INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments

Hassan Taherian[1], *Zhong-Qiu Wang*[1], and *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

taherian.1@osu.edu, {wangzhon,dwang}@cse.ohio-state.edu

Abstract

Despite successful applications of multi-channel signal processing in robust automatic speech recognition (ASR), relatively little research has been conducted on the effectiveness of such techniques in the robust speaker recognition domain. This paper introduces time-frequency (T-F) masking-based beamforming to address text-independent speaker recognition in conditions where strong diffuse noise and reverberation are both present. We examine various masking-based beamformers, such as parameterized multi-channel Wiener filter, generalized eigenvalue (GEV) beamformer and minimum variance distortion-less response (MVDR) beamformer, and evaluate their performance in terms of speaker recognition accuracy for i-vector and x-vector based systems. In addition, we present a different formulation for estimating steering vectors from speech covariance matrices. We show that rank-1 approximation of a speech covariance matrix based on generalized eigenvalue decomposition leads to the best results for the masking-based MVDR beamformer. Experiments on the recently introduced NIST SRE 2010 retransmitted corpus show that the MVDR beamformer with rank-1 approximation provides an absolute reduction of 5.55% in equal error rate compared to a standard masking-based MVDR beamformer.

Index Terms: Robust speaker recognition, beamforming, x-vector, deep neural network

1. Introduction

Smart speaker devices such as Amazon Echo and Google Home have gained popularity in recent years. These devices increasingly include a speaker recognition component that is used for authentication or personalized responses. However, the realization of robust speaker recognition is still a challenging task as distant speech signals are susceptible to distortions due to background noise and room reverberation. A natural way to enhance robustness against noise and reverberation is to utilize multi-channel speech enhancement techniques like beamforming as these devices are equipped with multiple microphones.

The focus of the speaker recognition community is mainly concerned with the monaural or single-channel case as speaker recognition has traditionally been applied to telephone speech. Major advances on increasing the robustness of speaker recognition include the introduction of the i-vector framework based on Gaussian mixture models (GMM/i-vector) [1], the Probabilistic Linear Discriminant Analysis (PLDA) back-end [2], Deep Neural Networks (DNNs) replacing the GMM component [3], and recently introduced x-vectors [4] for speaker embedding. X-vectors are of particular interest due to their use of inexpensive data augmentation for increasing robustness and utilization of the algorithms associated with i-vectors, e. g. PLDA for scoring. Monaural speech enhancement techniques have been investigated in this domain. Background noise is attenuated with supervised speech separation based on DNN in [5]. The methods in [6] and [7] use a deep autoencoder and long short-term memory (LSTM) respectively to estimate clean features and subsequently feed them to an i-vector based speaker recognition system.

Along a related direction, remarkable progress has been made in multi-channel speech enhancement. By introducing DNN based T-F masking to conventional beamforming, substantial improvements have been made on robust ASR [8], [9], [10]. The main reason behind impressive improvements is that monaural masking leads to more accurate estimation of speech and noise covariance matrices, the key components in the adaptive beamforming formulation. Motivated by its success in robust ASR, masking-based beamforming is also investigated in robust speaker recognition. In [11], Mošner *et al.* studied far-field speaker recognition in reverberant environments. In their study, multi-channel weighted prediction error (WPE) [12] is combined with masking-based beamforming to reduce the effect of reverberation. However this study only addresses reverberation effects. In real scenarios, reverberation and noise usually occur simultaneously and they have confounding effects that make speech enhancement substantially more challenging.

In this paper, we investigate multi-channel speaker recognition in adverse acoustic conditions where target speech is corrupted by strong diffuse noise and room reverberation. Specifically, we examine different masking-based beamforming methods and evaluate their performance on the conventional i-vector and state-of-the-art x-vector based speaker recognition systems. Our proposed system employs rank-1 approximation to construct speech covariance matrices for estimating steering vectors used in the MVDR beamformer. Consistent improvements are observed over other commonly used masking-based MVDR beamformers in terms of speaker verification error rate.

The rest of the paper is organized as follows. In Section 2, we describe the MVDR beamformer and how to derive the speech covariance matrix in different ways. We present our experimental setup in Section 3. We then provide evaluation results and comparisons in Section 4. Concluding remarks are given in Section 5.

2. Multi-Channel Speech Enhancement

2.1. MVDR Beamforming

The received signals to a microphone array can be formulated in the short-time Fourier transform (STFT) domain under the narrow-band assumption [13]:

$$\boldsymbol{y}(t,f) = \boldsymbol{c}(f)x(t,f) + \boldsymbol{h}(t,f) + \boldsymbol{n}(t,f) \qquad (1)$$

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1428

where $x(t,f)$ is the STFT value of the target signal at time t and frequency f, $\boldsymbol{c}(f)$ is the steering vector, and $\boldsymbol{c}(f)x(t,f)$, $\boldsymbol{h}(t,f)$, $\boldsymbol{n}(t,f)$, $\boldsymbol{y}(t,f)$ respectively represent the STFT vectors of the direct signal, its reverberation, noise, and received mixture.

MVDR seeks an optimal weight vector $\boldsymbol{w}(f)$ that can be applied to the received signals to suppress signals from non-target directions. Concretely, the output of MVDR minimizes the variance under the constraint that speech source signal will not be distorted [14]:

$$\begin{aligned} &\underset{w(f)}{\text{argmin}} \;\; \boldsymbol{w}(f)^H \boldsymbol{\Phi}_n(f) \boldsymbol{w}(f) \\ &\text{subject to} \;\; \boldsymbol{w}(f)^H \boldsymbol{c}(f) = 1 \end{aligned} \tag{2}$$

where $\boldsymbol{\Phi}_n(f)$ is the noise covariance matrix and $(.)^H$ is the conjugate transpose operator. The closed-form solution is:

$$\boldsymbol{w_{opt}}(f) = \frac{\boldsymbol{\Phi}_n(f)^{-1}\boldsymbol{c}(f)}{\boldsymbol{c}(f)^H \boldsymbol{\Phi}_n(f)^{-1}\boldsymbol{c}(f)} \tag{3}$$

Traditionally, $\boldsymbol{\Phi}_n(f)$ is obtained from the noise-only portions of the signal using voice activity detection (VAD). Recently, it is suggested that T-F masking can replace VAD to estimate $\Phi_n(f)$ more accurately [9], [10], [15]:

$$\boldsymbol{\Phi}_n(f) = \frac{\sum_t (1 - m(t,f)) \boldsymbol{y}(t,f) \boldsymbol{y}(t,f)^H}{\sum_t (1 - m(t,f))} \tag{4}$$

where $m(t,f)$ denotes the estimated T-F mask from the DNN at time t and frequency f. The well-established method for finding steering vector $\boldsymbol{c}(f)$ is by Generalized Cross Correlation with Phase Transform (GCC-PHAT) technique which estimates direction of arrival (DOA) of speech source [16]. Another way is to estimate $\boldsymbol{c}(f)$ as the principal eigenvector of the speech covariance matrix [17], eliminating the need for DOA estimation:

$$\boldsymbol{c}(f) = \mathcal{P}(\boldsymbol{\Phi}_x(f)) \tag{5}$$

where $\mathcal{P}(.)$ computes the principal eigenvector and

$$\boldsymbol{\Phi}_x(f) = \frac{\sum_t m(t,f) \boldsymbol{y}(t,f) \boldsymbol{y}(t,f)^H}{\sum_t m(t,f)} \tag{6}$$

is the estimated speech covariance matrix. In the ideal case, $\boldsymbol{\Phi}_x(f)$ is a rank-1 matrix. However this may not be valid in the presence of reverberation or imperfect mask estimation. In [15], the speech covariance matrix is estimated by subtracting the noise covariance matrix from the covariance matrix of noisy speech with the assumption that they are uncorrelated:

$$\boldsymbol{\Phi}_x(f) = \boldsymbol{\Phi}_y(f) - \boldsymbol{\Phi}_n(f) \tag{7}$$

$$\boldsymbol{\Phi}_y(f) = \frac{1}{T} \sum_t^T \boldsymbol{y}(t,f) \boldsymbol{y}(t,f)^H \tag{8}$$

where $\boldsymbol{\Phi}_y(f)$ is the spatial covariance matrix of noisy speech and T is the total number of frames. It has been observed in [17] that this derivation results in more accurate estimation of the steering vector.

2.2. Rank-1 Approximation of Speech Covariance Matrix

Originally proposed in [18] for multi-channel Wiener filtering and later applied to robust ASR [19], [20], the speech covariance matrix can be approximated using the decomposition technique:

$$\boldsymbol{\Phi}_x(f) = \boldsymbol{\Phi}_{r1}(f) + \boldsymbol{\Phi}_z(f) \tag{9}$$

where $\boldsymbol{\Phi}_{r1}(f)$ is a rank-1 matrix and $\boldsymbol{\Phi}_z(f)$ is the remainder matrix. Several solutions are suggested for estimating $\boldsymbol{\Phi}_{r1}(f)$, namely, first column decomposition, eigenvalue decomposition (EVD) and generalized eigenvalue decomposition (GEVD). In GEVD, we jointly diagonalize $\boldsymbol{\Phi}_x(f)$ (Eq. (6)) and $\boldsymbol{\Phi}_n(f)$:

$$\begin{cases} \boldsymbol{Q}(f)^H \boldsymbol{\Phi}_x(f) \boldsymbol{Q}(f) = \Lambda(f) \\ \boldsymbol{Q}(f)^H \boldsymbol{\Phi}_n(f) \boldsymbol{Q}(f) = \boldsymbol{I}_M \end{cases} \tag{10}$$

where $\boldsymbol{I}_M$ is an $M \times M$ identity matrix, M is number of microphones, and $\Lambda(f) = \text{diag}\{\lambda_1(f), \ldots, \lambda_M(f)\}$, assuming that eigenvalues are sorted in the descending order. Then, Eq. (9) can be written as:

$$\begin{aligned} \boldsymbol{\Phi}_x(f) &= \boldsymbol{Q}(f)^{-H} \text{diag}\{\lambda_1(f), \lambda_2(f), \ldots, \lambda_M(f)\} \boldsymbol{Q}(f)^{-1} \\ \boldsymbol{\Phi}_{r1}(f) &= \boldsymbol{Q}(f)^{-H} \text{diag}\{\lambda_1(f), 0, \ldots, 0\} \boldsymbol{Q}(f)^{-1} \\ \boldsymbol{\Phi}_z(f) &= \boldsymbol{Q}(f)^{-H} \text{diag}\{0, \lambda_2(f), \ldots, \lambda_M(f)\} \boldsymbol{Q}(f)^{-1} \end{aligned} \tag{11}$$

$\boldsymbol{\Phi}_z(f)$ can be interpreted as noise, or it can be considered as residual error and be ignored. Here, by ignoring $\boldsymbol{\Phi}_z(f)$, $\boldsymbol{\Phi}_x(f)$ simplifies to [19]:

$$\boldsymbol{\Phi}_{r1}(f) = \frac{tr(\boldsymbol{\Phi}_x(f))}{tr(\boldsymbol{q}_1(f)\boldsymbol{q}_1(f)^H)} \boldsymbol{q}_1(f) \boldsymbol{q}_1(f)^H \tag{12}$$

where $\boldsymbol{q}_1(f)$ is the first column of $\boldsymbol{Q}^{-H}(f)$ and $tr(.)$ is the trace operator. Applying Eq. (5) to the approximated rank-1 $\boldsymbol{\Phi}_x(f)$ yields an estimate of the steering vector $\boldsymbol{c}(f)$, which can be more accurate since the rank-1 assumption is valid.

2.3. T-F Mask Estimation

Accurate T-F mask estimation is pivotal in the performance of a masking-based beamformer. We choose the ideal ratio mask (IRM) as the training target in this study. The IRM is defined as the ratio between the energy of clean and noisy speech at each T-F unit [21]. The IRM can be defined in different T-F domains. We opt to use the IRM in the magnitude spectrogram domain:

$$\text{IRM}_i(t,f) = \frac{|c_i(f)x(t,f)|}{|c_i(f)x(t,f)| + |h_i(t,f) + n_i(t,f)|} \tag{13}$$

where i indicates the microphone index. In the IRM formulation, we treat the direct sound as the target signal and the remaining components as interference. The IRM estimation with supervised learning has been extensively used in monaural speech enhancement and robust ASR. See [14] for a recent review.

3. Speaker Verification and Experimental Setup

We conduct our experiments using the NIST retransmitted corpus [11]. This corpus consists of a subset of NIST SRE 2010 including 459 utterances from 150 female speakers with three

Table 1: *Results (%EER) for the real RIR dataset. First four rows are experiments on individual microphones and only "Best" and "Worst" performing microphones are reported. "Rank-1" indicates speech covariance matrix reconstructed by rank-1 approximation.*

SNR		**0 dB**		**5 dB**		**10 dB**		**15 dB**		**Average**	
		i-vector	x-vector	i-vector	x-vector	i-vector	x-vector	i-vector	x-vector	i-vector	x-vector
Unprocessed	Best Mic	35.01	25.47	24.74	17.51	17.82	12.05	12.79	8.28	22.59	15.83
	Worst Mic	42.24	34.59	36.48	28.72	29.77	22.22	23.17	16.98	32.92	25.63
Estimated IRM	Best Mic	29.98	23.27	22.12	16.67	15.30	11.43	11.43	8.28	19.71	14.91
	Worst Mic	37.63	32.39	31.66	27.04	26.00	21.28	20.65	16.04	28.99	24.19
BeamformIt		38.05	28.09	29.04	20.44	20.86	14.99	13.42	10.80	25.34	18.58
PMWF-0		35.85	25.47	26.42	18.13	18.03	13.10	14.05	9.12	23.59	16.45
PMWF-0 Rank-1		30.92	22.54	20.86	14.36	14.36	9.85	11.53	8.49	19.42	13.81
GEV-BAN		25.89	16.98	16.56	10.80	11.01	8.07	8.18	5.98	15.41	10.46
MVDR I (Eq. 6)		33.33	25.47	23.79	17.30	15.62	11.53	10.59	8.39	20.83	15.67
MVDR II (Eq. 7)		27.15	17.61	17.92	11.95	11.84	8.70	8.60	6.50	16.38	11.19
MVDR Rank-1		26.10	16.25	16.46	10.80	10.59	7.97	7.97	5.45	**15.28**	**10.12**

or five minute durations. The recordings are retransmitted by a loudspeaker in a highly reverberant environment[1]. Six microphones are placed for beamforming purposes. Their placement forms an ad-hoc microphone array with large inter-microphone distance in the range of 2.80 to 7.62 meters. In order to incorporate a typical microphone array with a small inter-microphone distance operating at reasonable reverberation times in our experiments, we create simulated room impulse responses (RIRs) using the Image method[2] and convolve them with the anechoic version of the NIST retransmitted dataset. We use the algorithm described in [22] to sample parameters of RIR simulation that is designed for one speaker and reverberation time (T60) in the range of 0.4 to 0.8 seconds for 6 linearly-arranged microphones with inter-microphone distance less than 0.09 meters. We denote the first dataset as the real RIR and second one as the simulated RIR.

To create diffuse babble noise, 10 speakers are first randomly selected from the TIMIT dataset, and then mixed together to generate 80-minute babble noise, which is split into two halves for training and testing. Following [23], babble noise is made diffuse under a spatial coherence constraint induced by the array geometry. Finally, we add the diffuse babble noise to the real and simulated RIR dataset. The sampling rate is 8 kHz.

We train a BLSTM for IRM estimation. The input feature is 129-dimensional log magnitude extracted using a frame length of 32 ms and a hop size of 8 ms. Global mean-variance normalization is performed on the features. The network includes 4 hidden layers, each with 300 units in each direction, and an output layer with 129 sigmoidal units. The cost function is mean squared error. A subset of NIST SRE 2008 is selected as training set which includes three- to five-minute telephone or interview conversations from female speakers. Same simulation procedure is used to create RIRs with T60 between 0.2 and 1 seconds. Diffuse babble noise is then added with signal-to-noise ratios (SNR) ranging from 0 dB to 15 dB. The total duration of the training data is 140 hours.

I-vectors [24] and x-vectors [4] serve as our baseline speaker recognition systems. Both of them are implemented in Kaldi [25]. I-vectors include a standard pipeline of feature extraction, a universal background model (UBM) based on GMM, i-vector extractor and PLDA. The i-vector training dataset contains 86,629 utterances from the PRISM dataset [26]. Twenty-dimensional MFCC feature is calculated every 10ms, based on a 20ms window length. Delta and acceleration features are also added to create a 60-dimensional feature. Cepstral mean normalization (CMN) is applied for a sliding window of 3s. The default energy-based VAD in Kaldi is applied. The number of full-covariance GMM components for UBM is set to 2048, and UBM is trained on a portion of training data (15,600 utterances). I-vectors with 600 dimensions are centered and projected to a 200 dimensional space by linear discriminant analysis (LDA) prior to PLDA scoring.

We use i-vector training data for x-vector training and augment it by adding two replicas from reverberation and babble noise. It is shown in [4] that data augmentation significantly improves speaker embedding performance. The reverberation replica is generated by convolving small or medium room RIRs from the OpenSRL dataset with clean training utterances. The babble noise replica is generated by 3-7 utterances from the MUSAN Babble dataset [27] and then mixed with training data with SNR in the range of 6-13 dB. We select 128,000 utterances randomly from both replicas and add to the x-vector training dataset. For training DNN embeddings, we remove those utterances that are less that 500 frames and those speakers that have less than 8 utterances. Totally, 169,660 utterances are generated and they include 5,565 speakers and 10,381 hours of data. Features are from a 23-channel Mel filterbank with the 25ms frame length and a 10ms frame shift. Similar to i-vectors, we apply CMN per utterance with the 3s sliding window and use the same VAD. X-vectors are subject to length-normalization before PLDA scoring.

4. Results and Discussion

Evaluation results for speaker verification are reported in terms of equal error-rate (EER). With the i-vector, we achieve 2.5% EER for clean test data (anechoic and without babble noise), and this number is reduced to 1.88% for the x-vector.

For a comprehensive comparison, we include other popular beamforming techniques in our experiments. Weighted delay-and-sum beamformer is implemented by using the BeamformIt toolkit [16], where the DOA estimate is obtained from GCC-PHAT. We also utilize parameterized multi-channel Wiener filter with parameter β (PMWF-β), which is describe in [28] and

[1]The authors did not report reverberation time (T60), but it can be inferred by listening.

[2]Available at https://github.com/ehabets/RIR-Generator.

Table 2: *Results (%EER) for the simulated RIR dataset. See Table 1 caption for notations.*

SNR		0 dB		5 dB		10 dB		15 dB		Average	
		i-vector	x-vector	i-vector	x-vector	i-vector	x-vector	i-vector	x-vector	i-vector	x-vector
Unprocessed	Best Mic	18.87	9.22	10.69	4.61	6.50	3.15	5.03	2.83	10.27	4.95
	Worst Mic	20.02	10.90	11.53	5.14	7.23	3.67	5.35	3.04	11.03	5.69
Estimated IRM	Best Mic	12.47	7.23	7.34	4.61	5.14	3.46	4.40	2.94	7.34	4.56
	Worst Mic	14.26	7.86	8.28	4.82	5.77	3.67	4.72	3.25	8.26	4.90
BeamformIt		18.45	10.06	10.59	5.14	6.92	3.25	5.03	2.62	10.25	5.27
PMWF-0		11.01	5.98	6.39	3.67	4.61	2.94	4.19	2.62	6.55	3.80
PMWF-0 Rank-1		6.92	4.30	4.30	2.83	3.35	2.31	3.35	2.20	4.48	2.91
GEV-BAN		6.39	4.19	4.09	2.83	3.46	2.31	3.25	2.10	**4.30**	**2.86**
MVDR I (Eq. 6)		10.06	5.24	5.24	3.46	4.09	2.73	3.56	2.73	5.74	3.54
MVDR II (Eq. 7)		7.44	4.51	4.61	3.35	3.77	2.73	3.56	2.62	4.85	3.30
MVDR Rank-1		6.60	4.61	4.40	3.04	3.46	2.52	3.25	2.41	4.43	3.14

later combined with T-F masking in [29], [30]. It is shown in [28] that by setting the parameter $\beta = 0$, PMWF-0 is tightly related to the MVDR beamformer. However, reference microphone selection is needed for PMWF-0. In this work, this is done by summating the estimated mask at each microphone over time and frequency and choosing the microphone with the largest summation [17]. Lastly, Generalized Eigenvalue beamformer with Blind Analytical Normalization post filter (GEV-BAN) described in [9], [31] is employed. All beamformers share the same mask estimator and we combine the estimated masks using median pooling before computing the speech and noise covariance matrices.

Table 1 shows the results of our experiments for the real RIR dataset. MVDR I and MVDR II refer to the MVDR beamformer which its speech covariance matrix derived using Eq. (6) and Eq. (7), respectively. In terms of the two speaker verification systems, the x-vector outperforms the i-vector in all cases. Note that although the x-vector is trained with 6-13 dB noisy utterances, the x-vector shows robustness to low SNR conditions.

We observed that applying ratio masking for monaural speaker verification leads to a consistent improvement. Nevertheless, error reduction of multi-channel speaker recognition is greater, especially at low SNRs. For example, using x-vector in 0 dB SNR, the result for the best performing microphone is 25.47% EER. Monaural speech enhancement reduces EER to 23.27%. On the other hand, using the rank-1 approximated MVDR beamformer reduces EER to 16.25%. On average, rank-1 approximated MVDR beamformer yields 22.47% and 32.12% relative improvement over the best monaural masking for the i-vector and x-vector, respectively.

We also observe that using rank-1 approximation for speech covariance matrix brings substantial improvements for both MVDR and PMWF-0 beamformers. Comparing the average improvements of rank-1 approximated MVDR and PMWF-0 with MVDR I and PMWF-0 for x-vector system, we observe an absolute EER reduction of 5.55% and 2.64%, respectively. Moreover, results obtained by MVDR II is better than MVDR I, supporting the argument that speech covariance matrix obtained by Eq. (7) leads to better steering vector estimation.

We report the results of the simulated RIR dataset in Table 2. The same improvement can be seen with rank-1 approximation. For the x-vector system, rank-1 approximated MVDR and PMWF-0 bring average improvement of 0.4% and 0.89% EER over MVDR I and PMWF-0, respectively. Note that improvements are smaller compared to real RIR dataset and it may be due to higher reverberation of real RIR dataset which is suppressed better by rank-1 approximation. It is worth noting that GEV-BAN, rank-1 approximated MVDR and PMWF-0 have comparable results for simulated RIR dataset. This is expected since beamformers are equivalent up to a scaling factor [13], [19]. However, the performance of rank-1 approximated PMWF-0 is worse than GEV-BAN and rank-1 approximated MVDR for real RIR dataset. This may be due to the fact that, inter-microphone distance is large and selecting reference microphone affects the optimal weights of the beamformer significantly. This indicates that choosing reference microphone can be the main reason for the poor performance of PMWF-0 and hence a better approach for selecting reference microphone is needed. Overall rank-1 approximated MVDR beamformer gives the best results in the real RIR dataset while GEV-BAN is performing better in the simulated RIR dataset.

5. Concluding Remarks

To our knowledge, this is the first study that introduces multi-channel speech enhancement to speaker recognition in order to deal with both diffuse noise and reverberation. The enhancement frontend is applied to conventional i-vector and state-of-the-art x-vector based speaker recognition systems. Different methods for extracting steering vectors from speech covariance matrices for the MVDR beamformer are explored. We have shown that speech covariance matrix reconstructed by rank-1 approximation gives the best result in the real RIR dataset. Combined with x-vector, rank-1 approximated MVDR reduces the EER by 5.55% absolutely compared to a standard version.

With straightforward implementation and robustness facilitated by data augmentation, x-vector based speaker recognition is a suitable substitute for i-vector based recognition. In addition, all the back-end techniques developed for i-vectors can be directly used for the DNN embeddings, like PLDA scoring and domain adaptation. Future research will explore training beamformers and DNN embeddings jointly, following the success of such joint training in robust ASR [32].

6. Acknowledgments

This research was supported in part by a National Science Foundation grant (ECCS-1808932) and the Ohio Supercomputer Center. The authors would like to thank L. Mošner and J. Černockỳ for providing data and discussions regarding [11].

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2012, pp. 4253–4256.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1695–1699.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[5] J. Chang and D. L. Wang, "Robust speaker recognition based on DNN/i-vectors and speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5415–5419.

[6] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5090–5094.

[7] M. Kolbœk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 305–311.

[8] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

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

[10] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 276–280.

[11] L. Mošner, P. Matějka, O. Novotnỳ, and J. H. Černockỳ, "Dereverberation and beamforming in far-field speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5254–5258.

[12] 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.

[13] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[14] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[16] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[17] Z.-Q. Wang and D. L. Wang, "Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5619–5623.

[18] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[19] 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.

[20] X. Sun, Z. Wang, R. Xia, J. Li, and Y. Yan, "Effect of steering vector estimation on MVDR beamformer for noisy speech recognition," in *IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[21] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[22] Z.-Q. Wang and D. L. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Proceedings of Interspeech*, vol. 2018, 2018, pp. 2718–2722.

[23] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[24] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 92–97.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[26] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 workshop*. Citeseer, 2011.

[27] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[28] 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.

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

[30] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, and S. Watanabe, "Multi-channel speech recognition: LSTMs all the way through," in *CHiME-4 workshop*, 2016, pp. 1–4.

[31] 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.

[32] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2632–2641.