INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

GEV Beamforming Supported by DOA-based Masks Generated on Pairs of Microphones

François Grondin, Jean-Samuel Lauzon, Jonathan Vincent, François Michaud

Université de Sherbrooke, Sherbrooke (Québec), Canada
{francois.grondin2,jean-samuel.lauzon,jonathan.vincent2,francois.michaud}@usherbrooke.ca

Abstract

Distant speech processing is a challenging task, especially when dealing with the cocktail party effect. Sound source separation is thus often required as a preprocessing step prior to speech recognition to improve the signal to distortion ratio (SDR). Recently, a combination of beamforming and speech separation networks have been proposed to improve the target source quality in the direction of arrival of interest. However, with this type of approach, the neural network needs to be trained in advance for a specific microphone array geometry, which limits versatility when adding/removing microphones, or changing the shape of the array. The solution presented in this paper is to train a neural network on pairs of microphones with different spacing and acoustic environmental conditions, and then use this network to estimate a time-frequency mask from all the pairs of microphones forming the array with an arbitrary shape. Using this mask, the target and noise covariance matrices can be estimated, and then used to perform generalized eigenvalue (GEV) beamforming. Results show that the proposed approach improves the SDR from 4.78 dB to 7.69 dB on average, for various microphone array geometries that correspond to commercially available hardware.

Index Terms: speech separation, GEV beamforming, direction of arrival, microphone array

1. Introduction

Distant speech processing is a challenging task, as the target speech signal is often corrupted by additive noise and reverberation from the environment [1]. Moreover, robust speech recognition often relies on sound source separation when dealing with the cocktail party effect. Speech separation methods can be divided in two main categories: blind speech separation and informed speech separation. Blind speech separation relies strictly on the mixture spectrogram to restore the individual sources, whereas informed speech separation uses additional information such as video, direction of arrival and speaker features.

Blind speech separation is particularly challenging as it needs to solve the permutation ambiguity. In fact, the order of the separated signals may differ from the order of the labels, which makes supervised learning difficult. To solve this issue, deep clustering (DC) uses contrastive embedding vectors and unsupervised clustering using k-means [2, 3, 4, 5, 6]. Alternatively, permutation invariant training (PIT) aims to find all possible permutations during training and keep the optimal one [7, 8, 9]. These methods aim to separate all sources in the mixture, though sometimes only a specific target source matters. In the latter case, using an objective function that emphasizes only on the target source leads to better performance [10].

Informed speech separation relies on additional information to perform separation. For instance, SpeakerBeam uses the speaker identification features to extract a specific speaker from a mixture [11]. When the video is also available, it is possible to solve the permutation issue by combining the audio signal and the motion of the lips [12, 13, 14]. Moreover, when dealing with multiple channels, it is often common to use the direction of arrival (DOA) of sound to solve permutation [15]. Beamforming is thus a special case of informed speech separation, as it exploits the spatial information to reconstruct the target source. This paper focuses on a beamforming approach that exploits the target source DOA information.

Delay and Sum (DS) and Minimum Variance Distortionless Response (MVDR) beamformers [16, 17, 18] improve the signal to noise ratio (SNR) of the target sound source, but relies on DOA of sound derived from the anechoic model for sound propagation, which often differs from the actual condition in a reverberant environment. On the otherhand, Generalized eigenvalue decomposition (GEV) beamforming maximizes the SNR using only the target and interfering signals covariance matrices [19]. Heymann et al. [20, 21, 22] show that these covariance matrices can be estimated with a bi-directional Long Short-Term Memory (BLSTM) network trained on noisy speech, and that blind analytic normalization (BAN) gain minimizes non-linear distortion for the separated signal. To be effective, this approach assumes that the interfering sounds differ from speech, which is a major limitation when dealing with the cocktail party effect. Chen et al. [23] propose to use the DOA of sound to estimate a time-frequency mask of a target source with a neural network, and then use this mask to compute the target and noise covariance matrices. This approach performs well but has one major drawback: the geometry of the microphone array needs to be known prior to training the neural network, which impacts considerably the versability of the system when dealing with microphone arrays of arbitrary shapes. Maldonado et al. [24] present a solution to deal with the arbitrary shape, but the time-frequency mask obtained from the microphone array DOA is essentially applied to a single channel to extract the target spectrum and no further beamforming is used during separation. Liu et al. [25] also propose to estimate a time-frequency mask based on the cross-spectrum between two microphones and the target time difference of arrival (TDOA), but their approach is limited to two microphones and the spacing between the microphones is fixed.

The method presented in this paper, called SteerNet, relies on a neural network trained on pairs of microphones with different spacing. This network generates a time-frequency soft mask for each pair of microphones for a set of target TDOAs, obtained from the DOA of the target source and the array geometry. These masks are combined and used to compute target and noise covariance matrices and to perform GEV beamforming. This method is appealing as it makes the best use of GEV beamforming using DOA to solve permutation, while being able to generalize to microphone arrays of arbitrary shapes.

http://dx.doi.org/10.21437/Interspeech.2020-2687

2. SteerNet

Figure 1 shows the SteerNet method to separate a target speech source using a microphone array with an arbitrary geometry. In this scenario, there are two speech sources, the target and interference, and it is assumed that these sources have different DOAs. SteerNet assumes that the DOA of the target speech is available and is obtained using sound source localization methods [26, 27, 28], or using a visual cue when both optical and acoustic images are properly aligned [29].

Using the target DOA, the idea is to generate a time-frequency mask using a neural network to capture the target source components. To deal with arbitrary geometries, the method breaks down the shape of the array in pairs of microphones, and use the TDOAs between microphones to generate multiple masks. Masks that put emphasis on the target source can be estimated when the target and interference TDOAs are different as the permutation is easily solved. This is the case with most pairs of microphones, yet some of them can have similar TDOAs. When both TDOAs are similar, SteerNet generates a mask that capture both sources, as permutation cannot be solved spatially. The overall target source mask is obtained by summing all the estimated masks amongst all pairs of microphones. This leads to a target mask that emphasizes the time-frequency region dominated by the target source due to the pairs of microphones that allow discrimination between sources. The noise mask is obtained as the complement of the target. The approach finally uses GEV-BAN beamforming, which relies on the target and noise covariance matrices (denoted as $\mathbf{\Phi_{XX}}$ and $\mathbf{\Phi_{NN}}$, respectively) obtained from the estimated masks.

2.1. Oracle pairwise ratio mask

Let's define the DOAs of the target and interference as $\boldsymbol{\theta}_t \in \mathcal{S}^2$ and $\boldsymbol{\theta}_i \in \mathcal{S}^2$ respectively, where $\mathcal{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 = 1\}$, holds unit vectors and $\|\dots\|_2$ stands for the Euclidean norm. Let's also define the set of indexes of microphone pairs as $\mathcal{Q} = \{(x, y) \in \mathcal{D}^2 : x < y\}$, where the set $\mathcal{D} = \{1, 2, \dots, D\}$ contains the microphone indexes, and D stands for the number of microphones. The TDOA for microphones $(u, v) \in \mathcal{Q}$ corresponds to the following expression:

$$\tau_{u,v} = \frac{f_S}{c}(\mathbf{r}_u - \mathbf{r}_v) \cdot \boldsymbol{\theta}_t, \tag{1}$$

where $\mathbf{r}_u$, $\mathbf{r}_v$, f_S and c stand for the positions of microphones u and v (in m), the sample rate (in sample/sec) and speed of sound (in m/sec), respectively. The steering vector in the direction of the target for a pair of microphones is defined as:

$$A_{u,v}(t, f) = \exp\left(j\frac{2\pi f \tau_{u,v}}{N}\right), \tag{2}$$

where N stands for the number of samples per frame in the Short Time Fourier Tranform (STFT), t the frame index, and f the frequency bin index.

Similarly, the difference between TDOAs for a given microphone pair associated to the target and interfering speech sources is estimated as:

$$\Delta\tau_{u,v} = \frac{f_S}{c}|(\boldsymbol{\theta}_t - \boldsymbol{\theta}_i) \cdot (\mathbf{r}_u - \mathbf{r}_v)|. \tag{3}$$

The steering vector aims to cancel the phase difference of the target source in the cross-spectrum between microphones u and v:

$$Y_{u,v}(t, f) = A_{u,v}(t, f)Y_u(t, f)Y_v(t, f)^*, \tag{4}$$

where the expression $\{\dots\}^*$ stands for the complex conjugate, and $Y_u(t, f)$ and $Y_v(t, f)$ stand for the spectra of microphones u and v, respectively. The gain $G_{u,v}$ is defined as a function of $\Delta\tau_{u,v}$, where the goal is to ensure it goes to a value of 1 when both TDOAs are similar, and goes to zero when they are different. To smooth the transition and control the sharpness, a sigmoid function is used (α is the steepness and β the offset):

$$G_{u,v} = \frac{\exp\{-\alpha(\Delta\tau_{u,v} - \beta)\}}{1 + \exp\{-\alpha(\Delta\tau_{u,v} - \beta)\}}. \tag{5}$$

This gain is then used to generate the ideal ratio mask for microphones u and v:

$$M_{\bar{u},v}(t, f) = \frac{|S_u(t, f)|^2 + G_{u,v}|I_u(t, f)|^2}{|S_u(t, f)|^2 + |I_u(t, f)|^2 + |B_u(t, f)|^2}, \tag{6}$$

$$M_{u,\bar{v}}(t, f) = \frac{|S_v(t, f)|^2 + G_{u,v}|I_v(t, f)|^2}{|S_v(t, f)|^2 + |I_v(t, f)|^2 + |B_v(t, f)|^2}. \tag{7}$$

When both the target and interference share a similar TDOA, the gain goes to one and the oracle mask captures both the target ($S_u(t, f)$ and $S_v(t, f)$) and interference ($I_u(t, f)$ and $I_v(t, f)$), and rejects the diffuse background noise ($B_u(t, f)$ and $B_v(t, f)$). On the otherhand, when discrimination between the target and interference is possible due to different TDOAs, the gain goes to 0 and the oracle mask captures only the target source. Finally, the mask for a pair of microphones (u, v) (denoted as $M_{u,v}(t, f)$) is obtained as follows:

$$M_{u,v}(t, f) = M_{\bar{u},v}(t, f)M_{u,\bar{v}}(t, f). \tag{8}$$

2.2. Mask estimation using BLSTM

To estimate the mask, the method first extracts the log absolute value $\mathbf{L}_{u,v} \in [0, +\infty]$ and the phase $\mathbf{P}_{u,v} \in [-\pi, +\pi]$ from the cross-spectrum $\mathbf{Y}_{u,v}$ as:

$$\mathbf{L}_{u,v} = \log(\|\mathbf{Y}_{u,v}\|_2^2 + \epsilon) - \log(\epsilon), \tag{9}$$

$$\mathbf{P}_{u,v} = \angle\mathbf{Y}_{u,v}, \tag{10}$$

where the constant ϵ holds a small value (here set to 10^{-20}) to avoid large negative values as the energy goes to zero and $\angle$ stands for the angle. Both features are then concatenated as:

$$\mathbf{C}_{u,v} = (\mathbf{L}_{u,v}, \mathbf{P}_{u,v}). \tag{11}$$

The ideal mask $\mathbf{M}_{u,v} \in \mathcal{U}^{T\times F}$ introduced in (8) is then estimated from $\mathbf{C}_{u,v} \in \mathbb{R}^{T\times 2F}$ using the following non-linear function:

$$g : \mathbb{R}^{T\times 2F} \rightarrow \mathcal{U}^{T\times F}, \tag{12}$$

where the set $\mathcal{U} = [0, 1]$ as the soft mask lies between 0 and 1, T stands for the number of frames and F for the number of frequency bins. For this task, the method uses a BLSTM [30] with two layers with a hidden size of $2H = 256$ and one dropout layer (with a probability of $p = 0.2$), as shown in Fig. 2. A batch norm layer is also added to speed up convergence while training. The BLSTM generates the estimated mask $\hat{\mathbf{M}}_{u,v}$ for each microphone pair at index (u, v):

$$\hat{\mathbf{M}}_{u,v} = g(\mathbf{C}_{u,v}). \tag{13}$$

During training, the loss function L corresponds to the mean square error weighted by the log absolute value of the cross-spectrum to give more weight to time-frequency regions dominated by speech [31] and ignore silence periods:

$$L = \|(\mathbf{M}_{u,v} - \hat{\mathbf{M}}_{u,v}) \odot \mathbf{L}_{u,v}\|_2^2, \tag{14}$$

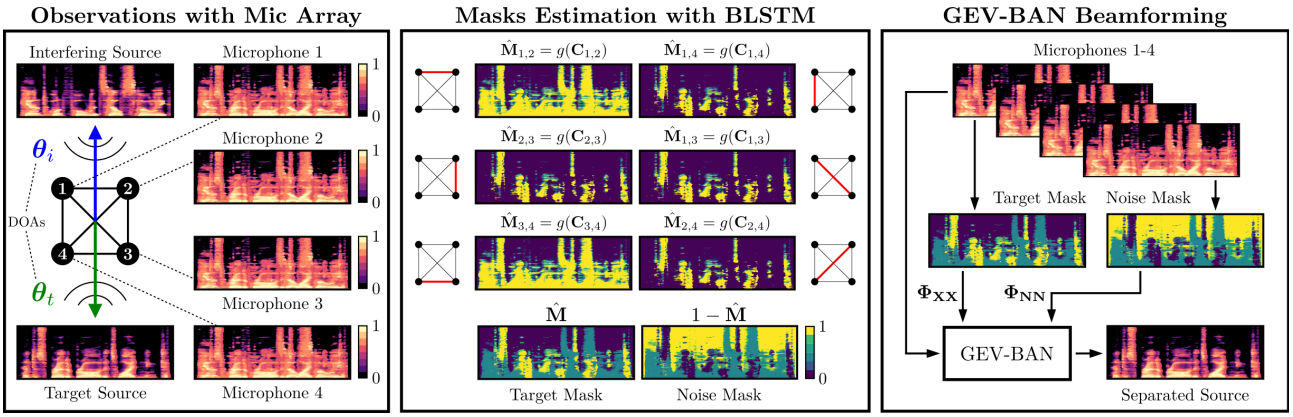


Figure 1: *Overview of SteerNet. In this example, the TDOAs of the target and interference are identical for the pairs (1,2) and (3,4), and thus the mask captures both sources, whereas the mask discriminates the target from the interfering source with other pairs. These masks are then used to compute the covariance matrices for the target ($\mathbf{\Phi_{XX}}$) and noise ($\mathbf{\Phi_{NN}}$) signals, and then GEV-BAN beamforming produces the separated source.*

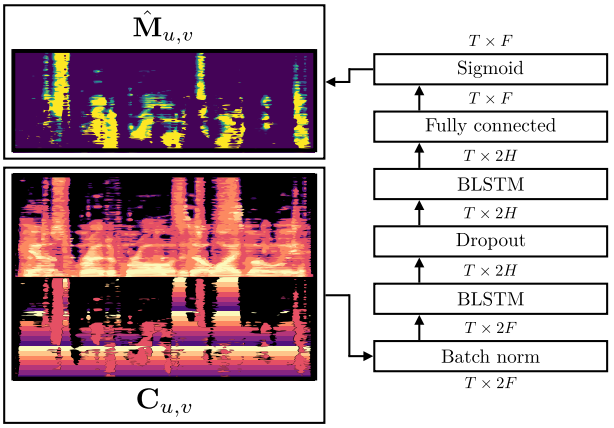


Figure 2: *Architecture of the BLSTM network. The expressions T, F and H stand for the number of frames, the number of frequency bins and the number of hidden states, respectively.*

where $\odot$ stands for the Hadamard product.

Once the BLSTM is trained, it is used to estimate the pairwise masks by inference, and the overall mask is obtained according to:

$$\hat{\mathbf{M}} = \frac{1}{|\mathcal{Q}|} \sum_{(u,v)\in\mathcal{Q}} \hat{\mathbf{M}}_{u,v}, \tag{15}$$

where $|\dots|$ stands for the cardinality of the set.

2.3. GEV-BAN beamforming

As suggested in [20], the target and noise covariance matrices can be estimated with a soft mask M_ν between 0 and 1:

$$\mathbf{\Phi}_{\nu\nu}(f) = \sum_{t\in\mathcal{T}} M_\nu(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^H, \tag{16}$$

where $\mathcal{T} = \{1,\dots,T\}$, $\nu \in \{\mathbf{X},\mathbf{N}\}$ (with $\mathbf{X}$ being the target, and $\mathbf{N}$ being interference and background noise), $\mathbf{Y}(t,f) \in \mathbb{C}^{M\times 1}$, $\{\dots\}^H$ is the Hermitian transpose and:

$$M_\nu(t,f) = \begin{cases} \hat{M}(t,f) & \nu = \mathbf{X} \\ 1-\hat{M}(t,f) & \nu = \mathbf{N} \end{cases}. \tag{17}$$

The vector $\mathbf{F}_{GEV}(f) \in \mathbb{C}^M$ then corresponds to the principal component of the following generalized eigenvalue decomposition:

$$\mathbf{F}_{GEV}(f) = \mathcal{P}\{\mathbf{\Phi_{NN}}(f)^{-1}\mathbf{\Phi_{XX}}(f)\}, \tag{18}$$

where $\mathcal{P}\{\dots\}$ stands for the principal component and $\{\dots\}^{-1}$ for the matrix inverse. Heymann et al. [20] also suggest using a blind analytic normalization (BAN) gain to cope with potential non-linear distortion of the target source, as follows:

$$g_{BAN}(f) = \frac{\sqrt{\mathbf{F}_{GEV}^H(f)\mathbf{\Phi_{NN}}(f)\mathbf{\Phi_{NN}}(f)\mathbf{F}_{GEV}(f)}}{\mathbf{F}_{GEV}^H(f)\mathbf{\Phi_{NN}}(f)\mathbf{F}_{GEV}(f)D^2}. \tag{19}$$

Finally, the reconstructed spectrogram for the target source $Z(t,f)$ can be obtained according to:

$$Z(t,f) = g_{BAN}(f)\mathbf{F}_{GEV}^H(f)\mathbf{Y}(t,f). \tag{20}$$

3. Dataset

To train the network, we generate a dataset of synthetic stereo speech mixtures in simulated reverberating rooms. The speech segments for training come from the LibriSpeech ASR corpus [32] that contains 360 hours of English text read by 482 men and 439 women, sampled at $f_S = 16000$ samples/sec. A simulator based on the image method [33] generates $10,000$ room impulse responses (RIRs).

For the network to generalize to various conditions, we sample the parameters in Table 1 according to a uniform distribution. Each RIR is defined by the room dimensions, reflection coefficient and speed of sound. The spacing between both microphones varies to generalize to arbitrary microphone array shapes, and the microphone pair is rotated randomly and positioned in the room by making sure there is a minimum distance between the microphones and all surfaces. Moreover, the sources are positioned randomly in the room in such a way that the distance between them and the microphones lie within a defined range.

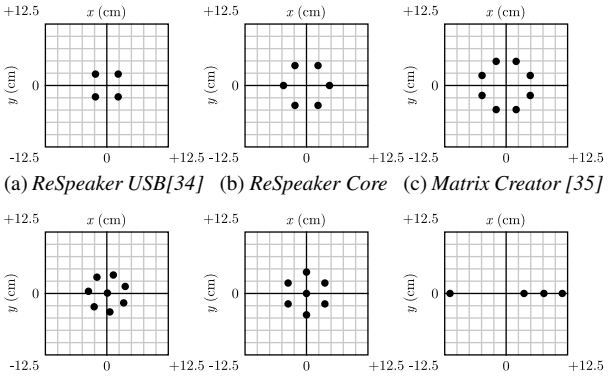
For each training sample, we convolve two speech segments of 5 seconds from Librispeech (one for the target and the other one for the interference) with one of the generated RIR. A randomly selected signal-to-noise ratio (SNR) then defines the gain

of each source. A random gain is also applied to each microphone, to cope with the potential gain mismatch between the microphones. Some diffuse white noise with random variance is then added to the mixture. Finally, all the signals are scaled by a common linear gain such that the signal range models scenarios with different volume levels. The STFT uses frames of $N = 512$ samples, spaced by $\Delta N = 128$ samples.

Table 1: *Simulation parameters.*

Parameters	Range
Room length (m)	$[5.0, 10.0]$
Room width (m)	$[5.0, 10.0]$
Room height (m)	$[2.0, 5.0]$
Surfaces reflection coefficient	$[0.2, 0.8]$
Speed of sound (m/s)	$[340.0, 355.0]$
Spacing between mics (m)	$[0.04, 0.20]$
Min. dist. between mics and surfaces (m)	0.5
Dist. between sources and mics (m)	$[1.0, 5.0]$
White noise variance	$[0.5, 2.0]$
Signal to noise ratio (dB)	$[-5.0, +5.0]$
Overall linear gain	$[0.01, 0.99]$

At test time, we use the test set from LibriSpeech, and convolve the sound segments with 1000 RIRs generated for each array geometry. We use the same simulation parameters as in Table 1, but ignore the spacing between microphones as the shapes correspond to the geometries of commercially available microphone arrays, as depicted in Figure 3. All these microphone arrays are planar, which means they span the xy-plane. Note that the target and interference sources are positioned such that there is at least one pair of microphones that leads to discriminative TDOAs.



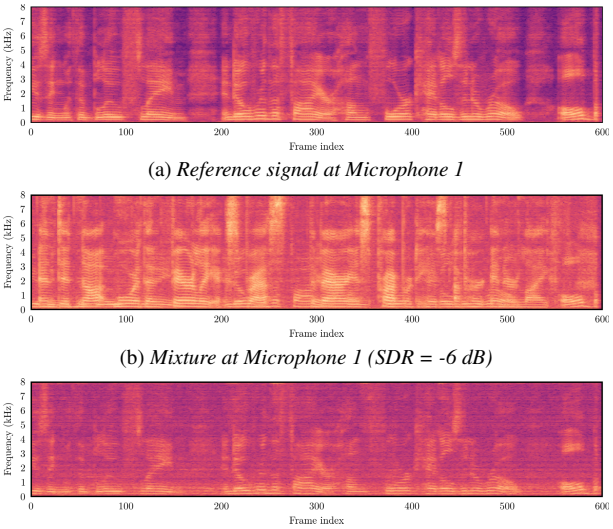
(a) *ReSpeaker USB[34]* (b) *ReSpeaker Core* (c) *Matrix Creator [35]*

(d) *Matrix Voice* (e) *MiniDSP UMA [36]* (f) *MS Kinect [37]*

Figure 3: *Microphone array geometries*

4. Results and Discussion

Results demonstrate that SteerNet performs efficient separation for a wide range of microphone array geometries and environmental conditions. For example, Fig. 4 shows the reference signal, the mixture and the separated signal with a Matrix Voice microphone array. Most features (formants, pitch, transient, etc.) are properly restored without any non-linear distortion, which is expected with GEV-BAN beamforming. However, there is some extra energy in the low frequencies, which is also expected as the Matrix Voice microphone array has a small aperture, making separation more challenging in low frequencies.



(a) *Reference signal at Microphone 1*

(b) *Mixture at Microphone 1 (SDR = -6 dB)*

(c) *Separated signal with SteerNet (SDR = +9 dB)*

Figure 4: *Example with the Matrix Voice 8-microphone array.*

The Signal-to-Distortion Ratio (SDR) is also computed for all the test samples using the BSS Eval toolbox [38]. Table 2 shows that the SDR improves with all microphone arrays, regardless of the shapes. This confirms that SteerNet enhances a target speech signal using its DOA and a trained BLSTM that generalizes for any pairs of microphones. It should be noted that the network is trained using the Adam optimizer with a learning rate of 0.001 and converges in around 20 epochs. The parameters to estimate the gain for the oracle mask during training in (5) are set to $\alpha = 10.0$ and $\beta = 1.0$. The Python code with audio samples is available online[1].

Table 2: *SDR improvement (more is better).*

Microphone Array	ΔSDR (dB)
ReSpeaker USB	+7.69
ReSpeaker Core	+5.63
Matrix Creator	+5.13
Matrix Voice	+4.78
MiniDSP UMA	+4.78
Microsoft Kinect	+6.83

The next step would be to optimize the hyperparameters for the proposed BLSTM architecture, and also investigate other neural network architectures. Moreover, SteerNet considers only one interfering source. This number could be increased to reflect more complex interaction scenarios. Background noise from various environments could also make the model more representative of real-life scenarios. Finally, it would be relevant to reduce the time context (currently set to 5 seconds) to adapt the approach to online processing with low latency.

[1] https://github.com/francoisgrondin/steernet

5. References

[1] H. Tang, W.-N. Hsu, F. Grondin, and J. Glass, "A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 2928–2932.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[3] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 4, pp. 787–796, 2018.

[4] 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," in *Proc. INTERSPEECH*, 2018, pp. 307–311.

[5] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. INTERSPEECH*, 2016, pp. 545–549.

[6] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE ICASSP*, 2018, pp. 1–5.

[7] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[8] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[9] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5739–5743.

[10] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Proc. IEEE ASRU Workshop*, 2017, pp. 437–444.

[11] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, 2019.

[12] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 109:1–109:11, 2018.

[13] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 3244–3248.

[14] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 6548–6552.

[15] Z. Chen, T. Yoshioka, X. Xiao, L. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Proc. IEEE ICASSP*, 2018, pp. 5384–5388.

[16] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 1, pp. 158–170, 2009.

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

[18] 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. IEEE ICASSP*, 2017, pp. 3246–3250.

[19] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[20] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE ASRU Workshop*, 2015, pp. 444–451.

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

[22] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *Proc. IEEE ICASSP*, 2017, pp. 5325–5329.

[23] 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 *Proc. IEEE SLT Workshop*, 2018, pp. 558–565.

[24] A. Maldonado, C. Rascon, and I. Velez, "Lightweight online separation of the sound source of interest through BLSTM-based binary masking," *arXiv preprint arXiv:2002.11241*, 2020.

[25] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *Proc. IEEE ICASSP*, 2018, pp. 6717–6721.

[26] F. Grondin and J. Glass, "SVD-PHAT: A fast sound source localization method," in *Proc. IEEE ICASSP*, 2019, pp. 4140–4144.

[27] ——, "Multiple sound source localization with SVD-PHAT," in *Proc. INTERSPEECH*, 2019, pp. 2698–2702.

[28] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics & Autonomous Systems*, vol. 113, pp. 63–80, 2019.

[29] F. Grondin and J. Glass, "Audio-visual calibration with polynomial regression for 2-D projection using SVD-PHAT," in *Proc. IEEE ICASSP*, 2020, pp. 4856–4860.

[30] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2016.

[31] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of IEEE ICASSP*, 2015, pp. 5206–5210.

[33] E. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[34] B. Sudharsan, S. P. Kumar, and R. Dhakshinamurthy, "AI vision: Smart speaker design and implementation with object detection custom skill and advanced voice interaction capability," in *Proc. ICoAC*, 2019, pp. 97–102.

[35] F. Haider and S. Luz, "A system for real-time privacy preserving data collection for ambient assisted living," in *Proc. INTERSPEECH*, 2019, pp. 2374–2375.

[36] A. Agarwal, M. Jain, P. Kumar, and S. Patel, "Opportunistic sensing with MIC arrays on smart speakers for distal interaction and exercise tracking," in *Proc. IEEE ICASSP*, 2018, pp. 6403–6407.

[37] L. Pei, L. Chen, R. Guinness, J. Liu, H. Kuusniemi, Y. Chen, R. Chen, and S. Söderholm, "Sound positioning using a small-scale linear microphone array," in *Proc. IPIN*, 2013, pp. 1–7.

[38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.