchronization[3] that roughly compensates for these offsets.

Assuming that the array signals are reasonable well synchronized, and assuming that the distances between the arrays are not too large and that the algorithms can deal with poor SNR values for some microphones, stacking all arrays to one big array can be advantageous due to the increased number of available channels. Furthermore, the GSS has more spatial information to distinguish between the sources. Also, beamfoming can be used to conduct an array selection. Note the used MVDR beamformer has a SNR based reference microphone selection and not a fixed reference microphone.

3.5. Back-End

We compare two hybrid acoustic models, both trained with the KALDI Toolkit [25] and the baseline Time Delay Neural Network (TDNN) recipe with i-vectors and Mel Frequency Cepstral Coefficients (MFCC) features.

The first Acoustical Model (AM) is trained on enhanced in-ear data. To cleanup the in-ear data we first stacked all 8 channels to one array (4 speakers each with 2 in-ear microphones) and ran the GSS two times. The first run was over the whole session and the resulting set of model parameters served as initialization for the second run, which was carried out for each utterance separately. With the assumption that each speaker has a fixed position relative to his in-ear microphones which is independent of his movements, it is reasonable to run the first parameter estimation on a complete session. The second, utterance-wise GSS adapts serves as fine-tuning towards the individual spatial configuration of an utterance.

To the in-ear data we did apply neither WPE nor beamforming. The enhanced data is simply the in-ear audio multiplied with the masks obtained from GSS.

The second AM is the provided baseline model (baseline), which is trained on the left channel of the target speaker's in-ear microphone and a random selection of array signals. This model was trained without enhanced training data.

No additional language model adaptation or rescaling is used.

4. Experimental evaluation

The proposed system is tested on the CHiME-5 data described in Section 2. For the front-end a STFT with a window size of 64 ms and a shift of 16 ms is used, whereas the back-end uses a STFT with window size 25 ms and a shift of 10 ms. For WPE we used the following parameters: 10 filter taps, a delay of 2 frames, beyond which correlation in the signal should be removed, and 3 iterations of WPE algorithm. For the clean speech Power Spectral Density (PSD) estimate a context of ± 1 frame was used. In the multi array track the delay is increased to 3 frames, and no PSD context is used. The following sections describe the improvements in WER achieved by the front-end components for the single and multi array scenario, respectively. All error rates given in the following are obtained on the development set of the microphone array data, except for the final WER with Source Activity Detector (SAD).

[3] https://github.com/chimechallenge/chime5-synchronisation

Table 1: *Overall WER in % on the development test set for single-array systems.*

Context in s	WPE	GSS	GSS Noise class	Beamforming	In-Ear TDNN w/o Masking	In-Ear TDNN w Masking	Baseline TDNN w/o Masking	Baseline TDNN w Masking
0	-	-		BeamformIt	96.98	-	81.69	-
				Sum Channels	93.94	-	81.13	-
15	-	✓	-	Channel 3	94.72	86.49	82.28	74.67
			✓			80.84		84.84
0	-	✓	-	MVDR + BAN	96.71	96.87	86.31	86.57
0			✓		94.65	95.09	81.98	82.56
2			-		89.08	86.40	78.16	76.58
2			✓		86.52	80.43	76.16	81.50
15			-		88.20	83.30	77.23	74.42
15			✓		84.86	79.89	75.11	84.76
2	✓	✓	-	MVDR + BAN	85.66	82.71	76.97	75.10
2			✓		82.69	77.15	74.82	78.87
15			-		85.11	79.35	74.42	**73.08**
15			✓		82.02	77.21	74.08	83.12

4.1. Single-array task

Table 1 gives a break-down of the benefits of the different enhancement steps in the single array scenario. Since the in-ear AM has not seen array data, the WER to start with is relative high (94.72 %), as expected. An enhancement with beamforming only (84.86 %) or by applying the mask to a single microphone channel without beamforming (80.84 %) achieves significant gains. The performance gain obtained by beamforming is less, probably because of the small spatial difference between overlapping speakers. A slightly higher gain of 2.78 % is achieved by using WPE as a preprocessing step and by using beamforming and masking together, i.e., beamforming followed by masking on the beamformed signal. WPE increases the effectiveness of the source separation system by decorrelating the observations. Compared to the initial WER of around 94.72 % a relative gain of 20 % is achieved.

When comparing these results with those obtained with the baseline AM the conclusions about the benefits of the individual components are the same. But the optimal parameter settings are different. This could possibly be explained with the sensitivity of the baseline AM to distortions introduced by masking in contrast to the in-ear AM which has seen those distortions during training. Whether masking introduces distortions depens on the noise class, without a noise class, at times when only one speaker is active, the observations are not affected by masking, because the posterior $\gamma_{t,f,k}$ for this speaker is equal to one. This may explain why the in-ear AM achives lower WER with a dedicated noise class in the GSS mixture model, the baseline AM results however suffer if a noise class is foreseen.

4.2. Multi array task

In Table 2 the gains from using multiple arrays are described for each system component. For all results a context of 15 s, see Section 3.2, and a noise class are used.

The beamformer profits from higher spatial resolution, which allows the separation of speakers standing only a small distance apart. Additionally, the beamformer probably takes advantage of choosing the reference microphone from all arrays instead of a single array. The improved separation can be seen in the approximately 5 % lower WER for both AMs when stacking all channels for beamforming and without additional spectral masking.

For the cACGMM a similar benefit from additional spatial information through stacked channels is expected. This expectation is met by the WER gain, which rises to 10 % for both AM, when using spectral masking and employing in-ear AM.

An additional gain is obtained when stacking the channels of all microphone arrays also for WPE. Quite surprisingly, now a back-end trained only on the in-ear data achieves a lower error rate than the back-end which is trained on mostly array data according to the challenge baseline setup. This clearly shows the strength of the model-based speech enhancement used here for a challenging ASR task. In total, a relative WER reduction of around 20 % was achieved compared to the baseline system. This was achieved without retraining of the AM and the use of a neural network in the front-end. With training on the enhanced In-Ear data the WER is even further reduced to 62.51 %.

We used the setup of the best systems from Tables 1 and 2 with a Source Activity Detector (SAD) Neuronal Network (NN)

Table 2: *Overall WER in % on the development test set for multi-array systems.*

Stack all channels WPE	GSS	MVDR + BAN	In-Ear TDNN w/o Masking	In-Ear TDNN w Masking	Baseline TDNN w/o Masking	Baseline TDNN w Masking
-	-	-	82.02	77.21	74.08	83.12
-	-	✓	77.19	77.36	69.98	83.17
-	✓	✓	71.95	66.24	65.50	78.78
✓	✓	✓	67.93	**62.51**	**64.41**	76.80

that estimates alignments $a_{t,k}^{\text{NN}}$. As input to the SAD we chose the array signal, a source activity mask estimated by the GSS and the time annotations given by the challenge organizers. The target alignments a ASR are the non slience alignments of an acoustic model. This resulted in a WER improvement by $1\,\%$.

At the moment we can only guess what the reason for the performance drop in WER from development to evaluation set in the multi-array case is because of the missing evaluation data. First, we used different acoustic models, In-Ear TDNN for multi-array and the Baseline TDNN for the single array track. Second, according to Figure 1, the eval data has less overlap than the development data. The In-Ear TDNN has shown its strength for overlap data, while the Baseline TDNN is better when only one speaker is active.

Table 3: *Results for the best system. WER (%) per session and location, together with the overall WER.*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	80.75	69.44	65.31	71.43
		S09	72.62	72.77	68.14	
	Eval	S01	82.63	63.15	79.46	69.60
		S21	74.75	59.70	64.79	
Multi	Dev	S02	68.65	65.86	56.39	61.73
		S09	58.72	60.97	60.67	
	Eval	S01	83.05	60.37	75.78	68.98
		S21	78.01	60.65	64.21	

5. Conclusions

We presented a speech enhancement system consisting of WPE, GSS and beamforming with spectral masking. The GSS employs a spatial mixture model that uses speaker time annotations to avoid the permutation problems. This front-end, which does not employ a neural network component, achieves a significant WER reduction on the challenging CHiME-5 dinner party scenario without the use of a neural network. Those WER improvements have been obtained using the acoustic model provided by the challenge organizers. Further significant gains are obtained if a stronger back-end and if system combination is used, see our companion paper [17]. An implementation of the described speech enhancement system without SAD is available on GitHub[4].

6. Acknowledgments

The work was in part supported by DFG under contract number Ha3455/14-1. Computational resources were provided by the Paderborn Center for Parallel Computing.

7. References

[1] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV Beamformer Front-End for the 3rd CHiME Challenge," in *Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, December 2015.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[4] `https://github.com/fgnt/pb_chime5`

[3] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, 12 2016.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 504–511.

[5] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, no. C, pp. 535–557, nov 2017. [Online]. Available: https://doi.org/10.1016/j.csl.2016.11.005

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[8] D. Yu, M. Kolbk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 241–245.

[9] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[10] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[11] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 241–244.

[12] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *European Signal Processing Conference (EUSIPCO),*. IEEE, 2016, pp. 1153–1157.

[13] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, May 2014. [Online]. Available: https://hal.inria.fr/hal-00922378

[14] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 85–88.

[15] ——, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[16] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[17] M. Kitza, W. Michel, C. Boeddeker, J. Heitkaemper, T. Menne, R. Schlüter, H. Ney, J. Schmalenstroeer, L. Drude, J. Heymann, R. Haeb-Umbach, and A. Mouchtaris, "The RWTH/UPB System Combination for the CHiME 2018 Workshop," 2018.

[18] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[19] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, 2018.

[20] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*. IEEE, 2014, pp. 268–272.

[21] J. T. Kent, "Data analysis for shapes and images," *Journal of statistical planning and inference*, vol. 57, no. 2, pp. 181–193, 1997.

[22] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[23] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[24] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[25] 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," Idiap, Rue Marconi 19, Martigny, Idiap-RR Idiap-RR-04-2012, Jan 2012.