Multichannel Speech Enhancement by Raw Waveform-Mapping Using Fully Convolutional Networks

Chang-Le Liu, Sze-Wei Fu, You-Jin Li, Jen-Wei Huang, Hsin-Min Wang, *Senior Member, IEEE*, and Yu Tsao, *Member, IEEE*

***Abstract*—In recent years, waveform-mapping-based speech enhancement (SE) methods have garnered significant attention. These methods generally use a deep learning model to directly process and reconstruct speech waveforms. Because both the input and output are in waveform format, the waveform-mapping-based SE methods can overcome the distortion caused by imperfect phase estimation, which may be encountered in spectral-mapping-based SE systems. So far, most waveform-mapping-based SE methods have focused on single-channel tasks. In this article, we propose a novel fully convolutional network (FCN) with Sinc and dilated convolutional layers (termed SDFCN) for multichannel SE that operates in the time domain. We also propose an extended version of SDFCN, called the residual SDFCN (termed rSDFCN). The proposed methods are evaluated on three multichannel SE tasks, namely the dual-channel inner-ear microphones SE task, the distributed microphones SE task, and the CHiME-3 dataset. The experimental results confirm the outstanding denoising capability of the proposed SE systems on the three tasks and the benefits of using the residual architecture on the overall SE performance.**

***Index Terms*—Multichannel speech enhancement, raw waveform mapping, fully convolutional network (FCN), inner-ear microphones, distributed microphones.**

I. Introduction

SPEECH-related applications for both human-human and human-machine interfaces have garnered significant attention in recent years. However, speech signals are easily distorted by additive or convolutional noises or recording devices, and such distortion constrains the achievable performance of these applications. To address this issue, numerous speech enhancement (SE) algorithms have been derived to improve the quality and intelligibility of distorted speech and are widely used as a preprocessor in speech-related applications, such as speech coding [1], [2], assistive hearing devices [3], [4], and automatic speech recognition (ASR) [5]. Generally speaking, SE methods can be divided into two categories. The first category adopts a single channel (also termed monaural) while the second category uses multiple microphones (also termed multichannel) to perform SE.

Traditional single-channel-based SE methods were derived based on the characteristics and statistical assumptions of clean speech and noise signals. Well-known approaches include spectral-subtraction [6], the Wiener filter [7], [8], and the minimum mean square error (MMSE) [9]. Another category of successful SE approaches is subspace-based methods, which aim to separate noisy speech into two subspaces, one for clean speech and the other for noise components. The clean speech is then restored based on the information in the clean-speech subspace. Notable subspace techniques include generalized subspace approaches with prewhitening [10], the Karhunen-Loeve transform [11], and principal component analysis (PCA) [12].

In recent years, machine-learning-based algorithms have been popularly used in the SE field. Unlike traditional methods, a machine-learning-based SE approach generally prepares a denoising model in a data-driven manner without imposing strong statistical constraints. Well-known machine-learning-based models include non-negative matrix factorization [13], compressive sensing [14], sparse coding [15], and robust principal component analysis (RPCA) [16]. More recently, deep learning models have been applied to the SE field. Owing to their outstanding nonlinear mapping capability, deep-learning-based SE methods have demonstrated notable performance improvements over traditional statistical methods and other machine-learning-based methods. Well-known deep-learning-based models include the deep denoising autoencoder (DDAE) [17], [18], deep fully connected networks [19]–[22], recurrent neural networks [23], [24], convolutional neural networks [25], [26], and long short-term memory [27]–[30].

Different from single-channel SE methods, the multichannel ones utilize information from plural channels to enhance the target speech signal. Among the multichannel SE methods, beamforming [31]–[33] is a popular method that exploits spatial

Manuscript received June 17, 2019; revised October 30, 2019 and January 2, 2020; accepted February 1, 2020. Date of publication February 26, 2020; date of current version June 26, 2020. This work was supported by the Ministry of Science and Technology, Taiwan (106-2221-E-001-017-MY2 and 107-2221-E-001-012-MY2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. A. W. H. Khong. *(Corresponding author: Yu Tsao.)*

Chang-Le Liu is with the Department of Electrical Engineering, National Taiwan University, 10617 Taipei, Taiwan (e-mail: b05901017@ntu.edu.tw).

Sze-Wei Fu, You-Jin Li, and Yu Tsao are with the Research Center for Information Technology Innovation at Academia Sinica, 11529 Taipei, Taiwan (e-mail: d04922007@ntu.edu.tw; jinliyou1991@iis.sinica.edu.tw; yu.tsao@citi.sinica.edu.tw).

Jen-Wei Huang is with the Department of Electrical Engineering, National Cheng-Kung University, 70101 Tainan, Taiwan (e-mail: jwhuang@mail.ncku.edu.tw).

Hsin-Min Wang is with the Institute of Information Science, Academia Sinica, 11529 Taipei, Taiwan (e-mail: whm@iis.sinica.edu.tw).

Digital Object Identifier 10.1109/TASLP.2020.2976193

information from multiple microphones to attenuate inference and noise signals. In addition to beamforming, other effective methods are based on a coherence algorithm that calculates the correlation of multiple input signals to estimate a filter to attenuate the interference components [34], [35]. Meanwhile, Li *et al.* proposed a method of using distributed-microphones for in-vehicle SE [36]. They argued the clean speech signals acquired by distributed-microphones are similar to each other while the noise signals acquired by distributed-microphones are irrelevant to each other. Therefore, the RPCA algorithm [16] is applied to the matrix formed by the acquired noisy signals from multiple channels to separate clean speech and noise components [36].

More recently, deep learning-based models also exhibit encouraging performance in multichannel SE tasks. Araki *et al.* showed that multichannel audio features can effectively improve the performance of the denoising auto-encoder (DAE) [37] based SE approach [38]. Wang and Wang proposed a deep learning-based time-frequency (T-F) masking SE method that estimates robust time delay of arrival over multiple singly-enhanced speech signals to obtain directional features and hence the beamformed signals. The enhancement is carried out by combining spectral and directional features [39]. Although the above-mentioned multichannel SE approaches have been able to provide satisfactory performance, they are performed in the frequency domain, i.e., they typically use the phase from the noisy input and require additional processing to convert the speech waveform into spectral features. To avoid imperfect phase estimation and reduce online processing, waveform-mapping-based audio signal processing methods have been developed. For example, in [41]–[44], a fully convolutional network (FCN) model was used to process the noisy waveform to generate an enhanced waveform, and in [45], [46], the FCN model was used to separate a singing voice from mono or stereo music.

In the present work, we propose a novel fully convolutional network that incorporates Sinc convolutional filters (termed SincConv) and dilated convolutional filters, to perform multichannel SE in the time domain. Therefore, the model is called Sinc dilated FCN (termed SDFCN). In addition, we derive an extended system from the SDFCN system. The extended system structures a residual architecture in which SDFCN is used to estimate and compensate for the residual components of the enhanced speech from a primary SE model. Therefore, it is named residual SDFCN (termed rSDFCN). We evaluate the proposed models on three multichannel SE tasks: inner-ear microphones (termed the IEM-SE task), distributed-microphones (termed the DM-SE task), and the CHiME-3 dataset [65]. For these tasks, the proposed SE models take inputs from multiple channels to generate a single-channel waveform with higher quality and intelligibility than individual noisy inputs. Two standardized metrics are used in the evaluation: short-time objective intelligibility (STOI) [47], [48] and perceptual estimation of speech quality (PESQ) [49]. In addition, we conduct subjective listening and speech recognition tests with the enhanced speech signals. Our experimental results confirm the outstanding denoising capability of the proposed SDFCN and rSDFCN models in all three multichannel SE tasks, demonstrating the benefits of using the residual architecture on the overall SE performance.

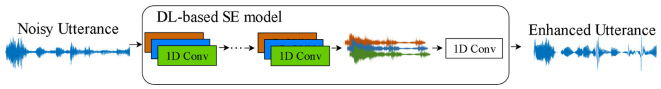

Fig. 1. A waveform-mapping-based SE system.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III presents the concept and architectures of the proposed SDFCN and rSDFCN models. Section IV presents the experimental setup and results. Finally, Section V concludes this work.

II. Related Works

Given a clean speech signal $\mathbf{x}$, the degraded signal can be formulated as $\mathbf{y} = g(\mathbf{x})$, where g denotes the degradation function. The goal of SE is to find a function that maps $\mathbf{y}$ to $\hat{\mathbf{x}}$ that approximates $\mathbf{x}$ as close as possible. In this section, we review related works, including the FCN-based waveform-mapping-based SE method, SincConv filters, and dilated convolutional filters.

A. Waveform-Mapping-Based SE

Previous studies have shown that the FCN model is suitable for waveform-mapping-based SE because the convolutional layers can more effectively characterize the local information of neighboring input regions [40]. FCN is a modified convolutional neural network (CNN) model in which the fully connected layers in CNN are completely replaced by the convolutional layers, as shown in Fig. 1. In FCN, the relation between each sample point $\hat{\mathbf{x}}_t$ of the output $\hat{\mathbf{x}}$ and the last connected hidden nodes $\mathbf{h_t} \in R^{L\times 1}$ can be represented by

$$\hat{\mathbf{x}}_t = \mathbf{v}^T \mathbf{h}_t + b, \tag{1}$$

where $\mathbf{v} \in R^{L\times 1}$ denotes a convolutional filter, b is a bias term, and L is the size of the filter. Note that $\mathbf{v}$ and b are shared in the convolution operation and are fixed for every output. Because the pooling step may reduce the precision of speech signal reconstruction, we did not apply any pooling operations (e.g., WaveNet [50]) to perform SE when using FCN. For more details about the structure of the FCN model applied to waveform-mapping-based SE, please refer to previous works [40], [41], [50].

B. SincConv Filters

As mentioned above, convolutional filters are often used to process raw-waveforms. When the CNN model is too deep or the training data is insufficient, the filters of the first few layers may not be well learned because of the vanishing gradient issue. To overcome this issue, Ravanelli *et al.* [51] recently proposed a novel convolutional architecture, called SincNet. Unlike conventional CNN models that learn all filters based on training data, SincNet predefines the filters of the first few layers to model the rectangular band-pass filter-banks in the frequency domain. Specifically, assuming that the filter function of the first

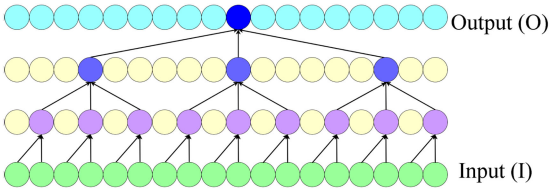

Fig. 2. Input (I) and output (O) with two-layered dilated convolutional filters.

layer is $\mathbf{v}$, which will be convolved with the input signal $\mathbf{y}$, then $\mathbf{v}$ can be written as follows:

$$\mathbf{v} = \mathbf{s} \circ \mathbf{w}$$

$$\mathbf{s}_t = 2f_{low}\text{sinc}(2\pi f_{low}t) - 2f_{high}\text{sinc}(2\pi f_{high}t)$$

$$\mathbf{w}_t = 0.54 - 0.46\cos\left(\frac{\pi t}{L}\right)$$

where $\circ$ is component-wise multiplication, L is the filter length, and f_{low} and f_{high} are the low and high cutoff frequencies learned during training, respectively. Obviously, this architecture is much more efficient because each filter in the first layer only consists of two coefficients rather than L (the original filter length) coefficients. In [51], it was shown that SincNet converged faster in training and performed better in testing than CNN on a speaker recognition task when the input was raw speech waveform.The smaller number of neurons enables SincNet to be well trained even on a dataset with a limited amount of training data [51].

C. Dilated Convolution

Previous works, such as WaveNet [50], Conv-TasNet [52], and WaveGAN [53] have shown that using a large temporal context window is important in waveform modeling. To efficiently take advantage of the long-range dependency of speech signals, dilated convolution was proposed in [54]. In [43], [50], [54], the effectiveness of the dilated convolutional layers was shown to expand the receptive field exponentially (rather than linearly) with depth. Fig. 2 shows an example that demonstrates the concept of dilated fully convolutional filters. The input signal (I) is processed by a dilated convolutional block to generate the output signal (O).

The input sequence has 18 points. When using a one-dimensional fully convolutional filter to process the input signal, the number of receptive fields is 18. On the other hand, when using a dilated fully convolutional block with filter sizes of 2, 3, and 3 and dilated rates of 1, 2, and 6, the receptive field is also 18. Compared to a single-layered FCN block, with the same size of receptive fields, the dilated fully convolutional block requires only half the number of parameters but four times the depth, suggesting that the dilated fully convolutional block can have a deeper architecture than the conventional fully convolutional filter when the total number of parameters is fixed.

III. The Proposed Multichannel SE System

In this section, we first introduce the proposed SDFCN multichannel SE system. Then, we explain the extended system, rSDFCN. The design concept and architectures of SDFCN and rSDFCN are presented in detail.

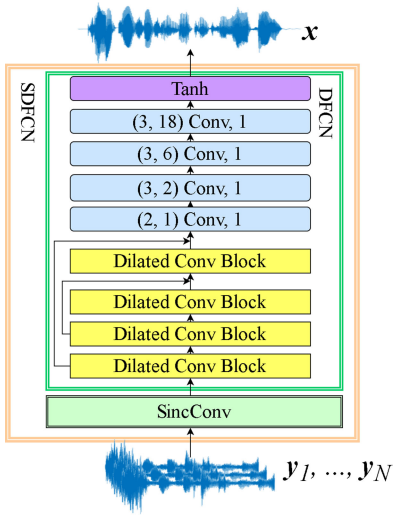

Fig. 3. Architecture of the SDFCN multichannel SE system. Each of four blue rectangles denotes one dilated convolutional layer, and the parameters are denoted as follows: ($p1$, $p2$) Conv $p3$, where $p1$ is the kernel size, $p2$ is the dilated rate, and $p3$ is the number of filters (channels).

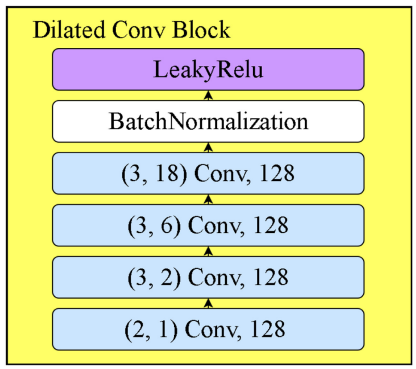

Fig. 4. Architecture of the dilated convolutional block in the SDFCN model.

A. The SDFCN System

Fig. 3 shows the architecture of the proposed SDFCN multichannel SE system, which consists of a SincConv layer and a dilated FCN (termed DFCN) module. The DFCN module consists of four layers of dilated convolutional blocks (Dilated Conv Block in Fig. 3), four dilated convolutional layers, and a tanh activation function layer. A skip-connection scheme is adopted to provide additional low-level information to the higher-level process. From our preliminary experimental results, we note that with such a skip-connection scheme, the SDFCN model can be trained more efficiently. Given the multichannel inputs: $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$, where N denotes the number of channels, we have

$$\hat{\mathbf{x}} = f_{DFCN}(f_{SincConv}(\mathbf{Y})), \tag{2}$$

where $f_{SincConv}(\cdot)$ and $f_{DFCN}(\cdot)$ denote the mapping functions of the SincCnov layer and the DFCN module, respectively. Fig. 4 shows the architecture of the dilated convolutional blocks

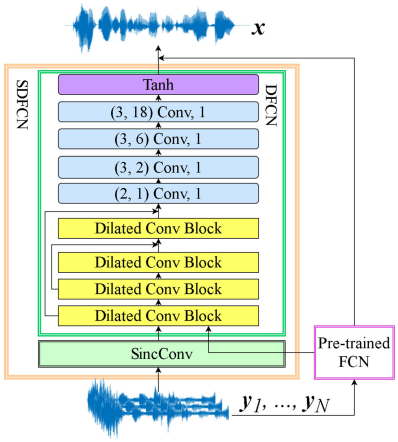

Fig. 5. Architecture of the rSDFCN multichannel SE system, which consists of a primary SE module (pre-trained FCN) and an SDFCN system.

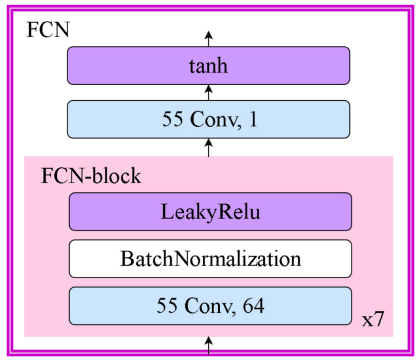

Fig. 6. Architecture of the FCN model that is used as the primary SE module in the proposed rSDFCN system. We use p_1 Conv p_2 to represent a convolutional layer with p_2 filters and kernel size of p_1.

(Dilated Conv Block in Fig. 3) in the SDFCN model. The block consists of four dilated convolutional layers (the four blue rectangles) followed by batch normalization and LeakyRelu. The receptive field of the dilated convolutional block is 54 ($2 \times 3 \times 3 \times 3$).

B. The Residual SDFCN (rSDFCN) System

Recently, residual structures have been popularly used in neural network models to attain better classification and regression efficacy. In speech signal generation tasks, residual connections also yield promising performance because the residual connection provides a linear shortcut, and the non-linear part of the network only needs to deal with the residuals (differences) of the estimated and reference signals, which are usually easier to model. In this work, we also explore the combination of the residual structures with SDFCN. This combined model is termed the residual SDFCN (rSDFCN). The architecture of an rSDFCN multichannel SE system is shown in Fig. 5.

As can be seen from the figure, an additional SE module (the pre-trained FCN in Fig. 5) is used. This SE module is treated as the primary SE module, and the output of the primary SE module is combined with the output of the SDFCN system to form the final enhanced output. The formulation of the rSDFCN can be represented as:

$$\hat{\mathbf{x}} = f_{DFCN}(f_{SincConv}(\mathbf{Y}), f_{Pr}(\mathbf{Y})) + f_{Pr}(\mathbf{Y}), \quad (3)$$

where $f_{Pr}(\cdot)$ is the mapping function of the primary SE module. When implementing the rSDFCN system, we first pre-train the primary SE module and then train the SDFCN system. In this way, the SDFCN system learns the residual components (or differences) of the clean reference and the enhanced output of the primary SE module. More specifically, the SDFCN system is trained with the aim of minimizing the following loss function:

$$\|f_{DFCN}(f_{SincConv}(\mathbf{Y}), f_{Pr}(\mathbf{Y})) - [\mathbf{x} - f_{Pr}(\mathbf{Y})]\|^2. \quad (4)$$

In this paper, we use a pre-trained FCN model as the primary SE module. Its architecture is shown in Fig. 6. The module consists of seven layers of convolution blocks, a convolutional layer, and a tanh activation function layer. Each convolution block consists of a convolutional layer (with length = 55 and channel = 64), batch normalization, and LeakyRelu. When implementing the rSDFCN, this pre-trained FCN can be prepared beforehand using a different training set. Please note that the architectures of the FCN, SDFCN, and rSDFCN presented above are designed based on the datasets used in this study. The parameters, including the numbers of layers and channel filters and the kernel size can be adjusted according to the target task.

IV. Experimental Setup And Results

In this section, we first introduce the experimental setup for the IEM-SE and DM-SE tasks[1]. Then, we present the results of the proposed SDFCN and rSDFCN systems for these two tasks. Finally, we discuss the performance of the rSDFCN system on several subsets of the CHiME-3 dataset with different subset size. For IEM-SE and CHiME-3 task, we also discuss the effectiveness of dilated convolution and SincConv layer.

A. Experimental Setup

We evaluated the SE performance in terms of two standard objective metrics: STOI [47], [48] and PESQ [49]. The STOI score ranges from 0 to 1, and the PESQ score ranges from 0.5 to 4.5. For STOI and PESQ, a higher score indicates that the enhanced speech signal has higher intelligibility and better quality, respectively, with reference to the speech signal recorded by the near-field high-quality microphone. In addition, we also conducted listening tests and evaluated the speech recognition performance of enhanced speech in terms of the Chinese character error rate (CER) using Google Speech Recognition [55]. For comparison, we implemented a DDAE-based multichannel SE system [17], [18]. In previous studies, the single-channel DDAE approach has shown outstanding performance in noise reduction [56], dereverberation [57], and bone-conducted speech enhancement [58]. Here, we extended the original single-channel DDAE approach to form a multichannel DDAE system. Fig. 7

[1]Speech samples and codes can be found via: [Online]. Available: https://yu-tsao.github.io/MCSE/

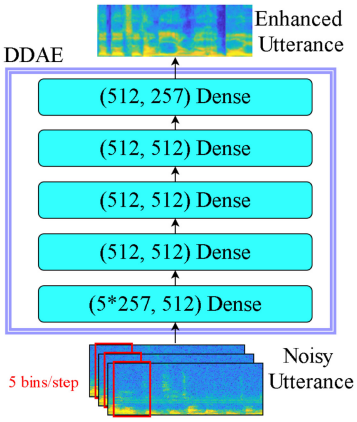

Fig. 7. Architecture of the DDAE multichannel SE system.

shows the architecture of the multichannel DDAE system, which consists of five dense layers. The input is multiple sequences of noisy spectral features (log-power spectrogram (LPS) in this study) from the multiple channels, and the output is a sequence of enhanced spectral features. The phase of one of the noisy speech utterances was used as the phase to reconstruct the enhanced waveform. All neural network models were trained using the Adam optimizer [59] with a learning rate of 0.001. The α value of LeakyReLU was set to 0.3.

B. The Inner-Ear Microphones-SE Task

When speech signals are recorded using inner-ear microphones, interference from the environment can be blocked, so that purer signals can be captured. However, owing to the different transmission pathways, the speech signals captured by the IEMs exhibit different characteristics from those recorded by normal air-conducted microphones (ACMs). Generally speaking, the high-frequency components of speech recorded by an IEM are suppressed, thereby notably degrading the speech quality and intelligibility. Moreover, owing to the loss of high-frequency components, IEM speech cannot provide a satisfactory ASR performance.

For the IEM-SE task, we intend to transform the speech signals captured by a pair of IEMs into ACM-like speech signals with improved quality and intelligibility. In the past, there have been some studies on IEM-to-ACM transformation. In [60], [61], bandwidth expansion and equalization techniques were used to map the IEM speech signals to the ACM ones. Because the mapping function between IEM and ACM is nonlinear and complex, traditional linear filters may not provide optimal performance. In the present study, we propose to perform multichannel SE in the waveform domain for IEM-to-ACM transformation.

Our recording condition is shown in Fig. 8. A male speaker sat in a sound booth (3 m $\times$ 5.2 m, 2 m in height) and wore a pair of IEMs and a near-mouth ACM. The three microphones simultaneously recorded speech signals spoken by the male speaker. The recording scripts were the Taiwan Mandarin Chinese version of Hearing in Noise Test (TMHINT) sentences [62]. There were 250 utterances for training and another 50 utterances for testing.

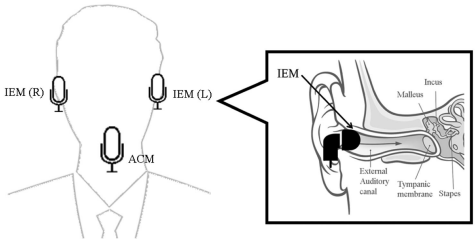

Fig. 8. Recording setting of IEM-SE task. There is a near-mouth microphone and two IEMs in both ears [73].

All utterances were recorded at a 16,000 Hz sampling rate then truncated to speech segments, with each segment containing 36,500 sample points (around 2.28 seconds).

Before discussing the results of the proposed SDFCN and rSDFCN systems, we first verified the effectiveness of dilated convolution and SincConv layer. There were totally four models trained in this set of experiments. We compared *FCN-55* with *DCN-54* to show the effect of dilated convolution. The benefits of SincConv layer were shown by comparing *FCN-251* with *SincFCN-251*. *FCN-55* is similar to regular FCN shown in Fig. 6, but it has only four layers in total. *DCN-54* was designed by replacing the last three Conv layers in *FCN-55* with dilated convolutional block shown in Fig. 4, where the kernel size is 55. The reason for using four-layer models is that models with less than four layers could not enhance utterances well in our preliminary experiments. As mentioned in the previous section, the receptive field of the dilated convolutional blocks was set to be 54 to approximate the kernel size used in FCN. *FCN-251* was designed by changing the kernel size of the first Conv layer in *FCN-55* from 55 to 251, and *SincFCN-251* was designed by replacing the first Conv layer in *FCN-251* with a SincConv layer. The reason that we changed the kernel size of the first layer was to make it have the same size as the original work [51]. For a fair comparison, the numbers of filters of all models trained and tested in the experiment are 30.

Table I lists the average STOI and PESQ scores of the original speech signals captured by the left and right IEMs (denoted as IEM (L) and IEM (R), respectively) and the enhanced speech signals by the four models mentioned above. The corresponding ACM speech was used as the reference to compute the scores. By comparing the results of the middle columns in Table I, we observe that the STOI and PESQ scores can be further improved by the dilated convolutional layer. The results in the last column in Table I show that the SincConv layer performs much better than the original convolutional layer.

Fig. 9 shows the learning curves of the four models in terms of the MSE scores. When computing the MSE scores, we have pre-processed each utterance by normalizing the waveform samples by the peak amplitude. From Table I and Fig. 9, we can see that although their losses (MSE) converge to a similar value, the training speed of *SincFCN-251* is much faster, and the corresponding STOI and PESQ scores are also higher than others. It is also noted that, *DCN-54* and *SincFCN-251* outperform *FCN-55* and *FCN-251* in terms of STOI and PESQ, respectively, which confirms the effectiveness of the dilated convolution block and

TABLE I
AVERAGE STOI AND PESQ SCORES OF *FCN-55*, *DCN-54*, *FCN-251*, *SINCFCN-251*, WITH SINGLE-CHANNEL/MULTICHANNEL INPUTS FOR IEM-SE TASK

No. of ch.	1	1	2	2	2	2
Model	IEM(L)	IEM(R)	*FCN-55*	*DCN-54*	*FCN-251*	*SincFCN-251*
STOI	0.694	0.694	0.801	**0.817**	0.727	**0.843**
PESQ	1.146	1.101	1.317	**1.360**	1.171	**1.476**

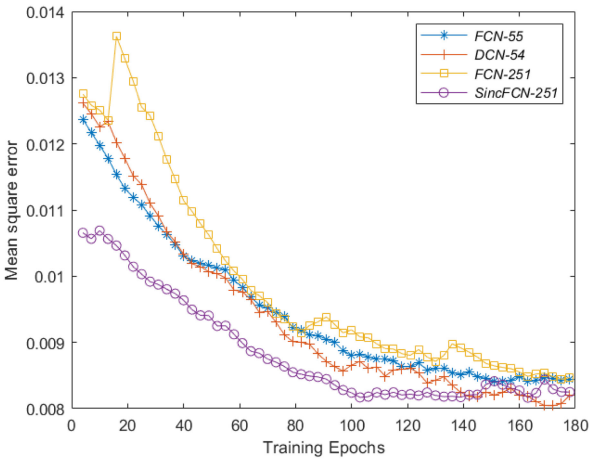

Fig. 9. MSE on the test set of *FCN-251*, *DCN-54*, *FCN-251*, and *SincFCN-251* over training epochs.

TABLE II
AVERAGE MSE SCORES OF *FCN-55*, *DCN-54*, *FCN-251*, AND *SINCFCN-251* FOR IEM-SE TASK

No. of ch.	2	2	2	2
Model	*FCN-55*	*DCN-54*	*FCN-251*	*SincFCN-251*
MSE	0.171	**0.165**	0.179	**0.175**

SincConv layer. Also, from Table I, we can observe that *FCN-55* outperforms *FCN-251*, which implies that the performance of FCN may not be improved by just increasing the kernel size of the convolutional layers. In Table II, we further list the average MSE scores of *FCN-55*, *DCN-54*, *FCN-251*, and *SincFCN-251* under the 180 training epoch condition in Fig. 9. The results in the table show that *DCN-54* and *SincFCN-251*, respectively, yield lower MSE scores as compared to *FCN-55* and *FCN-251*, again confirming the benefits of the dilated convolution and SincConv.

To visually compare FCN and SincConv, we plot the learned filters of *FCN-251* and *SincFCN-251* in Fig. 10 (for a clearer presentation, we only used 30 filters for both FCN and SincConv to plot Fig. 10). In the meanwhile, we plot the extracted features of an utterance from FCN and SincConv layers in Fig. 11 (for a clearer presentation, we only used seven filters for both *FCN-251* and *SincFCN-251* to obtain the features in Fig. 11). From Figs. 10 and 11, and Table II, we can note that our experiment results are quite consistent with those in the previous works [51], [68], [69]. From Fig. 10(b), we can see that the SincConv layer learns a filter bank containing more filters with high cut-frequencies compared to the traditional convolutional layer. The filters learned by FCN, as shown in Fig. 10(a), do not cover all the frequency ranges. We note that this phenomenon is due to the limited amount and coverage of training data, and the

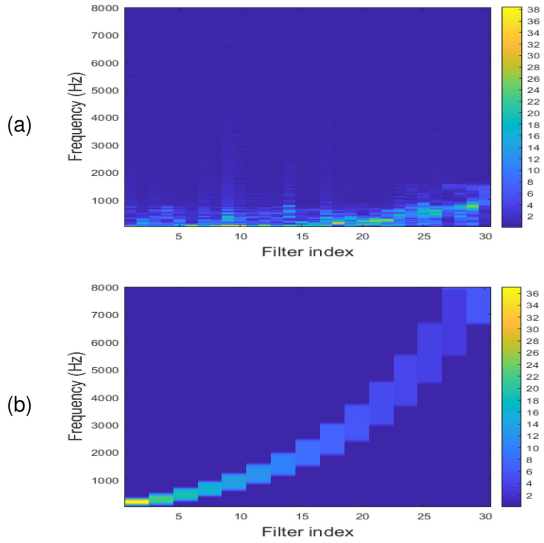

Fig. 10. Frequency responses of the learned filters of (a) the first layer of *FCN-251* and (b) the SincConv layer of *SincFCN-251*.

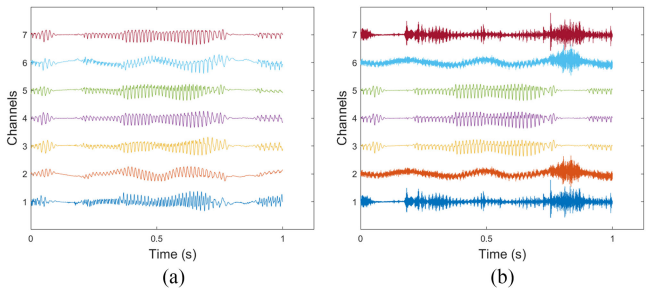

Fig. 11. Extracted features of an utterance audio from convolution by the filters of the first layers of (a) *FCN-251* and (b) *SincFCN-251*.

high frequency ranges become clearer when a sufficient amount and coverage of training data is available [41]. From Fig. 11, we can observe that the first-layered features of *SincFCN-251* contain more high frequency components than *FCN-251*. In addition, the results in Table II are consist with those in [69]: With dilated convolution, the network can more accurately model ground-truth waveforms in terms of MSE.

Furthermore, to investigate the effectiveness of using multiple (dual) channels, we also compared the SDFCN model trained with dual-channel input and that trained with single-channel input. The results are denoted as SDFCN (using dual-channel inputs), SDFCN(L) (using the left channel only) and SDFCN(R) in the left part of Table III. From the table, we first note that SDFCN(L) and SDFCN(R) achieve improved STOI and PESQ

TABLE III
AVERAGE STOI AND PESQ SCORES OF DIFFERENT SINGLE-CHANNEL/MULTICHANNEL SE MODELS FOR THE IEM-SE TASK

No. of ch.	1	1	2	2	2	2	2
Model	SDFCN(L)	SDFCN(R)	SDFCN	DFCN	FCN	DDAE	rSDFCN
STOI	0.861	0.824	**0.880**	0.867	0.834	0.773	**0.894**
PESQ	1.631	1.597	**1.643**	1.562	1.446	1.939	**1.986**

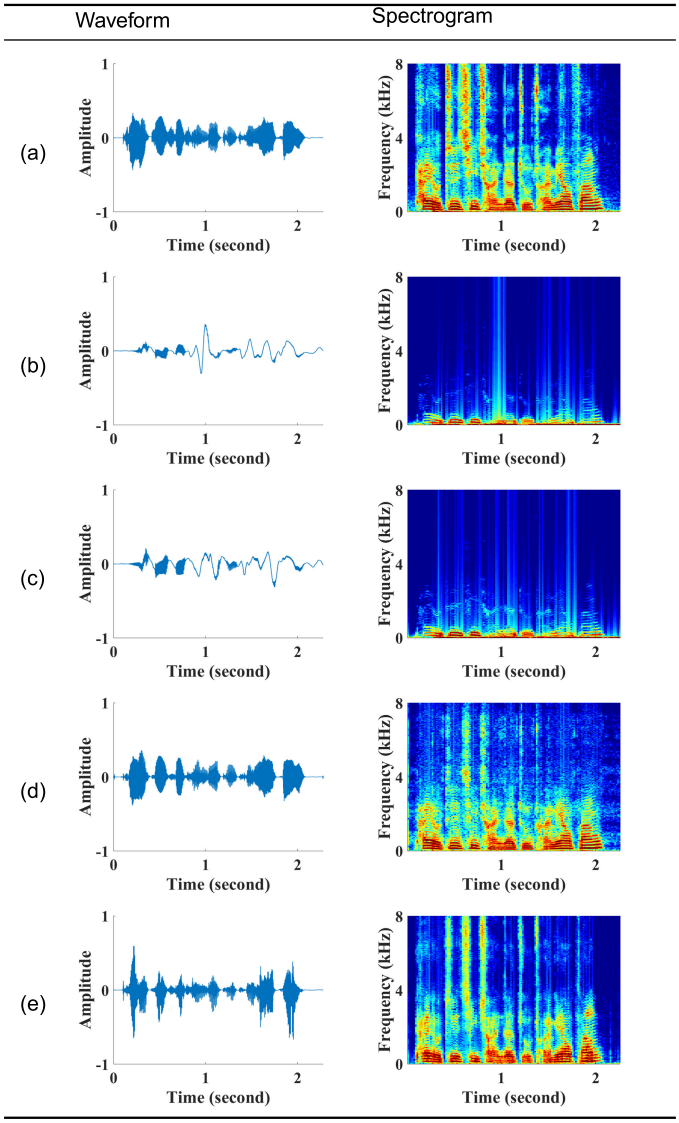

Fig. 12. Waveforms and spectrograms of an example utterance in the IEM-SE task: (a) recorded speech by near-mouth microphone; (b) and (c) recorded speech by right and left IEMs, respectively. (d) and (e) enhanced speech by rSDFCN and DDAE, respectively.

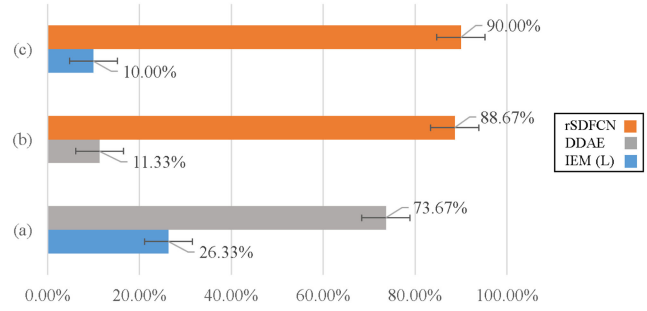

Fig. 13. Results of the AB preference test (with 95% confidence intervals) on speech quality compared between the proposed rSDFCN with IEM(L) and DDAE for the IEM-SE task.

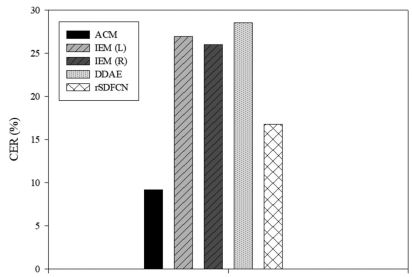

Fig. 14. ASR results achieved by different SE models for the IEM-SE task.

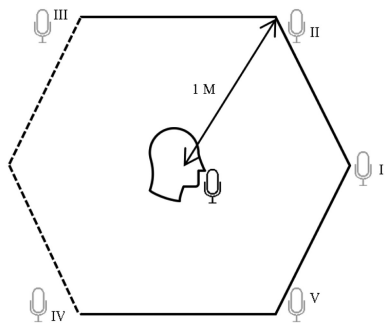

Fig. 15. Recording setting for the DM-SE task. There is a near-field high-quality microphone and five far-field lower-quality microphones. Distances of near-field microphone to far-field microphones are all 1 meter.

scores over IEM(L) and IEM(R), as shown in Table I, respectively. The results confirm the effectiveness of the proposed SDFCN system for single microphone SE. Next, we note that SDFCN outperforms both SDFCN (L)and SDFCN(R), confirming the advantage of the multichannel (dual-channel) mode over its single-channel counterparts.

Next, we report the results of rSDFCN in the right part of Table III. To confirm the effectiveness of SincConv, we replaced the SincConv layer in SDFCN with a normal convolutional layer, denoted as DFCN. FCN denotes the results of the pre-trained FCN module used in rSDFCN. Comparing the results of SDFCN and DFCN in Table III, we confirm the effectiveness of SincConv for the SE task. Comparing the results of SDFCN, FCN and rSDFCN in Table III, we confirm the effectiveness of the residual architecture for the SE task. Next, we note that both SDFCN and rSDFCN outperform the baseline DDAE system while rSDFCN outperforms SDFCN.

TABLE IV
AVERAGE STOI AND PESQ SCORES OF RSDFCN AND DDAE FOR THE DM-SE TASK.

AVG. STOI/PESQ	STOI			PESQ		
Input Microphone(s)	Unenhanced	DDAE	rSDFCN	Unenhanced	DDAE	rSDFCN
I	0.872	0.823	**0.932**	1.602	1.618	**1.648**
II	0.896	0.814	**0.930**	**1.736**	1.606	1.656
III	0.888	0.813	**0.931**	0.526	1.623	**1.644**
IV	0.881	0.813	**0.931**	1.495	1.581	**1.642**
V	0.893	0.816	**0.931**	**1.727**	1.581	1.646
I, II, V		0.823	**0.950**		1.655	**1.780**
I, II, III, IV, V		0.829	**0.954**		1.635	**1.826**

In addition to comparing the objective scores, we also conducted qualitative analysis. Fig. 12(a), (b), (c), (d), and (e) show the waveforms and spectrograms of the near-field ACM, IEM(L), and IEM(R) speech signals and the enhanced speech signals obtained by rSDFCN and DDAE, respectively. By comparing Fig. 12(a), (b), and (c), we can easily note that the IEM speech signals suffer from notable distortion, with high-frequency components being suppressed. Next, by comparing Fig. 12 (a) and (d), we note that the proposed rSDFCN multichannel SE approach can generate an enhanced speech signal similar to the ACM recorded speech signal. We can also observe that the DDAE-enhanced speech signal has a clearer structure in the high-frequency components while exhibiting some distortion in the low-frequency components.

To subjectively evaluate the perceptual quality of the enhanced speech, we conducted AB reference tests to compare the proposed rSDFCN with the original IEM speech (here IEM(L) was used since it gave slightly higher PESQ scores in Table I). For comparison, the DDAE enhanced speech was also involved in the preference test. Accordingly, three pairs of listening tests were conducted, namely rSDFCN versus IEM, DDAE versus IEM, and rSDFCN versus DDAE. Each pair of speech samples were presented in a random order. For each listening test, speech samples were randomly selected from the test set. 15 listeners participated in the listening test. Listeners were instructed to select the speech sample with better quality. The stimuli were played to the listeners in a quiet environment through a set of Sennheiser HD headphones at a comfortable listening level. The results of the AB reference tests are presented in Fig. 13. From the figure, it is clear that rSDFCN and DDAE outperform IEM with notable margins, confirming the effectiveness of these two SE approaches. Next, we note that rSDFCN yields a higher preference score compared to DDAE, showing that rSDFCN can more effectively enhance the IEM speech.

Finally, we tested the ASR performance in terms of the character error rate (CER). The results of the speech recorded by ACM, IEM(L), and IEM(R) and the enhanced speech by the rSDFCN and DDAE are shown in Fig. 14. The CER of the ACM-recorded speech is 9.2%, which can be regarded as the upper-bound. The CERs of the speech recorded by IEM(L) and IEM(R) and the enhanced speech by rSDFCN and DDAE are 26.9%, 26.0%, 16.8%, and 28.6%, respectively. From the results, we note that rSDFCN can improve the ASR performance over IEM(L) and IEM(R). Compared with IEM(L), CER decreased by 35.38% (from 26.0% to 16.8%). Comparing the results in Figs. 13 and 14 and Table III, we note that rSDFCN outperforms DDAE in terms of PESQ, STOI, subjective preference test scores, and ASR results, confirming the effectiveness of the proposed rSDFCN over the conventional DDAE approach for the IEM-SE task.

C. The Distributed Microphone-SE Task

For the DM-SE task, we also used the scripts of the TMHINT sentences to prepare the speech dataset. The layout of the recording is shown in Fig. 15. A high-quality near-field micro-phone (Shure PGA181 [63]) was placed right in front of the speaker and five lower-quality microphones (all of the same brand and model: Sanlux HMT-11 [64]) were located at the five vertices of the regular hexagon, 1 meter away from the speaker. The room size is 15.5 m $\times$ 11.2 m and 3.27 m in height. We labeled the lower-quality microphones in counterclockwise order from I to V starting from the microphone in front of the speaker. Herein, the goal was to generate an enhanced (high-quality) speech signal using the speech signals recorded by the distant and lower-quality microphones. To validate the effectiveness of using multiple channels for SE, we designed seven scenarios: five single-channel SE scenarios where the input consisted of the speech signal recorded by one of the five microphones [(I), (II), (III), (IV), or (V)] and the output was the enhanced speech signal, and two multichannel SE scenarios, where the input consisted of the speech signals recorded by three microphones (I, II, and V) and five microphones (I, II, III, IV, and V) and the output was the enhanced speech signal. For this set of experiments, we used 250 utterances for training and another 50 utterances for testing. All utterances were recorded at 16,000 Hz and then truncated to speech segments, with each segment containing 36,500 sample points (around 2.28 seconds).

It is worth noting that although both IEM- and DM-SE tasks are multichannel SE scenarios, there are clear differences between them. For the IEM-SE task, the high-frequency components of the IEM speech signals are suppressed. In other words, the IEM speech resembles the low-pass-filtered ACM speech. Meanwhile, for the DM-SE task, the speech signals recorded by microphones I, II, III, IV, and V were degraded versions of the speech recorded by the near-field microphone owing to lower-quality recording hardware, long-range fading, and room reverberation. As with the IEM-SE task, we tested the performance of rSDFCN and DDAE.

Table IV show the average STOI and PESQ scores of rSDFCN and DDAE under seven conditions. The scores of the speech recorded by the far-field microphone (using the corresponding speech recorded by the near-field microphone as a reference) are also listed for comparison. From the tables, we can easily

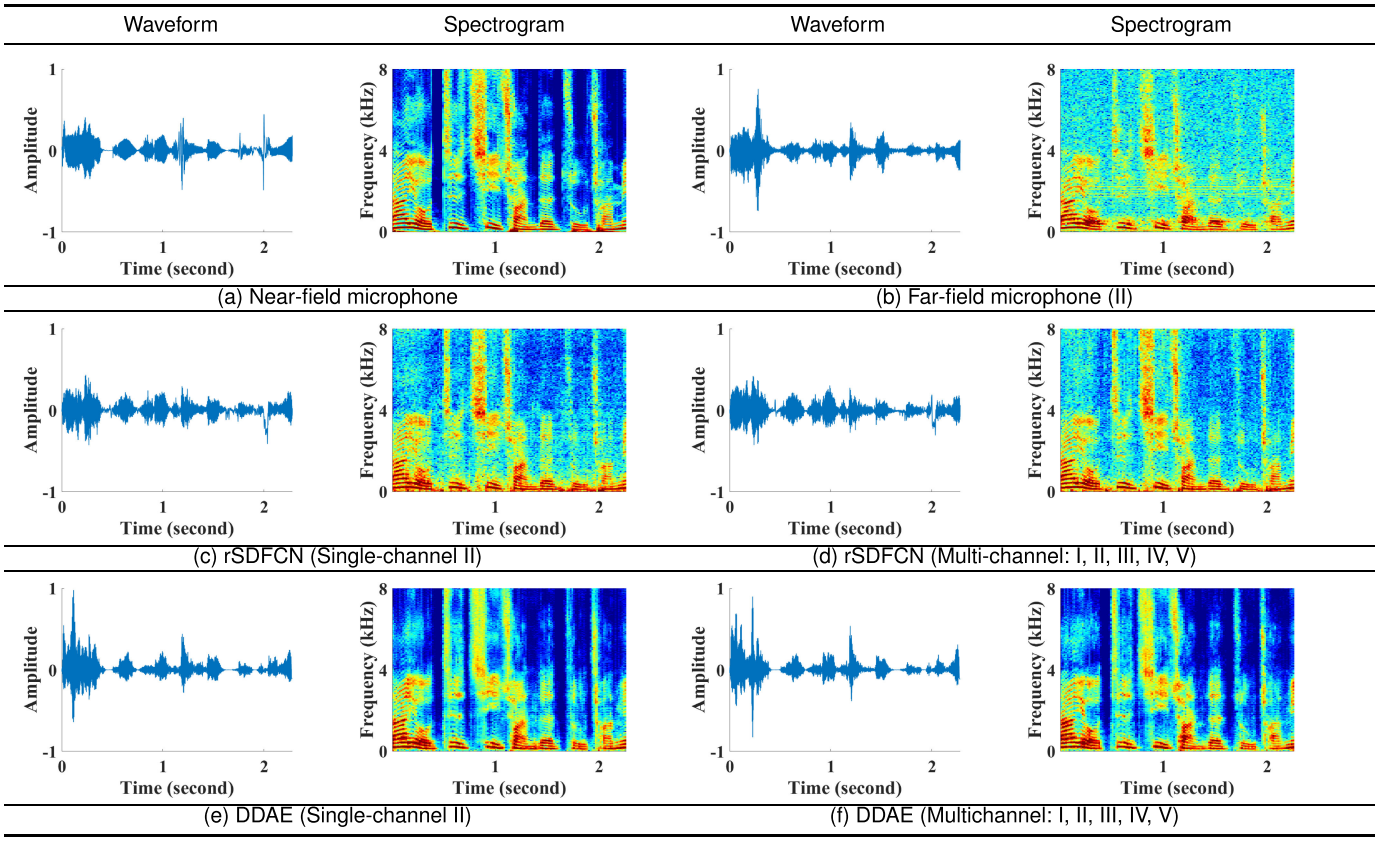

Fig. 16. Waveforms and spectrograms of an example utterance in the DM-SE task: (a) speech recorded by near-field microphone; (b) speech recorded by second far-field microphone (channel II); (c) and (e) enhanced speech by rSDFCN and DDAE with single-channel in-put; (d) and (f) enhanced speech by rSDFCN and DDAE with five channels of input.

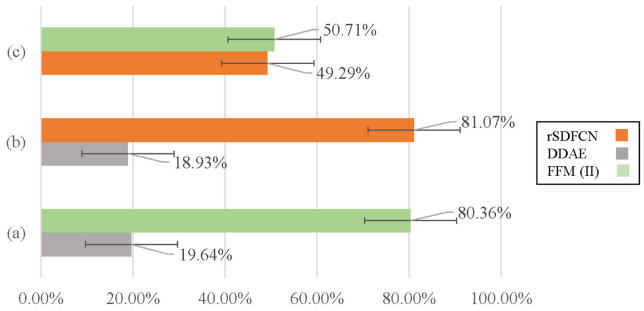

Fig. 17. Results of the AB preference test (with 95% confidence intervals) on speech quality compared between rSDFCN, FFM(II) and DDAE for the DM-SE task.

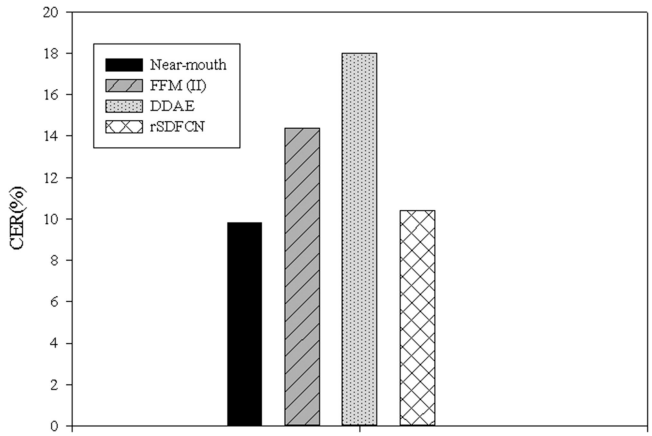

Fig. 18. ASR results achieved by different SE models for DM-SE task.

see that rSDFCN can improve the STOI and PESQ scores when multichannel inputs are used. When only one input is available (the task becomes a single-channel SE task), rSDFCN outperforms DDAE consistently across all of the five cases (single far-field microphone I, II, III, IV, and V). Meanwhile, for the multichannel task (I, II, V and I, II, III, IV, V), rSDFCN also outperforms DDAE. In addition, it is clear that the results of multichannel SE are superior to those of single-channel SE, implying that multichannel signals can provide useful information to more effectively enhance speech signals.

For qualitative analysis, the waveforms and spectrograms of a speech utterance recorded by the near-field microphone and the far-field microphone (channel II), along with the enhanced speech from rSDFCN and DDAE are shown in Fig. 16. For multichannel SE, we display the waveforms and spectrograms of the enhanced speech using five channels (I, II, III, IV, and V). From Fig. 16(d) and (f), we can observe that DDAE provided a relatively clear structure of restored spectrogram, and rSDFCN outperformed DDAE when comparing the waveform plots in contrast. This result is reasonable because DDAE aims to minimize the MSE of spectral magnitude, while rSDFCN aims to minimize the MSE in the waveform domain. Because

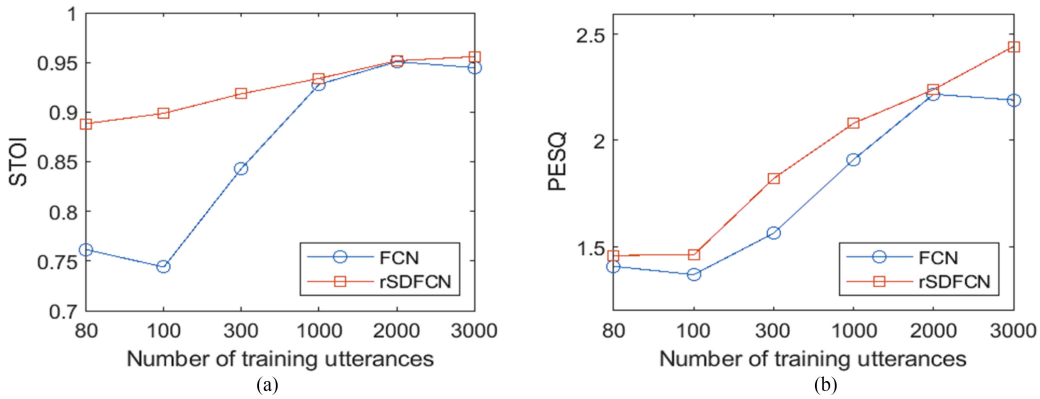

Fig. 19. Average (a) STOI and (b) PESQ scores of FCN and rSDFCN with different numbers of training utterances in the CHiME-3 dataset.

DDAE only enhances the magnitude spectrogram but not the phase information, it needs to borrow the phase information from the noisy speech when generating the speech waveforms. This may explain why DDAE performed worse than rSDFCN in terms of STOI and PESQ, as shown in Table IV, even though the spectrograms generated by DDAE were more similar to the ground-truth. This result is also consistent with those reported in previous works [70]–[72].

We also conducted listening tests on the enhanced speeches by rSDFCN, DDAE and the recorded speech by the second far-field microphone, termed FFM(II) (the channel II in Fig. 15, which achieved the highest PESQ score, as shown in Table IV). The results are shown in Fig. 17. From the figure, we note that DDAE cannot improve the speech quality effectively. A possible reason is that the distortions caused by distance does not affect the speech quality too much. Thus, although the DDAE approach can recover missing speech signal components, it may generate distortions and accordingly deteriorate the speech quality. In the meanwhile, we note that the rSDFCN can yield higher speech quality scores than the DDAE, confirming that rSDFCN is superior to DDAE in terms of subjective listening evaluations. Finally, we note that the rSDFCN enhanced speech and the one recorded by the second far-field microphone give comparable listening preference scores (50.71% versus 49.29%).

The recognition results using Google ASR are shown in Fig. 18. We report the performance of the speech recorded by the near-field microphone (as the upper-bound) and the second far-field microphone, namely, FFM(II) (channel II in Fig. 15, which achieved the best ASR results in our experiments) and the enhanced speech by DDAE and rSDFCN; the corresponding CERs are 9.8%, 14.4%, 18.0%, and 10.4%, respectively. From the CERs in Fig. 18, we first note a clear drop in ASR performance from near-field microphone speech to far-field microphone speech. Next, we note that the CER of the rSDFCN enhanced speech (10.4%) is much lower than that of the far-field microphone speech (14.0%) and close to that of the near-field microphone speech (9.8%). More specifically, the rSDFCN multichannel SE system reduced the CER by 27.8% (from 14.4% to 10.4%) compared to the unenhanced single-channel far-field microphone speech. Comparing the results in Figs. 17 and 18 and Table IV, we note that rSDFCN outperforms DDAE in terms of PESQ, STOI, subjective preference test scores, and ASR performance, confirming the effectiveness of the proposed rSDFCN over the conventional DDAE approach for the DM-SE task.

D. Speech Enhancement on the CHiME-3 Dataset

To further validate the effectiveness of using multichannel inputs for SE, we also tested our rSDFCN system on the CHiME-3 dataset [65]. As documented, the clean (reference) speech in the CHiME-3 training set was directly copied from the WSJ0 corpus [66], while the reference speech in the CHiME-3 testing set was generated from the booth recording. In this study, we directly used the clean speech as the reference to compute the STOI and PESQ scores of the enhanced speech. We tested our rSDFCN system on the simulated speech data of the CHiME-3 dataset. The simulated data is built by mixing clean speeches of the Wall Street Journal (WSJ0) corpus with four different real background noises: bus (**BUS**), cafeteria (**CAF**), pedestrian zone (**PED**) and street (**STR**). All the clean speeches and the noises are recorded by a 6-microphone array on a tablet. The total simulated set contains 7138 utterances, including 1728 of **BUS**, 1794 of **CAF**, 1765 of **PED**, and 1851 of **STR**. The goal is to use recorded six-channel noisy speeches as the input to generate enhanced speech. In our experiments, we trimmed all utterances to speech segments, each containing 36,500 sample points (around 2.28 seconds). Because the CHiME-3 dataset is far larger than the two datasets used in previous experiments, we also conducted experiments to explore the enhancement performance with respect to different numbers of training utterances. Note that in this experiment, we trained our model on utterances of **PED**, **STR** and **CAF**, and tested them on **BUS** because **BUS** was the most difficult for rSDFCN to archieve improvements over DDAE and FCN in our preliminary experiments. Fig. 19 shows the STOI and PESQ scores of FCN and rSDFCN with respect to different numbers of training utterances. From Fig. 19(a), we can see that rSDFCN, which contains the dilated and Sinc convolutional layers, achieves much higher STOI scores than FCN when the number of training utterances is limited. This implies that the benefits of the dilated and Sinc convolutional layers are more significant when the training set is small. Similar trends were also observed for the PESQ scores, as shown in Fig. 19(b).

TABLE V
AVERAGE STOI/PESQ SCORES OF RSDFCN AND DDAE EVALAUATED ON THE CHIME-3 DATASET, WHERE I, II, III, IV, V, AND VI DENOTE THE SINGLE CHANNEL RESULTS, CORRESPONDING TO PERFORMING SE USING CHANNEL 1, 2, 3, 4, 5, AND 6, RESPECTIVELY, AS SHOWN IN FIG. 1 OF THE CHIME-3 PAPER [65]; I-VI DENOTES THE RESULTS OF USING MULTICHANNEL (SIX-CHANNEL) INPUTS

AVG. STOI/PESQ	STOI			PESQ		
Input Microphone(s)	Unenhanced	DDAE	rSDFCN	Unenhanced	DDAE	rSDFCN
I	0.847	0.825	**0.878**	1.208	1.478	**1.592**
II	0.863	0.826	**0.880**	1.244	1.465	**1.604**
III	0.844	0.828	**0.880**	1.198	1.466	**1.620**
IV	0.884	0.824	**0.879**	1.308	1.462	**1.614**
V	0.893	0.827	**0.876**	1.337	1.469	**1.598**
VI	0.833	0.825	**0.880**	1.185	1.466	**1.628**
I– VI		0.853	**0.937**		1.621	**2.145**

Next, Table V shows the average STOI and PESQ scores of rSDFCN and DDAE with single-channel and multichannel inputs. Since there are four types of the background noises, we set utterances with one type of noise as the test set and use all utterances with the other three types of noises as the training sets in turn. This leave-one-out training and testing procedure repeated four times, and the average STOI and PESQ scores from the four sets of results were reported in Table V. Similar to the trends in the previous two datasets, Table V shows that the scores of multichannel-based rSDFCN are much higher than those of DDAE and single-channel-based rSDFCN.

V. CONCLUSION

In this paper, we proposed the SDFCN waveform-mapping-based multichannel SE system and an extended version, rSDFCN. We tested the proposed SE systems on three multichannel SE tasks: IEM-SE, DM-SE and CHiME-3. The experimental results for the three tasks confirmed the effectiveness of the proposed systems in achieving higher STOI and PESQ scores, as well as providing higher subjective listening scores and improved ASR performance. Meanwhile, the proposed waveform-based rSDFCN SE system outperformed the spectral-mapping-based DDAE SE system, which confirms that phase information is important for multichannel SE.

To the best of our knowledge, this study is one of the first works that adopt the concept of waveform mapping based on neural network models to enhance multichannel speech signals. In this work, both IEM-SE and DM-SE tasks simulated a "virtual" high-performance and near-field microphone to overcome the distortion caused by channel effects and spatial fading, and to attain improved speech quality (PESQ), speech intelligibility (STOI), subjective listening scores, and ASR performance. The proposed system also shows promising performance on the CHiME-3 dataset. Please note that different from the beamforming methods that require spatial and time-delay information, this study investigates the scenario where the speech signals are recorded by multiple microphones simultaneously. In the future, we will extend the proposed systems to multichannel tasks where multiple distortion factors including noise, interference, and reverberation are involved. Meanwhile, we will explore the possibility of combining the advantages of waveform-mapping and spectral-mapping-based multichannel SE methods to further improve our current systems.

REFERENCES

[1] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 663–678, 2018.

[2] J. Li *et al.*, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 3291–3301, 2011.

[3] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016, pMID: 27250154 PMCID: PMC5392064

[4] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1568–1578, Jul. 2017.

[5] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. San Diego, CA, USA: Academic Press, 2015.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[7] H. Krishnamoorthi, A. Spanias, V. Berisha, H. Kwon, and H. Thornburg, "An auditory-domain based speech enhancement algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4786–4789.

[8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] Loizou and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[11] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[12] R. Vetter, N. Virag, P. Renevey, and J.-M. Vesin, "Single channel speech enhancement using principal component analysis and mdl subspace selection," *Sixth Eur. Conf. Speech Commun. Technol.*, 1999, pp. 2411–2414.

[13] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013, *arXiv: 1709.05362*.

[14] J. Wang, Y. Lee, C. Lin, S. Wang, C. Shih, and C. Wu, "Compressive sensing-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2122–2131, Nov. 2016.

[15] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4758–4761.

[16] P. Huang, S. D. Chen, P. Smaragdis, and M. H.-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 57–60.

[17] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," *Fifteenth Ann. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 885–889.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *Interspeech*, 2013, pp. 436–440.

[19] M. Kolbaek, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[20] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," *Fifteenth Ann. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2685–2689.

[21] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan 2014.

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

[23] P. Campolucci, A. Uncini, F. Piazza, and B. D. Rao, "On-line learning algorithms for locally recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 10, no. 2, pp. 253–271, Mar. 1999.

[24] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3709–3713.

[25] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[26] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," *Interspeech*, 2016, pp. 3768–3772, [Online; accessed 2018-10-26]. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0211.html

[27] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 483–487.

[28] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Cham: Springer International Publishing, 2015, vol. 9237, pp. 91–99, dOI: 10.1007/978-3-319-22482-4_11. [Online]. Available: http://link.springer.com/10.1007/978-3-319-22482-4_11

[29] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. 16th Ann. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3274–3278.

[30] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," *Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.

[31] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multimicrophone noise reduction techniques for hands-free speech recognition-a comparative study," *Robust Methods Speech Recognit. Adverse Conditions (ROBUST-99)*, pp. 171–174, 1999.

[32] Q. Liu, B. Champagne, and P. Kabal, "Room speech dereverberation via minimum-phase and all-pass component processing of multi-microphone signals," in *Proc. IEEE Pacific Rim Conf. Commun., Comput., Signal Process.*, 1995, pp. 571–574.

[33] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[34] N. Yousefian and P. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 599–609, Feb. 2012.

[35] T. Kailath and T. J. Shan, "Adaptive beamforming for coherent signals and interference," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 3, pp. 527–536, Jun. 1985.

[36] X. Li, M. Fan, L. Liu, and W. Li, "Distributed-microphones based in-vehicle speech enhancement via sparse and low-rank spectrogram decomposition," *Speech Commun.*, vol. 98, pp. 51–62, Apr. 2018.

[37] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *25th International Conference*. Helsinki, Finland: ACM Press, 2008, pp. 1096–1103.

[38] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 116–120.

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

[40] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. - 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf., APSIPA ASC*, 2017, vol. 2018, pp. 6–12.

[41] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

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

[43] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5069–5073.

[44] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," *Interspeech*, 2017, pp. 2013–2017.

[45] E. M. Grais, H. Wierstorf, D. Ward, and M. D. Plumbley, "Multi-resolution fully convolutional neural networks for monaural audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Springer, 2018, pp. 340–350.

[46] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. 26th Eur. Signal Process. Conf.*, 2018, pp. 1577–1581.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[48] 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.

[49] 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 codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc. (Cat. no. 01CH37221)*, vol. 2, 2001, pp. 749–752.

[50] V. D. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[51] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *IEEE Spoken Lang. Tech. Workshop*, 2018, pp. 1021–1028.

[52] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency masking for speech separation," *IEEE/ACM Trans. audio, Speech, Lang. Process.*, vol. 27, no. 8, 2019, pp. 1256–1266.

[53] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," 2018, *arXiv:1802.04208*. [Online]. Available: http://arxiv.org/abs/1802.04208

[54] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[55] A. Zhang, *Speech Recognition (Version 3.8) [Software]. Available from https://github.com/Uberi/speech_recognition.*, 2017, original-date: 2014-04-23T04:53:54Z. [Online]. Available: https://github.com/Uberi/speech_recognition

[56] Y.-H. Lai *et al.*, "Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients:," *Ear Hearing*, vol. 39, no. 4, pp. 795–809, 2018.

[57] W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5454–5458.

[58] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Commun.*, vol. 104, pp. 106–112, Nov. 2018.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[60] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," *IEEE Int. Symp. Signal Process. Inf. Technol.*, 2006, pp. 426–431.
[61] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *J. Acoust. Soc. Amer.*, vol. 141, no. 3, pp. 1321–1331, 2017.
[62] L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the mandarin hearing in noise test (mhint):," *Ear Hearing*, vol. 28, pp. 70S–74S, 2007.
[63] [Online]. Available: https://www.shure.com/en-US/products/microphones/pga181
[64] "Sanlux hmt-11," [Online; accessed 2019-04-29]. [Online]. Available: http://www.sanyo.com.tw/s1504/sanyo_in_b.asp?model=2033
[65] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
[66] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.
[67] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1780–1792, 2018.
[68] S. Mittermaier, L. Kürzinger, B. Waschneck, and G. Rigoll, "Small-footprint keyword spotting on raw audio data with sinc-convolutions," 2019, *arXiv:1911.02086*.
[69] S. Gong, Z. Wang, T. Sun, Y. Zhang, C. D. Smith, L. Xu, and J. Liu, "Dilated fcn: Listening longer to hear better," *IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 254–258.
[70] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[71] J. Le Roux, "Phase-controlled sound transfer based on maximally-inconsistent spectrograms," *Signal*, vol. 5, p. 10, 2011.
[72] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, 2015.
[73] L. Chittka and A. Brockmann, "Perception space–the final frontier," *PLoS Biology*, Public Library of Science, vol. 3, no. 4, 2005.

Chang-Le Liu has been working toward the B.S. degree in electrical engineering with National Taiwan University, Taipei, Taiwan, since 2016. He was an intern as a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, and was involved in research in speech enhancement. His current research topic includes audio and image signal processing.

Sze-Wei Fu received the B.S. degree from the Department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei, Taiwan, in 2012, and the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University, in 2014. He is currently working toward the Ph.D. degree with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei. He is also a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. His research interests include speech processing, speech enhancement, machine learning, and deep learning.

You-Jin Li received the B.S. degree from the Department of Electronic Engineering, National Ilan University, Yilan, Taiwan, in 2014, and the M.S. degree from the Department of Electrical Engineering with Communications, National Ilan University, in 2016. He is currently working toward the Ph.D. degree with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwain. His research interests cover signal processing, speech enhancement, beamforming, deep learning, and multi-channel compression.

Jen-Wei Hunag received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2002 and 2009, respectively. He was a Visiting Scholar with IBM Almaden Research Center from 2008 to 2009, an Assistant Professor with Yuan Ze University from 2009 to 2012, and a Visiting Scholar with the University of Chicago in 2016. He is currently an Associate Professor with the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan. He is also the Director of Taiwanese Association of Artificial Intelligence. He was the Committee Member of IEEE CIS Member Activities committee from 2017 to 2018 and the Secretary of IEEE Tainan Section CIS Chapter from 2015 to 2017. His major research topics are data mining, machine learning, and artificial intelligence. Among these, social network analysis, spatial-temporal data mining, text mining and multimedia information retrieval are his special interests. In addition, some of his research are on FinTech and bioinformatics.

Hsin-Min Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow. He also holds a Joint Appointment as a Professor with the Department of Computer Science and Information Engineering, National Cheng Kung University. He is currently an Editorial Board Member of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *APSIPA Transactions on Signal and Information Processing*. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition. He was a General Co-Chair of ISCSLP2016 and ISCSLP2018 and a Technical Co-Chair of ISCSLP2010, O-COCOSDA2011, APSIPAASC2013, ISMIR2014, and ASRU2019. He was the recipient of the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA Distinguished Lecturer for the period of 2014–2015. He is a member of the International Speech Communication Association and ACM.

Yu Tsao (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. His research interests include speech and speaker recognition, acoustic and language modeling, audio coding, and bio-signal processing. He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *IEICE Transactions on Information and Systems* and a Distinguished Lecturer of APSIPA. He was the recipient of the Academia Sinica Career Development Award in 2017, the National Innovation Award in 2018 and 2019, and the Outstanding Elite Award, Chung Hwa Rotary Educational Foundation 2019–2020.