Multi-Channel Block-Online Source Extraction based on Utterance Adaptation

Juan M. Martín-Doñas[1*], *Jens Heitkaemper*[2*],
Reinhold Haeb-Umbach[2], *Angel M. Gomez*[1], *Antonio M. Peinado*[1]

[1]Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
[2]Department of Communications Engineering, Paderborn University, Germany
{mdjuamart,amgg,amp}@ugr.es, {heitkaemper,haeb}@nt.upb.de

Abstract

This paper deals with multi-channel speech recognition in scenarios with multiple speakers. Recently, the spectral characteristics of a target speaker, extracted from an adaptation utterance, have been used to guide a neural network mask estimator to focus on that speaker. In this work we present two variants of speaker-aware neural networks, which exploit both spectral and spatial information to allow better discrimination between target and interfering speakers. Thus, we introduce either a spatial pre-processing prior to the mask estimation or a spatial plus spectral speaker characterization block whose output is directly fed into the neural mask estimator. The target speaker's spectral and spatial signature is extracted from an adaptation utterance recorded at the beginning of a session. We further adapt the architecture for low-latency processing by means of block-online beamforming that recursively updates the signal statistics. Experimental results show that the additional spatial information clearly improves source extraction, in particular in the same-gender case, and that our proposal achieves state-of-the-art performance in terms of distortion reduction and recognition accuracy.

Index Terms: Source extraction, Multi-channel speech enhancement, Online beamforming, Robust speech recognition, Speaker adaptation

1. Introduction

In recent years, research on Automatic Speech Recognition (ASR) has led to an increase of recognition accuracy, mainly due to the rise of deep neural networks. One example of particular interest is ASR for conference scenarios, where multiple speakers have to be transcribed to get an accurate summary. However, the ASR performance can be severely affected by different types of distortion such as background noise, reverberation and, especially, overlapping speakers.

When the recording devices employ microphone arrays, beamforming techniques can be used as a front-end to reduce distortions. State-of-the-art techniques are based on time-frequency masks indicating speech or interference dominance [1, 2]. The estimated masks are used to obtain spatial speech and noise statistics, which in turn are needed to compute the coefficients of the beamformer. Among other approaches, neural networks have been applied to mask estimation, showing state-of-the-art results [3, 4], while still allowing for low-latency processing [5, 6, 7].

The performance of these neural network-based mask estimators degrades in the presence of multiple, simultaneously active speakers, as the model is unable to discriminate between them. Several techniques have been recently proposed to deal with this source separation problem, showing promising results: Deep Clustering (DC) [8], Deep Attractor Networks (DANs) [9] or permutation invariant training [10]. These techniques are designed to extract the signals from all speakers. However, in many scenarios we are only interested in one target speaker.

Other techniques focus on one target speaker by using context information to select the desired one [11, 12, 13]. Thus, in [14] Zmolikova *et al.* proposed a network structure called Speaker Beam (SB), which uses an adaptation utterance (AU) of the desired speaker to focus on its spectral characteristics for the separation task. However, the performance of SB degrades in open-speaker-sets in case of overlapping speakers having the same gender. Other works have defined a set of finite AUs and trained a mask estimator to focus on the speaker using one of these predefined utterances [15, 16]. These approaches rely on the fact that the AU is predefined and known in advance. In [17] a network is trained to focus on the direction of the target speaker, which is provided by oracle information. Lately, the information from an AU has been used directly in an ASR system to allow speech separation without any additional front-end [18, 19].

In this work we propose two novel multi-channel low-latency speech extraction systems, which retrieve spatial and spectral information from an AU to force a neural network to focus on the speech signal of a target speaker. In both systems, the AU does not need to be fixed but it can be any utterance from the target speaker at his/her target position. For the first proposal, a spatial pre-processing is applied to both the AU and the noisy speech signal before they are fed into the neural network mask estimator. For the second proposal, a set of features capturing spatial and spectral information are used for mask estimation. We assume that each speaker has recorded an initial utterance without other interfering speakers and that the speakers change their positions only slightly between the AU recording and the session from which the target speech is to be extracted. This requirement is mostly satisfied in a conference scenario. Nevertheless, if the target speaker moves, a short reinitialization can be applied to adapt the system to the new position. The advantage of the presented approach is that it does not depend on any specific AU, any specific noise condition during adaptation or the application of other additional information, while still exploiting the spectral and spatial properties of the target signal. Experimental results show that our proposal achieves good recognition accuracy and low distortion in comparison with other state-of-the-art approaches.

The reminder of the paper is organized as follows. The block-online beamformer procedure is explained in Section 2. In Section 3 the proposed mask estimators based on SB are introduced. The description of the experimental framework and the results are addressed in Section 4 and final conclusions are drawn in Section 5.

*Both authors contributed equally.

2. Block-Online Beamformer Estimation

We assume a multi-channel noisy speech signal in the short-time Fourier transform (STFT) domain,

$$\mathbf{Y}(t,f) = \mathbf{X}(t,f) + \mathbf{N}(t,f), \tag{1}$$

where $\mathbf{X}(t,f)$ and $\mathbf{N}(t,f)$ are the multi-channel target speech and the noise signal vector, respectively, t is the frame index and f the frequency bin index. The noise signal accounts for the background noise, late reverberation and interfering speakers. The time and frequency indices will be omitted where possible.

The estimation of the beamformer coefficients requires an estimation of the spatial correlation matrices (SCM) for both the clean speech $\mathbf{\Phi}_{XX}$ and the noise signals $\mathbf{\Phi}_{NN}$. To allow for low-latency processing, these matrices are recursively estimated in blocks of L frames for every frequency bin using a speech or a noise activity mask M_ν with $\nu = \{X, N\}$ [7],

$$\mathbf{\Phi}_{\nu\nu}(nL) = \beta_\nu \mathbf{\Phi}_{\nu\nu}((n-1)L) + (1-\beta_\nu)\widehat{\mathbf{\Phi}}_{\nu\nu}(nL), \tag{2}$$

$$\widehat{\mathbf{\Phi}}_{\nu\nu}(nL) = \sum_{l=0}^{L-1} M_\nu(nL-l)\mathbf{Y}(nL-l)\mathbf{Y}^H(nL-l), \tag{3}$$

where β_ν is the forgetting factor and n the block index. This way, we can estimate and apply the beamformer in each block of frames.

The aforementioned procedure needs an initialization of the speech and noise SCMs. For example, the work in [7] initializes the SCM by using an identity matrix for the noise and zero matrix for the speech. Better performance can be obtained with a proper initialization which makes use of prior information about the spatial characteristics of the signals. Thus, we can assume a diffuse noise field as initialization for the noise SCM with [20]:

$$\mathbf{\Phi}_{NN,\text{diffuse}}(f) = \phi_N(f) \cdot \operatorname{sinc}\left(2\pi f F_s \cdot \mathbf{d}/c\right), \tag{4}$$

where F_s is the sampling frequency, $\mathbf{d}$ is the matrix of distances between the microphones, c is the speed of sound and $\phi_N(f)$ is the noise Power Spectral Density (PSD). We calculate an estimate of the noise PSD from the first block of the distorted utterance. In Equation (4) the sinc operator has to be understood to be applied to each matrix element separately. The target speech SCM may be initialized with a matrix $\mathbf{\Phi}_{UU}$ corresponding to the SCM of the multi-channel AU signal $\mathbf{U}(t,f)$. This way the spatial information and the characteristics of the acoustic channels between the speaker and the microphones are exploited.

From the estimated SCMs a beamformer is computed for every block step. We use the rank-1 approximation [21] of the Minimum Variance Distortionless Response (MVDR) beamformer formulation presented in [22]:

$$\mathbf{F} = \frac{\mathbf{\Phi}_{NN}^{-1}\tilde{\mathbf{\Phi}}_{XX}}{\operatorname{tr}\{\mathbf{\Phi}_{NN}^{-1}\tilde{\mathbf{\Phi}}_{XX}\}}\mathbf{u}, \tag{5}$$

where $\mathbf{u}$ is a unit vector pointing to the reference microphone, $\operatorname{tr}\{\cdot\}$ is the trace operator and $\tilde{\mathbf{\Phi}}_{XX}$ is a rank-1 approximation of the speech SCM [21], defined as

$$\tilde{\mathbf{\Phi}}_{XX} = \mathbf{a}\mathbf{a}^H \cdot \operatorname{tr}\{\mathbf{\Phi}_{XX}\}/\operatorname{tr}\{\mathbf{a}\mathbf{a}^H\}, \tag{6}$$

where $\mathbf{a} = \mathbf{\Phi}_{NN}\mathcal{P}\left\{\mathbf{\Phi}_{NN}^{-1}\mathbf{\Phi}_{XX}\right\}$ with $\mathcal{P}\{\cdot\}$ standing for the principal component of a matrix.

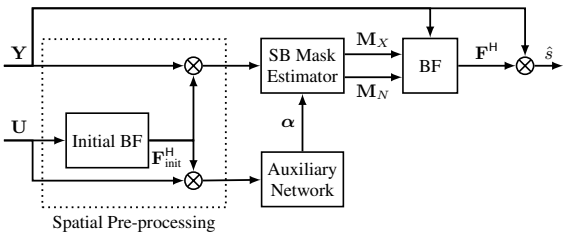

Figure 1: *System overview of the PreBF-based source extraction.*

3. SpatialBeam Speech Extraction

To estimate the masks M_ν for the beamforming vector computation, we develop a mask estimator based on the SB approach proposed in [14]. Disregarding the spatial pre-processing block (explained later), this SB approach corresponds to the right half of the diagram depicted in Figure 1. The approach consists of a neural network-based mask estimator, composed of a recurrent layer and several feedforward layers, provided with speaker information about the target speaker. This is achieved by introducing an auxiliary network fed with the AU. The auxiliary network obtains an output vector $\boldsymbol{\alpha}$, which will be referred to as speaker representation. On the other hand, one of the feedforward layers of the SB mask estimator is split into several sub-layers, whose individual outputs are combined by means of the weighting vector $\boldsymbol{\alpha}$. The auxiliary network and the mask estimator are jointly trained. This way, the mask estimator is adapted to focus on the target speaker.

The SB approach is adapted in this work for online mask estimation in a similar way to [7]. First, the bi-directional Long-Short-Term-Memory (LSTM) layer is replaced by a single LSTM layer of twice the output size, thus reducing the information available to the network to current and past frames. The utterance mean and variance offline normalization is replaced by a recursive mean normalization, since this has shown to perform well in the online case [7].

The SB approach shows a performance degradation when applied in a scenario with overlapping speakers with similar spectral characteristics, as observed in speakers of the same gender. In this case, the mask estimator is not able to separate the target speaker from the interfering ones. This problem is alleviated in cases where both speakers are part of the training data. However, the retraining option proposed in [14] is not feasible in an online scenario.

To solve the aforementioned problem, we propose to use spatial information obtained directly from the AU assuming a steady target speaker position. This allows a better separation between speakers with similar active frequencies and speech patterns. Thus, we propose two different approaches which exploit the spatial information provided by additional blocks.

The first proposal is to apply a spatial pre-processing to the noisy speech signal and the AU before its use in the SB mask estimator and the auxiliary network. The block diagram of this proposal is depicted in Figure 1 along with the SB mask estimator. We choose an MVDR beamformer as spatial pre-processor. The SCMs for this initial beamformer are the same ones used as initialization of the block-online beamforming presented in Section 2. Additionally, we apply the rank-1 approximation explained above to force the beamformer to concentrate on the spatial information in $\mathbf{\Phi}_{UU}$. Both the AU and the subsequent distorted utterance are enhanced with this initial beamforming vector, $\mathbf{F}_{\text{init}}$. The resulting single-channel signals are then fed

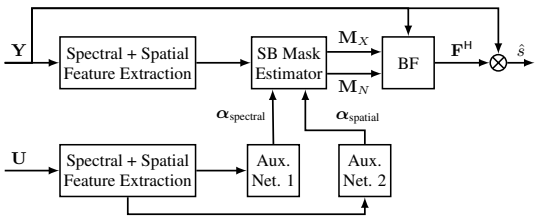

Figure 2: *System overview of the proposed online mask estimator based on SpeakerBeam and spatial features.*

into the SB network and the auxiliary network for target speaker and noise mask estimation. The final beamformer (BF in the figure) use these masks to extract the target signal in each block as described in the previous section. The advantage of this preprocessing, referred to as PreBF from now on, is that, although it does not provide a strong enough separation on its own for an ASR system, the overlapping speakers are suppressed so that a more accurate mask estimation is possible.

As an alternative we propose to use additional spatial features at the input of the auxiliary and mask estimation networks. Even if the magnitude spectral properties of the target speaker do not offer enough distinct information, the phase differences between different microphone signals provide another independent source of information to identify the target speaker. Therefore, the auxiliary network can calculate distinct speaker embeddings for speakers with similar speech by using this additional spatial information. Similarly, for speakers with small differences in their position with respect to the microphones, the spectral information may still offer a more conclusive speaker representation.

To enable the mask estimator to benefit from these more meaningful speaker embeddings, the mask estimator input has to be extended with these new spatial features. The block diagram of the extended SB mask estimator is depicted in Figure 2. As spatial features, we use the Interchannel Phase Difference (IPD) computed similarly to [23] as

$$\text{cosIPD}(t,f,p,q) = \cos\left(\angle y_{t,f,p} - \angle y_{t,f,q}\right), \quad (7)$$

$$\text{sinIPD}(t,f,p,q) = \sin\left(\angle y_{t,f,p} - \angle y_{t,f,q}\right), \quad (8)$$

where p, q are channel indices, $\angle$ is the phase operator and y can indicate either the noisy speech signal or the adaptation utterance (single-channel STFT domain). These features have been shown to result in a significant improvement for DC [23]. For two channel input signals, the spatial features are calculated as described in Equations (7) and (8), where one of the channels acts as reference channel. In the case of more than two channels, each pair of channels is treated like a two channel problem. The mean- and variance-normalized logarithmic spectrum of the AU is used at the input of the auxiliary network as spectral features.

To force the SB network to use both the spatial and spectral information, two independent auxiliary networks are trained, using either the spectral or the spatial features. The mean pooling at the output of each auxiliary network is carried out in both time and channel dimensions. The estimated speaker representation vector $\boldsymbol{\alpha}_{\text{spectral}}$ is used to weight half of the sub-layers of the SB adaptation layer and $\boldsymbol{\alpha}_{\text{spatial}}$ to weight the other half.

For the input of the SB mask estimator, the spectral and spatial features obtained from the multi-channel noisy speech signal are concatenated. To reduce the required size of the Recurrent Neural Network (RNN), a bottleneck in form of a feedforward layer is introduced at the input of SB.

4. Experiments

We evaluate the proposed approaches on a simulated database and compare their results to DANs and SB. To this end, we use the signal to distortion ratio (SDR) metric following the implementation presented in [24], which measures the performance of the enhancement procedure. Despite recent criticism expressed e.g. in [25], we chose this metric because of its comparability. Additionally, the system is evaluated in terms of the Short-Time Objective Intelligibility (STOI) metric [26] for speech intelligibility and Word Error Rate (WER) to test its performance in a conference scenario for ASR.

The multi-channel database is described in [27]. It consists of 30000 utterances for the training set, 500 utterances for the development set and 1500 utterances for the evaluation set. Each example is created by randomly choosing two utterances from the Wall Street Journal (WSJ) database and convolving the signals with six channel Room Impulse Responses (RIRs) simulated by the Image Method [28]. The utterances are previously downsampled to 8 kHz. The shorter of the two generated multi-channel signals is padded with zeros in order to match the duration of the other signal. This padding is done randomly at the start and end of the utterance. The observation utterance thus consists of the sum of both utterances plus additional white Gaussian noise with an SNR of 20 to 30 dB. Every speaker can only be found in one of the previous sets, which ensures different speakers in the training and evaluation of the system. Therefore, we characterize the database as an open-speaker-set database.

Note that the speaker position is assumed approximately fixed for the duration of an utterance. Thus, beamforming, whose coefficients are computed offline for the whole utterance, can be considered the best solution if low latency is not an issue.

For the STFT computation, a 512-point FFT is used with a Hann window and a 75% overlap, resulting in 257 frequency bins for each time frame. The mask estimator consists of an LSTM layer of 1024 units, two feedforward layers with 1024 units each and one output layer. The first feedforward layer is split into 30 sub-layers for the SB approach. The auxiliary network has two feedforward layers of 50 units and an output layer of 30 units, as in [14]. As loss function we chose the binary cross entropy between the estimated masks and ideal binary masks calculated from the reverberated clean speech signals.

Finally, for the block-online estimation we use blocks of five frames and a forgetting factor of $\beta_\nu = 0.95$.

4.1. Backend

The Acoustic Model (AM) of the ASR back-end is a Wide Residual Network as proposed in [29]. This back-end uses logarithmic mel filterbank input features and it consists of two LSTM layers. The AM is combined with a trigram language model from the WSJ baseline script provided by the KALDI toolkit [30]. All hyper-parameters were taken from [29]. The AM is trained on the artificially reverberated WSJ utterances without overlapped speech. The decoding is performed without language model rescoring. Note that the AM and the ASR engine operate offline for all experiments since we focus on the front-end processing. However, they may be replaced by an online version to obtain a fully online operating system.

4.2. Initialization Method Evaluation

First, we evaluate the performance of the different proposed online strategies when ideal binary masks are used. The results are shown in Table 1, where the offline method and the different

Table 1: *SDR, STOI and WER scores obtained for different initialization of the SCM estimation using ideal binary masks.*

Method	Initialization $\mathbf{\Phi}_{\mathbf{XX}}$	Initialization $\mathbf{\Phi}_{\mathbf{NN}}$	STOI	SDR dB	WER %
Offline	–	–	0.84	12.37	16.40
Online	Zeros	Identity	0.82	10.95	19.89
		Diffuse	0.82	**11.13**	19.60
	$\mathbf{\Phi}_{\mathbf{UU}}$	Identity	0.82	10.69	17.88
		Diffuse	**0.83**	11.10	**16.94**

Table 2: *SDR, STOI and WER scores obtained for different speaker extractors.*

BF	Extractor	STOI	SDR dB	WER %
Offline	SpeakerBeam	0.76	8.78	28.66
	DAN	0.78	**11.38**	23.70
	PreBF	**0.80**	10.00	**23.32**
	Spt. Features	**0.80**	9.70	23.50
Online	Online-PreBF	0.74	**5.54**	34.60
	Online-Spt. Features	**0.75**	5.09	**33.61**

initializations for the SCMs are compared in terms of STOI, SDR and WER. For the target speech SCM both an all zero initialization and the use of the SCM estimated on the AU is evaluated. On the other hand, the noise SCM is either initialized with the identity matrix or the SCM of a diffuse noise field. In terms of initialization, the best performance is achieved for the combination of diffuse noise SCM and target speaker spatial information extracted from the multi-channel adaptation utterance. It is observed that this combination is close to the offline beamformer in recognition accuracy, and it also obtains competitive results in distortion reduction and intelligibility. This shows that a proper initialization is helpful for beamformer convergence.

4.3. Speaker Extraction Evaluation

Next we compare in Table 2 different speaker extraction systems based on neural network mask estimation and beamforming. The results show that both the extractor using spatial pre-processing at its input and the network using spatial features achieve WER scores superior to the state-of-the-art approaches DANs and SB. In terms of signal distortion, the use of spatial information does not lead to any improvement. However, both systems outperform the state-of-art in speech intelligibility gain. This shows that both proposed systems achieve competitive results while they allow to focus on the target speaker. We also tested the PreBF proposal using only pre-processing but not the SB approach, obtaining a WER of $27.03\,\%$. This shows that the combination of the spatial pre-processing with the speaker information of the SB approach outperforms the independent systems.

The online versions of the proposed systems have a higher WER, mainly due to the block-online updating of the SCM statistics but also because of the use of a single LSTM layer. Nevertheless, the systems still achieve competitive results for online recognition, with the use of spatial features as the preferred approach for ASR.

Table 3: *SDR and WER scores obtained for the different speaker extractors. Results are separated for overlapped speaker of the same and different gender.*

Method	SDR (dB)		WER (%)	
	Differ.	Same	Differ.	Same
SpeakerBeam	10.17	7.25	23.13	34.82
PreBF	10.68	**9.24**	21.21	**25.67**
Spt. Features	**10.92**	8.49	**19.49**	28.52

As described in Section 3, SB struggles on utterances that contain mixed speech from speakers of the same gender. Therefore, we split the results into utterances with overlapped speakers of different and same gender to evaluate how our strategies perform in each scenario. The results for speech distortion and recognition accuracy are shown in Table 3. As can be seen, while the SB system performs well in the different gender case, it degrades in utterances with speakers of the same gender, increasing the final WER. The use of spatial features in our second approach improves the accuracy of the estimator but still underperforms on utterances with speakers of the same gender. This may be caused by the fact that the network has difficulties to learn both spectral and spatial characteristics for the separation task. On the other hand, the use of our PreBF approach is particularly effective in the same gender case, achieving similar results to the different gender case. This is especially true for the WER, where difference between different and same gender utterances is reduced from $11.69\,\%$ to $4.46\,\%$. The PreBF approach has the advantage that the input to the network is already processed, so the estimator exploits the more attenuated interfering speakers in the input signal to distinguish the target one.

5. Conclusions

In this paper we presented two novel systems for block-online multi-channel target speaker extraction, which exploit both spatial and spectral information of the target speaker obtained from an adaptation utterance. Also, we proposed an initialization of the spatial covariance matrices which was shown to be useful in online beamforming. The obtained results show that our systems outperform other state-of-the-art separation techniques, without the need of fixing an adaptation utterance in advance or relying on additional oracle information. Our experiments revealed that beamforming on the input of the mask estimator can reduce the separation error especially in utterances with speakers of the same gender, achieving low speech distortion and good recognition accuracy while allowing low-latency processing. In future work we will evaluate whether a combination of the proposed systems allows for further improvements.

6. Acknowledgements

The work was in part supported by DFG under contract number Ha3455/14-1, in part by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU under Grant FPU15/04161 and in part by the research stay program of the University of Granada.

7. References

[1] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5210–5214.

[2] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, September 2018, pp. 35–40.

[3] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2015, pp. 444–451.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[5] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 531–535.

[6] M. Zöhrer, L. Pfeifenberger, G. Schindler, H. Fröning, and F. Pernkopf, "Resource efficient deep eigenvector beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 3354–3358.

[7] J. Heitkaemper, J. Heymann, and R. Haeb-Umbach, "Smoothing along frequency in online neural network supported acoustic beamforming," in *ITG 2018, Oldenburg, Germany*, October 2018.

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

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

[10] D. Yu, M. Kolbaek, 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.

[11] K. Zmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, T. Nakatani, and J. Cernocký, "Optimization of speaker-aware multichannel speech extraction with ASR criterion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 6702–6706.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. Lopez Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv e-prints*, p. arXiv:1810.04826, October 2018.

[13] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," *arXiv e-prints*, p. arXiv:1807.08974, July 2018.

[14] K. Zmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. InterSpeech*, August 2017, pp. 2655–2659.

[15] Y. Kida, D. Tran, M. Omachi, T. Taniguchi, and Y. Fujita, "Speaker selective beamformer with keyword mask estimation," *arXiv e-prints*, p. arXiv:1810.10727, October 2018.

[16] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Proc. InterSpeech*, September 2018, pp. 2703–2707.

[17] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, December 2018, pp. 558–565.

[18] M. Delcroix, K. Zmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5554–5558.

[19] Y. Wang, X. Fan, I. Chen, Y. Liu, T. Chen, and B. Hoffmeister, "End-to-end anchored speech recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7090–7094.

[20] A. Schwarz, C. Hümmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2015, pp. 4380–4384.

[21] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37 – 51, 2018.

[22] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, 2007.

[23] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1–5.

[24] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "Mir_eval: a transparent implementation of common MIR metrics," in *Proc. of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, October 2014.

[25] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" in *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 626–630.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, 2019, Early Access.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. CHiME 2016 Workshop on Speech Processing in Everyday Environments*, September 2016, pp. 12–17.

[30] 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 *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2011.