Self-Attention for Multi-Channel Speech Separation in Noisy and Reverberant Environments

Conggui Liu and Yoshinao Sato
Fairy Devices Inc., Japan
E-mail: liu@fairydevices.jp, sato@fairydevices.jp

***Abstract*—Despite recent advances in speech separation technology, there is much to be explored in this field, especially in the presence of noise and reverberation. One of the significant difficulties is that locations where relevant context information is incorporated vary in the time, frequency, and channel directions. To overcome this problem, we investigated the use of self-attention for multi-channel speech separation with time-frequency masking. Our base model is a temporal convolutional network that is the same as Conv-TasNet, except it works in the frequency domain with the short-time Fourier transformation and its inverse. We combined this basis with a self-attention network. We explored nine different types of self-attention network for this purpose. To investigate the effects of the self-attention networks, we evaluated the performance of the proposed model, which we refer to as a confluent self-attention convolutional temporal audio separator network (CACTasNet), on a noisy and reverberant version of the wsj0-2mix dataset. We found that several different self-attention networks substantially improved the performance measured by scale-invariant signal-to-noise ratio and signal-to-distortion ratio. The results indicate that a self-attention mechanism can efficiently locate context information relevant to speech separation.**

I. Introduction

Speech separation aims to isolate each source's signal from given recordings in which all sources are overlapped. In recent years, many deep learning models for speech separation have been proposed and achieved significant progress. One of the most successful approaches is time-frequency masking, such as deep clustering [1][2], deep attractor [3][4], and permutation-invariant training [5][6]. More recently, "time domain" separation networks have been explored [7][8][9]. These methods perform masking in a latent space to/from which a waveform is transformed with a trainable encoder/decoder.

Despite the considerable progress made in recent years, speech separation is still challenging, especially in environments with background noise and reverberation. One of the significant difficulties is when context information relevant to separation exists in mixed speech signals. In estimating a mask at a certain time-frequency point, spectra far in time might play an essential role in some cases. In other cases, inter-channel intensity differences close in time might include crucial information. Moreover, relative locations of relevant information might be different for each time-frequency point. For efficient speech separation, relevant information must be located in the time, frequency, and channel directions. However, neither recurrent nor convolutional networks, which were used in previous studies, have successfully dealt with this problem.

To overcome this difficulty, we investigated the usage of a self-attention mechanism in this study. This mechanism was introduced for machine translation [10], where it enabled the network to pay attention to input words at different positions relevant to each translated output word. Self-attention mechanisms have also been used to estimate the importance of each frame in speech for emotion recognition [11] and speaker recognition [12][13][14]. Our aim of using self-attention is to enable an audio separator network to discover relevant context information to mask estimation at different locations in the time, frequency, and channel dimensions. A few previous studies have investigated self-attention for speech separation. For example, [15] used self-attention to process inter-channel information, while [16] utilized self-attention along the time direction for single-channel speech separation. However, these previous studies did not focus on how to apply the self-attention mechanism and how effectively it works for speech separation. This study explored nine types of self-attention networks for multi-channel speech separation under reverberant and noisy conditions: (1) time-wise attention, (2) channel-varying time-wise attention, (3) frequency-varying time-wise attention, (4) frequency-wise attention, (5) time-varying frequency-wise attention, (6) channel-varying frequency-wise attention, (7) channel-wise attention, (8) time-varying channel-wise attention, and (9) frequency-varying channel-wise attention.

To assess self-attention networks under realistic conditions, we simulated noisy and reverberant speech mixtures. Specifically, we mixed utterances from the WSJ0 corpus [17] with simulated room impulse responses (RIRs) and added noise from the MUSAN corpus [18]. We evaluated our models' performances measured using scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-distortion ratio improvement (SDRi) to find which types of self-attention networks provide significant improvement.

The remainder of this paper is organized as follows. Section II presents the self-attention networks and proposed method. The experimental setup and results are described in Section III. Finally, Section IV concludes the paper.

II. Methods

We suppose that the number of speakers S and number of microphones C are known and shared by all utterances in the training, evaluation, and testing.

A. Self-attention

In this section, we formulate the self-attention networks explored in this study. A self-attention mechanism attempts to guide a neural network to score the importance of input features in a sequence. Fig. 1 shows the structure of self-attention network proposed in [10]. After choosing an axis along which attention should be paid, we can reshape an input feature X of a data sample to have the shape of $(d_{\mathrm{u}}, d_{\mathrm{a}})$, where d_{a} and d_{u} denote the sizes of X in the attended and remaining dimensions, respectively. In other words, we assume

$$\dim X = (d_{\mathrm{u}}, d_{\mathrm{a}}). \tag{1}$$

In a self-attention network, an input feature is passed to three parallel layers: query, key, and value. These layers map an input features to a query Q, a key K, and a value V. We suppose that Q and K share the size in the embed space. Then, without loss of generality we can write

$$\begin{aligned} \dim Q &= (d_{\mathrm{e}}, d_{\mathrm{a}}) \\ \dim K &= (d_{\mathrm{e}}, d_{\mathrm{a}}) \\ \dim V &= (d_{\mathrm{v}}, d_{\mathrm{a}}). \end{aligned} \tag{2}$$

In other words, we map X to Q, K, and V as follows:

$$\begin{aligned} \text{query layer} &: X \in \mathbb{R}^{d_{\mathrm{u}} \times d_{\mathrm{a}}} \mapsto Q \in \mathbb{R}^{d_{\mathrm{e}} \times d_{\mathrm{a}}} \\ \text{key layer:} &: X \in \mathbb{R}^{d_{\mathrm{u}} \times d_{\mathrm{a}}} \mapsto K \in \mathbb{R}^{d_{\mathrm{e}} \times d_{\mathrm{a}}} \\ \text{value layer} &: X \in \mathbb{R}^{d_{\mathrm{u}} \times d_{\mathrm{a}}} \mapsto V \in \mathbb{R}^{d_{\mathrm{v}} \times d_{\mathrm{a}}}. \end{aligned} \tag{3}$$

We consider that each of a query layer, a key layer, and a value layer consists of a single fully-connected or convolutional layer. In the case of a fully-connected layer, we can represent Q, K, and V as

$$\begin{aligned} Q &= W_Q X \\ K &= W_K X \\ V &= W_V X, \end{aligned} \tag{4}$$

where W_Q, W_K, and W_V denote the respective weight matrices. In the case of a convolutional layer, Q, K, and V can be represented as

$$\begin{aligned} Q &= \mathrm{Conv}_Q(X) \\ K &= \mathrm{Conv}_K(X) \\ V &= \mathrm{Conv}_V(X), \end{aligned} \tag{5}$$

where $\mathrm{Conv}_Q, \mathrm{Conv}_K$, and Conv_V denote respective one-dimensional convolutions applied along the first axis of X. Calculating the similarity between the query and key, we get an attention map A, which represents the importance of the input feature in the attended dimension. In this study, we used the scaled dot-product attention [10]:

$$A = \mathrm{softmax}\left(\frac{Q^T K}{\sqrt{d_{\mathrm{e}}}}\right). \tag{6}$$

Note that the shape of the attention map is given by

$$\dim A = (d_{\mathrm{a}}, d_{\mathrm{a}}). \tag{7}$$

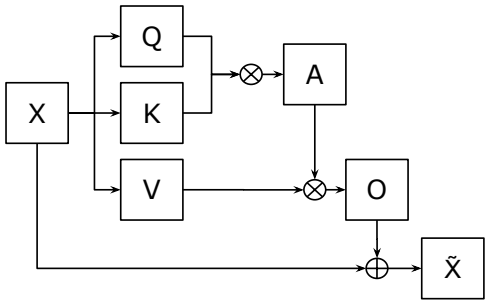

Fig. 1. Structure of a self-attention network.

Multiplying the attention map A by the value V, we get a weighted feature O^T:

$$O^T = V A^T \in \mathbb{R}^{d_{\mathrm{v}} \times d_{\mathrm{a}}}. \tag{8}$$

If we set $d_{\mathrm{v}} = d_{\mathrm{u}}$, we can introduce a residual connection to fuse the input feature and weighted feature. Consequently, we get the following output:

$$\tilde{X} = X + O^T. \tag{9}$$

In this study, we used a magnitude spectrogram as an input feature. Hence, X can be represented as follows:

$$X = \{X_{ftc} \,|\, 1 \le f \le F, 1 \le t \le T, 1 \le c \le C\}, \tag{10}$$

where F, T, and C denote the numbers of frequency bins, frames, and channels, respectively. There are many variants of attention networks depending on how the input feature space is decomposed and the choices of the query, key, and value layers. In this study, we explored nine types of self-attention networks for multiple-channel audio data, which are listed in Table I. Note that we set $d_{\mathrm{e}} = d_{\mathrm{v}} = d_{\mathrm{u}}$ although other options are possible. We refer to an attention network that estimates the importance of an input feature in the time direction as time-wise. In other words, a time-wise attention map is a square matrix of size T. Frequency-wise and channel-wise attention are analogously defined. If we deal with the input features as a function of the time frames, the query, key, value, and attention map can be considered as functions of the time frames. In contrast, the layers are time-independent. In other words, attention is performed at each time frame independently. We refer to such an attention network as time-varying. Frequency-varying and channel-varying are analogously defined. Note that the value of T is different between utterances. Therefore, if d_{u} depends on T, we must use convolution to fix the layer sizes.

Here, we explain some of the self-attention types in detail. The first example we consider is the time-varying channel-wise self-attention network. This network pays attention to the channel directions, i.e., microphones. Hence, the attention map is a square matrix of size C. The attention map is calculated using the spectra at each time frame independently. Therefore, the attention map can be considered as a function of a time frame. In other words, we can say that time-varying

channel-wise self-attention enables a network to estimate the importance of the channels at each time based on the spectra in all frequency bands. In the second example, we consider the frequency-varying time-wise self-attention network. This network pays attention to locations in the time direction, i.e., time frames. Hence, the attention map is a square matrix of size T. The attention map is calculated independently for each frequency bin using the spectral amplitudes stacked along the channel direction. Therefore, the attention map can be considered as a function of a frequency bin. In other words, we can say that frequency-varying time-wise self-attention enables a network to find relevant time frames based on the spectral amplitude of all channels for each frequency bin.

B. Model Structure

In this section, we describe the structure of our model. First, the base model is introduced. Then, we explain how the base model is harnessed by self-attention.

The base model used in this study is time-frequency masking with a temporal convolutional network (TCN). The network consists of the short-time Fourier transform (STFT), a TCN-based separator, and the inverse short-time Fourier transform (ISTFT). Both STFT and ISTFT can be implemented as a one-dimensional convolution layer with a fixed kernel function. The separator has the same structure as that of Conv-TasNet [8], except the STFT and ISTFT layers replace the encoder and decoder, respectively. It consists of R repetitions of M stacked one-dimensional convolutional blocks with dilaton factors $1, \ldots, 2^{M-1}$. In other words, the base model is a time-frequency version of Conv-TasNet, which was also investigated in [19] and [20]. As we use the global layer normalization, the separator lacks causality. However, it is possible to make the model causal by using cumulative layer normalization. Moreover, we use the SI-SDR loss in the time domain instead of the mean square error (MSE) loss in the frequency domain, as investigated in [19] and [20]. Note that while the "time-domain" models, such as TasNet [7], Conv-TasNet [8], and FurcaNext [9], perform separation in a latent space, our model does it in the time-frequency domain.

We added different self-attention networks to the base model, as illustrated in Fig. 2. We refer to the proposed model as a confluent self-attention convolutional temporal audio separator network (CACTasNet). The separator is divided into two paths. While the first path is not equipped with a self-attention network, the second path is at the entrance. We pass a single-channel spectrogram optionally along with inter-channel phase differences (IPDs) to the first path and a multi-channel spectrogram to the second path. Both paths consist of $R - r$ repetitions of M stacked one-dimensional convolution blocks. After the confluence of two paths with and without the self-attention network, r repetitions of M stacked one-dimensional convolution blocks follow. Then, all residual connections branched from the main paths are merged. Finally, the separator network outputs time-frequency masks. In this study, we set $R = 4$, $r = 2$, and $M = 8$.

We did not use a residual connection in the self-attention layer because it was not effective in our preliminary experiments. This can be understood as follows. In the proposed model, the flow of information extracted from an input feature and that extracted from the attention map join together inside the network. Therefore, introducing a residual connection in the self-attention layer is considered redundant.

III. Experiments

A. Experimental Setup

The number of speakers S and number of microphones C were supposed to be known and fixed. Explicitly, we set $S = 2$ and $C = 8$ in our experiments. Moreover, we assumed the arrangement of the microphones was fixed during an utterance but different between utterances.

We set the STFT window length and shift to 512 and 256, or equivalently 32 ms and 16 ms, respectively, while all remaining parameters were set to the same values provided in [8]. We picked up five microphone pairs to calculate sinIPD and cosIPD features. To reduce the computational complexity, we used only four microphones in the case of channel-wise self-attention. For the self-attention network, each of the convolutional layers had a filter of size 257.

B. Data

To assess the self-attention networks under realistic conditions, we simulated a noisy and reverberant version of the wsj0-2mix dataset [1]. We mixed utterances from the WSJ0 corpus [17] with background noise and reverberation. First, we generated RIRs with the image method [21]. The length, width, and height of a rectangular room were chosen from a uniform distribution from 5.0 m to 10.0 m, 5.0 m to 10.0 m, and 3.0 m to 4.0 m, respectively. The center of a virtual sphere whose radius was chosen from a uniform distribution from 0.075 m to 0.125 m was randomly placed within 0.2 m of the room center. Eight microphones were put on the surface so that any pair is no closer than 0.05 m. Each of two speakers, whose height was chosen from a uniform distribution from 1.5 m to 2.0 m, was randomly located more than 0.5 m away from the sphere center. We chose the speakers' locations so that their distances were each larger than 1.0 m. The reverberation time (T60) was chosen from a uniform distribution from 0.2 s to 0.6 s. Secondly, we chose a random pair of anechoic speech recordings by different speakers from WSJ0 and convolved them with a randomly chosen RIR to obtain a multi-channel reverberated audio mixture. We fixed the microphone arrangement during an utterance, while we used different ones for each utterance. Finally, we randomly added noise from the MUSAN dataset [18] as additive noise. The signal-to-noise ratio (SNR) was randomly chosen from a uniform distribution from 10 dB to 15 dB. Note that the noise we added did not include intelligible speech or music. In total, we generated 20,000, 5,000, and 3,000 multi-channel utterances for training, validation, and testing, respectively. The sampling rate of all data was 16 kHz.

TABLE I
NINE TYPES OF SELF-ATTENTION.
"FC" AND "CONV" REPRESENT FULLY-CONNECTED AND CONVOLUTIONAL LAYERS, RESPECTIVELY.

Type	Input	Dimension $d_{\mathrm{u}} = d_{\mathrm{e}} = d_{\mathrm{v}}$	Dimension d_{a}	Layer Query	Layer Key	Layer Value
time-wise	X	$F \cdot C$	T	FC	FC	FC
channel-varying time-wise	$X(c)$	F	T	FC	FC	FC
frequency-varying time-wise	$X(f)$	C	T	FC	FC	FC
frequency-wise	X	$T \cdot C$	F	Conv	Conv	Conv
channel-varying frequency-wise	$X(c)$	T	F	Conv	Conv	Conv
time-varying frequency-wise	$X(t)$	C	F	FC	FC	FC
channel-wise	X	$F \cdot T$	C	Conv	Conv	Conv
time-varying channel-wise	$X(t)$	F	C	FC	FC	FC
frequency-varying channel-wise	$X(f)$	T	C	Conv	Conv	Conv

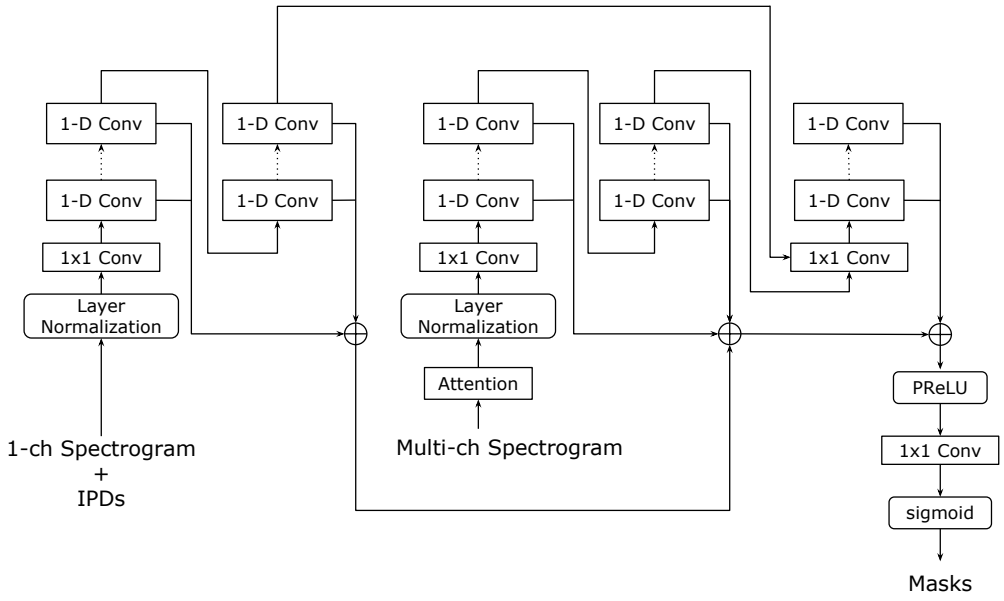

Fig. 2. Structure of the proposed CACTasNet model. "Attention" represents the self-attention network. "1x1 Conv" is a convolutional layer with kernel size 1, also known as point-wise convolution. "1-D Conv" denotes the one-dimensional convolutional block described in [8]. "PReLU" represents a parametric rectified linear unit.

C. Results

We conducted speech separation experiments with the proposed model on the noisy and reverberant dataset. Table II and III show the result without and with the IPD features, respectively. According to the experimental results, we found that several types of self-attention improve the performance substantially.

As shown in Table II, the time-varying channel-wise self-attention network showed the most significant improvement without the IPD features. Although the magnitude of the improvement was relatively small with the IPD features, the proposed model with several types of self-attention performed better than the base model. The frequency-varying time-wise self-attention network improved the performance most with the IPD features, as shown in Table III.

TABLE II
THE PERFORMANCE OF SPEECH SEPARATION WITHOUT IPDS

Model	Self-Attention Type	SI-SNRi	SDRi
base	-	5.8	6.5
proposed	time-wise	5.9	6.5
	channel-varying time-wise	6.6	7.2
	frequency-varying time-wise	6.7	7.3
	frequency-wise	6.1	6.7
	channel-varying frequency-wise	5.3	6.0
	time-varying frequency-wise	6.4	7.0
	channel-wise	5.8	6.5
	time-varying channel-wise	**6.8**	**7.4**
	frequency-varying channel-wise	6.4	7.0

D. Analysis

To further understand the self-attention mechanism for speech separation, we took an example utterance and analyzed

TABLE III
THE PERFORMANCE OF SPEECH SEPARATION WITH IPDS

Model	Self-Attention Type	SI-SNRi	SDRi
base	-	9.0	9.6
	time-wise	9.2	9.7
	channel-varying time-wise	9.1	9.7
	frequency-varying time-wise	**9.3**	**9.9**
	frequency-wise	9.1	9.7
proposed	channel-varying frequency-wise	9.2	9.7
	time-varying frequency-wise	9.2	9.8
	channel-wise	9.1	9.7
	time-varying channel-wise	9.2	9.8
	frequency-varying channel-wise	9.2	9.8

the attention maps. Fig. 3 shows the spectrograms of two clean speech utterances and background noise with a duration of 3.2 s, or 200 frames, at the head. In the following, we focus on the time-varying channel-wise attention, which achieved the best performance without the IPD features, and frequency-varying time-wise attention, which achieved the best performance with the IPD features.

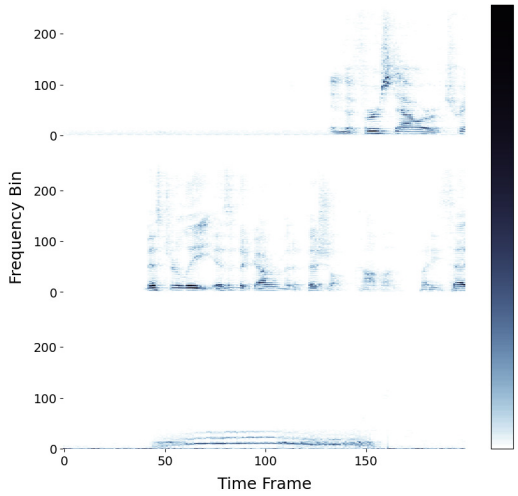

Fig. 3. Spectrogram of an example utterance. The upper panel shows speech by the first speaker. The middle panel shows speech by the second speaker. The bottom panel shows noise.

The time-varying channel-wise self-attention network utilizes inter-channel information, which is considered to be essential for multi-channel speech separation. Notably, this type of self-attention outperformed the channel-wise and frequency-varying channel-wise self-attention, as shown in Table II. Fig. 4 illustrates the attention map for the example utterance mentioned above. This superiority indicates that the time development of inter-channel information plays a significant role. In support of this, we found dynamical changes of the attention map for the example utterance mentioned above. Moreover, we did not observe non-trivial structure in the attention map during non-speech segments at the head of the utterance. The improvement of performance by the channel-wise self-attention became less significant but still remained with the IPD features, as shown in Table III. This change can be interpreted as follows: while the channel-wise self-attention contributes to speech separation in a similar manner to the IPDs, it can extract relatively more beneficial information.

The frequency-varying time-wise self-attention guides a network to find time frames where relevant information is incorporated. This type of self-attention outperformed the other types, as shown in Table III. Fig. 5 illustrates the attention map for the example utterance mentioned above. The attention maps showed different patterns in each frequency bin. This can be understood as different frequency bins putting weight on different points in the time direction. Further to that, we observed that some frames far in time were paid much attention. This means information located not only close in time but also far in time can be beneficial to speech separation. It is notable that this type of self-attention utilizes the spectral amplitude stacked along the channel direction. In other words, it utilizes inter-channel information in a different way from channel-wise attention networks to pay attention to the time direction.

To summarize, the proposed model with an appropriate type of self-attention outperformed the base model. This result indicates that the self-attention mechanism enables a network to find the locations where information relevant to speech separation is incorporated.

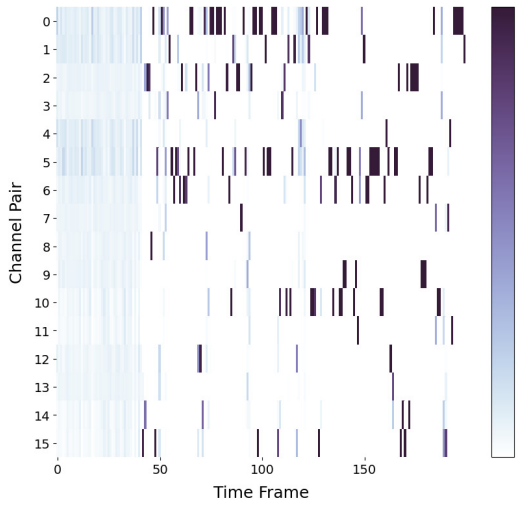

Fig. 4. Time-varying channel-wise attention map. At each time frame, a matrix of size $(4, 4)$ representing the channel-wise attention map is visualized as a 16-dimensional vector.

IV. CONCLUSIONS

In this study, we explored the use of self-attention for multi-channel speech separation. We proposed CACTasNet, a confluent self-attention extension of a convolutional temporal audio separator, and evaluated it with nine types of self-attention networks in a noisy and reverberant environment. Our experiments showed that several types of self-attention improved the performance of the base model significantly.

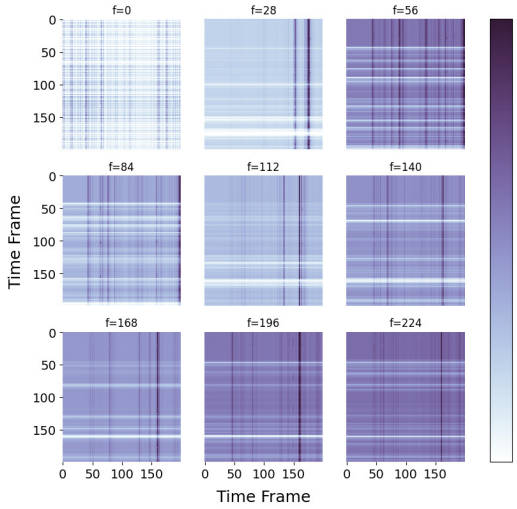

Fig. 5. Frequency-varying time-wise attention map at different frequency bins.

Specifically, time-varying channel-wise self-attention achieved the best performance without IPDs, while frequency-varying time-wise self-attention achieved the best performance with IPDs. The results indicate that a self-attention mechanism is beneficial to multi-channel speech separation in noisy and reverberant environments by giving audio separator networks the capability of discovering relevant context information in multi-channel mixed signals.

Finally, we discuss some of the future directions for this research. Future work will include the investigation of an online version of the proposed model. By using the cumulative layer normalization along with dropping future information in the self-attention network, the audio separator network is able to perform speech separation in a streaming manner. Another direction for future work is evaluating the performance measured by word error rate when combined with automatic speech recognition (ASR) systems. Studies in this direction would include training the audio separator network and an ASR network jointly. Further to that, many other ways to apply the self-attention mechanism for speech separation remain unexplored. This study should open an avenue to explore a wide variety of ways to utilize self-attention for speech separation and recognition of overlapping speech.

REFERENCES

[1] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 31-35, 2016.

[2] Z. Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2018.

[3] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 246-250, 2017.

[4] J. Wang et al., "Deep extractor network for target speaker recovery from single channel speech mixtures," 19th Annual Conference of the International Speech Communication Association (Interspeech), pp. 307-311, 2018.

[5] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 241-245, 2017.

[6] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing 25, no. 10, pp. 1901-1913, 2017.

[7] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 696-700, 2018

[8] 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 27, no. 8, pp. 1256-1266, 2019.

[9] Z. Q. Shi et al., "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," 26th International Conference on MultiMedia Modeling (MMM), pp. 653-665, 2020.

[10] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998-6008, 2017.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231, 2017.

[12] S. X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 171–178, 2016.

[13] F A R. R. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-Based Models for Text-Dependent Speaker Verification," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5359-5363, 2018.

[14] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," 18th Annual Conference of the International Speech Communication Association (Interspeech), pp. 1517–1521, 2017.

[15] D. M. Wang, Z. Chen, and T. Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," arXiv preprint arXiv:2004.13670, 2020.

[16] Y. L. Jin, C. J. Tang, Q. H. Liu, and Y. Wang, "Multi-Head Self-Attention Based Deep Clustering for Single-Channel Speech Separation," IEEE Access, vol. 8, pp. 100013-100021, 2020.

[17] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," Linguistic Data Consortium, 1993.

[18] D. Snyder, G. G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise corpus," arXiv preprint arXiv:1510.08484, 2015.

[19] F. Bahmaninezhad et al., "A comprehensive study of speech separation: spectrogram vs waveform separation," 20th Annual Conference of the International Speech Communication Association (Interspeech), pp. 574-4578, 2019.

[20] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6359-6363, 2020.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," The Journal of the Acoustical Society of America 65, no. 4, pp. 943-950, 1979.