Improvement of Spatial Ambiguity in Multi-Channel Speech Separation Using Channel Attention

Qian-Bei Hong*, Chung-Hsien Wu*†, Thanh Binh Nguyen†, and Hsin-Min Wang*
*Graduate Program of Multimedia Systems and Intelligent Computing,
National Cheng Kung University and Academia Sinica, Tainan, Taiwan
†Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
E-mail: {qbhong75, chunghsienwu, ntbbinh.ncku}@gmail.com, whm@iis.sinica.edu.tw

Abstract— **Multi-channel speech separation has been successfully applied in a complex real-world environment such as the far-field condition. The common solution to deal with the far-field condition is using a multi-channel signal captured by a structured microphone array and leveraging the inner difference between channels to enhance the speech separation performance. The spatial feature has been widely used in recent speech separation research. This feature appears to be insufficient when the location information becomes ambiguous. This is known as the spatial ambiguity problem. In order to deal with the spatial ambiguity problem, this study proposes an attention mechanism for the Temporal-Spatial Neural Filter (TSNF), in which the channel attention on merged features and the feature map of 1D convolution block in the temporal convolution network is proposed. The proposed method is evaluated on the multi-channel reverberant dataset which is built based on the WSJ0-2mix dataset. The dataset is simulated in the real-environment room by using the Room Impulse Response generator. In the experimental results, the proposed methods produced the SI-SNR improvement of about 1.2dB in close speakers' case, while a small decrease of 0.1dB in other cases.**

I. INTRODUCTION

The cocktail party problem [1] is the problem about real-world environments in which the sound is from several sources such as background noise, music from instruments, etc. In the cocktail party environment, human speech is overlapped and it is hard for a machine to focus the attention on only one of these sources [2]. Speech separation is considered as a task to extract or separate speech signals from a mixture of speech from several speakers and noises.

In the past few years, deep learning techniques have been introduced to solve the speech separation/enhancement problem and achieved promising results [3-4]. Most research focused on monaural speech, which used single channel for speech processing [5-6]. Recent research has proven that using information from the multi-channel signal could improve the accuracy of single channel systems and was more suitable for real-world conditions [7-8].

Multi-channel speech separation generally uses a microphone array to capture speech sound and leverage the difference of multi-channel signal from these microphones to obtain useful features for speech separation purposes. Spatial features such as interaural phase difference (IPD) [9] was quite popular in previous speech separation tasks and achieved good separation results. On the other hand, some studies [10-11] indicated that combining spatial features with spectral features such as logarithm power spectra (LPS) can provide effective information and improve speech separation performance. Due to the efficiency of the IPD feature in separation tasks, some research tried to explore and improve the limitation of IPD feature. Wang et al. [11] mentioned that deep clustering was conducted independently within each frequency band because of the IPD ambiguity. Chen et al. [12] showed that the disadvantage of the estimated phase difference is the potential phase wrapping in high frequency when using a small-space microphone array. However, when the direction of speech sounds is too close, especially the far-field overlapped speech, the spatial ambiguity will severely decrease the performance of speech separation using the spatial features.

As the spatial ambiguity has an effect on separation performance, one of the methods to deal with this problem is to consider the contribution of different subbands of spectral features and thus choose subband embedding selectively to achieve a better separation performance. In [12], the speech separation model used two subbands of different frequencies to consider the different effects on IPD features. However, when the distance between microphones is different, the experiment needs to employ different subband division to ensure the result. Therefore, instead of finding frequency subbands manually for different environmental condition, it is beneficial to construct a channel attention approach for automatic frequency bin selection.

II. SPEECH SEPARATION

Compared to the Conv-TasNet [6], in the proposed model, a channel attention approach is proposed to decide how much of the input is produced for transformation. We place channel attention to the bottleneck block and every convolution block in the Temporal-Spatial Neural Filter (TSNF) [13], inherited from Conv-TasNet. The proposed system framework is shown in Fig. 1.

A. Channel Attention Approach

Channel attention was adopted in the convolutional block attention module introduced in [14], in which the attention

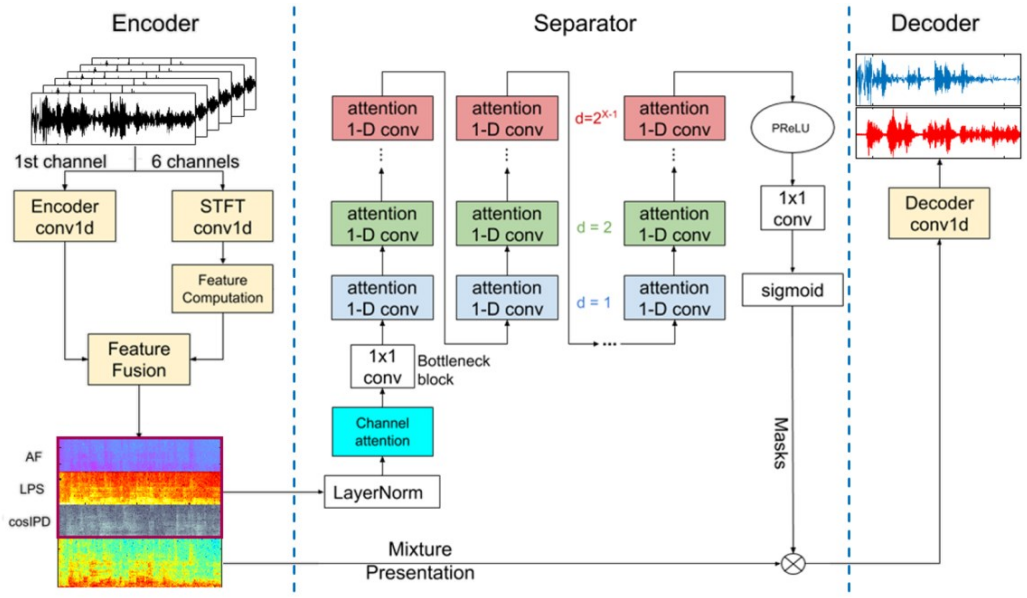

Fig. 1 The proposed system framework.

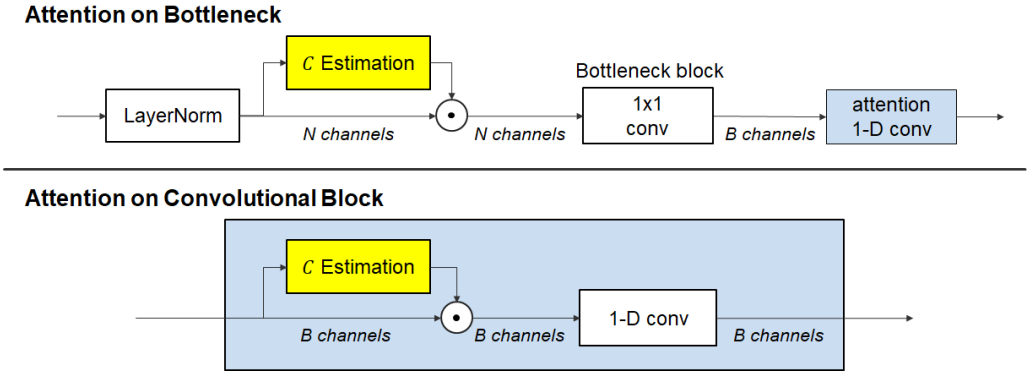

Fig. 2 Channel attention for the separator.

mechanism is used on both channel and spatial dimensions of a feature map in a convolution module. In this study, we apply the attention mechanism on the frequency dimension of features and the channel of the feature map in the convolution module. The input features (N -dimensional channels $\times K$ segments) are applied to the channel attention and the obtained output is used as the refined features for the separation module, which can be formulated as follows.

$$
\begin{aligned}
P_{avg} &= Avgpool(F) \\
P_{max} &= Maxpool(F) \\
C &= \sigma\left(H_2\left(H_1 P_{avg}\right) + H_2\left(H_1 P_{max}\right)\right) \\
\tilde{F} &= C \odot F
\end{aligned} \tag{1}
$$

where $\odot$ denotes element-wise multiplication, $\{F, \tilde{F}\} \in \mathbb{R}^{N\times K}$ denotes the input features and the corresponding refined features. In channel attention, $P_{avg} \in \mathbb{R}^{N\times 1}$ is the average pooling result of F, $P_{max} \in \mathbb{R}^{N\times 1}$ is the max pooling result of F, $H_1 \in \mathbb{R}^{(N/r)\times N}$ and $H_2 \in \mathbb{R}^{N\times(N/r)}$ are the shared Multi-Layer Perceptron (MLP) weights for the feature transformation , r denotes the ratio of dimensionality reduction of bottleneck feature, and σ denotes the sigmoid function. Two pooling outputs are fed into a shared MLP with one hidden layer. The hidden layer size is decided on the ratio r of the MLP. After that, the two output feature vectors corresponding to the two pooling outputs are merged together to form the channel attention vector C.

B. TSNF with Channel Attention

In this study, the channel attention mechanism was used on two parts as shown in Fig. 2. First, the weight from channel attention is applied to the encoder output before feeding to the bottleneck block, and the number of input channels is decreased from N channels to B channels after the bottleneck block transformation. Second, channel attention is also applied

TABLE I
HYPERPARAMETER DEFINITION FOR THE PROPOSED SYSTEM

Name	Description	Value
N	Number of filters in autoencoder	256
L	Length of the filters (in samples)	40
B	Number of channels in bottlenecks and the residual paths' 1x1-conv blocks	256
H	Number of channels in convolutional blocks	512
P	Kernel size in convolutional blocks	3
X	Number of convolutional blocks in each repeat	4
R	Number of repeats	4
n_fft	Number of FFT points	64

to every convolution block. The input and output sizes of channel attention are fixed to B channels for each convolutional block.

III. Experimental Results

A. Data Preparation

The dataset is created by following the description from the baseline TSNF [13], and the multi-channel dataset type is inherited from a single channel WSJ0-2mix, which was firstly introduced by the work in deep clustering methods [15].

WSJ0-2mix is generated based on the World Street Journal (WSJ0) corpus [16], which consists of read speech data in a corpus of WSJ news text. WSJ0-2mix consists of a training set of 30 hours (20,000 utterances), a validation set of 10 hours (5,000 utterances), and a test set of 5 hours (3,000 utterances). Every utterance is the two-speakers' mixture generated by randomly selecting two-speakers' signals from the WSJ0 corpus; the two signals are mixed at the signal-to-noise ratios from 0dB to 5dB.

B. Experimental Setup

The proposed system architecture is based on the Conv-TasNet and the hyperparameters are illustrated in TABLE I.

The STFT including 64-point Fast Fourier Transform (FFT), 2.5ms window length, and 1.25ms hop size is adopted to compute frequency domain features. After transformation, the 33 frequency values (including two DC values) are selected as input features, which corresponds to 0Hz to 4kHz for the signals with 8kHz sampling rate. The other features including LPS, cosIPD, sinIPD and angle feature (AF) [13] are used for mask estimation in this study, and the total 495 channels are shown in TABLE II.

In the following experiments, the scale-invariant source-to-noise ratio (SI-SNR) [6], are used for evaluation.

C. Spatial Ambiguity Problem

TABLE II
THE INPUT FEATURES FOR SPEECH SEPARATION

Features	Description	# Freq.
LPS	Spectrogram of reference channel	33
IPD	Phase difference from 6 microphone pairs (cosIPD and sinIPD)	6×33×2
AF	AF of target and interference speaker	2×33
Total		**495**

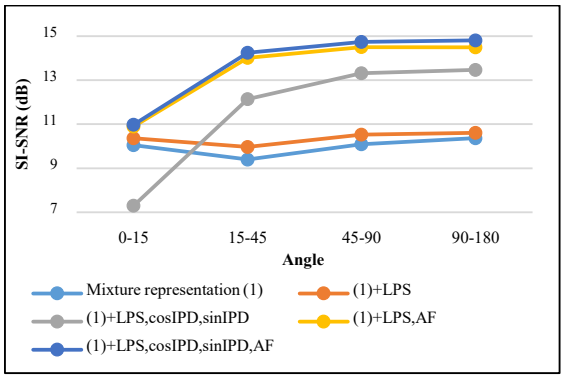

Fig.3 Speech separation performance as a function of features and angles.

In this experiment, the baseline systems [6][13] were implemented for spatial ambiguity analysis in multi-channel speech separation. For feature fusion, each feature was added individually to the model in order to evaluate its contribution in different feature setups. The model was evaluated on different angle ranges including 0º-15º, 15º-45º, 45º-90º and 90º-180º, and the corresponding percentages in the dataset were 16%, 29%, 26% and 29%, respectively.

The experimental results are shown in Fig. 3. The first model with mixture representation was the original single-channel Conv-TasNet trained on the multi-channel reverberant dataset, in which the first signal of the 6-channels waveforms was used to extract the temporal features. As Conv-TasNet was based on single-channel speech separation, the result showed the performance was irrelevant to angle, and achieved the average SI-SNR of 10.0dB for all angles. Furthermore, the LPS feature contributed equally to all angle difference ranges and increased the average SI-SNR of 0.4dB. Particularly, the spatial relevant features (including IPD and AF) obtained the performance improvement significantly on the angle larger than 15º, but in the case of 0º-15º, the spatial ambiguity undermined the speech separation performance, and the performance of adding IPD features was even worse than Conv-TasNet. In view of this, solving the spatial ambiguity problem for speech separation in real-life environments is very important, especially the far-field condition (small angle difference).

D. Evaluation on Different Ratio r of Channel Attention

In the proposed method, channel attention was used to consider the contributions of different frequency values to solve the spatial ambiguity problem. Firstly, the ratio r of channel attention needed to evaluate what dimension was reliable, thus we evaluated the performance on four different ratios (including 8, 16, 24 and 32). Evaluating channel attention in the case of all features (including mixture representation, LPS, IPD and AF), the improvement results of SI-SNR were shown in Fig. 4. Although we noticed that when the angle was larger than 15º, the SI-SNR was slightly degraded by about 0.1dB. But in order to improve the spatial ambiguity problem, a small performance drop in large angle conditions is acceptable. Overall, the experimental results achieved slightly difference between different ratios r. In the comparison of

TABLE III
PERFORMANCES OF CHANNEL ATTENTION WITH DIFFERENT FEATURES

Feature Setup	Channel Attention	SI-SNRi (dB)				
		0°-15°	15°-45°	45°-90°	90°-180°	Avg
Mixture representation (1)	-	10.1	9.4	10.1	10.4	10.0
(1) + LPS	No	10.3	10.0	10.6	10.8	10.4
	Yes	**11.0**	**10.6**	**11.2**	**11.2**	**11.0**
(1) + cosIPD	No	6.7	11.5	12.3	11.3	10.9
	Yes	**10.7**	**12.9**	**13.4**	**12.8**	**12.7**
(1) + cosIPD,sinIPD	No	7.2	12.0	12.9	13.1	11.8
	Yes	**9.4**	**12.3**	**13.2**	**13.2**	**12.4**
(1) + cosIPD,AF	No	10.2	13.9	14.5	14.6	13.7
	Yes	**11.9**	**13.9**	**14.5**	**14.6**	**14.0**
(1) + cosIPD,sinIPD,AF	No	10.7	**13.9**	**14.5**	**14.6**	13.8
	Yes	**11.9**	13.9	14.5	14.6	**13.9**
(1) + LPS,cosIPD,sinIPD,AF	No	11.0	**14.3**	**14.7**	**14.8**	14.1
	Yes	**12.2**	14.1	14.6	14.7	**14.1**

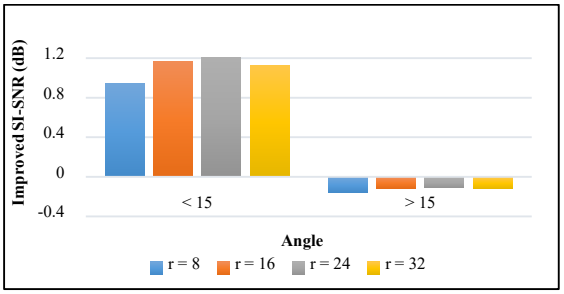

Fig.4 Evaluation on different ratio r of channel attention in the case of all features.

different angle ranges, the model with a ratio of 24 achieved the most performance improvement on small angle differences (<15°) and the least performance degradation on large angle differences (>15°). Therefore, the ratio of 24 was used in the following channel attention experiments.

E. Channel Attention Applied on Different Features

According to the description in Section III-C, even if different features were combined as input based on the variety of information to improve the speech separation performance, the spatial ambiguity problem still occurred in any case. Thus, this experiment compared the results of channel attention applied on different features, as shown in TABLE III. First, even though the LPS was irrelevant to angle, the performance of using channel attention were increased significantly in the all angles. This result proved that the frequency attention mechanism had a positive effect on frequency domain features. Second, the channel attention significantly improved the drawback of IPD in small angle ranges (0°-15°). In the case of adding cosIPD only, the SI-SNR was greatly increased from 6.7dB to 10.7dB. Finally, in the case of all features, channel attention could further improve the spatial ambiguity problem, thereby increase the SI-SNR from 11.0dB to 12.2dB. TABLE IV reported the performances of the proposed method with other state-of-the-art (SOTA) methods on far-field WSJ0-2mix. Our proposed method effectively avoids performance degradation caused by the spatial ambiguity. The results confirm again that the our proposed TSNF with channel attention outperformed the SOTA methods in far-field condition.

TABLE IV
PERFORMANCES OF TARGET SPEECH SEPARATION ON FAR-FIELD WSJ0-2MIX

Method	SI-SNRi (dB)		
	<15°	>15°	Avg
Mixture representation	10.1	10.0	10.0
Multi-band PIT [12]	8.3	11.7	11.2
ICD based MCSS [17]	8.1	13.2	12.4
Neural Spatial Filter [8]	4.9	10.0	9.1
Temporal-Spatial Neural Filter [13]	10.8	14.0	13.5
Proposed	**12.2**	**14.5**	**14.1**

IV. CONCLUSIONS

The main contribution of this study is to design an attention mechanism to deal with the spatial ambiguity problem in multi-channel speech separation tasks. We apply the channel attention into the convolutional process in TSNF by considering the contribution of every frequency band of input features and the channel of the feature map in the convolution module. In the experiments, the proposed channel attention system achieved the best SI-SNR of 14.1dB for all angle ranges; especially, the performance obtained significant improvement in close speakers' case. Therefore, this study effectively solved the spatial ambiguity problem and increased the reliable in far-field condition. In all cases, we assume that the speech is consistent, meaning that there is no interrupt or change in the number of speakers in one utterance. Thus, many scenarios still need to be further considered for practical applications in the future work.

ACKNOWLEDGMENT

This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project.

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875-1902, September 2005.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'CHiME'speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[3] Y.-M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40-63, January 2018.

[4] J.-C. Wang, Y.-S. Lee, C.-H. Lin, S.-F. Wang, C.-H. Shih, and C.-H. Wu, "Compressive sensing-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2122-2131, November 2016.

[5] C. Xu, X. Xiao, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid LSTM," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 6-10, April 2018.

[6] 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, May 2019.

[7] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Virtual Barcelona, pp. 6394-6398, May 2020.

[8] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: target speaker speech separation assisted with directional information," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association,* Graz, Austria, pp. 4290-4294, September 2019.

[9] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, pp. 558-565, December 2018.

[10] Z.-Q. Wang and D. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Proceedings INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, pp. 2718-2722, September 2018.

[11] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457-468, November 2018.

[12] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band PIT and model integration for improved multi-channel speech separation," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 705-709, May 2019.

[13] R. Gu and Y. Zou, "Temporal-spatial neural filter: direction informed end-to-end multi-channel target speech separation," *arXiv preprint arXiv:2001.00391*, 2020.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, pp. 3-19, September 2018.

[15] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 31-35, March 2016.

[16] L. D. Consortium, "CSR-I (WSJ0) Complete," https://catalog.ldc.upenn.edu/LDC93S6A (accessed 2020).

[17] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Virtual Barcelona, pp. 7319-7323, May 2020.