Deep Neural Network Based Multi-Channel Speech Enhancement for Real-Time Voice Communication Using Smartphones

Soonho Baek
Mobile Communications Business
Samsung Electronics
Suwon, Republic of Korea
soonho.baek@samsung.com

Myungho Lee
Mobile Communications Business
Samsung Electronics
Suwon, Republic of Korea
me.lee@samsung.com

Han-gil Moon
Mobile Communications Business
Samsung Electronics
Suwon, Republic of Korea
hangil.moon@samsung.com

2022 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-6654-4154-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCE53296.2022.9730152

***Abstract*— Recently, the performance of speech enhancement has been improved via deep neural networks. However, most of them are too heavy for voice communication using smartphones, and some are non-causal systems. In this paper, we introduce some effective techniques improving the performance even with light-weight models at causal system. We extract the input features by incorporating two kinds of beamformers. Furthermore, a normalization scheme is proposed to diminish the inter-channel variance between two beamformer outputs. The experimental results show the superiority of the proposed features. Moreover, the proposed method is extendable to any number of microphone systems without additional model training.**

Keywords*— speech enhancement, deep neural networks, real-time voice communication

I. INTRODUCTION

Voice communication via smartphones is widely being used regardless of place thanks to their easy portability. Depending on the situation, people on a voice call can be exposed to loud noisy environment that may incur performance degradation. To offer stable and clear speech quality regardless of situation, most smartphones in the market embed multiple microphones and adopt multi-channel speech enhancement solutions [1]-[5].

On the other hand, deep neural networks (DNNs) have become popular recently due to its excellent ability for non-linear modeling [6]-[21]. Various DNN models (FCN [7]-[9], [12], RNN [13][16], CRN [10][11], AE [6], VAE [15], etc.) have been attempted as well. Last but not least, various features (MFCC [9], LPS [12], etc.) and labels (spectral regression [12], IBM [7], IRM [8], cIRM [9], PSM [13] etc.) have been tried. However, most of them are non-causal systems not suitable for real-time voice communication system. In addition, some utilize only one microphone, although most smartphones have multiple microphones.

In [11], the dual channel solution based on DNNs is proposed for real-time speech enhancement at a close-talk scenario. It introduces a convolutional recurrent network (CRN) which has high computational efficiency. The inter-channel and intra-channel features are fed to DNNs, and it shows improved speech quality over the conventional solutions. However, there is no consideration about the channel variability depending on the configuration of microphone arrays. The mismatch between the channel environment of training database and that of the target device can incur performance degradation. Also, it is not feasible to be extended to three or more microphone systems.

In this paper, a new DNN based multi-channel speech enhancement method is introduced for real-time voice communication via smartphones. Considering the limited resources of smartphones, we propose effective features to improve the performance even at light-weight models. We employ two types of beamformers: a minimum variance distortionless response (MVDR) beamformer steering to the target speaker and a null-steering beamformer suppressing the target speech. In addition, a normalization concept is introduced to mitigate the inter-channel variance between two beamformer outputs. With this normalization concept, we can easily deal with the mismatch problem caused by the difference in acoustic properties between training database and target devices. Outputs obtained from two beamformers with normalization are fed to the DNN based spectral filter. The DNN filter is trained with noise-aware training (NAT) method thanks to capability capturing noise signal of the null-steering beamformer. The impact of the proposed features is analyzed by incorporating light-weight CRNs where the number of coefficients is 244K. The proposed method is compared with conventional DNN based approaches in two types of voice call scenarios: handset mode where the user's mouth is close to a bottom microphone of smartphones, and speakerphone mode where the smartphone is at a certain distance from the user's mouth. From experimental results, we demonstrate that the proposed features are effective to improve the performance significantly at a light-weight model. Furthermore, we show that the proposed method is extendable to three channel systems without additional model training.

The paper is organized as follows. Section 2 elaborates the proposed system including a normalization concept to minimize the inter-channel variance. A series of performance assessment results are presented in Section 3. Section 4 concludes this paper..

II. DNN BASED MULTI-CHANNEL SPEECH ENHANCEMENT

A block diagram of the proposed method is represented in Fig. 1. We employ a MVDR beamformer providing a speech reference [4][5] and a null-steering beamformer providing a noise reference by blocking speech components [2][3]. Note that the noise reference reflects the characteristics of the noise

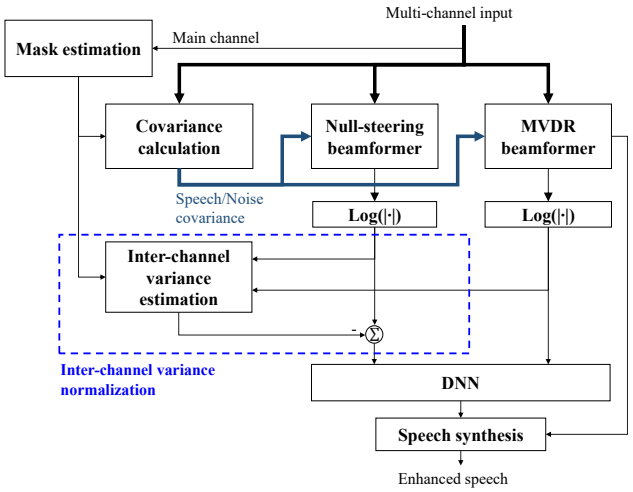
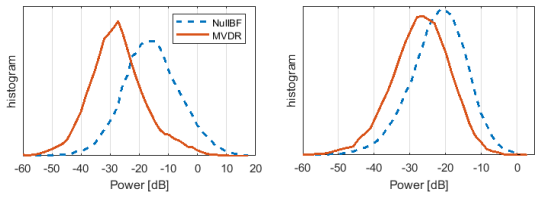


Fig. 1. Block diagram of the proposed method for multi-channel speech enhancement based on the deep neural network



(a) 0dB pub noise (b) 0dB street noise

Fig. 2. Histogram of the noise components in the MVDR beamformer output and in the null-steering beamformer output. The difference of noise power spectrum between two beamformers induces worse noise reduction performance.

components of the speech reference. However, there is a spectral power difference between the noise reference and the noise components of the speech reference due to the different beam patterns of two beamformers. In order to compensate the power difference, we propose the normalization scheme for noise reference, described in detail at the subsection 2.2. The log-spectral coefficients of the speech reference and the normalized noise reference are taken by the DNN model generating the gain values in the frequency domain. We employ the phase sensitive mask (PSM) as the training target, which incorporates the phase information [13]. As the DNN model, we adopt the convolutional recurrent network (CRN) architecture, which has the high computational efficiency [10][11]. The detail configuration for the CRN is described at Table I. Finally, the enhanced speech is obtained through multiplying the gain extracted from the DNN to the MVDR beamformer output.

The beamformers are described in the following subsections. In addition, the proposed normalization scheme, called ICVN, is introduced.

A. Beamformer

In order to obtain the beamformer coefficients for M channel, we employ the coefficients of the MaxSNR beamformer $\boldsymbol{w}_{MaxSNR}$ as follows [2][4][5]:

$$\boldsymbol{w}_{MVDR}(k,l) = \frac{\boldsymbol{w}_{MaxSNR}(k,l)}{\boldsymbol{w}_{MaxSNR}^{\mathrm{H}}(k,l)\boldsymbol{\Phi}_{\boldsymbol{N}}(k,l)\boldsymbol{w}_{MaxSNR}(k,l)} \cdot \frac{1}{\left(\boldsymbol{\Phi}_{\boldsymbol{N}}(k,l)\boldsymbol{w}_{MaxSNR}(k,l)\right)_1}, \quad (1)$$

$$\boldsymbol{w}_{nullBF}(k,l) = \left[I_M - \frac{\boldsymbol{w}_{MaxSNR}(k,l)\boldsymbol{w}_{MaxSNR}^{H}(k,l)\boldsymbol{\Phi}_{\boldsymbol{N}}(k,l)}{\boldsymbol{w}_{MaxSNR}^{H}(k,l)\boldsymbol{\Phi}_{\boldsymbol{N}}(k,l)\boldsymbol{w}_{MaxSNR}(k,l)}\right]_2, \quad (2)$$

where $\boldsymbol{\Phi}_{\boldsymbol{N}}(k,l)$ is the noise covariance matrix, and k and l are the frequency bin index and the frame index, respectively. The MaxSNR beamformer can be obtained from the dominant eigenvector of $\boldsymbol{\Phi}_{\boldsymbol{N}}^{-1}(k,l)\boldsymbol{\Phi}_{\boldsymbol{X}}(k,l)$ where $\boldsymbol{\Phi}_{\boldsymbol{X}}(k,l)$ is the speech covariance matrix [3]. In order to solve the generalized eigenvalue problem, we adopt the power iteration method as follows [2]:

$$\boldsymbol{w}_{MaxSNR}(k,l) = \frac{\boldsymbol{\Phi}_{\boldsymbol{N}}^{-1}(k,l)\boldsymbol{\Phi}_{\boldsymbol{X}}(k,l)\boldsymbol{w}_{MaxSNR}(k,l-1)}{\|\boldsymbol{\Phi}_{\boldsymbol{N}}^{-1}(k,l)\boldsymbol{\Phi}_{\boldsymbol{X}}(k,l)\boldsymbol{w}_{MaxSNR}(k,l-1)\|}. \quad (3)$$

The matrix inversion can be calculated in real-time through the matrix inversion lemma [2]. The covariance matrices are recursively estimated using a time-varying frequency-dependent smoothing parameter $\widetilde{\alpha_v}(k,l)$ as follows:

$$\boldsymbol{\Phi}_v(k,l) = \widetilde{\alpha_v}(k,l)\boldsymbol{\Phi}_v(k,l-1) + (1-\widetilde{\alpha_v}(k,l))\boldsymbol{Y}(k,l)\boldsymbol{Y}^T(k,l) \quad (4)$$

where $\boldsymbol{Y}(k,l)$ represents the signal from all M microphones and v may take $\boldsymbol{X}$ or $\boldsymbol{N}$. The smoothing parameters are determined by the mask $M(l,k)$ $(0 \le M(l,k) \le 1)$ indicating speech presence probability

$$\widetilde{\alpha_{\boldsymbol{X}}}(k,l) = \alpha + (1-M(l,k))(1-\alpha), \quad (5)$$

$$\widetilde{\alpha_{\boldsymbol{N}}}(k,l) = \alpha + M(l,k)(1-\alpha). \quad (6)$$

When target speech is present, $\widetilde{\alpha_{\boldsymbol{N}}}(k,l)$ is close to 1, thus freezing the update of noise covariance. In case a speech is absent, $\widetilde{\alpha_{\boldsymbol{X}}}(k,l)$ becomes 1. The value of α determines the tracking rate. In this paper, α is experimentally set to 0.97.

The mask can be estimated by applying DSP/DNN based single channel solutions to a main channel. In this paper, RNNoise, a light version of DNN based mask estimation, is adopted due to its low complexity and causality [21].

B. Inter-channel variance normalization

Let $\boldsymbol{Y}_s = \boldsymbol{X}_s + \boldsymbol{N}_s$ denote the speech reference provided by the MVDR beamformer and let $\boldsymbol{Y}_n = \boldsymbol{X}_n + \boldsymbol{N}_n$ denote the noise reference provided by the null-steering beamformer. The frequency bin index k and the frame index l are omitted. $\mathbf{X}_s$ and $\mathbf{X}_n$ represent the speech components in the speech reference and in the noise reference, respectively. $\boldsymbol{N}_s$ and $\boldsymbol{N}_n$ are the noise components in the speech reference and in the noise reference, respectively. Assuming that there is little speech leakage in the

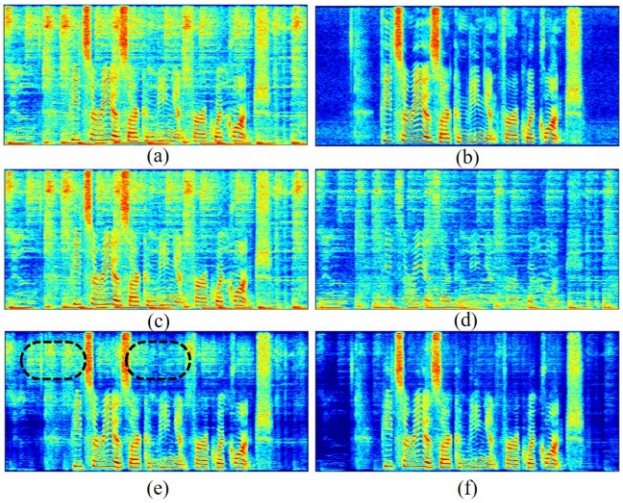


Fig. 3. Spectrograms of an utterance tested in handset mode at the music noise of 5dB: (a) noisy speech in the primary microphone, (b) clean speech, (c) speech reference, (d) noise reference, (e) DNN output without ICVN, and (f) DNN output with ICVN

output of the null-steering beamformer, the power of $\boldsymbol{X}_n$ is significantly small and it can be omitted as $\boldsymbol{Y}_n \approx \boldsymbol{N}_n$. It implies that the noise reference $\boldsymbol{Y}_n$ reflects the spectral properties of the noise components $\boldsymbol{N}_s$ of the speech reference. However, since the beam patterns of two beamformers are different, there is a spectral power gap between $\boldsymbol{N}_s$ and $\boldsymbol{N}_n$, called the inter-channel variance in this paper. Fig. 2 compares the distribution of noise power in the speech reference and in the noise reference in two different types of noise environments. It depends on the microphone array geometries and the spatial characteristics of the background noise or the interference. It involves not only the mismatch problem between the training phase and the enhancement phase, but also the non-consistent noise reduction performance.

To reduce this gap, we compensate the noise reference at the log-spectral feature domain as follows:

$$\log\left(\left|\tilde{\boldsymbol{Y}}_n\right|\right) = \log(|\boldsymbol{Y}_n|) - \boldsymbol{g}\,, \tag{7}$$

where $\boldsymbol{g}$ is the inter-channel variance. We are now looking for the value of $\boldsymbol{g}$ such that $\left|\tilde{\boldsymbol{Y}}_n\right|^2$ is similar with $|\boldsymbol{N}_s|^2$. It can be recursively estimated using a smoothing parameter $\tilde{\beta}(l,k)$ when target speech is not present

$$g(l,k) = \tilde{\beta}(l,k)g(k,l-1) + (1-\tilde{\beta}(l,k)) \cdot \{\log(|Y_n(l,k)|) - \log(|Y_s(l,k)|)\}. \tag{8}$$

In this paper, we set the smoothing parameter to the same value with $\widetilde{\alpha_N}(k,l)$.

III. Experimental Setup and Assessment

A. Dual channel training data preparation

We constructed the DNN training data collected from Galaxy S10 where two microphones have been embedded at bottom and top. The speech data was prepared by convolving the monaural speech signal with impulse responses from the mouth of the head dummy to each microphone of the device at the close-talk scenario. The monaural speech was obtained from two kinds of English corpuses: TIMIT and WSJ. The noise data was prepared by recording the various types of noise signals in the listening room. The noise samples were obtained from the freesound.org website. We tried to collect very large noise sets and experimentally summarized them into a compact data version which contains about 20 kinds. The noisy speech was synthesized at the several SNRs with respect to the primary channel. The SNRs were selected every 1dB in order from -5 to 10 dB. At the training phase, the covariance matrices and the inter-channel variance were estimated offline by averaging from each sentence with the ideal ratio mask [8].

TABLE I. LAYER CONFIGURATION CONSISTING OF 5 CNN BASED ENCODER LAYERS, GRU LAYER2, AND 5 CNN BASED DECODER LAYERS.
*(KERNEL SIZE, STRIDE, OUTPUT SIZE)

	Layer1	Layer2	Layer3	Layer4	Layer5
Encoder	(5,2,10)	(3,2,10)	(3,2,15)	(3,2,15)	(3,2,20)
GRU	140	140	-	-	-
Decoder	(3,2,15)	(3,2,15)	(3,2,15)	(3,2,15)	(3,2,15)

B. Experimental setup and assessment

The evaluation set was constructed from the recorded data through Galaxy S10. The 100 utterances selected from the testset of TIMIT corpus were recorded in handset mode and in speaker mode by using the dummy head. The speech signal was corrupted with the three types of unseen noise (mall, music, and pub) at three levels of SNR, i.e., 0dB, 5dB, and 10dB. In summary, the evaluation set is composed of 900 unseen noisy utterances. For objective evaluation, we used three different objective quality measurements: Perceptual Objective Listening Quality Analysis(POLQA) [22], Perceptual Evaluation of Speech Quality (PESQ) [23], and Short-Time Objective Intelligibility(STOI) [24].

We assume that all signals are sampled at 16 kHz. The frame length is 32 msec with the frame shift of 20 msec. All DNN models what we used are trained using the Adam optimizer [25] with a learning rate of 0.001. The mean square error was used as the objective function. The batch normalization was applied to the input layer. The minibatch size and the timesteps were set to 256 and 128, respectively. For training DNNs, we used Tensorflow [26], which is an open toolkit for machine learning provided by Google.

C. Impact of inter-channel variance normalization

We analyze the impact of the null-steering beamformer and that of ICVN from Table II. It shows that the performance is dramatically improved when the input features include the noise reference extracted from the null-steering beamformer. Note that there is one additional calculation of the equation (2) to obtain the coefficients of the null-steering beamformer when comparing to the MVDR beamformer. It implies that the performance of DNNs can be improved effectively just through simple signal processing. In addition, the further improvement

TABLE II. IMPACT OF THE NULL-STEERING BEAMFORMER BASED FEATURES AND THE INTER-CHANNEL VARIANCE NORMALIZATION.

Metrics	**Noisy**	**BF**	**BF+ 1ch DNN**	**BF+ 2ch DNN***	**BF+ 2ch DNN* +ICVN**
POLQA	2.510	2.757	3.309	3.623	**3.708**
PESQ	1.449	1.677	2.489	2.670	**2.819**
STOI	0.849	0.895	0.926	0.930	**0.938**

* DNN input features include the log-power spectrums of the null-steering beamformer output

is achieved through the ICVN. Fig. 3 shows the enhanced spectrograms from example speech utterances corrupted by music noises in SNR=5 dB. It is observed in Fig. 3 that the relative magnitude of noise components in speech reference to noise reference is different. This difference incurs the speech distortion or the residual noise. Fig . 3(f) shows that the residual noise found in Fig . 3(e) is reduced after applying ICVN. It implies that the proposed ICVN can prevent performance degradation caused by inter-channel variation.

D. Objective evaluation

Table III compares the proposed method with the conventional different techniques based on DNN in handset mode and in speakerphone mode, respectively. First, we compare it with a noise-aware training based single channel approach where the log-power spectrums of eleven consecutive frames and the estimated noise spectral information are fed to a three-layer feedforward DNN, each with 2048 units [12]. The corresponding output is generated through inferencing the pre-trained model with the MVDR beamformer output [27]. We find that the noise spectral information extracted from spatial filtering is noticeably effective to increase the performance of noise reduction based DNNs. In addition, although the proposed method uses only the current frame, it outperforms the conventional one ($\boldsymbol{DNN}_1$) which used a long context window including five future frames.

Secondly, the DNN based dual channel solution $\boldsymbol{DNN}_2$, proposed in [11] is compared with ours. It utilizes additional features extracted from the inter-channel and the intra-channel with two channel input. The speech reference and the noise reference without ICVN are used as the input of $\boldsymbol{DNN}_2$. In order to compare the performance fairly, the same DNN structure and the same training data were used. Table III shows the proposed method outperforms the conventional dual channel solution.

In summary, the proposed method achieves the best performance in all three objective measures in both handset mode and speakerphone mode.

E. Complexity analysis

In this subsection, the feasibility of real-time operation is analyzed. The proposed solution has been implemented in the fixed-point code to deploy it on the chipset used for voice call. The weights of CRN were quantized to 16 bits. The quantized model size is 488 Kbyte and it needs 537 thousand multiply-add operations for one frame. The real-time complexity on a 1.2 GHz ARM Cortex-A32 core is 23%. The solution delay is 12ms caused by frame overlap.

TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH VARIOUS DNN BASED APPROACHES IN THE HANDSET MODE AND IN THE SPEAKERPHONE MODE.

Mode	Metrics	**Noisy**	**Proposed method**	***DNN$_1$***	***DNN$_2$***
Handset	**POLQA**	2.510	**3.708**	3.471	3.516
	PESQ	1.449	**2.819**	2.724	2.671
	STOI	0.849	**0.938**	0.933	0.936
Speaker-phone	**POLQA**	2.122	**3.289**	2.949	3.142
	PESQ	1.487	**2.626**	2.330	2.397
	STOI	0.826	**0.914**	0.874	0.900

TABLE IV. PERFORMANCE EVALUATION ON THE CHANNEL MISMATCH ENVIRONMENT. THE EVALUATION WAS CONDUCTED THROUGH GALAXY S20 WHERE THREE MICROPHONES ARE EMBEDDED.

Metrics	Handset		Speakerphone	
	Noisy	Enh.	Noisy	Enh.
POLQA	2.26	**3.138**	2.484	**3.322**
PESQ	1.311	**2.277**	1.468	**2.536**
STOI	0.817	**0.915**	0.8	**0.876**

F. Evaluation on the channel mismatch environment

Table IV shows the performance on Galaxy S20 where three microphones are embedded at bottom, top, and back-side. The clean speech signals were obtained from 7 Korean speakers and it was corrupted by 7 types of noises at various SNRs (-5dB~10dB). Note that the number of microphones was increased comparing to that of Galaxy S10 used for training. Only by a little modification in beamformer modules to increase the channel size, we obtained the output though inferencing the DNN model trained with 2 channel data. Although there are channel mismatches including the number of microphones, the reasonable performance could be obtained without additional training. It implies that the proposed method is extendable easily to any kinds of multi-channel devices.

IV. CONCLUSIONS

In this paper, new effective features were proposed for DNNs based real-time speech enhancement for smartphones. The proposed input features for DNNs extracted from the MVDR beamformer and the null-steering beamformer. In addition, the normalization method called ICVN was proposed to reduce the inter-channel variance between two beamformer outputs. Experimental results showed that the performance could be improved effectively through adopting the normalized null-steering beamformer output as the input feature even at a light-weight model. In addition, the proposed method consistently outperformed other DNN based approaches. These results are remarkably meaningful for voice call solutions on smartphones because current smartphones have two or more microphones and its computing power is very limited. Furthermore, we showed that the proposed methods can be extended to multiple microphone configurations without additional retraining the model.

REFERENCES

[1] Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on signal processing* vol.47, no.10, pp.2677-2684 (1999).

[2] Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.19, no.1, pp. 206-219 (2010).

[3] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 493–1508, Sep. 2015.

[4] 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).

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. ICASSP* (2016).

[6] Lu, Xugang, Yu Tsao, Shigeki Matsuda, and Chiori Hori. "Speech enhancement based on deep denoising autoencoder," In *Interspeech*, pp. 436-440 (2013).

[7] Wang, Yuxuan, Arun Narayanan, and DeLiang Wang. "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing,* vol.22, no.12, pp.1849-1858 (2014).

[8] Narayanan and D. L.Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP* (2013).

[9] D.S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* vol.24, no.3 pp.483-492 (2016).

[10] K. Tan, and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech* (2018).

[11] K. Tan, X. Zhang, and D. Wang, "Real-time Speech Enhancement Using an Efficient Convolutional Recurrent Network for Dual-microphone Mobile Phones in Close-talk Scenarios," In *Proc. ICASSP (2019).*

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[13] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP* (2015).

[14] Pascual, Santiago, Antonio Bonafonte, and Joan Serra. "SEGAN: Speech enhancement generative adversarial network." *arXiv preprint arXiv:1703.09452* (2017).

[15] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. ICASSP* (2018)

[16] J. M. Martn-Donas, A. M. Gomez, I. L´opez-Espejo, and A. M. Peinado, "Dual-channel DNN-based speech enhancement for smartphones," in *Proc. MMSP* (2017).

[17] A.A. Nugraha, A. Liutkus, E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process*., vol. 24, no. 9, pp.1652–1664 (2016).

[18] Z-Q. Wang, et al., "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *Proc. ICASSP* (2018)

[19] J. Heymann, L. Drude and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc*. *ICASSP* (2016).

[20] S.-J. Chen, et al., "Building State-of-the-art Distant Speech Recognition Using the CHiME-4 Challenge with a Setup of Speech Enhancement Baseline," in *Proc. Interspeech* (2018).

[21] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement" in *Proc. MMSP* (2018).

[22] Perceptual Objective Listening Quality Assessment, ITU-T Rec. P.863, 2010.

[23] P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs, ITU-T Std. P.862.2, 2007.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[25] D. P. Kingma, and B. Jimmy, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).

[26] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[27] Y. Xu, DNN-for-speech-enhancement, GitHub repository, https://github.com/yongxuUSTC/DNN-for-speech-enhancement.