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

Guochang Zhang, Chunliang Wang, Libiao Yu, Jianqiang Wei

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

ABSTRACT

Speech quality is often degraded by background noise and reverberation. Usually, a dense prediction network is used to reconstruct clean speech. In this work, a novel backbone for speech dense-prediction is proposed. After adjusting part of the input and output, this backbone is used for multi-channel speech enhancement task in this paper. To improve the performance of the backbone, strategies such as multi-channel phase encoder, multi-scale temporal frequency processing, axial self-attention, and two-stage masking are designed. Our proposed method is evaluated based on the datasets of ICASSP 2022 L3DAS22 Challenge. The experimental results show that the proposed method outperforms previous state-of-the-art baselines by a large margin[1] and ranked second in L3DAS22 Challenge. The proposed backbone is also used for mono-channel speech enhancement and ranked first in both ICASSP 2022 AEC[2] and DNS Challenges(non-personal track)[3].

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

1. INTRODUCTION

Multi-channel speech enhancement aims to extract the target speech from the mixture of target speech, noise, reverberation, and interference speech captured by the microphone array. Compared to mono-channel speech enhancement, multi-channel algorithms can utilize additional spatial information, and have been shown to perform better in tasks such as noise suppression, dereverberation, and end-to-end speech recognition [1][2].

As deep neural networks(DNN) become mainstream in various audio processing fields, DNN has also been proven to be a very effective method in the field of multi-channel speech enhancement. The applications of DNN in multi-channel speech enhancement usually come in two forms. One is the combination of signal processing and DNN method, and the other is the full DNN method. In the first method, the time-frequency (T-F) mask is first estimated, and then the mask is used by various data-dependent beamformers, such as minimum variance distortionless response (MVDR), generalized eigenvalue (GEV), etc [3]. The full DNN method is either estimating the filter weights or directly mapping the target speech spectra from a multi-channel noisy spectra [4]. Since no prior assumptions are required, the second method usually has better performance when compared with the first.

In this paper, we present a full DNN scheme for estimating filter weights. This scheme is based on a novel speech dense-prediction backbone called multi-scale temporal frequency convolutional network with axial attention(MTFAA-Net). According to the characteristics of the speech signal, we have adopted the following strategies to improve the performance of the backbone:

- To achieve multi-scale modeling, the model is designed to perform dilated convolution in time dimension and up-down-sampling in frequency dimension.
- A lightweight axial self-attention(ASA) is introduced to express the dependence of features in time domain and frequency domain, respectively
- Inspired by traditional beamformer, we introduce a complex convolutional layer to encode phase information, which can be considered as a beamformer with learnable filter weights.
- The frequency band merging module according to equivalent rectangular bandwidth(ERB) is used to reduce redundant computation at high frequency bands.
- A two-stage masking mechanism is used to improve the length of filter weights.

The results on dev and blind test sets of L3DAS22 Challenge show that the proposed scheme outperforms official baseline(U-net) and the MVDR beamformer with ideal complex mask.

The rest of this paper is organized as follows. Section 2 presents problem formulation. Section 3 provides details of the proposed backbone for muti-channel speech enhancement. Results are reported in Section 4. Finally, we summarize our findings in Section 5.

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

[2]https://www.microsoft.com/en-us/research/academic-program/acoustic-echo-cancellation-challenge-icassp-2022/

[3]https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2022/

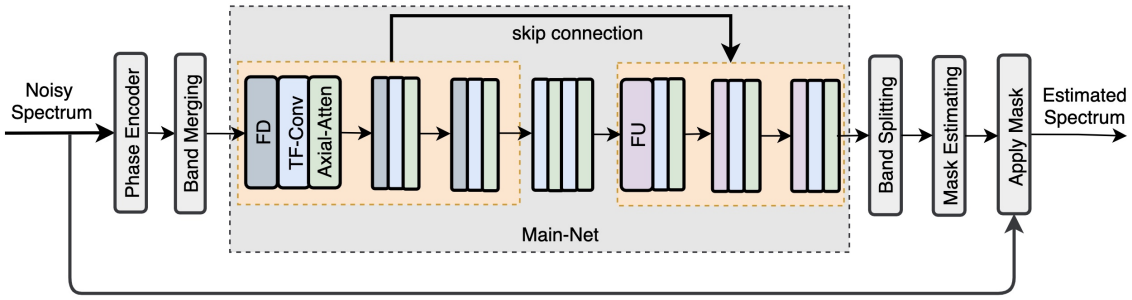

Fig. 1. Architecture of the proposed MTFAA-Net.

2. PROBLEM FORMULATION

This section describes the signal models for multi-channel signal capture and processing. Assume $y^m(t)$ with $m = 0 \cdots, M-1$ recorded by an array with M microphones. The clean speech is corrupted by noise and reverberation. The noisy speech $y^m(t)$ in the short-time fourier transform (STFT) domain can be given by:

$$\mathbf{Y}(t,f) = \mathbf{s}(t,f)\mathbf{H}(f) + \mathbf{N}(t,f), \tag{1}$$

where $\mathbf{Y}(f,t), \mathbf{N}(f,t), \mathbf{H}(f) \in \mathbb{C}^{M\times 1}$, $\mathbf{s}(t,f) \in \mathbb{C}^{1\times 1}$ denote M-channels noisy speech, M-channels noise, acoustic transfer function from speaker to microphones, and mono clean speech in STFT domain, respectively. t, f are indices of time and frequency, respectively. Typically, neural network beamformer method will estimate frame-wise multi-channel filter weights, and the output of neural network beamformers can be expressed as follows:

$$\widetilde{\mathbf{s}}(f,t) = \mathbf{W}^{\mathbf{H}}(t,f)\mathbf{Y}(t,f), \tag{2}$$

where $\mathbf{W}(t,f) \in \mathbb{C}^{M\times 1}$ denotes the M-channel filter weights estimated by the deep neural network. $^{\mathbf{H}}$ represents the conjugate transpose operator. Similarly, the proposed scheme also estimates filter weights.

3. PROPOSED BACKBONE FOR SPEECH DENSE-PREDICTION

In this section, we will show the details of the proposed backbone for speech dense-prediction. Fig.1 shows the overall network structure of MTFAA-Net. The MTFAA-Net consists of phase encoder, band merging and splitting modules, estimating and applying mask modules, and Main-Net module. The Main-Net includes several similar parts, each consisting of a frequency downsampling or upsampling, 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 [5]. The Main-Net of the proposed backbone is also a real network. To map complex spectral features to real, we design a phase encoder (PE) module. As shown in Fig.2.(a), PE contains a complex convolutional layer, a complex to real layer(complex modulo), and a feature dynamic range compression(FDRC) layer. The kernel size and the stride of complex convolutional layer are (3,3) and (1,1), respectively. FDRC is to reduce the dynamic range of speech features, which will make the model more robust. The complex convolutional layer can be viewed as a beamformer with learnable weights, with each out channel representing the output of a beamfomer.

3.2. Band Merging and Splitting

The distribution of speech valuable information is uneven in frequency dimension. Generally speaking, the redundant features of speech will be more at high frequencies. The features merging at high frequencies can reduce the redundancy of features and reduce the amount of computation. Band splitting(BS) is the inverse process of band merging(BM). In this paper, BM and BS bands are spaced according to the ERB scale [6][7].

3.3. TF-Convolution Module

We use 2D depthwise convolutions(D-Conv) instead of 1D D-Conv in temporal convolutional networks(TCN) [8]. The D-Conv is also designed as an dilated convolution in the time dimension, which can be seen as multi-scale modeling along the time dimension. The convolutional block used by the T-F convolution module(TFCM) is shown as 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 greatly improves the receptive field of the TFCM while small convolution kernels are used.

3.4. Axial Self-Attention

Self-attention can improve network ability to capture long-range relations between features. Unlike pixel or patch level

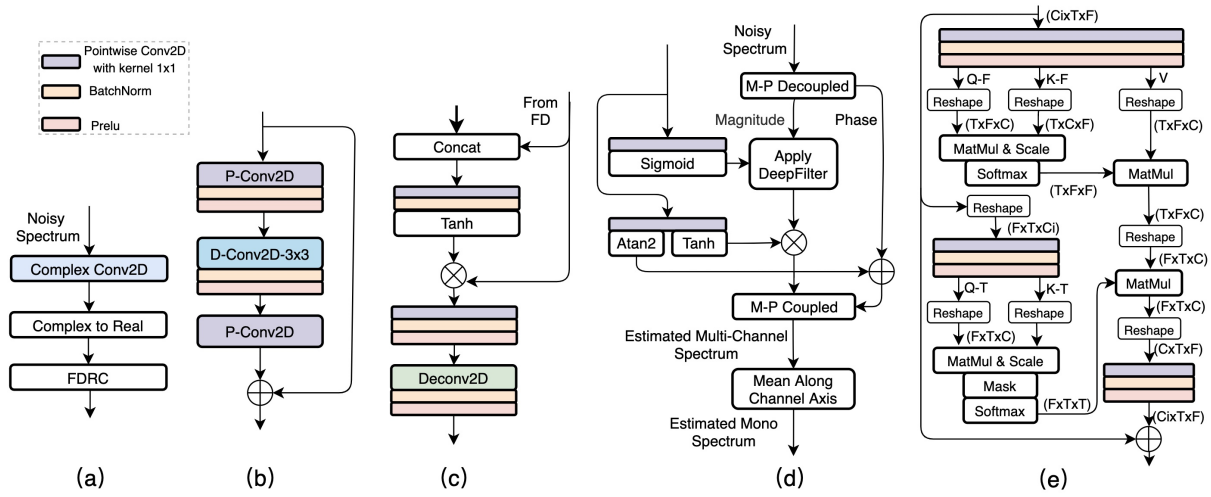

Fig. 2. Flow diagrams of the phase encoder(a), the convolutional block in TF-convolution module(b), frequency up-sampling module(c), mask estimating and applying module(d), and axial self attention module(e).

self-attention in computer vision [9][10], an ASA mechanism for speech is proposed in this paper. The proposed ASA can greatly reduce the need for memory and computation, which is more suitable for long sequences of speech signals. Fig.2.(e) shows the structure of the ASA, where C_i represents the input channel number and C represents the attention channel number. 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:

$$\mathbf{M}_F(t) = Softmax(\mathbf{Q}_f(t)\mathbf{K}_f^T(t)) \tag{3}$$

$$\mathbf{M}_T(f) = Softmax(Mask(\mathbf{Q}_t(f)\mathbf{K}_t^T(f))) \tag{4}$$

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 last dimension. $Mask(*)$ in T-attention is to adjust how long the timing dependency will be captured by the ASA.

3.5. Frequency Down and Up Sampling

Frequency downsampling(FD) and frequency upsampling(FU) are designed to extract multi-scale features. At each scale, 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 with kernel size of (3, 7) and stride of (1, 4), a batchnorm (BN) layer and a Prelu activation layer. FU is shown as Fig.2.(c), where Deconv2D is transpose convolution with kernel size of (3, 7) and stride of (1, 4). The gated mechanism is used in FU [5].

3.6. Mask Estimating and Applying

Fig.2.(d) shows the mask estimating and applying(MEA) module of multi-channel which 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 [11]. The second stage estimates the complex mask and applies them to both the magnitude and phase spectrum. Formally, the real $\mathbf{R}^{s2}(t,f)$ and image $\mathbf{I}^{s2}(t,f)$ part of multi-channel enhanced spectrum can be formulated as:

$$\mathbf{A}^{s1}(t,f) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} |\mathbf{Y}(t+v, f+u)| \cdot \mathbf{M}^{s1}(t,f,u,v) \tag{5}$$

$$\mathbf{R}^{s2}(t,f) = \mathbf{A}^{s1}(t,f) * \mathbf{M}_A^{s2}(t,f) * cos(\theta_{\mathbf{Y}}(t,f) + \mathbf{M}_\theta^{s2}(t,f)) \tag{6}$$

$$\mathbf{I}^{s2}(t,f) = \mathbf{A}^{s1}(t,f) * \mathbf{M}_A^{s2}(t,f) * sin(\theta_{\mathbf{Y}}(t,f) + \mathbf{M}_\theta^{s2}(t,f)) \tag{7}$$

where $\mathbf{M}^{s1}(t,f,u,v), \mathbf{A}^{s1}(t,f) \in \mathbb{R}^{(M\times 1)}$ denote the estimated mask and enhanced magnitude spectrum in stage1, respectively. $\theta_{\mathbf{Y}}(t,f)$ represents the phase spectrum of noisy speech. $\mathbf{M}_A^{s2}(t,f), \mathbf{M}_\theta^{s2}(t,f) \in \mathbb{R}^{(M\times 1)}$ denote magnitude and phase part of mask in stage2, respectively. Finally, the mono estimated spectrum is obtained using averaging along the channel dimension.

4. EXPERIMENTS

4.1. Datasets

In our experiments, we make use of the L3DAS22 dataset for training and evaluation, which contains more than 80 hours and 40000 virtual 3D audio environment of 8-channels audio recordings [12]. The L3DAS22 dataset is captured by two first-order A-format ambisonics microphones in office reverberant environment with 16kHz sampling rate. The dataset is

divided into three partitions: 80 hours train set, 3 hours dev test set, and 3 hours blind test set.

4.2. Implementation Details

We take the 8-channel STFT complex spectrum with hop size of 128 samples and a frame length of 1536 samples as input. 1/2 power compression is used for FDRC [13]. The input channel number of the complex convolutional layer in PE is 8, and the output channel number is 16. The output channel numbers of the three FDs are 80, 160, and 320. The number of convolution block in one TFCM is 6. To reduce the length of the feature sequence, we set the stride of the first FD to (2, 4). The attention channel number in the ASA is 1/4 of its input channel number. The number of ERB bands is set to 384. The real mask size in MEA is configured as (3, 1). Under this configuration, the number of multiply-accumulate operations (MACS) of MTFAA-Net is 5.5G per second.

The mean-squre-error(MSE) on power-law compressed spectrum with STFT consistency [14][15] is used as loss function. We take Adam optimizer with a learning rate of 0.0005 as the optimizer. We train the MTFAA-Net for 400 epochs. The batch size is set to 16.

4.3. Results

The evaluation metrics include PESQ, STOI, and word error rate(WER). Two pretrained models based on the conformer(large)[4] [16] and wav2vec[5] [17] architectures are used to compute the WER.

4.3.1. Ablation Study

We first evaluate the effectiveness of different modules of the MTFAA-Net based on dev test set of L3DAS22 Challenge. Table 1 shows the ablation results. After removing ASA, PESQ and STOI decreased by 0.14 and 0.004, respectively. When simultaneously removing ASA and setting the dilation of TFCM to 1, PESQ and STOI decreased by 0.33 and 0.11, respectively. After introducing ERB, we observe a 0.09 gain on PESQ and 0.003 gain on STOI(In order to compare with the same MACS, the frame size is set to 768). After adding stage1 in MEA, it improves PESQ by 0.05. The experimental results confirm that ASA, multi-scale modeling, BM&BS, and 2-stage mask can improve the performance of the backbone.

4.3.2. Comparison with other methods

We compare U-net(official baseline) [12], ideal complex mask MVDR(frame size of 1536, hop size of 128) and proposed MTFAA-Net on the L3DAS22 Challenge dev test set.

[4]https://github.com/wenet-e2e/wenet/blob/main/examples/librispeech/s0

[5]https://huggingface.co/facebook/wav2vec2-base-960h

Table 1. Ablation study on L3DAS22 Challenge dev test set.

Models	PESQ	STOI	WER-Conformer
Noisy	1.22	0.616	0.278
MTFAA-Net	**3.50**	**0.977**	**0.015**
-ASA	3.36	0.973	0.017
-ASA -DilatedTFConv	3.17	0.966	0.024
-ERB +FrameSize of 768	3.41	0.974	0.017
-Stage1 MEA	3.45	0.976	0.016

Table 2. Comparision with other methods on L3DAS22 Challenge dev test set.

Models	PESQ	STOI	WER-Conformer
Noisy	1.22	0.616	0.278
Offcial Baseline	1.69	0.877	0.180
Ideal Complex Mask MVDR	2.36	0.959	0.026
MTFAA-Net	**3.50**	**0.977**	**0.015**

Table 3. Comparision with other methods on L3DAS22 Challenge blind test set. T1 metric is a combination of STOI and WER(The higher is better).

Models	STOI	WER-Wav2Vec	T1 Metric
Offcial Baseline	0.878	0.212	0.833
MTFAA-Net	**0.975**	**0.025**	**0.975**

From the comparison results shown in Table 2, MTFAA-Net achieves a considerable STOI gain by 0.018, compared with ideal complex mask MVDR. More STOI gain (0.1) is achieved while compared with U-net.

Table 3 shows the results of the MTFAA-Net on the L3DAS22 Challenge blind test set. Compared with the official baseline, the WER of the proposed method has a relative reduction of 88%, and the STOI is improved by about 0.1. The proposed MTFAA-Net ranked second for the 3D speech enhancement task in the L3DAS22 Challenge.

5. CONCLUSION

In this paper, we propose a novel backbone for speech dense-prediction tasks and extend it to multi-channel speech enhancement task in noise and reverberation environment. The proposed method outperforms prior arts across PESQ, STOI, and WER. In the future, we will improve the ability of the proposed backbone in mono or multi-channel speech enhancement tasks and extend the backbone to other various tasks such as speaker separation, personal speech enhancement and so on.

6. REFERENCES

[1] Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech sep-

aration," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.

[2] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, and Michiel Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5075–5079.

[3] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 196–200.

[4] Andong Li, Wenzhe Liu, Chengshi Zheng, and Xiaodong Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," *arXiv preprint arXiv:2109.00265*, 2021.

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

[6] 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," in *ICASSP*, 2020.

[7] Jean-Marc Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.

[8] 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.

[9] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *Advances in Neural Information Processing Systems*, 2021.

[10] 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.

[11] 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.

[12] Marco Pennese Xinlei Ren Xiguang Zheng Chen Zhang Bruno Masiero Aurelio Uncini Danilo Comminiello Eric Guizzo, Christian Marinoni, "L3das22 challenge: Learning 3d audio sources in a real office environment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[13] 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.

[14] Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev, "Towards efficient models for real-time deep noise suppression," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 656–660.

[15] 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.

[16] Di Wu, Binbin Zhang, Chao Yang, Zhendong Peng, Wenjing Xia, Xiaoyu Chen, and Xin Lei, "U2++: Unified two-pass bidirectional end-to-end model for speech recognition," *arXiv preprint arXiv:2106.05642*, 2021.

[17] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.