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

Deep Neural Network-Based Generalized Sidelobe Canceller for Robust Multi-channel Speech Recognition

Guanjun Li[1,2], *Shan Liang*[1], *Shuai Nie*[1], *Wenju Liu*[1], *Zhanlei Yang*[3], *Longshuai Xiao*[3]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Huawei Technologies, China

{guanjun.li,sliang,shuai.nie,lwj}@nlpr.ia.ac.cn, {yangzhanlei1,xiaolongshuai}@huawei.com

Abstract

The elastic spatial filter (ESF) proposed in recent years is a popular multi-channel speech enhancement front end based on deep neural network (DNN). It is suitable for real-time processing and has shown promising automatic speech recognition (ASR) results. However, the ESF only utilizes the knowledge of fixed beamforming, resulting in limited noise reduction capabilities. In this paper, we propose a DNN-based generalized sidelobe canceller (GSC) that can automatically track the target speaker's direction in real time and use the blocking technique to generate reference noise signals to further reduce noise from the fixed beam pointing to the target direction. The coefficients in the proposed GSC are fully learnable and an ASR criterion is used to optimize the entire network. The 4-channel experiments show that the proposed GSC achieves a relative word error rate improvement of 27.0% compared to the raw observation, 20.6% compared to the oracle direction-based traditional GSC, 10.5% compared to the ESF and 7.9% compared to the oracle mask-based generalized eigenvalue (GEV) beamformer.

Index Terms: multi-channel speech enhancement, deep neural network, generalized sidelobe canceller, speech recognition

1. Introduction

Reducing noise or reverberation using multi-channel speech enhancement has been shown to improve the performance of automatic speech recognition (ASR) [1, 2, 3]. Traditional multi-channel speech enhancement methods often use signal level criteria and fail to guarantee the optimal ASR results [4, 5, 6].

With the booming of deep learning, there has been a trend to make the multi-channel speech enhancement front end a learnable module in the deep neural network (DNN) and optimize the enhancement module with ASR criteria. This end-to-end optimization manner enables the enhancement module to effectively improve the ASR performance [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. [7, 8, 9] use an attention network to combine the amplitude spectrums of different channels. But [7, 8, 9] do not exploit the spatial cues and fail to employ the traditional multi-channel signal processing knowledge.

The traditional multi-channel signal processing knowledge is instructive to the design of enhancement modules and makes the enhancement modules more interpretable. [10, 11] input the spectral and spatial cues of the multi-channel signals into DNN to directly estimate the coefficients of an adaptive beamformer, and apply beamforming in DNN. To help network training, [10, 11] need parallel clean data, which may not be accessible in some practical situations. [12, 13] guide DNN to estimate the coefficients of an adaptive beamformer using the well-studied beamforming designs, such as the minimum variance distortionless response (MVDR) beamformer [18] and the generalized eigenvalue (GEV) beamformer [19]. But [12, 13] require a certain amount of adaptation data to estimate the signal statistics, causing undesirable latency for real-time applications. Considering the real-time applications, [14, 15, 16, 17] design a fixed beamforming layer in the network to filter the multi-channel signals into several fixed beams pointing to different directions, and the resulting beams are combined thereafter. Compared with [14, 15], the elastic spatial filter (ESF) proposed in [16, 17] performs beamforming in the frequency domain, which reduces the computational complexity and enables DNN to utilize the knowledge of the traditional super-directive beamforming design [20].

However, in the ESF, the non-target beams other than the target beam from the target direction do not necessarily contain reference noise signals, causing the limited denoising ability when combining different fixed beams. The reference noise signals, which is often used in the traditional generalized sidelobe canceller (GSC) [21, 22] for speech enhancement in many practical applications, can help distinguish between the target speaker's signal and noise to better reduce noise from the target beam. Therefore, we can use the well-studied GSC structure to build a more powerful DNN-based speech enhancement front end.

In this paper, we propose a DNN-based GSC structure, which utilizes the traditional super-directive beamforming knowledge and the blocking technique to simultaneously perform localization and denoising. Moreover, we use an ASR criterion to optimize the entire network. More specifically, the multi-channel signals are first passed into a fixed beamforming layer, which is initialized using coefficients of the super-directive beamformer. The fixed beamforming layer filters the multi-channel signals into several fixed beams pointing to different directions, equally sampled in space. We further design an attention network to give more weight to the beam containing more target speaker's energy, and then weight and sum these fixed beams to get a target beam. It is worth noting that we can use the direction corresponding to the beam with the largest attention weight as an estimate of direction-of-arrival (DOA) of the target speaker. Next, we design a blocking layer, which inputs the estimated DOA and multi-channel signals to generate several reference noise signals that block the target signal. If there are M microphones, the blocking layer can generate $(M-1)$ reference noise signals like the traditional GSC. Besides, we use the coefficients of the traditional GSC's blocking matrix to initialize the blocking layer. Then we pass the target beam and reference noise signals into an active noise cancellation layer to obtain an enhanced spectrum, which is then input into an acoustic model to output the posterior probability. Fi-

http://dx.doi.org/10.21437/Interspeech.2020-1101

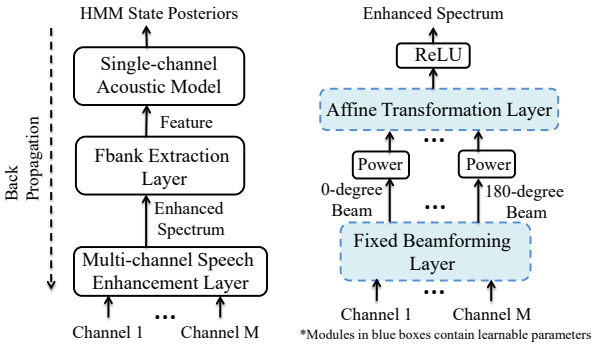

Figure 1: *(a) System architecture of joint multi-channel speech enhancement and acoustic model training. (b) ESF structure proposed in [16, 17]. The ESF cannot perceive the target beam and fails to use reference noise signals.*

nally, we backpropagate gradients from the acoustic model all the way back to the proposed GSC structure. Compared with the traditional GSC [21, 22], the proposed GSC is more suitable for ASR tasks, and it can automatically track the target speaker's direction without additional localization algorithms.

2. DNN-based GSC structure

2.1. System overview

Figure 1(a) shows the system architecture considered in this paper. This architecture guarantees that the multi-channel speech enhancement module can be directly optimized with ASR criteria. Under this architecture, we optimize our proposed DNN-based GSC, which is a kind of multi-channel speech enhancement layer. The multi-channel signals are first passed into the multi-channel speech enhancement layer to obtain a single-channel enhanced spectrum, which is then input into an acoustic model through a log-mel feature bank (Fbank) extraction layer. We optimize the entire network by computing the cross-entropy (CE) loss between the state labels predicted by the acoustic model and the target state labels.

Among the structures of implementing the multi-channel speech enhancement layer, the ESF structure [16, 17] (see Figure 1(b)) is one of the most popular structures proposed in recent years. However, as mentioned in the introduction, the ESF fails to utilize reference noise signals to reduce noise. To make full use of the reference noise signals, we design a DNN-based GSC structure (see Figure 2) to implement the multi-channel speech enhancement layer. The proposed GSC structure consists of 4 parts: fixed beamforming layer, attention network, blocking layer and active noise cancellation layer. These 4 parts will be covered in detail in the following sections.

2.2. Fixed beamforming layer

The fixed beamforming layer in this section was first proposed in [16, 17] and applied to the ESF structure. We will borrow this fixed beamforming layer in the proposed GSC structure and briefly introduce it below.

Suppose there are M microphones. The observation vector in short-time Fourier transform (STFT) domain is given by $\boldsymbol{y}(t,f) = [y_1(t,f), y_2(t,f), ..., y_M(t,f)]^T$, where t and $f \in [1, F]$ denote the time and frequency indices, respectively. The goal of the fixed beamforming layer is to decompose $\boldsymbol{y}(t,f)$ into several fixed beams pointing to D different

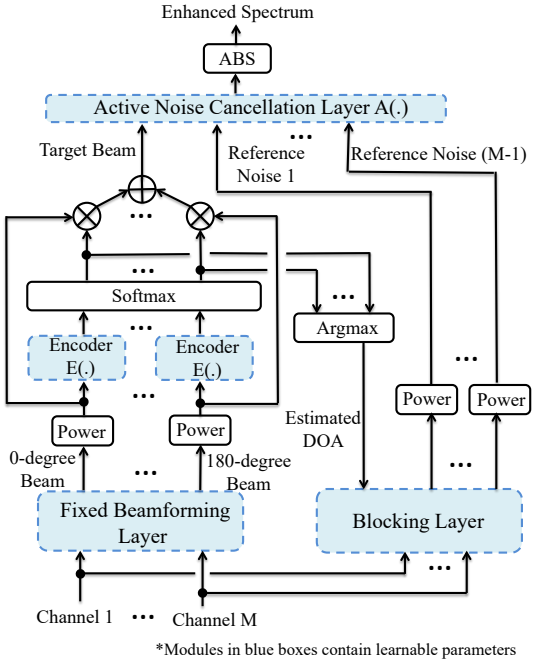

Figure 2: *The proposed GSC structure. It can automatically track the target speaker's direction and obtain the enhanced spectrum using the target beam and reference noise signals.*

directions,

$$\begin{bmatrix} x_1(t,f) \\ x_2(t,f) \\ \cdots \\ x_D(t,f) \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1^H(f)\boldsymbol{y}(t,f) \\ \boldsymbol{w}_2^H(f)\boldsymbol{y}(t,f) \\ \cdots \\ \boldsymbol{w}_D^H(f)\boldsymbol{y}(t,f) \end{bmatrix} + \boldsymbol{b}(f), \quad (1)$$

where $x_d(t,f)$ is the fixed beam pointing to the d-th direction, $\boldsymbol{w}_d(t,f)$ is the coefficient vector of the d-th fixed beamformer, $\boldsymbol{b}(f)$ is a bias vector and $(.)^H$ denotes the complex transposition operator. It is suggested in [16, 17] that $\boldsymbol{w}_d(t,f)$ can be initialized by the coefficients of the super-directive beamformer [20] to make DNN utilize the knowledge of traditional signal processing and generate D fixed beams with different directions, equally sampled in space.

Considering that the target speaker may come from an arbitrary direction, the target beam from the target speaker's direction may be any one of those fixed beams. The ESF in Figure 1(b) simply combines different fixed beams through an affine transformation, which cannot perceive the target beam and fails to use reference noise signals. To better enhance the target speaker's signal, we design the next three network modules.

2.3. Attention network

We first design an encoder $E(.)$ to encode those D fixed beams. The encoding results indicate the amount of the target speaker's energy. Then we can weight and sum those D fixed beams to obtain a target beam,

$$\boldsymbol{e}_d(t) = E\left(\boldsymbol{x}_d(t)\right), \qquad d = 1, 2, ..., D, \quad (2)$$

$$\alpha_d(t) = \frac{\exp\left(\boldsymbol{e}_d(t)\right)}{\sum_{d'=1}^{D} \exp\left(\boldsymbol{e}_{d'}(t)\right)}, \quad d = 1, 2, ..., D, \quad (3)$$

$$\boldsymbol{x}_{\text{tgt}}(t) = \sum_{d'=1}^{D} \boldsymbol{x}_{d'}(t)\alpha_{d'}(t), \quad (4)$$

where $\boldsymbol{x}_d(t) = \left[|x_d(t,1)|^2, |x_d(t,2)|^2, ..., |x_d(t,F)|^2\right]^T$ is the power vector of the d-th beam containing all frequencies and $\boldsymbol{x}_{\text{tgt}}(t) = \left[|x_{\text{tgt}}(t,1)|^2, |x_{\text{tgt}}(t,2)|^2, ..., |x_{\text{tgt}}(t,F)|^2\right]^T$ is the power vector of the target beam containing all frequencies. Eq. (4) can be regarded as selecting the target beam with the highest target speaker's energy from the D fixed beams because the softmax operation in Eq. (3) will make the attention weights sparse (see Figure 4 in the experimental part). In order to capture the sequence information of the spectrum, we use LSTM to implement $E(.)$ and the specific parameters of $E(.)$ are given in the experimental part. The design of the attention mechanism is inspired by the work of multi-channel speaker extraction [23].

Although we have obtained a target beam, it still contains residual noise due to the limited denoising ability of the fixed beamformer. Inspired by the traditional GSC, we design a blocking layer to obtain reference noise signals to further reduce noise from the target beam.

2.4. Blocking layer

The first step of implementing the blocking layer is to identify the target speaker's direction, which can be obtained by reusing the attention weights in Eq. (3), i.e.,

$$d_s(t) = \arg\max_d \boldsymbol{\alpha}_d(t), \tag{5}$$

where $d_s(t)$ is the estimated DOA of the target speaker at time t. Note that the attention weights in Eq. (3) are used not only to obtain the target beam, but also to estimate DOA.

Given $d_s(t)$, we can use a blocking matrix $\boldsymbol{N}_{d_s}(f) \in \mathbb{C}^{M\times M-1}$ to perform blocking the target speaker's signal and generating reference noise signals,

$$\boldsymbol{n}(t,f) = \boldsymbol{N}_{d_s}^H(f)\boldsymbol{y}(t,f) + \boldsymbol{b}_{d_s}(f), \tag{6}$$

where $\boldsymbol{n}(t,f) = [n_1(t,f), n_2(t,f), \ldots, n_{M-1}(t,f)]^T$ is a reference noise signals vector with dimension of $(M-1)\times 1$ and $\boldsymbol{b}_{d_s}(f)$ is a bias vector corresponding to direction $d_s(t)$. We experimentally found that adding bias items in Eq. (1) and Eq. (6) helped improve performance.

The blocking matrix $\boldsymbol{N}_{d_s}(f)$ in Eq. (6) is expected to span the null space of direction $d_s(t)$. To help the network converge, we initialize $\boldsymbol{N}_{d_s}(f)$ using the traditional GSC's blocking matrix $\boldsymbol{N}_{d_s}^{\text{GSC}}(f)$,

$$\boldsymbol{N}_{d_s}^{\text{GSC}}(f) = \begin{bmatrix} -g_{2,d_s}^*(f) & -g_{3,d_s}^*(f) & \ldots & -g_{M,d_s}^*(f) \\ 1 & 0 & \ldots & 0 \\ \vdots & \ldots & \ddots & \\ 0 & 0 & \ldots & 1 \end{bmatrix}, \tag{7}$$

where $g_{m,d_s}(f)$ is the m-th entry of the steering vector corresponding to direction $d_s(t)$ under far-field assumption [24] and $(.)^*$ represents complex conjugate.

2.5. Active noise cancellation layer

After obtaining the target beam and the reference noise signals by Eq. (4) and Eq. (6), the enhanced spectrum can be obtained by an active noise cancellation layer $A(.)$, which performs affine transformation and is expected to further remove noise from the target beam,

$$\boldsymbol{s}_{\text{pow}}(t) = A\left(\left[\boldsymbol{x}_{\text{tgt}}^T(t), \boldsymbol{n}_1^T(t), ..., \boldsymbol{n}_{M-1}^T(t)\right]^T\right), \tag{8}$$

where $\boldsymbol{s}_{\text{pow}}(t) \in \mathbb{R}^{F\times 1}$ is the enhanced spectrum at time t and $\boldsymbol{n}_m(t) = \left[|n_m(t,1)|^2, |n_m(t,2)|^2, \ldots, |n_m(t,F)|^2\right]^T$ is the power vector of the m-th reference noise signal containing all frequencies. The active noise cancellation layer here is similar to the active noise cancellation part [25] of the traditional GSC. Note that both $\boldsymbol{n}(t,f)$ in Eq. (6) and $\boldsymbol{n}_m(t)$ in Eq. (8) contain reference noise signals, but their formats are different. $\boldsymbol{n}(t,f)$ contains signals of all channels at frequency f while $\boldsymbol{n}_m(t)$ contains signals of all frequencies at channel m. The output of Eq. (8) is in the form of an amplitude spectrum, which is sufficient for ASR tasks. Outputting complex spectrum to improve speech intelligibility for humans will be our future work.

The enhanced spectrum undergoes an absolute (ABS) function[1] to avoid negative values and then is input into the Fbank extraction layer and the acoustic model. Finally, the entire network is optimized by using CE loss.

3. Experimental results

3.1. Datasets

We generated 90 hours of reverberant and noisy training data by convolving clean utterances with room impulse responses (RIRs) simulated by the image method [27, 28]. The clean utterances were selected from the WSJ0 corpus [29]. We adopted a linear microphone array containing 4 microphones with spacing of 0.05 m. The target speaker was randomly located in angles from 0° to 180°. The reverberation time was randomly sampled from 0.2 s to 0.6 s. Six types of common additive noise ('bus', 'cafeteria', 'office', 'hallway', 'living' and 'kitchen') from the DEMAND corpus [30] were added to the training data at a signal-to-noise ratio (SNR) randomly sampled from 0 dB to 20 dB. We used the same configuration as the training set to generate a 15-hour validation set. As for the test data, we generated three 8-hour test sets with different SNRs. The SNR in each test set was low (0$\leq$SNR$<$5), medium (5$\leq$ SNR$\leq$15), and high (15$<$SNR$\leq$25). The speakers in the test sets were different from the training set. Moreover, the additive noises in the test sets were different audio segments from the same datasets used for training. For the 2-channel experiments, we picked 2 microphones with spacing of 0.1 m out of 4 sensors.

3.2. Settings

We compared the proposed GSC with (1) the raw noisy observation (denoted as RAW), (2) the oracle-direction based traditional GSC (denoted as O-GSC), (3) the oracle direction-based super-directive beamformer (denoted as O-SD), (4) the oracle mask-based GEV beamformer (denoted as O-GEV) and (5) the ESF proposed in [16, 17]. O-GEV was implemented based on the open source implementation [31]. RAW, O-GSC, O-SD and O-GEV used the same baseline acoustic model, which was trained using the first channel of the training data, to perform ASR. The ESF and the proposed GSC were trained in a stage-wise manner. We first trained the acoustic model and then jointly optimized the multi-channel speech enhancement layer and single-channel acoustic model. We used a weighted finite-state transducer (WFST) with a 3-gram language model as the decoder. The decoder was implemented in Kaldi [32]. We kept the parameters of the Fbank extraction layer fixed during training. The STFT frame size was 400 with 50% overlap and

[1]We experimentally found that the ABS function was more helpful for convergence of both ESF and the proposed GSC than the ReLU [26] function. For consistency, we replaced ReLU in Figure 1(b) with ABS.

Table 1: *WER (%) as a function of the input SNR (dB) for the 2-channel test sets.*

Method	SNR<5	5≤SNR≤15	15<SNR	Avg.
RAW	59.73	37.80	26.28	41.29
O-GSC	55.29	34.96	25.34	38.53
O-SD	56.11	34.72	24.88	38.57
O-GEV	57.93	36.43	26.57	40.31
ESF	55.09	34.49	25.21	38.26
Proposed	**51.51**	**31.36**	**23.83**	**35.57**

Table 2: *WER (%) as a function of the input SNR (dB) for the 4-channel test sets.*

Method	SNR<5	5≤SNR≤15	15<SNR	Avg.
RAW	59.73	37.80	26.28	41.29
O-GSC	53.25	34.75	25.91	37.97
O-SD	50.53	31.52	23.84	35.32
O-GEV	44.21	29.46	24.54	32.74
ESF	47.77	29.74	23.50	33.67
Proposed	**41.17**	**26.92**	**22.30**	**30.13**

64-dimensional Fbank features were used. The acoustic model consisted of 3 LSTM layers with 512 cells followed by one fully connected layer with 1984 outputs. $E(.)$ in Eq.(2) consisted of one LSTM layer with 30 cells followed by one fully connected layer with one output. $A(.)$ in Eq.(8) consisted of one fully connected layer with input size of $201 \times M$ and output size of 201. We set the number of fixed beams in Eq. (1) to $D = 7$ (corresponding to 0° beam, 30° beam, etc.) because we experimentally found that the system performance did not necessarily improve when $D > 7$. All the networks were trained using PyTorch [33] and Adam [34] was used as the optimizer in all the experiments.

3.3. Results

We systematically evaluated the word error rate (WER) of the proposed GSC in 2-channel (Table 1) and 4-channel (Table 2) scenarios. Regarding the performance of the traditional speech enhancement methods, in low SNR (SNR < 5), the adaptive beamformer (O-GSC in the 2-channel scenario and O-GEV in the 4-channel scenario) performed better because of the strong noise reduction capabilities of the adaptive beamformer. In high SNR (SNR>15), the fixed beamformer (O-SD) performed better because the adaptive beamformer will bring more speech distortion than the fixed beamformer while reducing noise in high SNR. However, without oracle information, the proposed GSC not only outperformed the adaptive beamformer in low SNR but also outperformed the fixed beamformer in high SNR. In the 4-channel scenario, the advantages of O-GSC were no longer obvious, which may be due to the well-known signal cancellation problem [35] of the traditional GSC. Unlike the traditional GSC, the proposed GSC can avoid the signal cancellation problem by directly being optimized according to the ASR criterion. It can be seen from Table 2 that ESF is already a strong baseline whose performance is better than O-GSC and O-SD but slightly worse than O-GEV. However, the proposed GSC can further reduce WER. In the 4-channel scenario, the proposed GSC on average can achieve a relative WER reduction of 27.0% compared to RAW, 20.6% compared to O-GSC, 14.7% compared to O-SD, 7.9% compared to O-GEV and 10.5% compared to ESF.

To further illustrate that the proposed GSC worked as expected, we showed some intermediate results of the proposed GSC. Figure 3 shows the beampatterns of the fixed beamforming layer and the blocking layer learned by the network in the

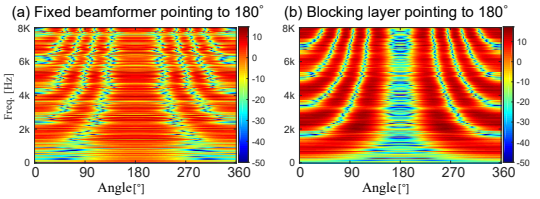

Figure 3: *The beampatterns of the fixed beamforming layer and blocking layer in 2-channel case. Only the beampatterns pointing to* 180° *are shown for space limitation.*

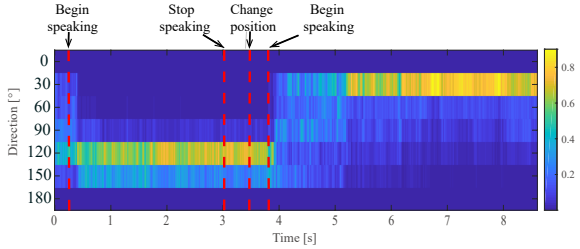

Figure 4: *An example of the attention weights in Eq. (3) changing with target speaker's direction (2-channel case). The input SNR=5dB. The target speaker is first located at* 125° *and his position changes to* 35° *after* 3.5*s.*

2-channel scenario. There were a total of $D = 7$ pairs of beampatterns (pointing to 0°, pointing to 30°, etc.), but we only used one pair (pointing to 180°) for example to save space. It can be seen from Figure 3 that the fixed beamforming layer attempted to preserve the signal from a given direction, while the blocking layer attempted to block the signal from a given direction to generate reference noise signals. Considering a practical situation, we plotted the change of the attention weights in Eq. (3) with the target speaker's direction (see Figure 4). According to Eq. (5), the index of the maximum value of attention weights can represent the target speaker's direction. Figure 4 shows that the proposed GSC can track the target speaker's direction in real time. Moreover, the attention weights took about 1 second to become stable when the target speaker's direction suddenly changed. Note that the localization accuracy of the proposed GSC may not be as high as that of the localization algorithm (i.e., GCC-PHAT [36]), because the localization accuracy is limited by the beams number D. However, considering the mainlobe width of the beamformer, a rough direction is sufficient for the proposed GSC to perform speech enhancement.

4. Conclusions

In this paper, we utilize the knowledge of traditional GSC to build a DNN-based GSC structure, which is optimized by using an ASR criterion. Compared with the traditional GSC, the proposed GSC performs localization and noise reduction simultaneously, and is more suitable for ASR tasks. Experimental results show that the proposed GSC outperforms not only the traditional speech enhancement methods using oracle information but also the ESF proposed in recent years. In the future, we plan to improve the robustness of the proposed GSC to the microphone array geometry mismatch.

5. Acknowledgments

This work was supported in part by the National Key R&D Plan of China (No. 2016YFB1001404) and the China National Nature Science Foundation (No. 61971419, No. 61573357, No. 61503382). We would like to thank Prof. Weimin Zhang from Tsinghua University for helpful writing suggestions.

6. References

[1] M. Wölfel and J. W. McDonough, *Distant speech recognition*. Wiley Online Library, 2009.

[2] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons, 2012.

[3] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[4] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 489–498, 2004.

[5] M. L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2008, pp. 104–107.

[6] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.

[7] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, "Multi-channel attention for end-to-end speech recognition," *2018 Interspeech*, pp. 0–0, 2018.

[8] S. Kim and I. Lane, "Recurrent models for auditory attention in multi-microphone distance speech recognition," *arXiv preprint arXiv:1511.06407*, 2015.

[9] S. Kim, I. Lane, S. Kim, and I. Lane, "End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition." in *Interspeech*, 2017, pp. 3867–3871.

[10] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[11] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," 2016.

[12] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5325–5329.

[13] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[14] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5075–5079.

[15] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[16] K. Kumatani, W. Minhua, S. Sundaram, N. Ström, and B. Hoffmeister, "Multi-geometry spatial acoustic modeling for distant speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6635–6639.

[17] M. Wu, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[18] 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, 2009.

[19] 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.

[20] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007.

[21] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Transactions on antennas and propagation*, vol. 34, no. 3, pp. 311–319, 1986.

[22] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[23] G. Li, S. Liang, S. Nie, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Direction-aware speaker beam for multi-channel speaker extraction." in *Interspeech*, 2019, pp. 2713–2717.

[24] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[25] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer handbook of speech processing*. Springer, 2008, pp. 945–978.

[26] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

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

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

[29] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 81–84.

[30] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[31] "https://github.com/fgnt/nn-gev/."

[32] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *The Journal of the acoustical society of America*, vol. 54, no. 3, pp. 771–785, 1973.

[36] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.