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

End-to-End Multi-Channel Speech Enhancement Using Inter-Channel Time-Restricted Attention on Raw Waveform

Hyeonseung Lee, Hyung Yong Kim, Woo Hyun Kang, Jeunghun Kim, Nam Soo Kim

Department of Electrical and Computer Engineering,
Seoul National University, Seoul, South Korea
{hslee,hykim,whkang,jhkim}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

This paper describes a novel waveform-level end-to-end model for multi-channel speech enhancement. The model first extracts sample-level speech embedding using channel-wise convolutional neural network (CNN) and compensates time-delays between the channels based on the embedding, resulting in time-aligned multi-channel signals. Then the signals are given as input of multi-channel enhancement extension of WaveUNet which directly outputs estimated clean speech waveform. The whole model is trained to minimize modified mean squared error (MSE), signal-to-distortion ratio (SDR) cost, and senone cross-entropy of back-end acoustic model at the same time. Evaluated on the CHiME-4 simulated set, the proposed system outperformed state-of-the-art generalized eigenvalue (GEV) beamformer in terms of perceptual evaluation of speech quality (PESQ) and SDR, and showed competitive results in short time objective intelligibility (STOI). Word-error-rates (WERs) of the system's output on simulated sets were comparable to that of bidirectional long short-term memory (BLSTM) GEV beamformer. However, the system showed relatively high WERs on real sets, achieving relative error rate reduction (RERR) of 14.3% over noisy signal on real evaluation set.

Index Terms: multi-channel speech enhancement, acoustic beamforming, deep neural network, robust speech recognition

1. Introduction

As the number of microphones per device has steadily increased in the last decade, it became more important to develop a method that fully exploits rich information of the multi-channel signals, which leads to the improved performance of multi-channel speech enhancement [1]. The goal of multi-channel speech enhancement, or beamforming, is to recover a single speech source from a mixture of background signals obtained from multiple microphones, both for better perceptual quality and improved accuracy of back-end automatic speech recognition (ASR) system. Most current methods for multi-channel speech enhancement are variants of filter-and-sum beamforming approach, where estimated finite impulse response (FIR) filters are convoluted with each channel signal separately and then the filtered signals are summed up into a single signal.

In delay-and-sum (DAS) beamforming [2], discrete delta-function filters are calculated using estimated time difference of arrival (TDOA) between microphones, then applied to each channel signal to delay it. This simple technique enhances the signal from target direction while attenuates other directions, but lacks in an ability to perfectly reduce noises. In minimum variance distortionless response (MVDR) [3, 4, 5, 6] and generalized eigenvalue (GEV) [7, 8, 9] beamforming, the most popular and high-performance beamforming approaches, spatial covariance matrices of speech or noise signals are estimated first, then the filter coefficients are solved using different derivations. MVDR and GEV beamformer have shown state-of-the-art results in terms of perceptual quality and back-end ASR accuracy.

Moreover, beamformers that directly predicts the filter coefficients were proposed, both for time-domain [10] and frequency-domain [11] filters. In these methods, raw waveform or spectrogram is given as an input of filter prediction network which outputs filter coefficients, then the filters are applied to input signals. Although these methods introduced end-to-end filter prediction networks, the networks were optimized with back-end acoustic model only to minimize ASR costs, not speech enhancement losses. Besides filter-and-sum beamformers, there were some remarkable studies of machine learning-based multi-chanel speech separation [12, 13].

On the other hand, in single-channel speech enhancement and source separation, recently there has been dramatic advances in performance of end-to-end approaches. SEGAN [14] and WaveUNet [15] are wav-to-wav models which have fully 1D-convolutional neural network (CNN) auto-encoder architecture and directly output waveform without introducing any linear filters, trained to minimize euclidean distance between clean and estimated wav or also minimize adversarial loss. EHNet [16] is a spectrogram-to-spectrogram model that employs convolutional recurrent neural network structure which transforms given noisy real spectrum to clean real spectrum, trained to minimize mean-squared-error (MSE) cost, achieving state-of-the-art performance in many enhancement metrics.

Despite success of end-to-end single-channel speech enhancement models, some obstacles exist to extending those models to the multi-channel scenario. Firstly, as inputs are multi-channel noisy speeches which are not in-time, the model should be capable of considering and utilizing time-delays between channels. Secondly, because either power differences or time delays between multi-channel signals and target clean signal are very difficult to predict, just minimizing euclidean distance between clean and estimated waveform may leads to relatively low performance of the enhancement model.

Motivated by this observation, we propose a novel end-to-end multi-channel speech enhancement model that has an inter-channel time-delay compensation front-end, employing time-restricted attention on multi-channel signals. It is followed by a wav-to-wav multi-channel enhancement back-end model which is an extension of basic WaveUNet. Both the front-end and the back-end are jointly optimized to minimize modified MSE and maximize SDR, and also to minimize ASR cost to boost the intelligibility of estimated clean speech.

2. Proposed system

Overall structure of the proposed system is shown in Fig. 1. In inter-channel time-delay compensation stage, sample-level em-

http://dx.doi.org/10.21437/Interspeech.2019-2397

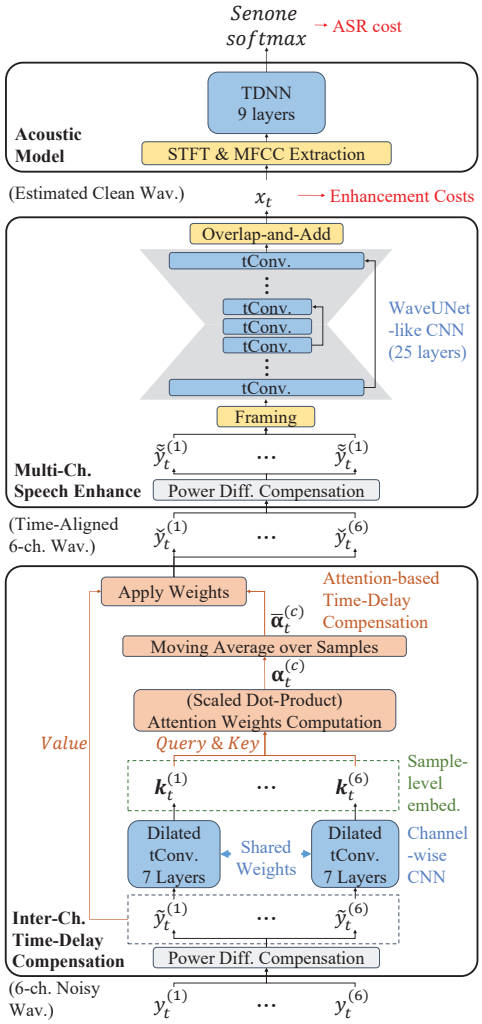

Figure 1: *Overall structure of the proposed model*

bedding is extracted by channel-wise CNN that works independently and equally on each channel, then time-aligned signals are constructed by applying inter-channel time-restricted attention on the embedding. Next in multi-channel speech enhancement stage, single enhanced waveform is obtained by feeding the time-aligned signals through the multi-channel extension of basic WaveUNet. Models in both stages are optimized with enhancement costs and ASR cost.

2.1. Inter-channel time-delay compensation

2.1.1. Inter-channel power difference compensation

Signals of different microphones have different powers, which reflects distances from sources to microphones and room characteristic. However, we experimentally found the power difference makes the model optimization harder because of different dynamic ranges of signals in each utterance, resulting in a degraded performance. To solve this issue, we normalized the signals to have same power in utterance level by adjusting the power of each signal to the power of reference signal:

$$P(s_{1:T}) := \sqrt{\sum_{t=1}^{T}(s_t - \sum_{t'=1}^{T} s_{t'})^2} \tag{1}$$

$$\tilde{y}_t^{(c)} = y_t^{(c)} \times P(y_{1:T}^{(r)}) / (P(y_{1:T}^{(c)}) + \epsilon) \tag{2}$$

where $y_t^{(c)}$ is a t-th sample in c-th channel signal, P is a square root of time-averaged power function, r is an index of the reference channel, T is length of signals, and $\epsilon = 10^{-7}$ is a small positive constant.

2.1.2. Attention-based time-delay compensation

In order to compensate time-delays between channels, first sample-level speech embedding is extracted using channel-wise CNN that has a large receptive field. The channel-wise CNN consists of 7 dilated convolutional layers, where the number of filters of every layer is 20, kernel size is 3, and the dilation rates are 1, 3, 9, 27, 81, 243, 819 from the bottom to the top. LeakyReLU [17] activation function is used for all layers except the last layer with no activation function, and dropout [18] rate of 0.5 is applied during training phase. It shares parameters over all channels and operates independently on each channel:

$$\boldsymbol{k}_t^{(c)} = \text{ChannelWiseCNN}(\tilde{y}_{t-\gamma_w:t+\gamma_w}^{(c)}) \tag{3}$$

where $\boldsymbol{k}_t^{(c)}$ is sample-level speech embedding vectors and $\gamma_w = 1228$ is half-size of the receptive field. Intuitively, what the channel-wise CNN does is making two embedded points close if their input samples contain similar speech content, otherwise far apart. Regarding all the $\boldsymbol{k}_t^{(c)}$ as keys and $\boldsymbol{k}_t^{(r)}$ as a query, time-restricted attention [19] weights are calculated:

$$\boldsymbol{q}_t = \boldsymbol{k}_t^{(r)} \tag{4}$$

$$\boldsymbol{K}_t^{(c)} = \left[\boldsymbol{k}_{t-\tau_d}^{(c)}\ \boldsymbol{k}_{t-\tau_d+1}^{(c)}\ \cdots\ \boldsymbol{k}_{t+\tau_d}^{(c)}\right] \tag{5}$$

$$\boldsymbol{e}_t^{(c)} = {\boldsymbol{K}_t^{(c)}}^{\mathrm{T}} \boldsymbol{q}_t\ /\ \sqrt{2\tau_d + 1} \tag{6}$$

$$\boldsymbol{\alpha}_t^{(c)} = \text{softmax}(\boldsymbol{e}_t^{(c)}) \tag{7}$$

$$\bar{\boldsymbol{\alpha}}_t^{(c)} = \frac{1}{2\tau_{ma}+1} \sum_{t'=t-\tau_{ma}}^{t+\tau_{ma}} \boldsymbol{\alpha}_{t'}^{(c)} \tag{8}$$

where $\boldsymbol{q_t}$ is a query vector, $\boldsymbol{K}_t^{(c)}$ is a key matrix, τ_d is the maximum possible delay between multi-channel signals, $\boldsymbol{e}_t^{(c)}$ is an unnormalized score vector, and $\boldsymbol{\alpha}_t^{(c)}$ is an attention weight vector. The attention weight is smoothed by applying rectangular moving average filter of half-width $\tau_{ma} = 50$, mitigating the noisy nature of sample-level attention weights. Lastly, regarding the inputs as values, the smoothed attention weight $\bar{\boldsymbol{\alpha}}_t^{(c)}$ is convoluted with each channel signal, aligning it to the reference signal.

$$\boldsymbol{v}_t^{(c)} = \left[y_{t-\tau_d}^{(c)}\ y_{t-\tau_d+1}^{(c)}\ \cdots\ y_{t+\tau_d}^{(c)}\right] \tag{9}$$

$$\check{y}_t^{(c)} = {\boldsymbol{v}_t^{(c)}}^{\mathrm{T}} \bar{\boldsymbol{\alpha}}_t^{(c)} \tag{10}$$

where $\boldsymbol{v}_t^{(c)}$ is a value row vector and $\check{y}_t^{(c)}$ is the time-aligned signal.

2.2. Multi-channel speech enhancement

2.2.1. Inter-channel power difference compensation

For the same purpose as in Section 2.1.1, time-aligned signals are normalized to have same power for better performance of enhancement model:

$$\tilde{\check{y}}_t^{(c)} = \check{y}_t^{(c)} \times P(\check{y}_{1:T}^{(r)}) / (P(\check{y}_{1:T}^{(c)}) + \epsilon) \tag{11}$$

2.2.2. Multi-channel WaveUNet

Power compensated time-aligned signals are fed to the multi-channel WaveUNet that directly outputs estimated clean signal:

$$\hat{x}_{1:T} = \text{MultiChannelWaveUNet}(\tilde{y}_{1:T}^{(1:C)}) \quad (12)$$

where $C = 6$ is the number of input channels. This is an modified version of basic WaveUNet [15], where the number of channels is modified from 1 to 6 for input tensor, and from more-than-1 to 1 for output tensor, since the task we focus on is not single-channel multi-source separation but multi-channel single-source enhancement. The number of units in upsampling layers are tripled to fit the increased input channels.

The utterance-level enhancement function MultiChannel-WaveUNet consists of three parts: framing, CNN auto-encoder, and overlap-and-add. In framing part, each signal is divided into frames with window length of 16384 and hop size of 8192, and signals of all channels in every frame are concatenated into a single tensor, which is then given as an input of the CNN auto-encoder. In CNN auto-encoder part, the CNN consists of total 26 convolutional layers, which includes 12 downsampling layers with stride 2 and kernel size of 15, a middle layer, 12 upsampling layers with stride 2 and kernel size of 5, and an output layer with 1 filter and kernel size of 1. The numbers of filters in upsampling layers are 72, 144, 216, 288, 360, 432, 504, 576, 648, 720, 792, 864 from the bottom to the middle, and the middle layer has 936 filters. The numbers of filters in downsampling layers are 288, 264, 240, 216, 192, 168, 144, 120, 96, 72, 48, 24 from the middle to the top. To boost the training speed, every upsampling layer also receives additional input which is an activation of downsampling layers at symmetrical position, i.e. activation of i-th downsampling layer from the bottom is fed to i-th upsampling layer from the top. Each convolutional layer includes layer normalization [20] right after the affine transform, followed by LeakyReLU activation except the output layer with no activation function. Dropout rate of 0.5 is applied to all hidden layers during training. Finally in the overlap-and-add part, output signals of the CNN auto-encoder are multiplied with a hanning window of size 16384, then merged into a single utterance by adding the windowed outputs frame-by-frame with 50% overlap. The output is an estimation of clean speech, which is optimized to minimize costs in the following section.

2.3. Objective functions

The whole network described so far is trained to minimize the mixed cost function:

$$L_{total} = L_{MESPN-MSE} + \lambda_{SDR} L_{SDR} + \lambda_{ASR} L_{ASR} \quad (13)$$

where $L_{MESPN-MSE}$ is min-error-shift power-normalized MSE cost, L_{SDR} is SDR cost, and L_{ASR} is senone cross-entropy cost. λ_{SDR} and λ_{ASR} are respectively set to 0.0001 and 0.3, adjusting dynamic ranges of three costs to be similar.

2.3.1. Min-error-shift power-normalized MSE (MESPN-MSE)

As mentioned in Section 1, power differences and time delays between multi-channel signals and target clean signal are very difficult to predict accurately, because of which MSE-trained model shows degraded performance or even fails to converge. To overcome this problem, we used MESPN-MSE for training:

$$f(s_{1:T}^{ref}, s_{1:T}^{deg}) := \frac{1}{T}\sum_{t=1}^{T}\left(s_t^{ref} - \frac{P(s_{1:T}^{ref})}{P(s_{1:T}^{deg}) + \epsilon} s_t^{deg}\right)^2 \quad (14)$$

$$g(s_{1:T}^{ref}, s_{1:T}^{deg}) := \min_{-\tau_{dit} \le b \le \tau_{dit}} f(s_{1+|b|:T-|b|}^{ref}, s_{1+|b|:T-|b|}^{deg}) \quad (15)$$

$$L_{MESPN-MSE} = g(x_{1:T}, \hat{x}_{1:T}) \quad (16)$$

where s^{ref} and s^{deg} means reference and degraded signal, and τ_{dit} is the maximum possible delay between input speech signals and a target speech.

2.3.2. Signal-to-distortion ratio (SDR)

As it is known that an enhancement model optimized using both MSE and SDR costs can show outperform a MSE-trained model in terms of perceptual qualities [21], we employed SDR cost as a secondary objective. The SDR cost can be summarized as follows after some derivation [22]:

$$\text{SDR}(s_{1:T}^{ref}, s_{1:T}^{deg}) := 10 log_{10} \frac{||s_{1:T}^{\perp}||^2}{||s_{1:T}^{deg} - s_{1:T}^{\perp}||^2} \quad (17)$$

$$s_{1:T}^{\perp} = \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} s_{1:T}^{deg} \quad (18)$$

$$L_{SDR} = -\text{SDR}(x_{1:T}, \hat{x}_{1:T}) \quad (19)$$

where $\boldsymbol{A}$ is a $T \times G$ non-symmetric Toeplitz matrix with a constant G. Its the first column and row are defined as follows:

$$A_{i,1} = \begin{cases} s_i^{ref}, & \text{if } 1 \le i \le T \\ 0 & \text{otherwise} \end{cases}, \quad A_{1,j} = \begin{cases} s_j^{ref}, & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

2.3.3. Senone cross-entropy

ASR cost is used to improve the intelligibility of estimated clean speech. Firstly, 40-dimensional mel-frequency cesptral coefficients (MFCC) are extracted from the estimated clean speech with frame length 400 and hop size 160. It is given as an input of time-delay neural network (TDNN) acoustic model. The model consists of 9 layers, which has time-delay lists of [-2,-1,0,1,2], [-1,0,1], [0], [-1,0,1], [0], [-3,0,3], [-3,0,3], [-6,-3,0], [0] from the bottom to the top. All hidden layers have 750 units and include layer-normalization followed by LeakyReLU activation without dropout. The ASR cost is cross-entropy between target senone one-hot vector and estimated senone softmax:

$$L_{ASR} = -\frac{1}{T_f}\sum_{t=1}^{T_f}\sum_{i=1}^{Q} q_{t,i} log(\hat{q}_{t,i}) \quad (21)$$

where T_f is the number of frames in the utterance, $q_{t,i}$ and $\hat{q}_{t,i}$ respectively means target senone posterior (0 or 1) and estimated senone posterior probability for t-th frame and i-th senone class, and Q is the number of senone classes.

3. Experimental Setup

3.1. Database

We evaluated the proposed system on CHiME-4 16-kHz 16-bit PCM database that consists of two groups: real and simulated set. The real set contains speeches recorded in four noisy environments: bus, cafe, pedestrian area, and street. The simulated set contains artificially mixed signals simulated from clean speeches and noises recorded in the same four environments. Every utterance is recorded or simulated to form six-channel noisy signals and target clean signal. The numbers of utterances in training, development, and evaluation set are 8738 (1600 real, 7138 simu.), 3280 (1640 real, 1640 simu.), and 2640 (1320 real,

Table 1: *speech enhancement scores on CHiME-4 simulated development set*

Enhancement Method	PESQ	STOI	SDR
None (score avg. over channels)	2.01	0.83	4.06
BeamformIt	2.31	0.88	5.48
BLSTM GEV	2.42	0.88	6.03
DNN MTS-IRM GEV (in lit. [9])	2.62	**0.95**	11.35
Proposed	2.77	**0.95**	**15.87**
(w/o time-delay compensat.)	2.54	0.92	13.85
(w/o ASR cost, i.e. $\lambda_{ASR} = 0$)	**2.81**	0.89	15.11

Table 2: *speech enhancement scores on CHiME-4 simulated evaluation set*

Enhancement Method	PESQ	STOI	SDR
None (score avg. over channels)	1.99	0.81	5.06
BeamformIt	2.20	0.86	6.22
BLSTM GEV	2.45	0.87	5.26
DNN MTS-IRM GEV (in lit. [9])	2.64	**0.95**	12.17
Proposed	2.69	0.94	**15.86**
(w/o time-delay compensat.)	2.44	0.92	13.90
(w/o ASR cost, i.e. $\lambda_{ASR} = 0$)	**2.73**	0.88	14.92

1320 simu.). 16-bit PCM signals were converted to float and divided by 2^{16-1}, resulting in signals $y_t^{(c)}$ and x_t in range [-1,1]. Considering maximum distance between array microphones in CHiME-4 is less than 22 cm, τ_d was set to 14. τ_{dit} and G were set to 320, assuming maximum delay between input speeches and clean speech is under 20 ms. Reference channel index r was set to 5, since the 5-th is the closest one to the speaker.

3.2. Training procedure

All 8738 utterances in the training set were used to train the proposed model for 30 epochs with learning rate of 0.001 and mini-batch size of 2 utterances. Adam optimizer [23] with β_1=0.9 and β_2=0.999 was utilized for the training. All network weights are randomly initialized using Glorot's method [24]. Among the models came out after every epoch, the one scored smallest total cost L_{total} on simulated development set was chosen as the test model. Target senone posterior $q_{t,i}$ was obtained before the training by forced alignment on the training set, using triphone Gaussian mixture model-hidden Markov model (GMM-HMM) system of Kaldi [25] advanced baseline script for CHiME-4.

4. Evaluation Results

4.1. Speech Enhancement Experiments

We compared speech enhancement performances of the proposed model with three baseline systems: the most widely-used delay-and-sum beamformer (BeamformIt) [2], a bi-directional long short-term memory (BLSTM) GEV [8], and a deep neural network multi target-specified ideal ratio mask (DNN MTS-IRM) GEV [9] which have shown state-of-the-art performance on CHiME-4 database in both enhancement and recognition scores. Three metrics are employed to measure the quality of enhanced waveforms: perceptual evaluation of speech quality (PESQ) [26], short-time objective intelligibility measure (STOI) [27], and speech-to-distortion ratio (SDR) [28]. The close-talking microphone recordings (channel 0) are considered as clean speech signals. The enhancement scores evaluated on simulated sets are presented in Table 1 and Table 2. The proposed model showed superior enhancement scores over the baseline systems, except that STOI of DNN MTS-IRM GEV on simulated evaluation set is slightly higher than proposed model. The proposed model without inter-channel time-delay compensation stage showed relatively low performance, proving it is a crucial component in the whole system. The proposed model trained without ASR cost showed lower STOI than the proposed model, indicating that ASR cost positively affects intelligibility of the enhanced speech while lowering PESQ a little.

Table 3: *WERs (%) for enhancement systems on CHiME-4*

Enhancement Method	Development		Evaluation	
	simu	real	simu	real
None (ch#5)	5.43	6.36	6.75	12.12
BeamformIt	4.48	3.64	7.32	7.03
BLSTM GEV	**3.11**	**2.89**	**3.99**	**4.01**
Proposed	3.33	4.98	4.26	10.39

4.2. Speech Recognition Experiments

Word-error-rates (WERs) of the enhancement system outputs were measured with Kaldi advanced baseline script for CHiME-4, where a TDNN acoustic model was trained with lattice-free maximum mutual information (LF-MMI) criterion. All 6-channel signals and enhanced signal were used to train the acoustic model. The WERs are reported in Table 3. The proposed system showed WERs comparable to BLSTM GEV on simulated sets. However, WERs of proposed model on real data was worse than BLSTM GEV with a large performance gap. The relative error rate reduction (RERR) of proposed system over 5-th channel noisy signal on real evaluation set was 14.3%, which is significantly smaller than 36.9% RERR on simulated evaluation set. Our hypothesis is that the proposed model tend to generate distortion in output when the input is a real data because of the real recorded signals in CHiME-4 database containing considerable amounts of unpredictable close-source noises such as tap sound and breathing. The model just trained to minimize the difference between target signals and model outputs with no explicit constraints about speech distortion.

5. Conclusions

We proposed an end-to-end multi-channel speech enhancement model that includes two novel modules: a time-restricted attention-based inter-channel time-delay compensation and a multi-channel enhancement extension of basic WaveUNet. The whole model is trained to minimize modified MSE, SDR cost, and ASR cost at the same time. On CHiME-4 simulated sets, the system outperformed state-of-the-art GEV beamformer in terms of PESQ and SDR, and showed comparable STOI score. On CHiME-4 real evaluation set, the system achieved 14.3% RERR over noisy waveform. For future work, we could investigate to reduce RERR gap between simulated set and real set.

6. Acknowledgements

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1701-04.

7. References

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

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

[3] J. Capon, "High resolution frequency-wavenumber spectrum analysis," in *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[4] 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. Interspeech*, 2016, pp. 1981-1985.

[5] 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 *Proc. ICASSP*, 2017, pp. 276-280.

[6] Z. Q. Wang and D. Wang, "All-neural multi-channel speech enhancement," in *Proc. Interspeech*, 2018, pp. 3234-3238.

[7] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529-1539, 2007.

[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] W. Jiang, F. Wen, and P. Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," in *IEEE Access*, vol. 6, pp. 52385-52392, 2018.

[10] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, 2016, pp. 1976-1980.

[11] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. ICASSP*, 2017, pp. 271-275.

[12] Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. L. Roux, "Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information," in *Proc. Interspeech*, 2017, pp. 2461-2465.

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

[14] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642-3646.

[15] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th ISMIR Conference*, 2018, pp. 334-340.

[16] H. Zhao, S. Zarar, I. Tashev, and C. H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, 2018, pp. 2401-2405.

[17] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *JMLR*, vol. 15, no. 1, pp. 1929-1958, 2014.

[19] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khundapur, "A time-restricted self-attention layer for ASR," in *Proc. ICASSP*, 2018, pp. 5874-5878.

[20] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*.

[21] J. Kim, M. El-Kharmy, and J. Lee, "End-to-end multi-task denoising for joint SDR and PESQ optimization," *arXiv preprint arXiv:1901.09146*.

[22] H. Nakajima, Y. Takahashi, K. Kondo, and Y. Hisaminato, "Monaural source enhancement maximizing source-to-distortion ratio via automatic differentiation," *arXiv preprint arXiv:1806.05791*.

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

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249-256.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motliicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU Workshop*, 2011.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codes," in *Proc. ICASSP*, 2001, pp. 749-752.

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

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.