# NEURAL NETWORK BASED TIME-FREQUENCY MASKING AND STEERING VECTOR ESTIMATION FOR TWO-CHANNEL MVDR BEAMFORMING

*Yuzhou Liu*[1,3] *Anshuman Ganguly*[2,3] *Krishna Kamath*[3] *Trausti Kristjansson*[3]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Electrical Engineering, The University of Texas at Dallas, USA
[3]Amazon Lab126, USA

## ABSTRACT

We present a neural network based approach to two-channel beamforming. First, single- and cross-channel spectral features are extracted to form a feature map for each utterance. A large neural network that is the concatenation of a convolution neural network (CNN), long short-term memory recurrent neural network (LSTM-RNN) and deep neural network (DNN) is then employed to estimate frame-level speech and noise masks. Later, these predicted masks are used to compute cross-power spectral density (CPSD) matrices which are used to estimate the minimum variance distortion-less response (MVDR) beamformer coefficients. In the end, a DNN is trained to optimize the phase in the estimated steering vectors to make it robust for reverberant conditions. We compare our methods with two state-of-the-art two-channel speech enhancement systems, i.e., time-frequency masking and masking-based beamforming. Results show the proposed method leads to 21% relative improvement in word error rate (WER) over other systems.

***Index Terms***— Two-channel speech enhancement, MVDR beamforming, steering vector, neural networks

## 1. INTRODUCTION

Automatic speech recognition (ASR) has been growing rapidly in recent years due to the introduction of deep neural networks (DNNs) for acoustic modeling (AM) and language modeling (LM). However, it is still very challenging to recognize speech from the far field, where speech signal is severely corrupted by noise and reverberation. Given multi-channel recordings, many speech enhancement methods have been proposed as front-end processing for far-field ASR, and they lead to substantial performance gain over unprocessed speech [8, 11, 14, 19]. This paper is concerned with far-field speech enhancement when only two-channel recordings are available.

Conventional two-channel speech enhancement methods can be broadly categorized into two groups: non-linear and linear. Non-linear techniques suppress noise non-linearly based on the statistical information of speech or noise. One of the most popular non-linear approaches is time-frequency (T-F) masking [27]. In T-F masking, cross-channel features, e.g., level differences, time differences, cross-correlation and phase differences, are first computed at each T-F unit of the input utterance. A T-F level binary or ratio mask that attenuates noise and retains speech is then estimated from the features, using thresholding [1, 14], DNNs [5, 13, 18], or clustering [9, 23]. T-F masking based approaches work well in terms of noise suppression, but they may introduce artifacts in the processed speech, which degrades the performance of ASR with multi-condition DNN-AMs.

Two-channel linear approaches are usually based on beamforming, where complex-valued linear filters are applied to multi-channel input, and all the channels are then added together to enhance the target speech. As one of the most widely-used beamforming techniques, minimum variance distortion-less response (MVDR) beamforming tries to minimize the power of the beamformed signal, while keeping unity gain at the look direction. To obtain MVDR filter coefficients [2, 7], we need to estimate the spatial statistics of the target speech and noise, including steering vectors of the look direction which can be estimated from speech cross-power spectral density (CPSD) matrices, and noise CPSD matrices.

Recently, much effort has been made to combine T-F masking with beamforming [4, 8, 9, 19, 28]. The general idea is to estimate speech and noise spectral masks, and apply them to the noisy speech so that spatial statistics and beamforming coefficients can be easily derived. In [9], a complex Gaussian mixture model (CGMM) is built to estimate masks for noise. Speech and noise CPSD matrices are then estimated from the masked signals, and steering vectors are estimated using the principal eigenvectors of speech CPSD matrices. Finally, the enhanced speech is obtained using MVDR beamforming. In [8], Heymann et al. propose to use short time Fourier transform (STFT) features and bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) to estimate speech and noise masks for each channel. A median mask across all the channels is then obtained for computation of beamforming coefficients. In [28], Xiao et al. extend Heymann et al.'s approach by including ASR cost function and iterative processing in mask estimation. In [19], BLSTM based masks are utilized to initialize spatial clustering based masks, leading to improved generalization in mask estimation. To sum up, masking based beamforming does not require the prior knowledge of microphone array geometry, and has been shown to be robust in real noisy environments.

In this paper, we extend Heymann et al.'s masking based beamforming [8] for two-microphone setup in three major aspects. First, to better utilize information in the two channels for mask estimation, we stack two-channel STFT as a T-F feature map. A large neural network concatenated by a convolutional neural network (CNN), BLSTM and DNN takes the feature map as input to predict the mean masks of the two channels. Second, we include several cross-channel features to enrich the feature map. Lastly, we observe that the steering vectors estimated by principal eigenvectors of speech CPSD matrices fail to match the target look direction if speech is severely corrupted by reverberation. To address this issue, we use estimated steering vectors and masks as features, and train a DNN to optimize the phase in the estimated steering vector.

In the remainder of this paper, after reviewing existing masking-based MVDR beamforming in Section 2, the proposed method is described in Section 3. In Section 4, we present experimental results and comparisons. A conclusion is given in Section 5.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

## 2. REVIEW OF MASKING BASED MVDR BEAMFORMING

In this section, we briefly describe conventional MVDR beamforming and Heymann et al.'s masking based MVDR beamforming.

Noisy and reverberant speech signals received at the two microphones are denoted as:

$$\begin{aligned} Y_m(t,f) &= h_m(f)S(t,f) + N_m(t,f) && (1)\\ &= X_m(t,f) + N_m(t,f), \qquad \text{for} \quad m = 1,2 && (2)\end{aligned}$$

with $S(t,f)$ denoting the clean speech STFT at time frame $t$ and frequency channel $f$, $h_m(f)$ denoting the room impulse response between the speaker and the $m^{th}$ microphone, $Y_m(t,f)$, $X_m(t,f)$ and $N_m(t,f)$ denoting the STFT of noisy reverberant speech, reverberant speech and noise received at the the $m^{th}$ microphone. The variables can also be written in vector notation by omitting the microphone index: $\mathbf{Y}(t,f)$, $\mathbf{X}(t,f)$, $\mathbf{N}(t,f)$ and $\mathbf{h}(f)$.

The goal of MVDR beamforming is to recover the clean speech with a linear filter $\mathbf{w}(f)$,

$$\widetilde{S}(t,f) = \mathbf{w}^{\mathrm{H}}(f)\mathbf{Y}(t,f) \tag{3}$$

such that the energy of beamformed signal is minimized, and unity gain is maintained at the look direction. The close form solution to the optimization problem is:

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_{\mathrm{NN}}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^{\mathrm{H}}(f)\mathbf{\Phi}_{\mathrm{NN}}^{-1}(f)\mathbf{d}(f)} \tag{4}$$

where $\mathbf{d}(f)$ is the steering vector of the microphone array, and $\mathbf{\Phi}_{\mathrm{NN}}(f)$ is the noise CPSD matrix.

Conventionally, the steering vector can be estimated using the time difference of arrival (TDOA) between the two microphones [2]. Taking the first microphone as the reference microphone, we have:

$$\mathbf{d}(f) = [1 \quad e^{-j2\pi f\tau}] \tag{5}$$

where $\tau$ is the TDOA. Noise CPSD matrix can be calculated on noise-only frames estimated by voice activity detection (VAD).

Unlike conventional methods, masking based MVDR beamforming [8] first estimates speech and noise masks of the mixture with BLSTM, and then calculates MVDR coefficients using masked signals. The details are described in the following two subsections.

### 2.1. BLSTM based Mask Estimation

In [8], the mask estimator consists of multiple BLSTM-RNNs with shared weights, one for each channel. The input feature of each BLSTM is a single frame of 513-dimensional STFT magnitude of one channel, with 16-kHz sampling rate, 64-ms frame size, 16-ms frame shift, and 1024 FFT size.

The target binary masks for speech and noise are defined as:

$$\mathrm{IBM}_{\mathrm{X}}(t,f) = \begin{cases} 1, & \text{if } \frac{|X(t,f)|}{|N(t,f)|} > 10^{\mathrm{th}_{\mathrm{X}}(f)}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathrm{IBM}_{\mathrm{N}}(t,f) = \begin{cases} 1, & \text{if } \frac{|X(t,f)|}{|N(t,f)|} < 10^{\mathrm{th}_{\mathrm{N}}(f)}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where $\mathrm{th}_{\mathrm{X}}(f)$ and $\mathrm{th}_{\mathrm{N}}(f)$ are thresholds for speech and noise masks at each frequency bin.

The BLSTM-RNN consists of three hidden layers: a 256-unit BLSTM layer followed by two 513-unit feedforward layers with the ReLU activation function [6]. The output layer has 1026 sigmoid units estimating frame-level speech and noise masks simultaneously. Batch normalization [12] and dropout [10] are applied throughout the network. The cross-entropy cost function, utterance-level backpropagation through time, and Adam optimization algorithm [15] are used during training.

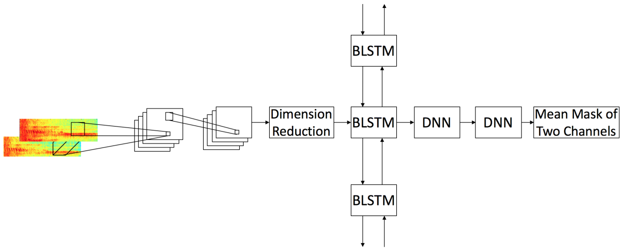


**Fig. 1**: Diagram of the CLDNN based mask estimator.

After estimating masks at each channel, a median speech mask and noise mask of all the channels, $M_{\mathrm{X}}(t,f)$ and $M_{\mathrm{N}}(t,f)$, are obtained for beamforming coefficients computation.

### 2.2. MVDR Coefficients Computation

The speech and noise CPSD matrices can be calculated as:

$$\mathbf{\Phi}_{\mathrm{VV}}(f) = \frac{\sum_{t=1}^{T} M_{\mathrm{V}}(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} M_{\mathrm{V}}(t,f)} \tag{8}$$

where $\mathrm{V} \in \{\mathrm{X}, \mathrm{N}\}$. The steering vector can be estimated as the principal component of the estimated speech CPSD matrix:

$$\mathbf{d}(f) = P\{\mathbf{\Phi}_{\mathrm{XX}}(f)\} \tag{9}$$

In the end, we calculate MVDR filter coefficients using Eq. (4) and apply the filters to two-channel recordings to get the enhanced speech.

## 3. PROPOSED SYSTEM

Although Heymann et al.'s masking based beamforming [8] works well for realistic noise and can be readily applied to 2-microphone setup, it still has limitations: it does not make full use of multichannel information and does not work well in reverberation. In this section, we propose three extensions to [8] to improve its performance for noisy and reverberant speech.

### 3.1. Extension One: Dual-channel Neural Network

The mask estimator in [8] only takes single-channel STFT as input, which is suboptimal as there is far richer information in two-channel inputs. To address this issue, we propose to stack two-channel STFT coefficients as a feature map and use a CNN-BLSTM-DNN (CLDNN) based neural network to predict the mean ideal masks of the two channels. The diagram of network is given in Fig. 1.

CLDNN based acoustic model has been explored in [24], and has been shown to outperform LSTM based AM as it takes advantage of the complementarity of different neural networks. In this paper, the CNN part of the CLDNN consists of two convolutional layers and a dimensionality reduction layer. The input to the first convolutional layer is stacked 2-channel STFT magnitude with a context window size of 25 frames, 12 frames before and 12 frames after the current frame. The two convolutional layers both have 32 feature maps. A 9x9 T-F filter is used for the first convolutional layer, followed by a 4x3 filter for the second convolutional layer. The pooling strategy is non-overlapping max pooling in frequency with a pooling size of 3, and pooling is only performed for the first hidden layer. After the second convolutional layer, a 513-unit linear layer is employed to reduce the dimensionality of the feature map. Batch normalization is included in all CNN layers. All other details, including the BLSTM and DNN part of the CLDNN, and training recipes, follow those in Section 2.1.

### 3.2. Extension Two: Cross-channel Features

To fully make use of information in all channels, more cross-channel features should be added to the input of the mask estimator. Many studies have investigated cross-channel features, e.g., level differences, time differences, cross-correlation and phase differences, for two-channel T-F masking [1, 13, 14] and direction of arrival estimation [17]. However, since cross-correlation based features work poorly on close microphone distances, we decide to use phase differences and CPSD as cross-channel features for mask estimation.

The phase difference between the two channels is computed as: $PD(t,f) = \varphi(Y_1(t,f)Y_2^{\mathrm{H}}(t,f))$, where $\varphi$ denotes the phase of a complex number. We also include the cosine value of phase difference $CPD(t,f)$ in the feature as a smoother version of $PD(t,f)$. Phase difference based features are useful for separating sources from different directions in less reverberant conditions.

CPSD based cross-channel features are computed as:

$$RCPSD(t,f) = \log(\text{abs}(\text{real}(Y_1(t,f)Y_2^{\mathrm{H}}(t,f)))) \qquad (10)$$

$$ICPSD(t,f) = \log(\text{abs}(\text{imag}(Y_1(t,f)Y_2^{\mathrm{H}}(t,f)))) \qquad (11)$$

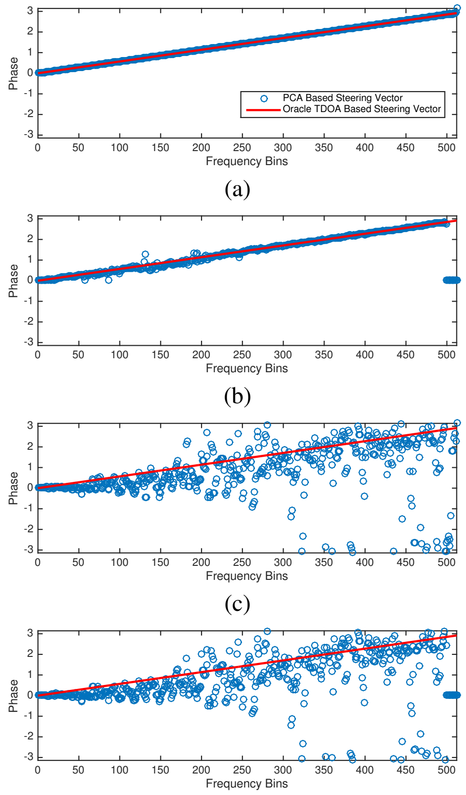
where $RCPSD$ and $ICPSD$ correspond to the real and imaginary part of CPSD, respectively. We apply the absolute and logarithm operation to compress the dynamic range of both features. Since the imaginary part of ideal diffuse noise CPSD is always close to 0 [22], $ICPSD$ tends to have larger values at speech-dominant T-F units in diffuse noise, which makes it very effective for separating speech from diffuse noise.

All proposed cross-channel features have the same dimension as single-channel STFT, thus can be readily stacked into the feature map. The final feature map of the CLDNN has a depth of 6.

### 3.3. Extension Three: Post-Processing for Steering Vectors

In conventional TDOA based steering vectors (see Eq. (5)), the phase difference between the two channels $\varphi(\mathrm{d}_2(f)/\mathrm{d}_1(f)) = -2\pi f\tau$ is proportional to the frequency index and TDOA. On the other hand, steering vectors in masking based MVDR [8] correspond to the principal components of speech CPSD matrices.

In Fig. 2, we compare $\varphi(\mathrm{d}_2(f)/\mathrm{d}_1(f))$ in TDOA based steering vectors (Eq. (5)) and PCA based steering vectors (Eq. (9)) in several different scenarios. The horizontal axis corresponds to frequency bins, and the vertical axis corresponds to phase difference. The TDOA based steering vectors in all subfigures are calculated using the oracle TDOA. If clean anechoic speech arrived at each channel is used to compute PCA based steering vector, $\varphi(\mathrm{d}_2(f)/\mathrm{d}_1(f))$ estimated by the two approaches almost perfectly matches, as shown in Fig. 2(a). If ideal-binary-masked anechoic noisy speech is used for the PCA based method, the phase differences by the two approaches become a little different, but the gap is still very small as in Fig. 2(b). However, as shown in Fig. 2(c) and 2(d), if reverberant speech is used instead, the resulting $\varphi(\mathrm{d}_2(f)/\mathrm{d}_1(f))$ in PCA based steering vectors starts to deviate a lot from that in TDOA based steering vectors, in other words, the resulting PCA based steering vector does not match the desired look direction anymore.



(a)

(b)

(c)

(d)

**Fig. 2**: Phase difference between the two channels in steering vectors: (a) PCA based steering vectors estimated from clean speech. (b) PCA based steering vectors estimated from ideal-binary-masked noisy anechoic speech (-5 dB AC noise). (c) PCA based steering vectors estimated from reverberant speech (400 ms $T_{60}$). (d) PCA based steering vectors estimated from ideal-binary-masked noisy reverberant speech (-5 dB AC noise, 400 ms $T_{60}$).

DNN based direction of arrival estimation has been extensively explored recently. Some studies analyze frequency-domain features [3, 25], while some use T-F masking as front-end preprocessing [20]. Inspired by these studies, we propose to use a DNN to predict the TDOA, and fix the scattered phase difference in PCA based steering vector accordingly. Two features are used in the DNN. The first is the original $\varphi(\mathrm{d}_2(f)/\mathrm{d}_1(f))$ calculated from PCA based steering vector. The second is $\sum_{t=1}^{T} M_{\mathrm{X}}(t,f)$, which corresponds to the relative power of speech in each frequency bin. We normalize the two features, and concatenate them to get a 1026-dimensional feature vector for each utterance. There are two hidden layers in the DNN, each with 80 ReLU units. The output of the DNN estimates the oracle TDOA of the speech. Based on the distance between the two microphones, we linearly discretize all possible TDOAs into 30 classes, thus a 30-unit softmax output layer is used. During test, the estimated TDOA $\widetilde{\tau}$ is given by the inner product of softmax outputs and mean TDOAs of each class.

After the estimating $\widetilde{\tau}$, we keep the absolute value of PCA based steering vectors $|\mathrm{d}_1(f)|$ and $|\mathrm{d}_2(f)|$, and only change their corresponding phase to $\varphi(\mathrm{d}_1(f)) = 1$ and $\varphi(\mathrm{d}_2(f)) = -2\pi f\widetilde{\tau}$. The resulting steering vector is then used to compute filter coefficients.

**Table 1**: WER comparison (%) on the simulated data set.

| Speech Enhancement | Development | Evaluation |
|---|---|---|
| Unprocessed Channel 1 | 35.84 | 41.31 |
| T-F Masking | 31.72 | 38.39 |
| Masking based MVDR | 24.69 | 30.06 |
| Extension 1 | 21.10 | 25.16 |
| Extension 1+2 | 20.58 | 24.19 |
| Extension 1+2+3 | 19.00 | 23.67 |
| Extension 1+2+Oracle TDOA based Steering Vector | 18.59 | 21.63 |

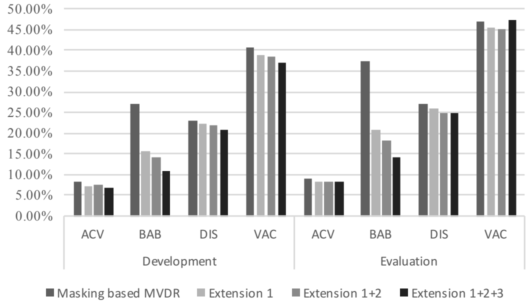


**Fig. 3**: WER comparison in terms of noise type.

## 4. EVALUATION RESULTS AND COMPARISONS

To evaluate the proposed methods in very noisy and reverberant conditions, we create a simulated corpus based on the CHiME-4 dataset [26]. There are 3 subsets in the corpus, namely training, development and evaluation set. 7138, 100, and 100 clean utterances respectively are to used to create mixtures in each subset. All the clean utterances are from the clean Wall Street Journal recordings in CHiME-4's simulated data. We use the fast image-source method [16] to simulate room impulse response. The room dimension, microphone-array location, speaker location and noise location are randomly generated for each utterance in the corpus. The distance between the two microphones is set to 2.2 cm to prevent aliasing and phase wrapping. Four types of noise are used, namely AC (ACV), babble (BAB), dish washer (DIS) and vacuum cleaner (VAC), with the first two as diffuse noise and the other two as point noise. In the training set, each clean utterance is mixed with a random noise at a random signal to noise ratio (SNR) in -5, 0, 5 dB, and at a random $T_{60}$ in 200 and 400 ms. In the development and evaluation set, each clean utterance is systematically mixed with 4 noise types at all SNRs in -5, 0, 5 dB, and all $T_{60}$ in 200 and 400 ms. In total, we have 7138 simulated mixtures in the training set, and 2400 simulated mixtures in the development and evaluation set, each.

The speech recognizer used in this paper follows the CHiME-4 ASR baseline in Kaldi [21], i.e., a DNN acoustic model and a RNN language model in addition to a 5-gram language model. The DNN-AM is trained using the first-channel mixtures in the training set.

We compare the proposed speech enhancement methods with T-F masking and masking based MVDR [8]. For T-F masking, the feature map and CLDNN in Section 3 are used to predict a soft speech mask. The estimated mask is then directly applied to noisy STFT. For masking based MVDR, we implement Heymann et al.'s system [8] and match their reported results on CHiME-4. The network is then retrained for the new data set. Table 1 summarizes the WER obtained in the experiments. Each extension to masking based MVDR incrementally improve the ASR performance on both development and evaluation set. The final system (extension 1+2+3) leads to more than 21% of WER reduction comparing with baseline approaches. In addition, we combine the first two extensions with oracle TDOA based steering vector, and report its WER in the last row. Results indicate that extension 3 substantially reduces the performance gap between PCA based and oracle TDOA based steering vectors.

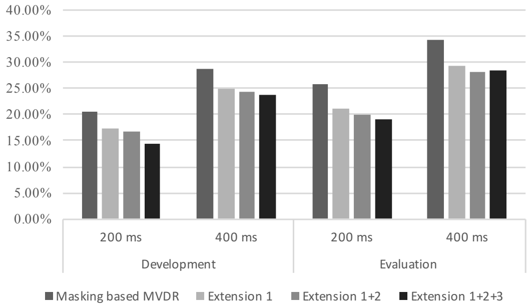


**Fig. 4**: WER comparison in terms of $T_{60}$.

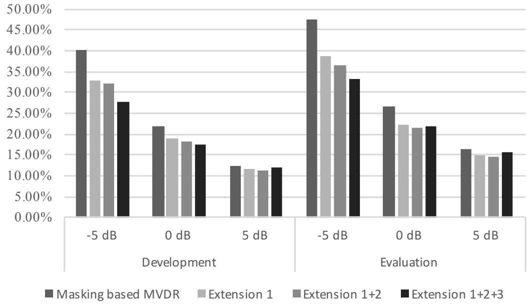


**Fig. 5**: WER comparison in terms of SNR.

Fig. 3, 4 and 5 compare the proposed methods with Heymann et al.'s masking based MVDR in terms of different noise types, $T_{60}$s and SNRs. As observed in the figures, the proposed extensions systematically and incrementally reduce WER in almost all noise types and $T_{60}$s. The largest improvement comes from diffuse babble noise, where WER is cut by more than half. WER slightly increases for vacuum cleaner noise in the evaluation set when extension 3 is employed, probably because the noise is very broadband so that the derived PCA based steering vector is too noisy to make an accurate TDOA estimation. In Fig. 5, it is shown that the proposed extensions work especially well for very low SNRs, and match the baseline system in high-SNR conditions.

## 5. CONCLUSION

We have proposed three extensions, namely dual-channel neural networks, cross-channel features and steering vector post-processing, for masking based MVDR beamforming. Experimental results show that the proposed two-channel speech enhancement algorithm greatly improves ASR performance. It should be mentioned that the proposed system only includes linear beamforming. Non-linear masking based approaches can be utilized as post-filters to further reduce the residual noise in the beamformed signal.

## 6. REFERENCES

[1] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, pp. 1763–1773, 2004.

[2] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. J. T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, p. 61, 2015.

[3] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," *arXiv:1705.00919*, 2017.

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

[5] N. Fan, J. Du, and L. R. Dai, "A regression approach to binaural speech segregation via deep neural network," in *Proc. ISCSLP*, 2016, pp. 1–5.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011, pp. 315–323.

[7] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[9] T. Higuchi, N. Ho, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/ offline ASR in noise," in *Proc. ICASSP*, 2016, pp. 5210–5214.

[10] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[11] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. L. Roux, V. Mitra, and S. Watanabe, "Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend," *Computer Speech and Language*, 2017.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[13] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 2112–2121, 2014.

[14] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *Proceedings of Interspeech*, 2009, pp. 2495–2498.

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[16] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1429–1439, 2010.

[17] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proceedings of Interspeech*, 2015, pp. 160–164.

[18] J. M. Martn-Donas, A. M. Gomez, I. López-Espejo, and A. M. Peinado, "Dual-channel DNN-based speech enhancement for smartphones," in *Proc. MMSP*, 2017.

[19] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proc. ICASSP*, 2017, pp. 286–290.

[20] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. ICASSP*, 2017, pp. 6125–6129.

[21] D. Povey, A. Ghoshal, and G. Boulianne, "The kaldi speech recognition toolkit," in *Proc. IEEE ASUR*, 2011.

[22] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Signal Processing*, vol. 89, pp. 703–709, 2009.

[23] S. Rickard, *Blind Speech Separation*. New York: Springer, 2007, ch. The DUET Blind Source Separation Algorithm, pp. 217–237.

[24] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, 2015, pp. 4580–4584.

[25] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. ICASSP*, 2016, pp. 405–409.

[26] 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," in *Computer Speech and Language*, 2016.

[27] D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.

[28] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *Proc. ICASSP*, 2017, pp. 3246–3250.