Time Domain Speech Enhancement With Attentive Multi-scale Approach

Chen Chen*, Nana Hou*, Duo Ma† and Eng Siong Chng*‡
* Nanyang Technological University, Singapore
† National University of Singapore, Singapore
‡ Temasek Laboratories, Nanyang Technological University, Singapore
E-mail: CHEN1436@e.ntu.edu.sg

***Abstract*— Speech enhancement aims to suppress the additive noise from noisy speech signals to improve the speech quality. It is believed that multi-scale temporal information learned from the speech inputs strengthens the mask prediction or noise suppression especially for the encoder-mask-decoder like structure in time-domain speech enhancement techniques. In this paper, we propose a multi-scale encoding and decoding scheme that captures multiple temporal resolutions for improving speech quality. We also propose an attention module to capture the global temporal information of each-scale embedding in the encoding layer. The experiments show that the proposed approach achieves 9.0% and 2.9% relative improvements over the best baseline in terms of perceptual evaluation of the speech quality (PESQ) and signal-to-distortion ratio (SDR), respectively.**

I. Introduction

Speech enhancement improves speech intelligibility and quality by reducing noise from noise-corrupted speech signal [1]. It is widely applied as a pre-processing module in many real-world applications, such as automatic speech recognition (ASR) [2][3], teleconferencing [4], and hearing-aids [5][6].

Speech enhancement methods can be generally grouped into frequency-domain techniques [7][8][9] and time-domain techniques [10][11][12]. Frequency-domain methods usually utilize short-time Fourier transform (STFT) to convert raw audio signals into frequency-domain features. These features are then enhanced and finally reverted back to signals by the inverse short-time Fourier transform (iSTFT). Phase information is usually ignored in this processing [13]. Unlike frequency-domain approach, time-domain methods attempt to address this problem with a convolution encoder to extract spectrum-like features and a convolutional decoder to reconstruct signals from the processed features, which avoids decomposing the signals into magnitude and phase [14]. The encoding and decoding scheme is key to improving the speech quality for time-domain approaches [15].

Speech has a rich temporal structure over multiple time scales to realize the various phonemic, prosodic and linguistic content [16][17]. The speech encoder and decoder with a single window or filter length captures only one specific feature [18], this restricts the model to learn the multi-scale temporal structure from the input speech. To overcome this, previous work [19] explored STFT analyses with different window lengths to obtain multi-resolution acoustic features for the speech recognition task. Likewise, recent works [20] proposed a multi-scale octave convolution layer to learn robust representations from the input speech signal. These two works produced significant improvement over the single-resolution-based baselines on the speech recognition task. Additionally, it was shown that speech analysis of multiple temporal resolutions leads to better speech understanding [21]. The prior studies are the source of inspiration for this work.

To learn the rich temporal structure of speech, in this paper, we propose an attentive multi-scale time-domain speech enhancement approach (AMS-SE) to improve the mask prediction. Specifically, AMS-SE is composed of three network components: an attentive multi-scale speech encoder that encodes the time-domain speech signals into different scaled spectrum-like feature representation [22] with attention weights we named as attentive coefficients, a mask predictor to estimates a receptive mask for each attentive embedding coefficient, and a speech decoder that reconstructs the enhanced speech by modulating the receptive mask with the attentive embedding coefficients of the noisy inputs. Experimental results show that the proposed AMS-SE outperforms Conv-TasNet baseline [23] in terms of perceptual evaluation of the speech quality (PESQ) and signal-to-distortion ratio (SDR).

The rest of this paper is organized as follows. In section 2, we describe the proposed AMS-SE architecture. In section 3, experimental settings and results are presented. Section 4 concludes the study.

II. AMS-SE Architecture

Suppose that a noisy signal $y(t)$ is the mixture of the clean speech $s(t)$ and the background noise $n(t)$, we have

$$y(t) = s(t) + n(t), t = 1, \ldots, T \tag{1}$$

where T is the numbers of the samples of the noisy signal $y(t)$. During the inference at run-time, given a noisy signal $y(t)$, the speech enhancement approach is expected to estimated $\hat{s}(t)$ that is close to $s(t)$ subject to an optimization criterion.

A. AMS-SE network

We now introduce the proposed AMS-SE network, which consists of three modules: the attentive multi-scale speech

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-100E-2018-006) and Air Traffic Management Research Institute of Nanyang Technological University.

encoder, the mask predictor and the speech decoder, as illustrated in Figure 1.

1) The attentive multi-scale speech encoder: As shown in Figure 1(b), the proposed speech encoder includes multiple parallel 1-D convolutional layers followed by the rectified linear unit (ReLU) activation functions and the corresponding self-attention module. In this work, we utilize three parallel 1-D convolutional layers with short (l_1), middle (l_2), long (l_3) filter lengths, respectively, to extract the multi-scale embedding coefficients from the noisy waveforms $y(t)$. Specifically, the convolutional filters with long filter length can learn the course-grained feature of speech, while the convolutional filters with short filter length can capture the fine-grained features. To concatenate the embeddings across different scales, we align them by keeping the same stride, $l_1/2$, and the numbers of filters are set to N.

Then, such multi-scale embedding coefficients are fed into the self-attention module [24] to learn the attention map for each-scale embedding coefficient. As shown in Figure 2, the three 1×1 convolutional layers are first applied to transform the embedding coefficients to the latent representations, which is named Q, K and V respectively. We compute the attention map as:

$$Attention(Q,K,V) = softmax(QK^T)V \tag{2}$$

where the softmax activation function is performed along the product of Q and K^T. Such three attentive embedding coefficients are then concatenated along the channel dimension and fed into the mask predictor for receptive mask estimation.

2) The mask predictor: The TCN-based mask predictor is designed to suppress the additive noise in attentive embedding coefficients. Similar to Conv-TasNet [23], the mask predictor module consists of a temporal convolutional network (TCN) [25]. As shown in Figure 3, the concatenated attentive embedding coefficients is firstly normalized by its mean and variance on channel dimension scaled by the trainable bias and gain parameters. Then, a 1×1 convolutional layer with B filters adjusts the number of channels for the inputs as a bottleneck layer. To capture the long-range temporal information of the speech with a manageable number of parameters, dilated depth-wise convolutional layers "d-conv" are stacked in several temporal convolutional blocks (TCB) by exponentially increasing the dilation factor $[2^0,\ldots,2^{X-1}]$. In this work, we form X TCBs as a batch and repeat the batch for R times in the TCN-based mask predictor. To keep the TCN-based mask in a consistent dimension with the input features, three 1×1 convolutional layer (with N filters and 1×1 kernel size) are applied with a sigmoid activation function for ensuring that the estimated mask ranges within [0, 1]. We obtained the enhanced attentive embedding coefficients $\hat{E}_i$ for each scale $i = 1,2,3$ by applying the receptive mask M_i on the attentive embedding coefficients E_i of the noisy signal in each scale,

$$\hat{E}_i = M_i \bigotimes E_i, i = 1,2,3 \tag{3}$$

where $\bigotimes$ is an operator for element-wise multiplication. E_i is the multi-scale attentive embedding coefficients.

3) The speech decoder: The decoder reconstructs the time-domain speech signal from the enhanced attentive embedding coefficients. Attentive embedding coefficients at each scale lead to a corresponding enhanced output. We reconstruct the multi-scale attentive enhanced embedding coefficients into time-domain signals $(\hat{s}_1,\hat{s}_2,\hat{s}_3)$ with the multi-scale de-convolutional layers in the speech decoder.

B. The optimization strategy

During training, we calculate a multi-scale scale-invariant signal-to-distortion ratio (SI-SDR) [23] loss, defined as L, that aims to minimize the signal reconstruction error,

$$L = \alpha_1\rho(\hat{s}_1,s) + \alpha_2\rho(\hat{s}_2,s) + \alpha_3\rho(\hat{s}_3,s) \tag{4}$$

where s is the clean speech. $\hat{s}_1,\hat{s}_2,\hat{s}_3$ are the reconstructed time-domain signals from each-scale enhanced attentive embedding coefficient. α_1, α_2, and α_3 denote the respective weights for each reconstructed signals. We use the SI-SDR loss, denoted as $\rho()$, as the measure of reconstruction error.

$$\rho(\hat{s},s) = 10\log_{10}\left(\frac{\|\frac{\langle\hat{s},s\rangle}{\langle s,s\rangle}s\|^2}{\|\frac{\langle\hat{s},s\rangle}{\langle s,s\rangle}s-\hat{s}\|^2}\right) \tag{5}$$

where $\hat{s}$ and s are the enhanced and clean signals of the noisy inputs, respectively. $\langle,\rangle$ is the inner product. To ensure scale invariance, the signals $\hat{s}$ and s are normalized to zero-mean prior to the SI-SDR calculation.

The calculation of multi-scale SI-SDR L loss is required only during training and not at run-time inference. At run-time inference, we evaluate the quality of the signals reconstructed at multiple scales individually, i.e., $\hat{s}_1,\hat{s}_2,\hat{s}_3$, and collectively as a weighted summation as $\hat{s} = \alpha_1\hat{s}_1 + \alpha_2\hat{s}_2 + \alpha_3\hat{s}_3$.

III. Experiment

A. Database

We conduct experiments on a publicly available database [26] that is widely used in speech enhancement. Specifically, the noisy set includes 11,572 utterances from 28 speakers and is mixed by 10 different types with four SNRs levels [0 dB, 5 dB, 10 dB, 15 dB] at sampling rate of 16 kHz. The test set contains 5 type of unseen noise in SNRs range [2.5 dB, 7.5 dB, 12.5 dB, 17.5 dB]. The unseen noise represents a major source of mismatch between training and test data.

B. Experimental setup

1) Network configuration: During the training stage, the noisy waveform was cut into several segments with a duration of 1 second each for batch training. The detailed configuration is listed in Table I. To concatenate the attentive embedding coefficients, we set $l_1/2$=10 as common stride across different scales and we align three scale attentive embedding coefficients by zero-padding operation. The network was optimized by the Adam algorithm [27]. The learning rate started from 0.001 and was halved when the loss increased on the development set for at least 3 epochs. An early stopping

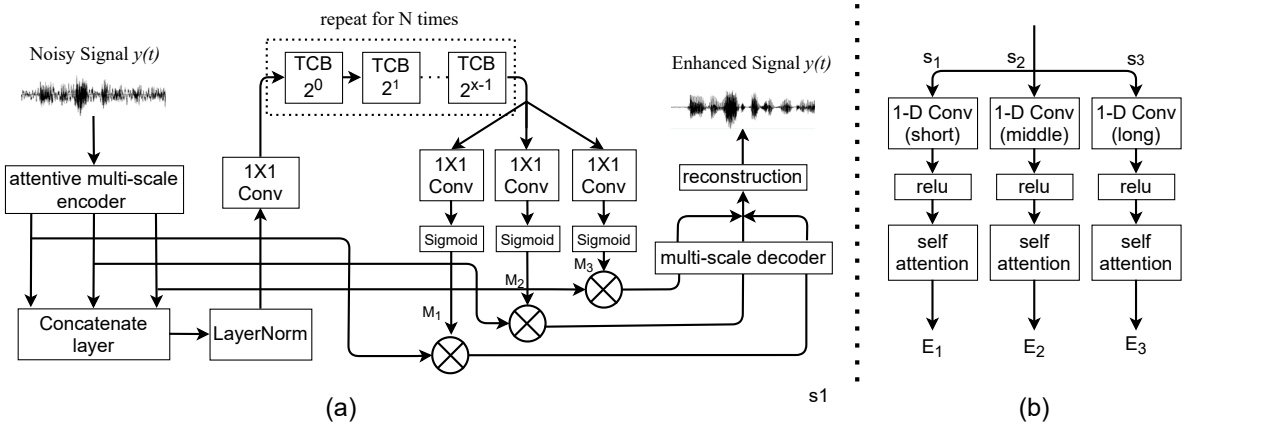

Fig. 1. The block diagram of (a) the proposed AMS-SE structure, that consists of (b) the attentive multi-scale speech encoder, the mask predictor and the decoder. "1-D Conv" is the 1-D convolutional operation and "1×1 Conv" is the convolutional operation with 1×1 convolutional filters. $\otimes$ refers to the element-wise multiplication. "relu" and "sigmoid" are the rectified linear unit (ReLU) and sigmoid functions. The structure "TCB" block is similar to Conv-TasNet as shown in Figure 3.

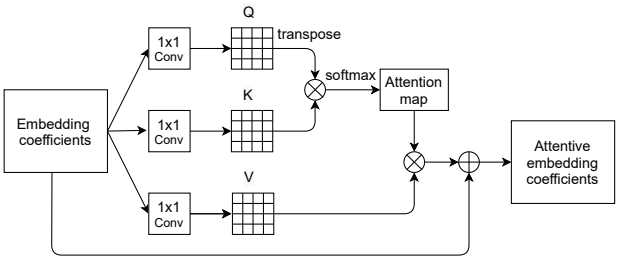

Fig. 2. Block diagram of the self-attention module. "1×1 Conv" is the convolutional operation with 1×1 convolutional filters. $\otimes$ refers to the element-wise multiplication, and $\bigoplus$ is the residual connection. "softmax" is performed on each row.

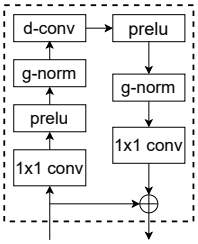

Fig. 3. Block diagram of the temporal convolutional block (TCB). "d-conv" is the dilated depth-wise convolutional layers stacked in several TCBs to exponentially increase the dilation factors. $\bigoplus$ is the residual connection.

scheme was applied as soon as the loss increased on the development set for 5 epochs.

2) Evaluation metrics: We report the performances in terms of the following metrics. PESQ [28] stands for perceptual evaluation of the speech quality, ranging from -0.5 to 4.5. Three objective metrics that approximate mean opinion scores (MOSs) [29]: CSIG, CBAK, and COVL. They are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Signal-to-distortion ratio (SDR) is also conducted for measuring speech quality. Short-time objective intelligibility (STOI) [30] reflects the improvement of speech intelligibility. Higher scores are better for all metrics.

TABLE I

THE NETWORK CONFIGURATION FOR THE PROPOSED AMS-SE ARCHITECTURE.

Symbol	Description	Numbers
l_1	short filter length of the speech encoder and decoder	20
l_2	middle filter length of the speech encoder and decoder	80
l_3	long filter length of the speech encoder and decoder	160
N	Number of filters in encoder and decoder	256
B	Number of channels in bottleneck	256
H	Number of channels in TCNs	512
P	Kernel size in TCN block	3
X	Number of TCBs in each repeat	8
R	Number of repeats	4

IV. RESULTS

A. Effect of the multi-scale speech encoder and decoder

We first analyze and summarize the performances with the proposed multi-scale speech encoder and decoder. The self-attention module is not utilized in this experiment in the Table II. We observe that the proposed AMS-SE with $\alpha_1 = 0.8, \alpha_2 = 0.1, \alpha_3 = 0.1$ achieves 3.4% and 1.4% relative improvement in terms of PESQ and SDR. We obtain the best performances of AMS-SE with $\alpha_1 = 0.4, \alpha_2 = 0.3, \alpha_3 = 0.3$.

We also observe that the parameters of AMS-SE are not increased significantly. Learning the multi-scale temporal information in the speech encoder and decoder has improved the quality of the speech.

TABLE II

PESQ, COVL, CBAK, CSIG, AND SDR(DB) IN A COMPARATIVE STUDY OF THE MULTI-SCALE SPEECH ENCODER AND DECODER. "#PARAS" DENOTES THE NUMBER OF PARAMETERS IN THE MODEL. α_1, α_2, AND α_3 DENOTE THE RESPECTIVE WEIGHTS FOR EACH RECONSTRUCTED SIGNALS.

Model	Param	$(\alpha_1, \alpha_2, \alpha_3)$	PESQ	COVL	CbAK	CSIG	SDR
Noisy	-	-	1.97	2.63	2.44	3.34	8.54
Conv-TasNet [23]	7.62M	-	2.67	3.30	3.31	3.94	19.68
AMS-SE	8.01M	(0.8,0.1,0.1)	2.76	3.40	**3.47**	**4.04**	19.95
	8.01M	(0.6,0.2,0.2)	2.77	3.40	3.41	4.01	19.97
	8.01M	(0.4,0.3,0.3)	**2.80**	**3.41**	3.45	4.03	**19.97**

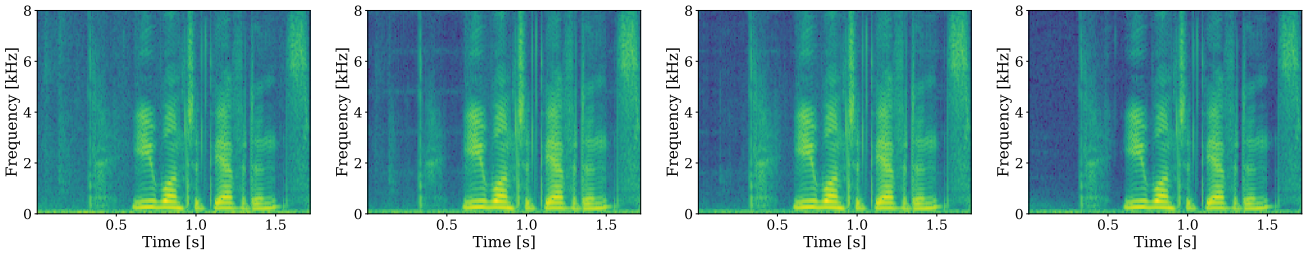

Fig. 4. The spectrograms of a set of samples (p252_396.wav) in the test set for (a) noisy input, (b) the best baseline Conv-TasNet, (c) enhanced result of AMS-SE and (d) clean signal (ground-truth).

B. Effect of the attentive multi-scale speech encoder

We report the effect of the attentive multi-scale speech encoder shown in Table III. We apply the self-attention module (SA) for the proposed AMS-SE with the different configurations of $\alpha_1, \alpha_2, \alpha_3$. We observe that the proposed self-attention module significantly improves the speech quality without approximate similar parameters of the models. We obtain the best performances with $\alpha_1 = 0.6, \alpha_2 = 0.2, \alpha_3 = 0.2$ in terms of the PESQ and SDR.

TABLE III

PESQ, COVL, CBAK, CSIG, AND SDR(DB) IN A COMPARATIVE STUDY OF THE ATTENTIVE MULTI-SCALE SPEECH ENCODER AND DECODER. "#PARAS" DENOTES THE NUMBER OF PARAMETERS IN THE MODEL. α_1, α_2, AND α_3 DENOTE THE RESPECTIVE WEIGHTS FOR RECONSTRUCTED SIGNAL. "SA" DENOTES THE SELF-ATTENTION MODULE.

Model	SA	Paras	$(\alpha_1, \alpha_2, \alpha_3)$	PESQ	COVL	CBAK	CSIG	SDR
AMS-SE	×	8.01M	(0.8,0.1,0.1)	2.76	3.40	3.47	4.04	19.95
	√	8.20M	(0.8,0.1,0.1)	2.81	3.48	3.48	4.10	20.08
	×	8.01M	(0.6,0.2,0.2)	2.77	3.40	3.41	4.01	19.97
	√	8.20M	(0.6,0.2,0.2)	**2.91**	**3.52**	3.45	4.13	**20.25**
	×	8.20M	(0.4,0.3,0.3)	2.80	3.41	3.45	4.03	19.97
	√	8.20M	(0.4,0.3,0.3)	2.86	3.51	**3.51**	**4.15**	20.12

TABLE IV

PESQ, COVL, CBAK, CSIG, AND STOI IN A COMPARATIVE STUDY OF OTHER COMPETITIVE TECHNIQUES.

Method	Domain	PESQ	STOI	COVL	CBAK	CSIG
Noisy	-	1.97	0.91	2.63	2.44	3.34
Winer	-	2.22	-	2.67	2.68	3.23
SEGAN [31]	T	2.16	0.93	2.80	2.94	3.48
MMSE-GAN [32]	F	2.53	0.93	3.14	3.12	3.80
Conv-TasNet [23]	T	2.67	0.93	3.30	3.31	3.94
AMS-SE	T	**2.91**	**0.94**	**3.52**	**3.45**	**4.13**

C. AMS-SE vs. other competitive methods

Table IV summarizes the comparison between the proposed AMS-SE and other competitive techniques in terms of PESQ, CSIG, CBAK, COVL, and STOI. We observe that the proposed AMS-SE obtained the best performances. Comparing with the best baseline Conv-TasNet method, the AMS-SE achieves the 9.0% relative improvements over the best baseline Conv-TasNet in terms of PESQ.

To further show the contribution of the AMS-SE approach, we select a set of samples from the test set, which contains noisy signal, clean signal, enhanced signal by Conv-Tasnet baseline, and enhanced signal by AMS-SE. The respective magnitude spectrograms are shown in Figure 4. We can see that the AMS-SE can produce more clear spectrum under same conditions.

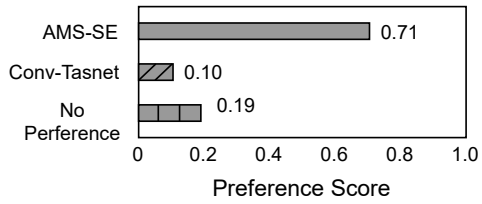

Fig. 5. The A/B preference test result of the enhanced speech between the proposed AMS-SE and the best baseline Conv-TasNet. We conducted t-test using a significance level of $p < 0.05$, which is depicted with error bars.

D. Subjective evaluation

Since the Conv-TasNet presents the best baseline performances in the objective evaluation as shown in Table IV, we only conduct an A/B preference test between the Conv-TasNet and the proposed AMS-SE to evaluate the signal quality and intelligibility by subject listening. We randomly selected 20 pairs of listening examples and invited 10 subjects to choose their preference.

The percentage of the preferences is shown in Figure 5. We observe that the listeners clearly preferred the proposed AMS-SE with a preference score of 71% to the best baseline Conv-TasNet with a preference score of 10%. Most subjects significantly preferred the enhanced waveforms by AMS-SE with a significance level of p < 0.05. Some listening examples are available at Github [1].

V. CONCLUSIONS

We propose an attentive multi-scale time-domain speech enhancement framework (AMS-SE) to learn the rich multiple temporal resolution information from the speech. Experiment results show that AMS-SE outperforms the best baseline Conv-TasNet in terms of all evaluation metrics. The proposed self-attention module also improves the speech quality by capturing the long-range dependencies of the attentive embedding coefficients. Furthermore, the subjective evaluation shows that the AMS-SE is significantly preferred over the Conv-TasNet.

[1] https://chrisole.github.io/APSIPA-2021/

REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[3] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[4] A. Wang, K. Yao, R. E. Hudson, D. Korompis, F. Lorenzellii, S. Soli, and S. Gao, "Microphone array for hearing aid and speech enhancement applications," in *Proceedings of International Conference on Application Specific Systems, Architectures and Processors: ASAP'96*. IEEE, 1996, pp. 231–239.

[5] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[6] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "Tinylstms: Efficient neural speech enhancement for hearing aids," *arXiv preprint arXiv:2005.11138*, 2020.

[7] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "A comparative study of time and frequency domain approaches to deep learning based speech enhancement," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[8] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6628–6632.

[9] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1246–1251.

[10] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain." in *Interspeech*, 2018, pp. 1136–1140.

[11] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7098–7102.

[12] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[13] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9458–9465.

[14] 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.

[15] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder." in *INTERSPEECH*, 2019, pp. 86–90.

[16] B. Zellner, "Pauses and the temporal structure of speech," in *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.* John Wiley, 1994, pp. 41–62.

[17] X. Teng, X. Tian, and D. Poeppel, "Testing multi-scale processing in the auditory system," *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.

[18] M. S. Lewicki, "Efficient coding of natural sounds," *Nature neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[19] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on timit," *PloS one*, vol. 13, no. 10, p. e0205355, 2018.

[20] J. Rownicka, P. Bell, and S. Renals, "Multi-scale octave convolutions for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7019–7023.

[21] S. Greenberg, "A multi-tier framework for understanding spoken language," in *Listening to Speech*. Psychology Press, 2012, pp. 411–433.

[22] 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.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[25] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[26] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, 2016, pp. 352–356.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

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

[31] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[32] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel mmse discogan for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop, Google Office, Hyderabad, India*, 2018, pp. 1–3.