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§, Yu Tsao‡,
†Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
‡Research Center for Information Technology Innovation at Academia Sinica, Taipei, Taiwan
§Department of Electrical Engineering, National Cheng-Kung University, Tainan, Taiwan

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 paper, 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, 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 de-noising autoencoder (DDAE) [17], [18], deep fully connected networks [19], [20], [21], [22], recurrent neural networks [23], [24], convolutional neural networks [25], [26], and long short-term memory [27], [28], [29], [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], [32], [33] is a popular method that exploits spatial 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 two 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 single-enhanced speech signals to obtain directional features and hence the beam-formed 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 [25], [41], [42], [43], [44], a fully convolutional network (FCN) model was used to enhance on 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 SincNet) 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 both IEM-SE and DM-SE tasks and on CHiME-3 dataset, demonstrating the benefits of using the residual architecture on the overall SE performance.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 presents the concept and architectures of the proposed SDFCN and rSDFCN models. Section 4 presents the experimental setup and results. Finally, Section 5 concludes this work.

II. RELATED WORKS

Given a clean speech signal $x$, the degraded signal can be formulated as $y = g(x)$, where $g$ denotes the degradation function. The goal of SE is to find a function that maps $y$ to $\hat{x}$ and approximates $x$ 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 $\hat{x}_t$ of the output $\hat{x}$ and the last connected hidden nodes $h_t \in R^{L \times 1}$ can be represented by

\[
\begin{align*}
\hat{x}_t &= v^T h_t + b,
\end{align*}
\]

where $v \in R^{L \times 1}$ denotes a convolutional filter, $b$ is a bias term, and $L$ is the size of the filter. Note that $v$ and $b$ are shared in the convolution operation and is 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, assume that the filter function of the first layer is $v$, which will be convolved with the input signal $y$, then $v$ can be written as follows:

\[
\begin{align*}
  v &= s \circ w \\
  s_t &= 2f_{\text{low}} \text{sinc}(2\pi f_{\text{low}} t) - 2f_{\text{high}} \text{sinc}(2\pi f_{\text{high}} t) \\
  w_t &= 0.54 - 0.46\cos\left(\frac{\pi t}{L}\right)
\end{align*}
\]
where ◦ is component-wise multiplication, \( L \) is the filter length, and \( f_{\text{low}} \) and \( f_{\text{high}} \) 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 is raw speech waveform. A lower number of neurons enables SincNet to be well trained even on a limited training dataset [51].

C. Dilated Convolution

Previous works, such as WaveNet [50], Conv-TasNet [52], and WaveGAN [53] showed that using a large temporal context window is important in waveform modeling. To efficiently take advantage of the long-range dependence of speech signals, dilated convolution is 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. Both the design concept and architectures of SDFCN and rSDFCN are presented in detail.

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: \( Y = [y_1, y_2, \ldots, y_N] \), where \( N \) denotes the number of channels, we have

\[
\hat{x} = f_{\text{DFCN}}(f_{\text{SincConv}}(Y)),
\]

where \( f_{\text{SincConv}}(\cdot) \) and \( f_{\text{DFCN}}(\cdot) \) denote the mapping functions of the SincConv layer and the DFCN module, respectively. Fig. 4 shows the architecture of the dilated convolutional blocks (Dilated Conv Block in Fig. 3) in the SDFCN model. The block consists of four layers of dilated convolutional layers (the four blue rectangles) followed by batch normalization and LeakyRelu. The receptive field of the dilated
Figs. 5: Architecture of rSDFCN multichannel SE system, which consists of primary SE module (pre-trained FCN) and SDFCN system.

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 provide 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 a residual SDFCN (rSDFCN). The architecture of an rSDFCN multi-channel 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:

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

(3)

where \( f_{\text{Pr}}(\cdot) \) 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:

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

(4)

In this paper, we use a pre-trained FCN model 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 \( p_1 \).

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 \( p_1 \).

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we first introduce the experimental setup for the IEM-SE and DM-SE tasks. 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 subsets of CHiME-3 dataset with different subset size. For IEM-task and CHiME-3 dataset, 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 multi-channel 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 multi-channel DDAE system. Fig. 7 shows the architecture of the
multi-channel 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 and the $\alpha$ value of LeakyReLU 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, the 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 multi-channel SE in the waveform domain for IEM-ACM transformation.

Our recording condition is shown in Fig. 8. A male speaker sat in a sound booth (3m × 5.2m, 2m 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. All utterances were recorded at 16,000 Hz and then truncated to speech segments, each 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 the dilated convolution and SincConv layers. There were totally four models trained in this preliminary experiment: 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 block 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 SincConv layer. The reason that we changed the kernel size of the first layer was to make it have the same size as that used in 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 signal 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 original convolutional layer.

Fig. 9 shows the learning curves of the four models in terms of 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
TABLE I: Average STOI and PESQ Scores of FCN-55, DCN-54, FCN-251, SincFCN-251, and Single-channel/Multi-channel SDFCN Models for IEM-SE Task.

<table>
<thead>
<tr>
<th>No. of ch.</th>
<th>Model</th>
<th>IEM(L)</th>
<th>IEM(R)</th>
<th>FCN-55</th>
<th>DCN-54</th>
<th>FCN-251</th>
<th>SincFCN-251</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STOI</td>
<td>0.694</td>
<td>0.694</td>
<td>0.801</td>
<td>0.817</td>
<td>0.727</td>
<td>0.843</td>
</tr>
<tr>
<td>1</td>
<td>PESQ</td>
<td>1.146</td>
<td>1.101</td>
<td>1.317</td>
<td>1.360</td>
<td>1.171</td>
<td>1.476</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

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 SincConv layer. Also, from Table I, we can observe that FCN-55 outperforms FCN-251, which implies that the performance of FCN can 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 the 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 found in the previous works [51], [68], [69]. From Fig. 10, 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 phenomena is due to the limited training data, and the high frequency ranges become clearer when a sufficient amount of training data is available [41]. From Fig. 11, we can observe that the first-layered features of a testing utterance generated by SincFCN-251 contain more high frequencies than FCN-251. In addition, the results shown in Table II are similar to that in [69]: With dilated convolution, the network modeled ground-truth waveforms more accurately in terms of MSE.

Furthermore, to investigate the effectiveness of using mul-
tiple (dual) channels, we also compared the proposed 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) 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 scores over IEM (L) and IEM (R), 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 multi-channel (dual channel) 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 tested the results of replacing the SincConv layer in SDFCN with a normal convolutional layer; the results are listed 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.

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 notable distortion, with high-frequency components being suppressed. Next, by comparing Fig. 9 (a) and (d), we note that the proposed rSDFCN multi-channel 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 as shown in Table I). For comparison, the DDAE enhanced speech were 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 randomized order. For each listening test, speech samples were randomly selected from the test set, and 15 listeners participated. 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, both rSDFCN and DDAE clearly outperform IEM with notable margins, confirming the effectiveness of these two SE approaches. Next, we note that rSDFCN yield a higher preference score as 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 low-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 15.5m×11.2m and 3.27m in height. We labeled the low-quality microphones in counterclockwise order from I to V starting from the
TABLE III: Average STOI and PESQ Scores of Different Single-channel/Multi-channel SE Models for the IEM-SE Task.

<table>
<thead>
<tr>
<th>No. of ch.</th>
<th>1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>SDFTC (L)</td>
<td>SDFTC (R)</td>
<td>SDFTC</td>
</tr>
<tr>
<td>STOI</td>
<td>0.861</td>
<td>0.824</td>
<td><strong>0.880</strong></td>
</tr>
<tr>
<td>PESQ</td>
<td>1.631</td>
<td>1.597</td>
<td><strong>1.643</strong></td>
</tr>
</tbody>
</table>

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. (d) and (e) enhanced speech by rSDFCN and DDAE, respectively.

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 low-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 multi-channel 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, each containing 36,500 sample points (around 2.28 seconds).

It is worth noting that although both IEM- and DM-SE tasks are multi-channel 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 low-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 see that rSDFCN can improve the STOI and PESQ scores when multi-channel 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 multi-channel SE are superior to those of single-channel SE, implying that multi-channel signals can provide useful information to more effectively enhance speech signals.

For qualitative analysis, the waveforms and spectrograms of

![Fig. 15: Recording setting for the DM-SE task. There is a near-field high-quality microphone and five far-field low-quality microphones. Distances of near-field microphone to far-field microphones are all 1 meter.](image)
TABLE IV: Average STOI and PESQ Scores of rSDFCN and DDAE for the DM-SE Task.

<table>
<thead>
<tr>
<th>Input Microphone(s)</th>
<th>STOI</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unenhanced</td>
<td>DDAE</td>
</tr>
<tr>
<td>I</td>
<td>0.872</td>
<td>0.823</td>
</tr>
<tr>
<td>II</td>
<td>0.896</td>
<td>0.814</td>
</tr>
<tr>
<td>III</td>
<td>0.888</td>
<td>0.813</td>
</tr>
<tr>
<td>IV</td>
<td>0.884</td>
<td>0.813</td>
</tr>
<tr>
<td>V</td>
<td>0.893</td>
<td>0.816</td>
</tr>
<tr>
<td>I, II, V</td>
<td>0.823</td>
<td>0.950</td>
</tr>
<tr>
<td>I, II, III, IV, V</td>
<td>0.829</td>
<td>0.954</td>
</tr>
</tbody>
</table>

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 multi-channel 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 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 using DDAE were more similar to the ground-truth. This result is also consistent with those reported in previous works [70], [71], [72].

We also conducted listening tests to compare the proposed rSDFCN method with the DDAE and 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 did 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 multi-channel 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 Fig. 17 and 18 and Table IV, 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 DM-SE task.

D. Speech Enhancement on the CHiME-3 dataset

To further validate the effectiveness of using multi-channel inputs for SE, we also tested our rSDFCN system on the
CHiME-3 dataset [65]. As reported in the work, the clean (reference) speech in the CHiME-3 training set were directly copied from the WSJ0 corpus, while the reference speech in the CHiME-3 testing set were generated from the booth recording. It is not fair to use the booth-recorded data 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 [66] 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, 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 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, CAF and tested them on BUS because BUS was the most difficult for rSDFCN to get a significant improvement compared 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 the proposed 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.

Next, Table V shows the average STOI and PESQ scores of rSDFCN and DDAE with single-channel and multi-channel 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. As similar trends in the previous two datasets, Table V shows that the scores of multi-channel-based rSDFCN are much higher than that 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, to further improve the performance. We tested the proposed SE systems on three multichannel SE tasks: IEM-SE,
Fig. 19: Average (a) STOI and (b) PESQ scores of FCN and rSDFCN with different numbers of training utterances in the CHiME-3 dataset.

<table>
<thead>
<tr>
<th>AVG. STOI/PESQ</th>
<th>STOI</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Microphone(s)</td>
<td>Unenhanced</td>
<td>DDAE</td>
</tr>
<tr>
<td>I</td>
<td>0.847</td>
<td>0.825</td>
</tr>
<tr>
<td>II</td>
<td>0.863</td>
<td>0.826</td>
</tr>
<tr>
<td>III</td>
<td>0.844</td>
<td>0.828</td>
</tr>
<tr>
<td>IV</td>
<td>0.884</td>
<td>0.824</td>
</tr>
<tr>
<td>V</td>
<td>0.893</td>
<td>0.827</td>
</tr>
<tr>
<td>VI</td>
<td>0.833</td>
<td>0.825</td>
</tr>
<tr>
<td>I– VI</td>
<td>0.853</td>
<td>0.853</td>
</tr>
</tbody>
</table>

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 results. The proposed system also shows promising performance on the standardized 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.

ACKNOWLEDGMENT

The authors would like to thank the financial support provided by Ministry of Science and Technology, Taiwan (106-2221-E-001-017-MY2 and 107-2221-E-001-012-MY2).

REFERENCES


**Hsin-Min Wang** (S’92–M’95–SM’04) 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 in the Department of Computer Science and Information Engineering at National Cheng Kung University. He currently serves as 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 received 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 2014–2015. He is a member of the International Speech Communication Association and ACM.

**Yu Tsao** (M’09) 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. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include speech and speaker recognition, acoustic and language modeling, audio coding, and bio-signal processing. He is currently an Associate Editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and IEICE transactions on Information and Systems. Dr. Tsao received the Academia Sinica Career Development Award in 2017, National Innovation Award in 2018 and 2019, and Outstanding Elite Award, Chung Hwa Rotary Educational Foundation 2019-2020.