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

SkipConvNet: Skip Convolutional Neural Network for Speech Dereverberation using Optimally Smoothed Spectral Mapping

Vinay Kothapally[1], *Wei Xia*[1], *Shahram Ghorbani*[1], *John H.L. Hansen*[1]
Wei Xue[2], *Jing Huang*[2]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, TX, USA
[2]JD AI Research, JD.com

{vinay.kothapally, wei.xia, shahram.ghorbani,john.hansen}@utdallas.edu
{xuewei27, jing.huang}@jd.com

Abstract

The reliability of using fully convolutional networks (FCNs) has been successfully demonstrated by recent studies in many speech applications. One of the most popular variants of these FCNs is the 'U-Net', which is an encoder-decoder network with skip connections. In this study, we propose 'SkipConvNet' where we replace each skip connection with multiple convolutional modules to provide decoder with intuitive feature maps rather than encoder's output to improve the learning capacity of the network. We also propose the use of optimal smoothing of power spectral density (PSD) as a pre-processing step, which helps to further enhance the efficiency of the network. To evaluate our proposed system, we use the REVERB challenge corpus to assess the performance of various enhancement approaches under the same conditions. We focus solely on monitoring improvements in speech quality and their contribution to improving the efficiency of back-end speech systems, such as speech recognition and speaker verification, trained on only clean speech. Experimental findings show that the proposed system consistently outperforms other approaches.

Index Terms: fully convolutional networks, speech dereveberation, speech recognition, speaker verification

1. Introduction

Recent years have seen an exponential rise in the need for effective systems for distant capture of naturalistic speech [1, 2, 3] to improve the experience of human-machine interactions. These systems find their applications in many consumer devices today as personal assistance. Speech captured by these devices in confined spaces such as conference rooms, lobby, cafeteria, etc. faces two major challenges: (a) reverberation: self-distortion due to reflections which greatly reduce the intelligibility of the speech, and (b) background-noise: speech from multiple overlapping speakers, music or other acoustical sounds picked up from the environment. These two challenges are well-known and have been dealt with various signal processing and deep neural network (DNN) based speech enhancement strategies.

In earlier days, statistical signal enhancement methods played a crucial part in front-end for many speech processing pipelines [4, 5, 6, 7]. However, over the past few years, DNN based approaches for speech enhancement showed promising results in enhancing distorted speech. While DNN strategies for time-domain and frequency-domain processing [8, 9] were developed, most approaches prefer to operate on short-time fourier transform (STFT) of distorted speech, to enhance the log-power spectrum (LPS) and reuse the unaltered distorted phase signal to restore a cleaner time-domain signal. As reverberation has its effects spread over time and frequency, sequence-to-sequence learning strategies like recurrent neural networks (RNNs) and long short-term memory (LSTM) [10] have been explored to capture and leverage the temporal correlations for speech dereverberation. Besides the extreme capabilities of these networks to capture temporal correlations in speech, they fail to capture the spectral structure of formants encoded in the short-time fourier transform (STFT). Therefore, researchers have moved to convolutional neural networks (CNNs) [11] which learn the dependencies from a group of neighboring time-frequency pixels. In a conventional CNN, the spectral structure learned using 2-D convolutions is compromised due to the presence of fully connected layers. For this reason, researchers used fully convolutional networks (FCNs) which substitute the fully connected layers with 1x1 convolutions to prevent the loss of spectral structure information. In past couple of years, many FCN architectures like U-Net, ResNet, DenseNet etc. were adopted from computer vision for various speech applications [12, 13, 14, 15, 16]. Since these networks have shown significant success, exploration of various network architectures that further improve the system performance in speech related tasks has been a part of the research. In this study, we propose modifications to one such FCN architecture, U-Net, specifically designed for speech dereveberation task. We also show that using pre-processed LPS for training such networks improves the efficiency significantly.

The rest of this paper is organized as follows. Section 2 briefly introduces to the problem statement. Section-3, provides insights on the optimal smoothing proposed to be used as a preprocessing step. Section-4, describes the proposed 'SkipConvNet' for single-channel speech dereverberation. Details on the experimental setup and results are presented in Section-5. Finally, we conclude our work in Section 6.

2. Problem Formulation

For a given a room impulse response (RIR), a reverberant speech signal received by an omni-directional microphone can be modeled as:

$$x(t) = \sum_{t=0}^{L} s(t) * h(L-t) + n(t) \tag{1}$$

$$X(t,f) = S(t,f)H(t,f) + N(t,f) \tag{2}$$

where, $x(t)$ is the signal as observed by a distant microphone, $s(t)$ is the clean speech signal from the source, $h(t)$ is the room impulse response (RIR), and $n(t)$ is background additive noise. The relation in frequency domain can be represented as Eq-(2), where $X(t,f), S(t,f), H(t,f)$ and $N(t,f)$ represent the STFT of observed reverberated speech, clean speech, RIR and

http://dx.doi.org/10.21437/Interspeech.2020-2048

background noise respectively. In this study, the noise levels-$n(t)$ are considered to be lower compared to target speaker to have the analysis focused solely on evaluating and suggesting strategies to reverse the impacts of reverberation. The early and late reflections in an RIR creates a smearing effect both in time and frequency of a speech spectrogram, see Fig-1(a). The emphasis of this study is on learning a non-linear function using FCNs that maps the LPS of a reverberant speech $X(t, f)$ to a corresponding clean speech $S(t, f)$. Later, the estimated enhanced LPS from the network is combined with the unaltered reverberant phase response to reconstruct the enhanced speech.

3. Optimal Smoothing based Pre-processing

Estimation of the effects of late reflections statistics play an important role in the design of a robust speech dereverberation system. We use a minimum statistics based approach [17] originally used to estimate the noise PSD from a noisy speech, to estimate late reflection's PSD given a reverberant speech. The approach is based on the assumption that energy across all frequencies will theoretically tend to be zero during silent periods in speech, unless affected by stationary-noise or reverberation or both. Thus a robust PSD estimate for late reflections can be modeled from a reverberant speech by monitoring the minimum energy in every frequency bin over time, particularly during silent speech regions. Although this approach is focused on estimating PSD of noise (in our case reverberation), we only use only the smoothed speech LPS computed in the process for training our proposed system.

$$P(t, f) = \alpha_{opt}(t, f)P(t-1, f) + (1 - \alpha_{opt}(t, f))|X(t, f)|^2, \quad (3)$$

$$\alpha_{opt}(t, f) = \frac{1}{1+[P(t-1,f)/\sigma_n^2(t,f)-1]^2}. \quad (4)$$

In general, a fixed smoothing parameter 'α' is used in speech applications to obtain a robust PSD, $P(t, f)$. It is known that having a fixed value often comes with trade-off issues in estimation. A sliding window smoothing mechanism is robust for a higher value of 'α' but blurs the speech activity and silence boundaries. On the contrary, abrupt changes in speech activity can be recorded with a lower value of 'α' at the cost of a less reliable estimate of PSD. To address this issue, a time-varying and frequency-dependent smoothing parameter is used to get an accurate estimate of speech PSD as shown in Eq-(3). The optimal smoothing parameter is computed using Eq-(4), where $\sigma_n^2(t, f)$ is the variance of noise. We suggest referring [17] for a detailed derivation of this optimal smoothing parameter. In addition, the smooth PSD values below '-80 dB' for all training samples are clipped to have a constant dynamic range, see Fig-1.

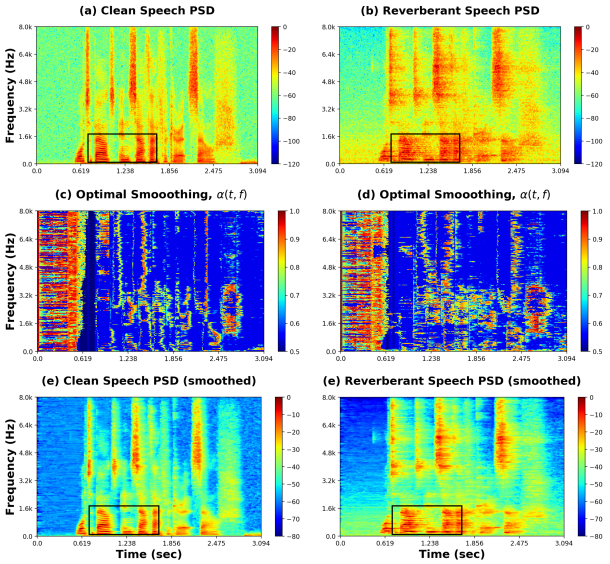

Figure 1: *Optimal Smoothing as Preprocessing*

The original and smoothed versions of LPS for a reverberant and a corresponding clean speech utterances are shown in Fig-1. The optimal smoothing parameter adapts itself accordingly for active and silent (late reflections in our case) regions of speech. It is clear from the highlighted regions in Fig-1 that optimal smoothing helps retrieve the lost formant structure in reverberant speech. Thus, we propose to use this smoothing strategy to pre-process the reverberant speech before being fed to the proposed system for training purposes.

4. Skip Convolutional Neural Network

In this section, we start with a formal description of standard U-Net and then explain the modifications we propose to make it a 'SkipConvNet'. U-Net is an encoder-decoder based image-2-image translation network. The encoder stage in the architecture extracts spectral and temporal features from the LPS of the input reverberant speech, and the decoder constructs an enhanced LPS from the encoded features. The encoder and decoder networks constitute of multiple layers of convolutions followed by down-sampling and up-sampling, respectively. With increasing numbers of layers in the network, each neuron's re-

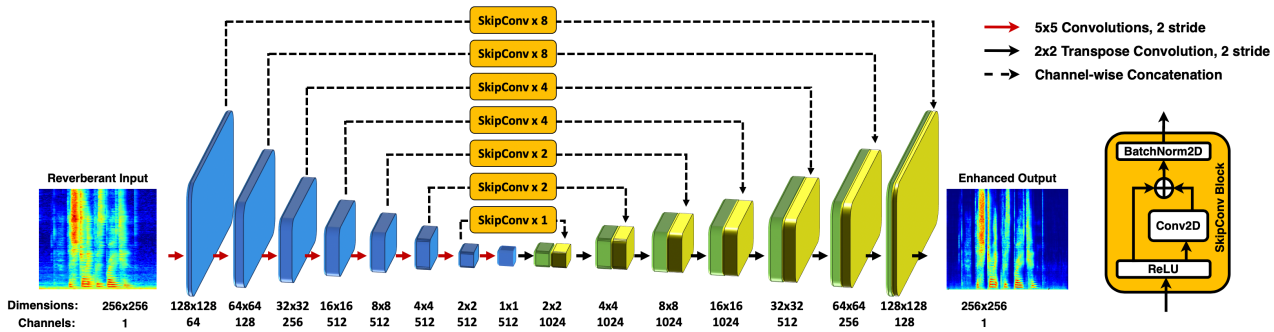

Figure 2: *SkipConvNet: UNet with convolution modules in skipped connections trained on optimally smoothed PSDs*

ception field increases. Alternatively, the dimension of the processed encoded features decreases. Similar to [18], we use enough layers in the architecture such that the encoder down-samples a given input to a single pixel. This ensures that the decoder uses all spectral and temporal features learned by the encoder from the input to construct an enhanced output. The skip connections play the most crucial role in the U-Net architecture. A skip connection is a link between the encoder and decoder used to share learned features, represented by dotted lines in Fig-2. A skip connection in each layer concatenates the output from encoder before down-sampling with the decoder in a corresponding layer, with the assumption that both input and the output PSD have a similar structure. These concatenated features from the encoder and the previous layers of the decoder not only helps to preserve the information from being lost during down-sampling in the encoder, but also guides the decoder towards the reconstruction of the enhanced output.

Although the skip connections have proven to be efficient in building a robust system, a recent study by [19] analyzes a probable semantic gap in features exchanged between encoder and decoder. For instance, the first layer of the encoder extracts low-level local spectral and temporal features. These features are concatenated with the final layer of the decoder which receives highly processed features from its previous layers. Merging these two incompatible sets of features might limit the learning abilities of the FCNs. Addition of a few convolutional layers within each skip connection can compensate for the incompatibilities by transforming the features from the encoder to be more intuitive to the decoder. We believe that, with the minimized differences within the features at each layer of the decoder, the learning ability of FCN's can potentially be maximized.

Unlike [19], which uses convolutions with varying kernels in parallel, we use standard convolutions followed by normalization and non-linear activation for the encoder and the decoder respectively. However, we use multiple '*skipconv*' blocks in series for each skip connection in the architecture. A *skipconv* block constitutes of a non-linear activation followed by a 5x5 convolution with a residual connection and does not alter the dimensions of the features set. The features are then normalized before being shared with either another *skipconv* block or the decoder. The number of *skipconv* blocks used in a particular skip connection is made to vary inversely with the depth of the layer in the encoder it is associated with, see Fig-2. For instance, skip connection associated with the final layer of the encoder has only one *skipconv* block whereas the first layer in the encoder has a total of eight *skipconv* blocks. This is based on the assumption that the deeper layers of the network deal with high-level information and require minimal transformation to reduce the semantic gap within feature sets at the decoder.

5. Experimental Results

All experiments were run on the REVERB challenge corpus [20, 21]. The outcomes of SkipConvNet are compared with the outcomes of a standard U-Net [18] trained on LPS of reverberant and clean speech utterances with and without the proposed pre-processing and a widely used statistical dereverberation algorithm, Weighted prediction error (WPE) [22, 23]. We also test the enhanced speech utterances with back-end speech automatic speech recognition (ASR) and speaker verification (SV) models trained using only clean speech.

5.1. Dataset

The Reverb Challenge corpus is a collection of simulated and real recordings of speech sampled at 16kHz in different rooms with varying levels of reverberation and a background noise at 20dB SNR. The simulated data is generated by convolving clean speech utterances from WSJCAM0 [24] and room impulse responses (RIRs) collected from three different rooms (small, medium and large sizes) and two different microphone placements (near, far) using a single microphone, 2-channel and an 8-channel microphone arrays. The multi-channel real recordings were drawn from MC-WSJ-AV corpus [25]. The corpus is divided into *train*, *dev* and *eval* sets. The *train* set consists of 7,861 simulated utterances which are used to train the systems being compared in this study. However, the *dev* and *eval* set contain both simulated and real recordings.

5.2. Pre-Processing and Network architecture

We compute STFT with a frame length of 512 samples and an overlap of 384 samples for a given speech utterance. We then compute LPS of a speech signal from an optimally smoothed LPS of the speech using Eq-3 & 4. We only consider the lower half of the , since the STFT is symmetric. Later, the LPS of each utterance is divided into batches with 256 consecutive frames to form spectral images of size 256x256. We re-use the U-Net architecture proposed in [18] as our baseline. Fig-2 gives an overview of the proposed 'SkipConvNet' architecture with 'SkipConv block(s)' replacing all skip connections. All convolutions in the encoder and skipconv blocks use a kernel size of 5x5 with a stride of 2. Similarly, all convolutions in the decoder use transposed convolution with a kernel size and stride of 2. We train our network using a total of 62,888 spectral images (corresponding to 7,861 utterances) from *train* set of the corpus with Adam optimizer.

The network is trained to minimize the mean square error (MSE) between the estimated and corresponding clean LPS. We use a batch-size of 8 and train the network for 10 epochs. Finally, estimated LPS from the network is combined with the unaltered noisy phase to reconstruct the enhanced speech. We report the improvements seen on the *eval* set of the corpus.

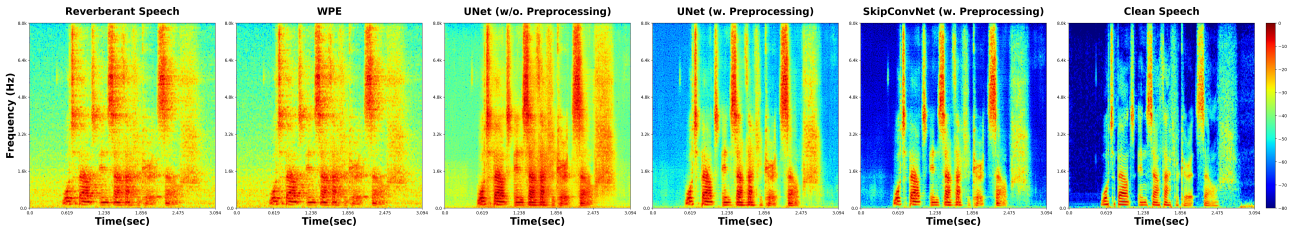

Figure 3: *Reverberant, Enhanced and Clean Speech Spectrograms*

Table 1: *Improvements in Speech Quality Measures for SimData and RealData*

	Simulated												Real
	CD			LLR			FWSegSNR			SRMR			SRMR
Room	#1	#2	#3	#1	#2	#3	#1	#2	#3	#1	#2	#3	
Far Microphone													
Reverb	2.65	5.08	4.82	0.38	0.77	0.85	6.75	0.53	0.14	4.63	2.94	2.76	3.51
WPE	2.42	5.04	4.76	0.35	0.79	0.83	7.34	0.80	0.34	4.87	3.15	2.99	3.85
UNet	2.18	3.65	3.42	0.26	0.60	0.56	7.19	2.16	2.13	4.54	3.49	3.30	4.40
UNet+Pre-Processing	2.28	3.27	3.07	0.25	0.49	0.48	10.47	7.73	6.93	4.84	4.42	4.08	5.51
SkipConvNet	**2.12**	**3.06**	**2.82**	**0.22**	**0.46**	**0.46**	**11.80**	**8.88**	**8.16**	**5.10**	**4.76**	**4.25**	**6.87**
Near Microphone													
Reverb	1.96	4.58	4.20	0.34	0.51	0.65	8.10	3.07	2.32	4.37	3.67	3.66	4.05
WPE	1.82	4.53	4.12	0.33	0.52	0.62	8.66	3.44	2.69	4.51	3.89	3.92	4.42
UNet	2.14	3.06	2.79	0.28	0.39	0.38	7.31	4.41	5.24	4.34	3.91	4.06	4.68
UNet+Pre-Processing	2.05	2.82	2.71	0.22	0.35	0.36	11.68	9.62	8.87	4.67	4.50	4.40	5.87
SkipConvNet	**1.86**	**2.57**	**2.45**	**0.19**	**0.30**	**0.35**	**13.07**	**10.96**	**10.22**	**4.99**	**4.75**	**4.56**	**7.27**

Table 2: *SV and ASR performance on simulated and real recordings for models trained on clean speech*

	Speaker Verification						Speech Recognition					
	EER (%) X-vector PLDA-trained on Clean Speech						WER (%) Acoustic Model-trained on Clean Speech					
	SimData			RealData			SimData			RealData		
Method	1ch	2ch	8ch	1ch	2ch	8ch	1ch	2ch	8ch	1ch	2ch	8ch
Reverb	8.21	-	-	6.14	-	-	34.92	-	-	93.52	-	-
WPE	8.51	5.88	1.72	6.14	5.26	4.09	30.32	18.76	**5.79**	90.19	78.45	53.78
UNet+Pre-Processing	2.63	2.01	1.86	5.85	5.56	4.39	10.67	9.26	7.19	48.33	49.12	46.79
SkipConvNet	**2.20**	**2.01**	**1.48**	**5.26**	**3.80**	**3.51**	**8.99**	**7.54**	5.81	**35.73**	**34.22**	**30.99**

5.3. Results

We begin our presentation of experimental results with a second look at the optimal smoothed PSD's from Fig-1. From Fig-1(a),(c) & (e), we see that optimal smoothing helps in preserving the formant structure during the speech frames by having a low smoothing parameter while assigning the regions with reverberant contents with a higher smoothing parameter.

We then measure the relative enhancement achieved by each system using several speech quality measures, as shown in Table-1. A FCN based U-Net and the proposed 'SkipConvNet' performed consistently better compared to the widely used statistical dereverberation algorithm, WPE. However, we observed a 39.19% relative improvement in the performance of the baseline U-Net by solely introducing the proposed pre-processing. This shows that the proposed pre-processing helps all FCN networks and is not biased to the proposed 'SkipConvNet'. However, the proposed 'SkipConvNet' consistently performed the best compared to U-Net and the U-Net trained on pre-processed inputs with an average of 54.45% and 10.40% relative improvements over all quality metrics respectively. Consistent improvements in 'SRMR' and 'FWSegSNR' in addition to 'CD' ensure the reduction of reverberation and background noise in enhanced speech utterances without any processing artifacts/distortions.

Finally, we test the improvements in back-end automatic speech recognition (ASR) and speaker verification systems (SV) achieved with proposed system for single and multi-channel streams, see Table-2. For multi-channel streams of data, individual channels are enhanced with different dereveberation techniques discussed in the study and then spatially combined using BeamformIT [26, 27] beamforming strategy. Since the proposed pre-processing enhanced the performance of traditional U-Net, we compare the proposed system's achievements with only WPE and U-Net trained on pre-processed spectral images. For an ASR system, we use TDNN based acoustic model [28] trained on single channel clean speech of the REVERB CHALLENGE corpus. Similarly, for a speaker verification system, we train a X-vector model [29] on Voxceleb 1 & 2 corpus [30, 31] and a PLDA backend on the in-domain single-channel clean speech of the corpus. Three utterances from each speaker from both simulated and real recordings of *eval* are considered in the enrollment set and the rest in the evaluation set. We see a relative improvement of 35.03% and 16.42% in speaker verification performance using X-vectors averaged over simulated and real recordings compared to WPE and U-Net trained on pre-processed spectral images. Similarly, we see a 48.15% & 23.94% relative improvements in the performance of an automated speech recognition (ASR) system averaged over simulated and real recordings compared to WPE and U-Net trained on pre-processed spectral images. For interested readers, check out: `https://vkothapally.github.io/SkipConv/`

6. Conclusions

In this study, we presented 'SkipConvNet' an encoder-decoder based FCN with convolutional modules introduced in skip connections which enhanced the learning ability to map reverberant speech to its corresponding clean speech. We have proposed the use of optimal smoothing of PSD as a preprocessing step for training the network which has shown considerable improvements in the network's performance. With the proposed modifications, we achieved significant improvements in speech quality for both real and simulated data from REVERB CHALLENGE corpus in comparison with traditional U-Net and also widely-used WPE dereverberation algorithm. We have also shown that the proposed system also improves the performance of single multi-channel back-end speech systems like speech recognition and speaker verification. To summarize, the addition of convolutions in skip connections reduces the incompatibilities within the feature sets received at each layer of the decoder which improves the learning capabilities of the network. We believe that this work can be extended to other complex-FCN architectures that have recently been researched for speech enhancement.

7. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines," 2018.

[2] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio." in *INTERSPEECH*, 2019, pp. 1851–1855.

[3] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon." in *INTERSPEECH*, 2018, pp. 2758–2762.

[4] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[5] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *IEEE Workshop on Acoustic Echo and Noise Control*. Citeseer, 1999, pp. 27–30.

[6] E. A. P. Habets, "Single and multi-microphone speech dereverberation using spectral enhancement," 2007.

[7] V. Kothapally and J. H. Hansen, "Speech detection and enhancement using single microphone for distant speech applications in reverberant environments." in *INTERSPEECH*, 2017, pp. 1948–1952.

[8] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6865–6869.

[9] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[10] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] W. Xia and K. Koishida, "Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation," in *INTERSPEECH*, 2019, pp. 3629–3633.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[18] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 390–394.

[19] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[20] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013, pp. 1–4.

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

[22] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[23] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 1995, pp. 81–84.

[25] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 357–362.

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

[27] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 346–358.

[28] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement," in *Proc. REVERB Workshop*, 2014, pp. 1–8.

[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.