ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746610

MULTI-SCALE TEMPORAL FREQUENCY CONVOLUTIONAL NETWORK WITH AXIAL ATTENTION FOR SPEECH ENHANCEMENT

Guochang Zhang, Libiao Yu, Chunliang Wang, Jianqiang Wei

Department of Speech Technology, Baidu Inc, Beijing, 100085, China

ABSTRACT

Speech quality is often degraded by acoustic echoes, background noise, and reverberation. In this paper, we propose a system consisting of deep learning and signal processing to simultaneously suppress echoes, noise, and reverberation. For the deep learning, we design a novel speech dense-prediction backbone. For the signal processing, a linear acoustic echo canceller is used as conditional information for deep learning. To improve the performance of the speech dense-prediction backbone, strategies such as a microphone and reference phase encoder, multi-scale time-frequency processing, and streaming axial attention are designed. The proposed system ranked first in both AEC and DNS Challenge (non-personal track) of ICASSP 2022. In addition, this backbone has also been extended to the multi-channel speech enhancement task, and placed second in ICASSP 2022 L3DAS22 Challenge[1].

Index Terms— speech dense-prediction, speech enhancement, multi-scale, axial attention

1. INTRODUCTION

In speech communication applications, such as voice interactions or video conferencing systems, speech quality is often degraded by acoustic echoes, background noise, and reverberation. To suppress the acoustic echo, an audio-processing component called a linear acoustic echo canceller (LAEC) [1] can be employed.

However, due to the presence of nonlinear distortions and vibration effects of the speaker, the performance of LAEC severely degrades. For this reason, a residual echo suppressor (RES) based on signal processing (SP) [2] or deep neural network (DNN) is usually required to further suppress acoustic echo. The DNN based RES can achieve better performance than the SP based method [3]. In addition, DNNs have also achieved remarkable results in removing background noise and suppressing reverberation [4].

In this work, we present a system for simultaneous denoising, dereverberation, and echo cancellation. The system is a combination of SP and DNN. The SP part consists of a simple time delay compensator (TDC) based on generalized correlation [5] and a LAEC based on the two-echo-path model [6] with PNLMS adaptive filter [7]. For the DNN part, we propose a novel backbone for speech dense-prediction called multi-scale temporal frequency convolutional network with axial self-attention (MTFAA-Net). In this work, our contributions include:

- For the echo cancellation, we design a novel combination of SP and DNN. Unlike the previous concatenation of LAEC and DNN, we only use LAEC as conditional information for DNN, which avoids introducing the distortion caused by LAEC into the estimated target speech.
- A backbone for various speech dense-prediction tasks is proposed. The Phase encoder (PE), multi-scale temporal frequency processing, and streaming axial self-attention (ASA) are designed to improve the performance of the backbone. The frequency band merging module according to equivalent rectangular bandwidth (ERB) is applied after PE to process full-band signals with low computational complexity.

The results on the evaluation set and blind test sets of ICASSP 2022 AEC Challenge [8] and ICASSP 2022 DNS Challenge [9] show that the proposed scheme achieves impressive performance on echo cancellation, denoising, and dereverberation.

The rest of this paper is organized as follows. Section 2 presents problem formulation. Section 3 provides details of the proposed backbone for speech enhancement. Section 4 shows datasets and experimental results. Finally, we draw conclusions in Section 5.

2. PROBLEM FORMULATION

Let us consider the signal models in the short time Fourier transformation (STFT) domain. The microphone signal $Y(t,f)$ is composed of echo $E(t,f)$, background noise $N(t,f)$, and near-end speech with reverb $s(t,f)H^e(f) + s(t,f)H^l(f)$. We denote this model as:

$$Y(t,f) = s(t,f)H^e(f) + s(t,f)H^l(f) + E(t,f) + N(t,f), \quad (1)$$

where $s(t,f)H^e(f)$, $s(t,f)H^l(f)$ are the near-end speech which is convolved with the early part of the room impulse

[1] https://www.l3das.com/icassp2022/results.html

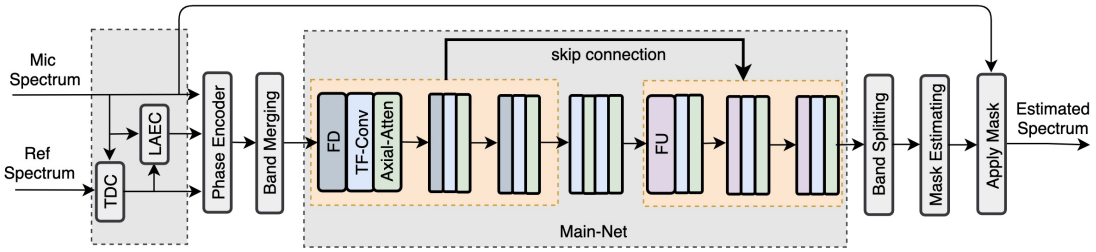

Fig. 1. Architecture of the proposed MTFAA-Net.

response (RIR) $H^e(f)$ and with the late reflections $H^l(f)$, respectively. t, f are indices of time and frequency, respectively. $s(t,f)H^e(f)$ will be taken as the target to be estimated.

The output of LAEC $Y_{laec}(t,f)$ can be regarded as a mixture of clean speech, residual echo, reverberation, and background noise. Different from the previous scheme [3] that only input $Y_{laec}(t,f)$ and far-end reference signal into the network, the network proposed in this paper also takes $Y(t,f)$ as input, which can avoid the performance degradation caused by the distortion introduced by LAEC, and it also helps the network to identify which T-F regions have been suppressed by LAEC due to the presence of echoes.

3. PROPOSED BACKBONE FOR SPEECH DENSE-PREDICTION

In this section, we will show the details of the proposed architecture. Fig.1 shows the overall structure of MTFAA-Net with LAEC and TDC. The MTFAA-Net consists of PE, band merging (BM) and band splitting (BS) modules, mask estimating and applying (MEA) modules, and Main-Net module. The Main-Net includes several similar parts, each consisting of a frequency downsampling (FD) or frequency upsampling (FU), a T-F convolution, and an ASA. With a few tweaks, the MTFAA-Net can be applied to various speech dense-prediction tasks.

3.1. Phase Encoder

Real speech enhancement networks are easier to implement and achieve state-of-the-art results on many datasets [4]. The main part of the backbone is also a real network. To map complex spectral features to real, we design a PE module, which is shown in Fig.2.(a). In the PE module, there are three complex convolutional layers, which receive the microphone signal, the LAEC output and the far-end reference signals, respectively. The kernel size and the stride of the complex convolutional layers are (3,1) and (1,1). In addition, PE also contains a complex to real layer (complex modulo), and a feature dynamic range compression (FDRC) layer. The FDRC is to reduce the dynamic range of speech features, which will make the model more robust.

3.2. Band Merging and Splitting

The distribution of speech valuable information is uneven in frequency dimension, especially for the full-band signal. There are a lot of redundant features in the high frequency bands. The features merging at high frequencies can reduce the redundancy. BS is the inverse process of BM. In this paper, BM and BS bands are spaced according to the ERB scale [10].

3.3. TF-Convolution Module

We use 2D depthwise convolutions (D-Conv) instead of 1D D-Conv in temporal convolutional networks (TCN) [11]. The D-Conv is also designed as a dilated convolution in the time dimension, which can be seen as multi-scale modeling along the time domain. The convolutional block used by the T-F convolution module (TFCM) is shown in Fig.2.(b), which consists of two pointwise convolutional (P-Conv) layers and one D-Conv layer with a kernel size of (3,3). B convolution blocks with dilations from 1 to 2^{B-1} are concatenated together to form a TFCM. Multi-scale modeling improves the receptive field of the TFCM while small convolution kernels are used.

3.4. Axial Self-Attention

Self-attention can improve the network's ability to capture long-range relations between features. Unlike pixel or patch level attention in computer vision [12], an ASA mechanism for speech is proposed in this paper. The ASA can reduce the need for memory and computation, which is more suitable for long sequence signals such as speech. Fig.2.(d) shows the structure of the ASA, where C_i and C represent the input and the attention channel numbers, respectively. Attention score matrices of ASA are calculated along the frequency and time axis, which are called F-attention and T-attention, respectively. Score matrices can be represented as:

$$\mathbf{M}_F(t) = Softmax(\mathbf{Q}_f(t)\mathbf{K}_f^T(t)) \quad (2)$$

$$\mathbf{M}_T(f) = Softmax(Mask(\mathbf{Q}_t(f)\mathbf{K}_t^T(f))) \quad (3)$$

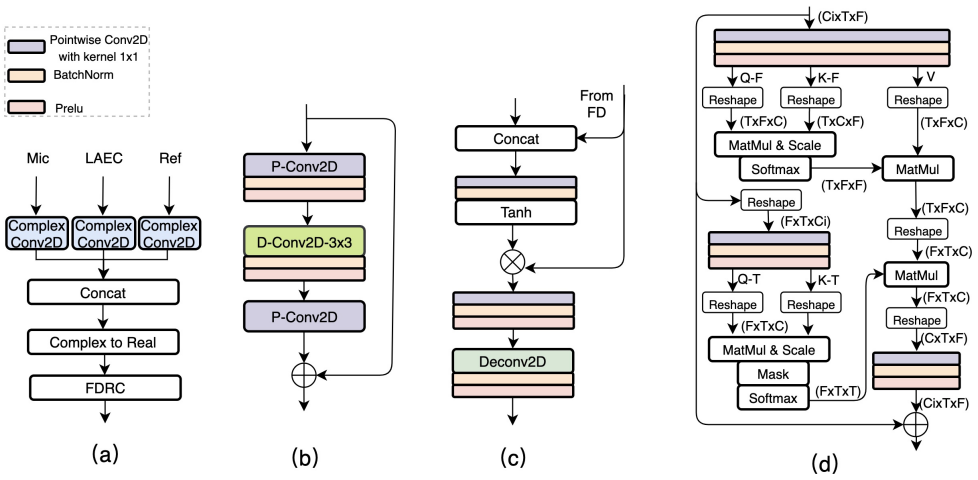

Fig. 2. Flow diagrams of the PE (a), the TFCM (b), the FU (c), and the ASA (d).

where $\mathbf{Q}_f(t), \mathbf{K}_f(t) \in \mathbb{R}^{T\times C}, \mathbf{M}_F(t) \in \mathbb{R}^{F\times F}$ denote key, query, and score matrix of F-attention at frame t. $\mathbf{Q}_t(f)$, $\mathbf{K}_t(f) \in \mathbb{R}^{F\times C}$, $\mathbf{M}_T(f) \in \mathbb{R}^{T\times T}$ denote key, query, and score matrix of T-attention at band f. T, F denote frame numbers and frequency band numbers, respectively. $Softmax$ will be computed along the last dimension. $Mask(*)$ in T-attention is to adjust how long the timing dependency will be captured by the ASA. For MTFAA-Net-Streaming, masking upper triangular part of the input matrix is used, which results in causal ASA.

3.5. Frequency Down and Up Sampling

FD and FU sampling are designed to extract multi-scale features. At each scale, the TFCM and ASA are used for feature modeling, which will improve the network's ability to describe features. FD is a convolution block, which contains a Conv2D layer, a batchnorm (BN) layer, and a Prelu activation layer. FU is shown in Fig.2.(c), where Deconv2D is transpose convolution. Kernel size, stride, and groups of Conv2D and Deconv2D are (1, 7), (1, 4), and 2, respectively. The gated mechanism is used in FU [4].

3.6. Mask Estimating and Applying

The mask estimating and applying (MEA) module consists of two stages. The first stage estimates the real mask of size $(2V+1, 2U+1)$ and applies it to the magnitude spectrum in the form of deep-filter [13]. The second stage estimates the complex mask and applies them to both the magnitude and phase spectrum. Formally, the real $R^{s2}(t,f)$ and image $I^{s2}(t,f)$ part of the enhanced spectrum can be formulated as:

$$A^{s1}(t,f) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} |Y(t+u, f+v)| \cdot M^{s1}(t,f,u,v) \quad (4)$$

$$R^{s2}(t,f) = A^{s1}(t,f) * M_A^{s2}(t,f) * cos(\theta_Y(t,f) + M_\theta^{s2}(t,f)) \quad (5)$$

$$I^{s2}(t,f) = A^{s1}(t,f) * M_A^{s2}(t,f) * sin(\theta_Y(t,f) + M_\theta^{s2}(t,f)) \quad (6)$$

where $M^{s1}(t,f,u,v), A^{s1}(t,f)$ denote the estimated mask and enhanced magnitude spectrum in stage-1, respectively. $\theta_{\mathbf{Y}}(t,f)$ represents the phase spectrum of noisy speech. $M_A^{s2}(t,f), M_\theta^{s2}(t,f)$ denote magnitude and phase part of mask in stage-2, respectively.

4. EXPERIMENTS

4.1. Datasets

Both training and evaluation sets are synthesized using clean speech, background noise, echo, and RIR sets. We use speech and noise clips of DNS4[2] for training. VCTK corpus[3] and DEMAND [14] are used as evaluation speech and noise set, respectively. The ICASSP 2022 AEC Challenge training and dev far-end single talk clips are used as training and evaluation echo sets [8]. For RIR, we use the image source method to obtain 100,000 and 1,000 pairs of RIRs with reverberation time from 0.1s to 0.8s [15] for training and evaluation, respectively. All sets are sampled at 48kHz. Signal to noise ratio (SNR) and signal to echo ratio (SER) are set to [-5, 15]dB and [-10, 10]dB, respectively, for training, and [0, 10]dB and [-5, 5]dB, respectively, for evaluation.

4.2. Implementation Details

We use the STFT complex spectrum with hop length of 8ms and a frame length of 32 ms as input. 1/2 power compression is used for FDRC [16]. The output channel numbers of the complex convolutional layers in PE are 4. The output channel numbers of the three FDs are 48, 96, and 192. The number of convolution block in one TFCM is 6. The attention

[2]https://github.com/microsoft/DNS-Challenge
[3]https://datashare.ed.ac.uk/handle/10283/2651

channel number in the ASA is 1/4 of its input channel number. The number of ERB bands is set to 256. The real mask size in MEA is configured as (3, 1). For the MTFAA-Net -Streaming, the convolutional layers and the ASAs are also configured as causal, and the total system latency is 40ms. The weighting function of target speech RIR is configured to be the same as in [17].

The mean-square-error on power-law compressed spectrum with STFT consistency [18] is used as a loss function. We take Adam with a learning rate of 5e-4 as the optimizer. We train the MTFAA-Net for 300k steps with a batch size of 16.

4.3. Results

4.3.1. Ablation Study

We first evaluate the effectiveness of different modules of the MTFAA-Net. Table 1 shows the ablation results. After removing ASA, the performance of the model decreased on all three tasks, and the PESQ decreased by 0.12 on the echo task. When simultaneously removing ASA and setting the dilation of TFCM to 1, the PESQ decreased by 0.26 on the echo task. By introducing additional conditional information from LAEC, the performance of the model on the echo task can be further improved. However, if the LAEC and the model are simply concatenated together, the performance of the system will be degraded because of the distortion introduced by the LAEC.

Table 1. Ablation study on dereverberation (Rervb), Echo cancellation (Echo), and denoising (Noise) tasks.

	PESQ			STOI		
	Rervb	Echo	Noise	Rervb	Echo	Noise
Noisy	2.9	1.21	1.28	95.8	73.3	78.7
MTFAA-Net -Streaming	**3.74**	**2.67**	**2.65**	**98.0**	**92.1**	**91.1**
-ASA	3.70	2.55	2.55	97.8	91.9	90.5
-ASA -DilatedTFConv	3.61	2.41	2.39	97.5	91.4	89.8
-Conditional LAEC	3.74	2.61	2.64	98.0	92.0	91.1
-Conditional +Concatenated LAEC	3.53	2.51	2.53	97.8	91.4	90.8

4.3.2. Comparison with the State-of-the-Art

Table 2 and Table 3 show the subjective and word accuracy (WAcc) results, which are provided by the AEC and DNS Challenge organizers, respectively. One can find that the proposed scheme outperforms the other methods by a large margin in subjective evaluation. For AEC Challenge, one can observe a 0.072 gain on subjective-MOS, compared with Team 4. For DNS Challenge, our system achieves a considerable gain on BAK-MOS by 0.47, compared with Team14. The system ranked 1st in the two challenges, which proves the robust performance of the proposed backbone.

Table 2. Comparison with other methods on AEC full-band blind test set.

Method	Subjective-MOS	WAcc	Final Score
Baseline	4.100	0.659	0.752
Team4	4.528	0.799	0.866
MTFAA-Net-Streaming	**4.600**	0.814	**0.883**

Table 3. Comparison with other methods on DNS full-band blind set.

Method	SIG-MOS	BAK-MOS	OVRL-MOS	WAcc
Noisy	4.29	2.15	2.63	**0.72**
Baseline	3.62	3.93	3.26	0.63
Team14	4.26	4.27	3.89	0.69
MTFAA-Net-Streaming	**4.30**	**4.70**	**4.13**	0.70

Table 4. Comparison with other methods on DNS wide-band non-blind test set. The look ahead is ∞ to indicate that the model is non-streaming.

Method	Look Ahead	WB-PESQ	NB-PESQ
Noisy	-	1.58	2.45
PoCoNet [19]	∞	2.75	-
FullSubNet [20]	32ms	2.78	3.31
SN-Net [4]	∞	3.39	-
MTFAA-Net	∞	**3.52**	**3.76**
MTFAA-Net-Streaming	40ms	3.32	3.63

We also removed the SP part and conducted a comparative evaluation on the DNS wide-band non-blind test set. The training and evaluation sets are the same as SN-Net [4]. The results are shown in Table 4. The MTFAA-Net outperforms all of the others by a large margin.

We also evaluate the inference time. Under configuration on Section 4.2, the number of multiply-accumulate operations of MTFAA-Net is about 2.4G per second. The real-time factor of the proposed system with Python implementation is about 0.6 (on MacBook Pro with Intel Core i5 core), which satisfies the real-time-processing requirement.

5. CONCLUSION

This paper presents MTFAA-Net, a new backbone for speech dense-prediction tasks. After introducing the conditional information of LAEC, the MTFAA-Net achieves the state-of-the-art performance in both AEC and DNS Challenge of ICASSP 2022. We hope that MTFAA-Net's robust performance will encourage unified modeling more speech dense-prediction tasks. In the future, we will improve the ability of the proposed backbone and extend the backbone to other various tasks such as personal speech enhancement, source separation, etc.

6. REFERENCES

[1] Jacob Benesty, Tomas Gänsler, Dennis R Morgan, M Mohan Sondhi, Steven L Gay, et al., "Advances in network and acoustic echo cancellation," 2001.

[2] Amit S Chhetri, Arun C Surendran, Jack W Stokes, and John C Platt, "Regression-based residual acoustic echo suppression," in *Proc. IWAENC*, 2005, vol. 5.

[3] Jean-Marc Valin, Srikanth Tenneti, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *ICASSP*. IEEE, 2021, pp. 7133–7137.

[4] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu, "Interactive speech and noise modeling for speech enhancement," *AAAI*, 2020.

[5] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[6] Kazuo Ochiai, Takashi Araseki, and Takasi Ogihara, "Echo canceler with two echo path models," *IEEE Transactions on Communications*, vol. 25, no. 6, pp. 589–595, 1977.

[7] Donald L Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on speech and audio processing*, vol. 8, no. 5, pp. 508–518, 2000.

[8] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, and Robert Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP*, 2022.

[9] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sergiy Matusevych, Sebastian Braun, Emre Sefik Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner, "Icassp 2022 deep noise suppression challenge," in *ICASSP*, 2022.

[10] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," 2020.

[11] Yi Luo and Nima 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.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *International Conference on Computer Vision (ICCV)*, 2021.

[13] Wolfgang Mack and Emanuël AP Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[14] J Thiemann, N Ito, and E Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments (2013)," *URL http://parole. loria. fr/DEMAND*.

[15] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "gpuRIR: A Python Library for Room Impulse Response Simulation with GPU Acceleration," *Multimed Tools Applx*, 2020.

[16] Andong Li, Chengshi Zheng, Renhua Peng, and Xiaodong Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[17] Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev, "Towards efficient models for real-time deep noise suppression," in *ICASSP*. IEEE, 2021, pp. 656–660.

[18] Sebastian Braun and Ivan Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

[19] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," in *INTERSPEECH*, 2020.

[20] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP*. IEEE, 2021, pp. 6633–6637.