参赛学生姓名：白旭鹏 叶德言

中学：北京市第一0一中学

省份：北京市

国家/地区：中国

指导老师姓名：孔祥宇 周宇辰

指导老师单位：北京信息科技大学

北京市第一0一中学

论文题目：Array-Agnostic Multi-Channel Speech Separation

Abstract

With the surge of robotics, video conferencing, smart vehicles, smart home appliances, metaverse, etc., providing high-quality speech signals and accurate recognition have become essential to our daily life, where speech separation plays a key role in the complex scenarios. Microphone array enable high quality speech separation with spatial information well captured. However, most of the existing methods highly depend on the geometric parameters of specific type of microphone array including shape, size and number of microphones. Therefore, the model trained by using the dataset from specific type of array can't be applied to process data from other type of array with different geometry parameters, blocking the application in complex real-life scenarios. In this project. we propose an array-agnostic multi-channel speech separation method for microphone generalization by leveraging iterative joint full-band and sub-band learning with self-attention based multi-channel encoding. Our method employs comprehensive analysis of conjoint full-band and sub-band models to both capture the global spectral context and the long-distance crossband dependencies, and modeling signal stationarity and attending the local spectral pattern. Furthermore, we enable iterative learning to improve the performance. The experiment results demonstrate the proposed method can adapt to various type of microphone array and achieve higher performance than baseline models, while consuming less time and computing resources and training. We hope it can be utilized as a novel approach in wider scope of multi-channel relevant tasks for more effective utilization of spatial and frequency domain information.

Keywords—speech separation, microphone array generalization, iterative sub-band and full-band learning, self-attention

Table of Contents

Array-Agnostic Multi-Channel Speech Separation

Xupeng Bai
Beijing 101 High School

Deyan Ye
Beijing 101 High School

1 INTRODUCTION

With the recent surge of video conferencing, metaverse, smart vehicles, smart home appliances, service robots, etc., providing high-quality speech signals and accurate recognition have become essential to our daily life. People join the online meetings remotely from diverse clients including laptop, mobile phone, conferencing station, etc. While the acoustic scenario is more complex in the mixed audio stream because of the nature of sound. In real life, it is difficult to find a completely undisturbed environment, and some sound sources will disturb the speaker from time to time with interfering speech and noise, and there are also the cases that two or more people speak simultaneously. In such cases, speech separation is critical to improve quality of voice to ensure the efficiency of the voice interaction with diverse client devices. The similar requirements are also raised widely from the area of video conferencing, voice based control of smart vehicles, smart home appliances, service robots, etc.

In traditional single-channel speech processing, classic challenges include the "cocktail party" problem, where a target speaker is followed while other speakers and background noise are ignored, with reverberation removal by which the interference of additive noise from other sources and reverberation from surface reflections, localization of multiple sources, and source separation[30], as illustrated in Figure 1. Microphone array based speech separation has be shown more effective than using single microphone, because it can provide additional spatial hints of speakers' locations and the corresponding speech components, hence this can ease challenging separation task significantly. Microphone array has been widely applied in modern devices, which consists of a set of microphones positioned in a way that the spatial information is well captured.

The choice of microphone array configuration depends on the specific use case and desired objectives. This selection process ensures better utilization of spatial information obtained from the application, leading to improved performance of the microphone array. The linear array is used primarily employed on mobile phones and headphones, particularly in models with two microphone units，positioned at the top and bottom of the device. The Planar microphone array, which includes various two-dimension arrangement and number of microphone units such as four microphone circular array, eight microphone circular array, are widely used in smart speakers for oral communication and voice interaction. More microphones result in finer spatial partitioning and better recognition of distant scenes.

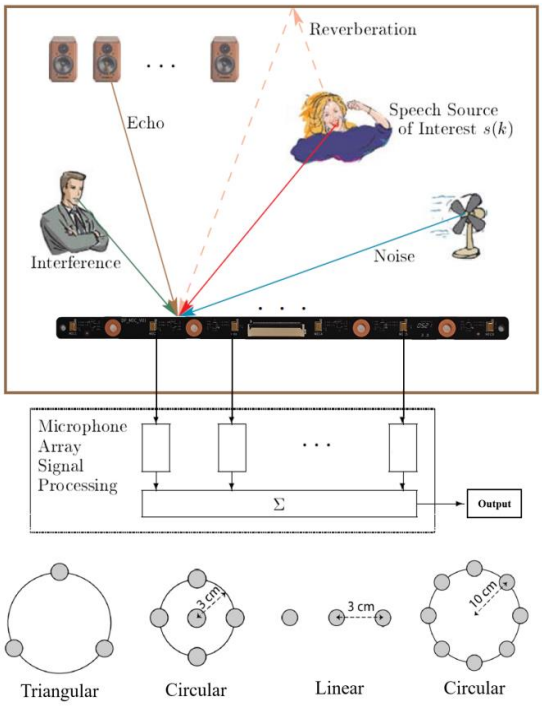

Figure 1. Microphone arrays with various shapes and geometric properties support tasks of multi-channel localization, denoising, reverberation removal, speech enhancement and separation, etc.

A lot of work on microphone array based speech separation achieves good results, including signal processing methods based on Blind Source Separation (BSS)[1] and beam forming[2], as well as deep learning based methods such as NBC[18], SpatialNet[29] . But most of them can only support array with fixed number of microphones, depending on the number of channels and spatial information correlated with number and geometric parameters of the microphone array. They are difficult to handle the tasks to apply the model trained via data from specific type of microphone array to those via others. But when the conditions and situations of application changes, the array need to be adjusted or transform the physical shape to meet the various needs, or even with mixed usage of different types of microphone array, it obviously requires lots of efforts and resources. For instance, the model trained on array with 8 microphones can't be directly applied to the one with 4 or 16 microphones. Therefore, the array-agnostic method will play an important role in the field of speech separation. Speech separation

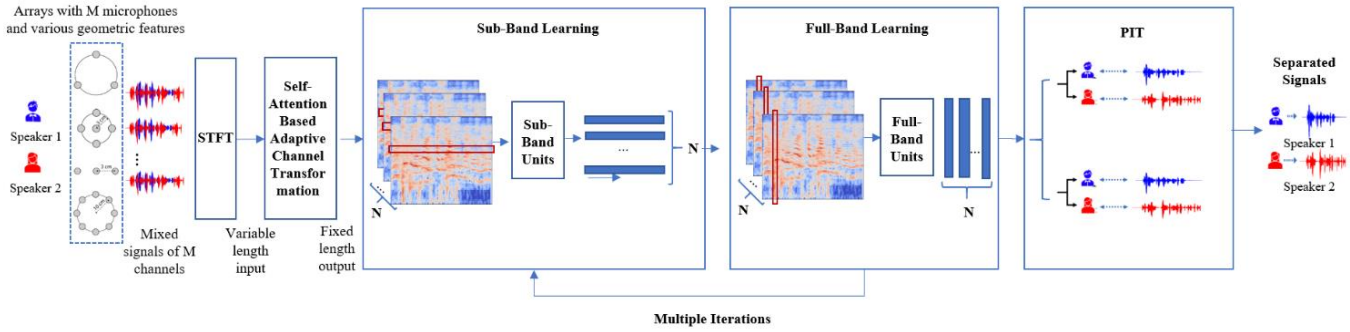

Figure 2. The propose method for array agnostic speech separation

model that can be trained once and generalized to multiple types of microphone arrays become critical for application of speech separation in more complex real-life scenarios. There are very few prior efforts for microphone array agnostic speech separation, such as VarArray[34], a computationally expensive conformer based method relying on large volume dataset, without code public available for reproduction. There are also a few attempts in related domain, such as array geometry agnostic multi-channel personalized speech enhancement[31], but it is not easy to apply such technologies to speech separation because they only focus on speech enhancement or dereverberation that are easier than speech separation from perspective of complexity and difficulty of the task.

Motivated by the above intuition, we propose an array-agnostic multi-channel speech separation method for microphone generalization by leveraging iterative joint full-band and sub-band learning with self-attention based multi-channel encoding. Our approach involves a thorough analysis of combined full-band and sub-band models to capture both the overall spectral context and the correlations across different frequency bands， while also considering signal stationarity and local spectral patterns. Moreover， we facilitate iterative learning to enhance the performance. The experiment results demonstrate the proposed method can adapt to various type of microphone array and achieve higher performance than baseline models，while consuming less time and computing resources and training. We hope it can be utilized as a novel approach in wider scope of multi-channel relevant tasks for more effective utilization of spatial and frequency domain information.

Our main contributions can be summarized as follows:

- Iterative joint full-band and sub-band learning to better learn the global and local speech feature, which was rarely investigated before.
- Self-attention to support variable length input from different types of microphone array
- The experiment results show that our methods have a strong microphone array generalization capability compared to baseline models
-

2 RELATED WORK

In the conventional single channel speech separation field, the typical methods are based on Blind Source Separation (BSS)[1] and beam forming[2]. Without dedicated enhancement for speech, conventional signal processing method could not achieve ideal result in complex scenarios. Many speech processing models have gradually shifted from signal processing-based, decomposition, and rule-based methods to learning-based approaches driven by data, with deep learning being a prominent representative. Researchers found it is more efficient and accurate to use deep learning approaches to the single channel speech separation. From the perspective of signal representation, the learning based methods can be categorized into two areas, speech separation based on frequency-domain and speech separation based on time-domain.

Progress in the field of frequency-domain based speech separation are widely used, by utilizing Short-Time Fourier Transform (STFT)[3] to convert speech signals into spectrograms, which leverages the slow time-varying nature of speech signals. Masking method has been proposed, using an Ideal Binary Mask (IBM)[4] to generate masks. Due to the uncertainty of separation results, previous deep learning methods could only perform speaker-dependent speech separation, where the training and testing speech data were from the same speaker. In order to explore models for speaker-independent speech separation, Hershey developed Deep Clustering (DC)[5], where two-dimensional features are mapped to a three-dimensional space to make the input mixed features more discriminative. Nevertheless，this approach is not end-to-end and lacks high computational efficiency. Additionally, the objective function of DC method compares the difference between mixed embeddings and target embeddings, rather than comparing the speech itself, which can degrade the performance of speech separation. To achieve end-to-end speech separation, Yu proposed the permanent invariant training (PIT)[26]approach, which helps address the issue of uncertain output order of neural networks. Subsequently, researchers have improved upon the original PIT method and proposed algorithms such as full-band PIT[6], which further consider the order of frequencies. Additionally, sub-band methods were also used to assistant the process, some methods have been proposed to address various challenges in speech separation. In 2021, Xiang Hao and

Xiangdong Su et al. proposed a single-channel voice enhancement model that combines Full-band and Sub-band, called Full-Sub Net[7]. This model combines the advantages of Full-band and Sub-band to achieve excellent real-time voice enhancement. However, it is only suitable for single channel, which cannot be widely used in life scenes for speech separation or handle multiple speaker tasks at the same time.

In parallel with frequency-domain methods, time-frequency-domain based speech separation methods have also been developed. Luo et al. proposed TasNet[8] utilizing LSTM networks. In subsequent research, they introduced Convolutional Time-domain Audio Separation Network (Conv-TasNet)[9], which employs a fully convolutional network model. Conv-TasNet utilizes encoded features as input, uses the TCN structure as the separation module, and finally performs multiplication with the encoder's output to obtain the final separation features. However, Conv-TasNet can only focus on information related to the length of segmented speech, failing to integrate information from entire sentences. To cope this problem, Luo et al. developed the Dual-Path Recurrent Neural Network (DPRNN)[10]. This model is able to capture global information from longer segments of speech. In response to phase mismatch issues and the suboptimal performance of speech separation in the frequency domain, where separated speech still contains interference sources, Cunhang Fan et al. proposed an end-to-end post-filtering method called End-to-End Post-Filter (E2EPF)[11]. This method allows the separated features to dynamically focus on the speech being separated.

When multiple microphones are present， spatial information can be utilized along with the spectral pattern of speech. Traditional approaches include Inter-Channel Phase Difference (IPD)[12], automatically learned spatial features[13], directly predict the spatial filters[14][15], or to estimate the filter parameters using the separated speech signals of deep learning based methods, such as in[16][17]. In recent years, models with better performance continues to emerge. NBC[18] and NBC2[32] proposed multi-channel speech separation methods with narrow-band Conformer

A number of efforts have been made for microphone array generalization. [23] designed a system that can generalize across various array geometries (e.g., linear, circle, and ad-hoc array) and provide reliable recognition performance in a wide range of real-world settings, even under adverse acoustic conditions. [20]incorporated a set of fixed beamformers and an attention network to select the dominant beam using the target speaker embedding. The speaker embeddings have also been employed for speech separation. A perceptually motivated PSE model with low complexity was proposed in[24]. [23]introduced two real-time PSE models and tested with reverberant target speech corrupted by both noise and interfering speech. In[15], the authors proposed a Transform-Average-Concatenate (TAC) layer for multi-channel speech separation that was invariant to the order of the microphones. An inter-channel processing layer based on self-attention was proposed in [23]. The work by [21] used deep symmetric sets layers based on a Siamese network for speech dereverberation. [22]proposed a model to process a variable number of microphone pairs for speech enhancement. Hassan et al. improved single-channel Perceptual Sound Equalization (PSE) to meet the demands of multi-channel applications. They attempted microphone array migration using a method called the virtual microphone generation algorithm[31]. However， the aforementioned approaches may result in the loss of spatial information and the inefficient use of available resources. In the existing methods of microphone array-based speech separation, the spatial information can't be fully utilized and preserved during modeling, and the method needs to be improved.

3 OUR METHOD

We start from generating datasets Room Impulse Response(RIR) dataset, and design an array-agnostic multi-channel speech separation method based on iterative joint full-band and sub-band learning, self-attention based adaptive multi-channel encoding and PIT.

3.1 Dataset Generation

For training and evaluation of the microphone generalization capability of the models, we take great efforts to generate the large datasets from different type of microphone array. The training and test data are synthesized from two sources, clean speech dataset as the spatialized version of the WSJ0-2mix[25], and RIR dataset generated by using gpuRiR[28].

WSJ0-2mix is a speech recognition dataset of speech mixtures using utterances from the Wall Street Journal (WSJ0) corpus. In this experiment, the speech pairs are overlapped in the manner used in [10], in which the tail part of one signal is overlapped with the head part of the other signal. The overlap ratio is uniformly sampled in the range of [10%, 100%]. The resulting utterances mixed are set to 4 seconds in all the items in the dataset .

The room impulse responses are generated through simulation with the gpuRIR package， which utilizes the GPU for implementing the image method. We expand the RIR generation process, and create diversified microphone array layouts. By meticulously modifying the implementation to generate RIR data for various configurations of microphone, we successfully constructed rich datasets to support more complex and practically

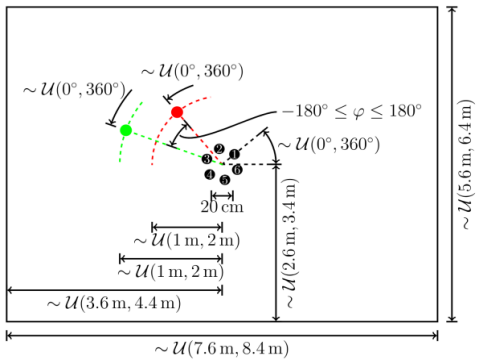

Figure 3. Target scenario of RIR dataset generation [28]

relevant to model training and testing scenarios. The typical scenario of RIR generation is shown in Figure 3.

During the synthesis process, the number of microphones in the RIRs determines the number of microphones in the synthesized speech generated by wsj0mix and RIRs, independent of the clean speech from wsj0mix. We modified portions of the code of gupRIR to produce RIRs with varying numbers of microphones. In the specified setting， there are two speakers acting as sources of sound，with the room's dimensions(length， width， and height)randomly generated between 3-8 meters for length and width， and 3-4 meters for height. The reverberation time is set randomly between 0.1 and 1 seconds. At first， the dimensions of the room and the reverberation time are established to confirm their accuracy. Following this, the position and angle of the microphone array are randomly chosen, ensuring that the microphones are evenly distributed on the array. Then, a sound source location is randomly selected, and the position of the second sound source is calculated based on its angular relationship with the first speaker and the center of the microphone array, ensuring its legality. This approach achieves randomization in the position and angle of the microphone array, enhancing the robustness of the trained model.

3.2 Sub-Band Learning

We enable sub-band learning for speech separation, and take NBSS as the baseline model. The network architecture is shown in Figure 4 and the data flow is shown in Figure 5. During sub-band learning, frequency wise temporal evolution of the STFT magnitude is informative due to the non-stationary nature of speech against the stationary of noise. It can capture longer distance time dependent information and have better frequency continuity. As well, less parameters training data and training time will be required.

The input speech signals are represented as:

$$\boldsymbol{X'}_{\boldsymbol{m}} \in \mathbb{R}^{T\times F} \quad (1)$$

where $m \in \{1, ..., M\}$ denotes microphone channel. $\boldsymbol{X_m}$ denotes the $\boldsymbol{m^{th}}$ speech signal with time T and frequency F. By utilizing STFT, we extract the multi-channel signals in the frequency domain:

$$X_{f,t}^{m} = \sum_{n=1}^{N} Y_{f,t}^{n,m} \quad (2)$$

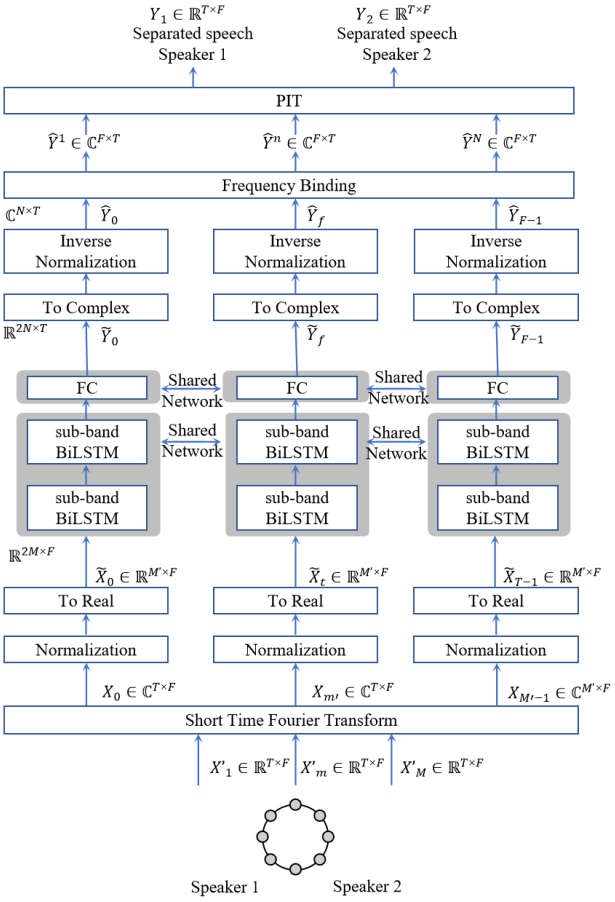

Figure 4. The network architecture of sub-band learning for speech separation

The indices f， t， and n represent the frequency， time frame， and speaker， where, f $\in$ {0, ..., $F-1$}, t $\in$ {1, ..., T}， and $n \in$ {1, ...N} respectively. $\boldsymbol{X_{f,t}^{m}}$ and $\boldsymbol{Y_{f,t}^{n,m}}$ are the complex-valued STFT coefficients of the microphone signals and reverberant spatial image of speech sources. Our target is to recover the reverberant spatial image of multiple speakers at a reference channel, e.g. we have $\boldsymbol{Y_{f,t}^{n,r}}$, where r represents reference channel index.

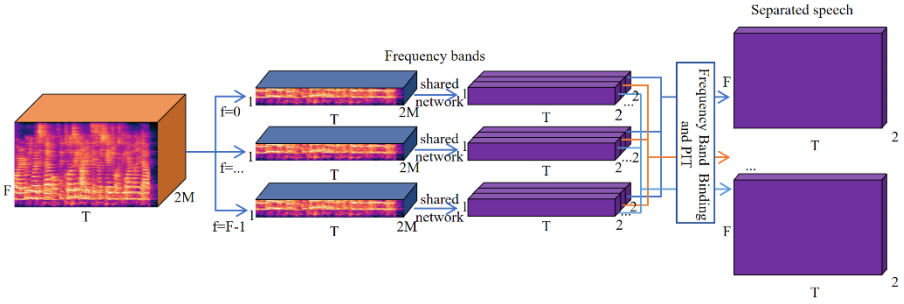

Figure 5. The data flow of sub-band learning based speech separation

The network takes the STFT coefficients of multi-channel mixture signals directly as the input feature. For one TF bin, the STFT coefficients are concatenated along channels as $X_{f,t}$:

$$X_{f,t} = [X^1_{f,t}, \ldots, X^M_{f,t}]^T \in \mathbb{C}^{M\times 1} \quad (3)$$

Considering the $X_{f,t}$ from the whole time-axis for one frequency band which is also the input sequence of the sub-band network, there is:

$$X_f = (X_{f,1}, \ldots, X^M_{f,T}) \in \mathbb{C}^{M\times T} \quad (4)$$

As the network can only process real numbers, the complex-valued input sequence should be converted into real-valued sequence. This is done by simply replacing the complex number with its real and imaginary parts. Through normalization that is performed as $X_f/\bar{X}_f$ and realization, the outcome real-valued sequence is denoted by $\tilde{X}_t \in \mathbb{C}^{2M\times F}$. The output here is $\tilde{Y}_f \in \mathbb{R}^{2N\times T}$ that consists of the real part and imaginary part of the (normalized) spatial image at the reference channel of N speakers. $\tilde{Y}_f \in \mathbb{R}^{2N\times T}$ is the input of PIT that will be mentioned in section 3.5.

Nevertheless，sub-band learning is unable to capture the overall spectral pattern and leverage long-range cross-band dependencies. In particular，for sub-band signals with an extremely low signal-to-noise ratio(SNR)，the sub-band model may struggle to restore the clean speech，but it may become achievable with the assistance of full-band dependencies. On the other hand, the full-band model is helpful to preserve speaker information, trained to learn the regression between the high dimensional input and output, lacking a mechanism dedicated to the sub-band information, such as the signal stationarity.

In model described above, the sampling rate is 16 kHz. STFT is performed by using a hanning window of length 512 samples (32ms) with a hop size of 256 samples. The numbers of hidden units of the first and second BiLSTM layers are set to 256 and 128 respectively. The dimensions of the input and output of the full connection layer are 256 and 4 respectively.

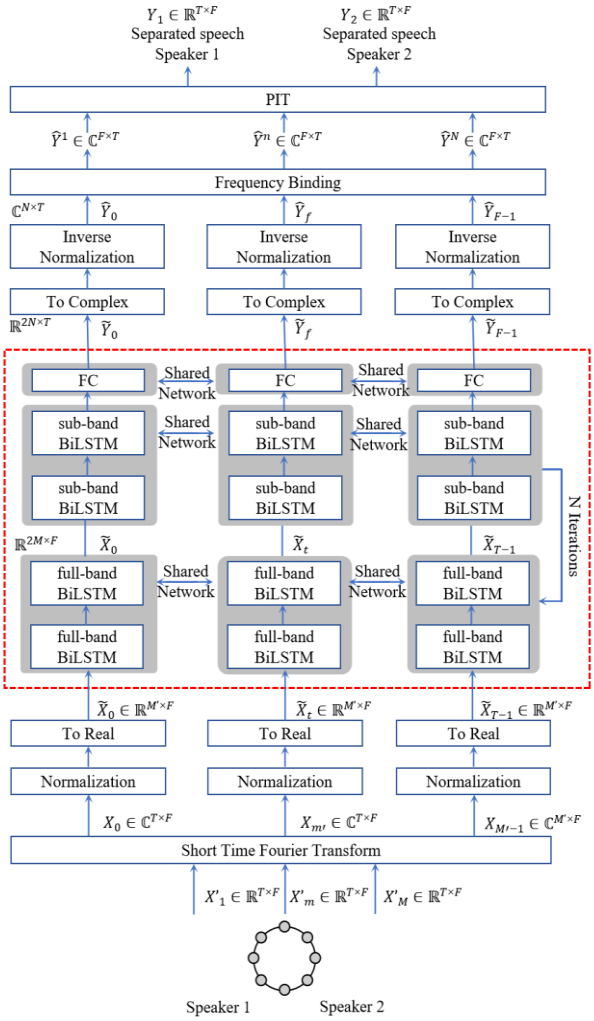

Figure 6. The network architecture of iterative joint full-band and sub-band learning

3.3 Iterative Joint Full-Band and Sub-Band Learning

In the full-band learning, where each frequency band goes under the STFT process, it is capable of capturing long term global information, as well as functions effectively on capturing cross-band harmonic structure and long distance frequency dependence.

To insert full-band learning module before sub-band learning blocks, we need to adjust the format in form of dimensions of the

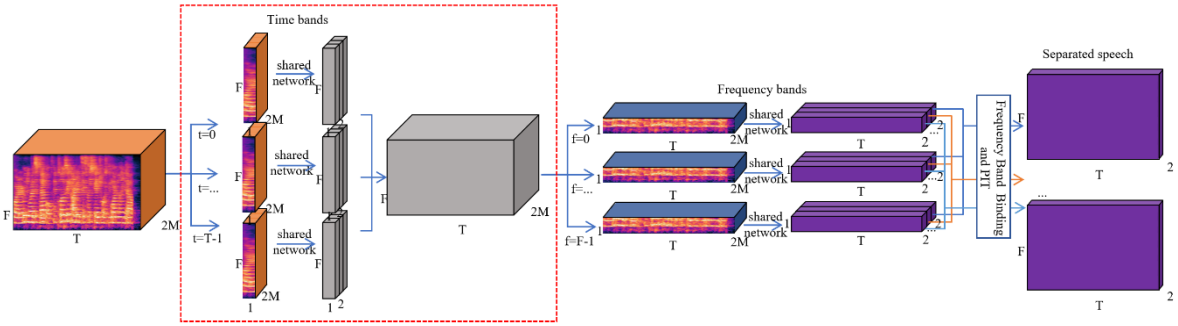

Figure 7. The data flow of iterative joint full-band and sub-band learning

signal after STFT and full-band learning. To adjust the dimensions of tensor as input to the network, we perform STFT on frequency-axis. Considering the $X_{f,t}$ from the whole frequency-axis for one time band which is the input sequence of the full-band network, there is:

$$X_t = (X_{1,t}, \ldots, X_{F,t}^M) \in \mathbb{C}^{M \times F} \quad (5)$$

After full-band processing, we reshape the output $\tilde{X}_t \in \mathbb{R}^{2M \times F}$ into the input form of sub-band module, $\tilde{X}_f \in \mathbb{R}^{2M \times T}$. We integrate full-band and sub-band modules sequentially to fully leverage the advantages of both. It could learn both the global and local feature of the speech information more comprehensively. Iterations of full-sub-band could repeat above procedure with the similar sub-band and full-band modules respectively.

As shown in Figure 5, the full-band units extract the T-F bin and learn the features of each frequency. Then the concatenated features will be passed to full-band unites for learning cross-band features. Above process are repeated for several rounds.

3.4 Self-Attention Based Adaptive Multi-Channel Encoding

We utilize self-attention to enable adaptive multi-channel encoding before iterative joint full-band sub-band learning, with network architecture shown in Figure 8 and data flow shown in Figure 9, boasting its core advantage in capturing dynamic dependencies among elements within the data, enabling deep internal interaction and understanding without external supervision. In audio processing, particularly for the analysis of microphone array spectrograms, the incorporation of self-attention is particularly crucial, not only to enhance the exchange of information between different frequencies and timestamps on the spectrogram, but also foster interactions among signals captured by microphones in the spatial dimension.

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (6)$$

The multi-head self-attention mechanism enhances the model's ability to capture features by independently analyzing the relationships within the input sequence across different heads. To address the varying number of microphones (mics) in our setup, we have designed and integrated this mechanism to adapt to the diversity in microphone arrays and their positions.

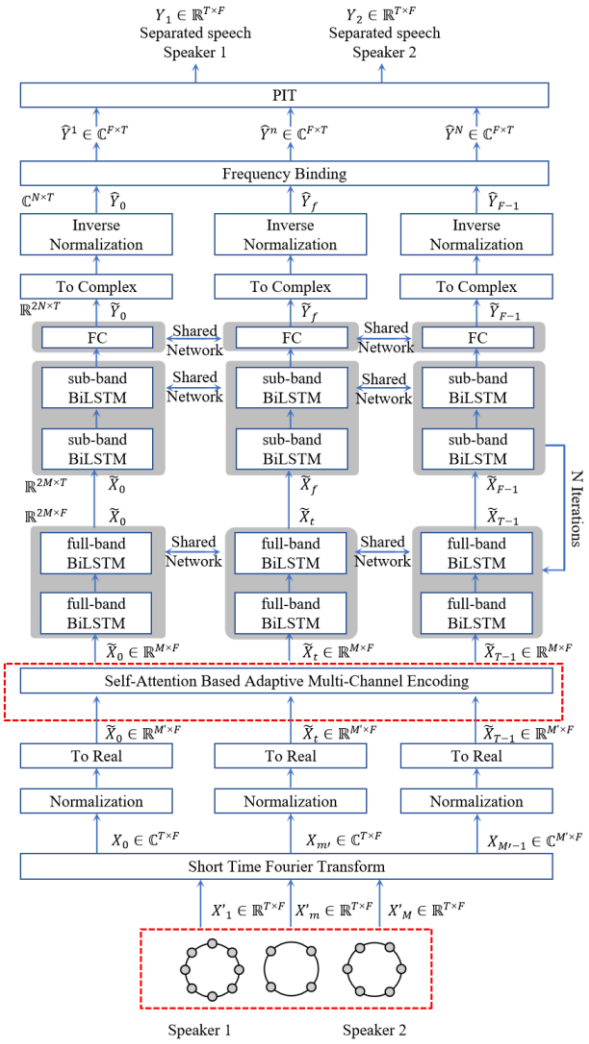

Figure 8. The network architecture of our method with self-attention based adaptive multi-channel encoding

After converting the audio signals into spectrograms using the STFT, we obtain input tensor with the shape $[B, M \times 2, F \times TF]$, where B represents the batch size. Unlike traditional Multi-Head Self-Attention methods, where data is directly split across the heads, we create unique Q, K (key) and V weight matrices for each head.

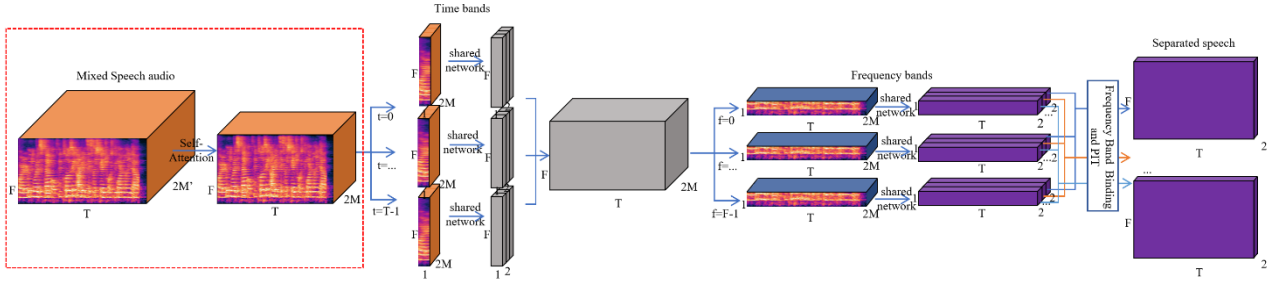

Figure 9. The data flow of our method with self-attention based adaptive multi-channel encoding

This allows each head to process the entire input independently, generating M computed outputs.

To enable the flexibility number of microphone while reducing memory usage, we focus on specific sections of the output data - the last two steps along the microphone dimension. This simplifies the data dimensions to $[B, M \times 2, F \times TF]$, reducing both computational load and memory requirements. At the same time, it preserves as much of the original spatial information as possible during the mic count transition. Notably, we optimized the Q matrix by retaining only the last two steps of its second dimension for computation. This strategy ensures efficient resource utilization and minimizes memory consumption without compromising critical information extraction.

By adjusting the self-attention module mentioned earlier，the model can acquire more intricate and adaptable feature representations，effectively managing the difficulties brought by variations in the number of microphones. No matter the size of the input from the microphone array，the model can dynamically adjust its internal weight distribution to ensure effective integration and utilization of information. This helps prevent performance degradation due to fluctuations in the number of microphones. This high degree of flexibility and adaptability enables our model to exhibit stronger robustness and accuracy.

3.5 Permutation Invariant Training

We utilize Permutation Invariant Training (PIT) to enable speaker independent multi-talker speech separation. Traditional speech separation algorithms require prior knowledge of the correspondence between each speaker and their respective speech components, i.e., which parts of the audio belong to which speaker.

Nevertheless，in practical scenarios，this association is frequently undisclosed. PIT tackles this problem by introducing a loss function that is permutation invariant. Specifically, PIT computes all possible combinations of speaker assignments during training and optimizes the model by selecting the permutation with the smallest loss (MSE). This ensures that the model's output is not compromised by misaligned speaker labels, preventing training failures. The key advantage of PIT is that it allows the model to automatically match the output based on the speech features, thereby improving the effectiveness of multi-speaker speech separation. This approach is often combined with deep learning models and has shown significant performance improvements, especially in scenarios involving two or more speakers.

However, considers it a separation problem by minimizing the separation error. To be more specific，PIT initially identifies the optimal output-target pairing and subsequently reduces the error based on said pairing. In contrast to the multi-class regression technique and deep clustering(DPCL)technique，our approach will focus on directly minimizing the separation error. This strategy, which is directly implemented inside the network structure, elegantly solves the long-lasting label permutation problem that has prevented progress on deep learning based techniques for speech separation. Once the relationship between the outputs and source streams are determined for each output meta-frame, the separated speech can be estimated, considering all meta-frames by, for example, averaging the same frame across meta-frames.

4 EXPERIMENT

4.1 Datasets

The models of both our methods and other methods for comparison are trained and tested by using the dataset generated as described in section 3.1. The WSJ0-2mix dataset contains 20000, 5000 and 3000 speech pairs with varying lengths for training, validation and test respectively and a volume of 32.17G. The RIR dataset also comprises 20,000 training samples, 5,000 for validation, and 3,000 for testing correspondingly. With configuration with 8 microphones as a typical example, it spends about 2 hours to generate a RIR dataset with a volume of 14.77GB by using dual RTX 3090 GPUs. The microphone array is circular, with a sampling rate of 16000 Hz. To facilitate the transition among different numbers of microphones, we process the RIRs to convert them configurations among 4, 8 and 16 microphones, for generating the datasets for training, validation and test correspondingly.

4.2 Comparison Methods

From the related work introduced in section 2, we selected two representatives methods, NBSS and NBC2, from related work for comparison experiments. NBSS is utilized as our baseline model. Because implementation of the only method we found for the similar task with our method, VarArray, a conformer based method, is not public available. We select NBC2, which is also a conformer based method, as another typical method for comparison.

NBSS is an end-to-end narrow-band speech separation network. A long short-term memory (LSTM) network is designed to take as input the STFT coefficients of multi-channel mixture signals for one frequency, and predict the STFT coefficients of multiple speech sources for the same frequency.

NBC2 is a narrow-band conformer network facilitating the combination of transformer and convolution network to focus on exploiting the rich information present in narrowband. It is an optimized version of NBC, with a novel hidden-layer normalization method, i.e. group batch normalization (GBN), outperforming the original NBC by achieving superior performance.

Both these two methods originally can only support dataset with fixed number of channels. During the experiments, we trained the models with 8 channels, and then in test phase we customize the generated datasets in runtime to feed the datasets with 4 and 16 channels into the trained models to evaluate the capability of microphone generalization.

4.3 Evaluation Metrics

We selected following evaluation metrics for the experiments.

SDR, which stands for Signal-to-Distortion Ratio[27], is a widely common metrics to evaluate source separation systems [21], which requires to know both the clean signal and the enhanced signal. It is an energy ratio in dB, between the energy of the target signal contained in the enhanced signal and the energy of the errors coming from the interfering speakers and artifacts.

Table 1. Results of Comparison Experiment on datasets for 4 and 16 microphones

Number of Microphone	Model	SDR	SI-SNR	NBPESQ	WBPESQ
4	NBSS	3.966868114	2.789995583	2.321677382	1.713327104
	NBC2	7.060799832	6.047250829	2.471098240	1.901377514
	Ours	**11.85791569**	**11.11448656**	**3.035696774**	**2.539441137**
16	NBSS	5.976771644	1.497891834	2.824686552	2.325496179
	NBC2	5.909863798	1.574737890	2.687947484	2.190612630
	Ours	**7.192643164**	**2.653528307**	**3.191332539**	**2.749928918**

SI-SNR, also known as SI-SDR[35], is a commonly used evaluation method in the literature for assessing the quality of signal separation or enhancement tasks. It measures the similarity between a generated signal vector and a reference signal vector by considering their relative directions and magnitudes. A higher SI-SNR value indicates a better alignment between the generated and true signal vectors, suggesting a more accurate signal reconstruction or enhancement. Conversely, a lower SI-SNR value suggests a larger deviation in the direction of the vectors, indicating poorer performance.

PESQ, an abbreviation for Perceptual Evaluation of Speech Quality，is a standardized approach to objectively evaluate the quality of speech signals transmitted across different communication channels[33]. It simulates the human auditory system to provide a quantitative measure of speech intelligibility and perceived quality. The principle behind PESQ is to compare a reference speech signal (usually the original, undistorted speech) with a test signal (the speech that has undergone some form of processing or transmission). PESQ calculates a score based on the differences between these two signals, considering factors such as distortions, noise, and other degradations. There are two types of PESQ used in the experiments, Narrow-Band PESQ (NBPESQ), and Wide-band PESQ (WBPESQ).

4.4 Experiment Results

We trained models of NBSS, NBC2 and our method by using the same dataset of generated in process described in section 4.1. Next，we input the datasets from 4 microphones and 16 microphones into the models. As shown in Table 1, the performance of NBSS and NBC are not good on this task. Although NBC has shown its adaptive ability in some level on dealing with various kinds of arrays, the PESQ and SI-SNR is still highly-decreased comparing to its original performance on fixed array. Our proposed model shows better generalization ability when processing data from different kinds of microphone arrays. That shows without requiring to change the model architecture for individual arrays, a single model can be shared between multiple arrays with different shapes and the number of microphones. Also, training with different geometries has an augmentation effect which improves the robustness compared with fixed geometry training.

5 CONCLUSION

In this project, we propose a multi-channel speech array agnostic separation model with iteratively joint full-band and sub-band learning. The frequency permutation problem is solved by PIT. The proposed self-attention based adaptive multi-channel encoding method could learn to transform various number of channels to fixed array. To enhance the quality of separated speech, it integrates the advantages of the full-band and sub-band models which means it can capture both the global spectral information and the long-distance cross-band dependencies, meanwhile retaining the ability to modeling signal stationarity and attending the local spectral pattern. In this way, fewer prior knowledge is required to separate speeches and the same trained model can be used on the speech separation task despite different kinds of microphone array. The comparison experiments and ablation study demonstrate that the proposed method consistently outperforming geometry-dependent method and prove the capability of microphone array generalization.

REFERENCES

[1] Choi, Seungjin, et al. "Blind source separation and independent component analysis: A review." Neural Information Processing-Letters and Reviews 6.1 (2005): 1-57

[2] Gurbuz A C, McClellan J H, Cevher V. A compressive beamforming method, 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008: 2617-2620

[3] Griffin, Daniel, and Jae Lim. "Signal estimation from modified short-time Fourier transform." IEEE Transactions on acoustics, speech, and signal processing 32.2 (1984): 236-243.

[4] G. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., 2001, pp. 79–82.

[5] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. Int. Conf. Acoust., Speech Signal Process., 2016, pp. 31–35.

[6] Quan C, Li X. Multi-channel narrow-band deep speech separation with full-band permutation invariant training[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 541-545.

[7] Hao X, Su X, Horaud R, et al. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6633-6637.

[8] Luo Y, Mesgarani N. TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 696-700.

[9] Luo Y, Mesgarani N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[10] Luo Y, Chen Z, Yoshioka T. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation",ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 46-50.

[11] Fan C, Tao J, Liu B, et al. End-to-End Post-filter for Speech Separation with Deep Attention Fusion Features[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2020: 1303-1314.

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

[13] 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", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 7319–7323.

[14] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing," in ASRU, 2019, pp. 260–267.

[15] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-Channel Speech Separation" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , May 2020, pp. 6394–6398.

[16] T.Ochiai, M.Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 6384–6388.

[17] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiMEchallenge," in ASRU, 2015, pp. 444–451.

[18] C. Quan, X. Li, "Multichannel Speech Separation with Narrow-band Conformer," in Interspeech, September 2022, pp.5378–5382

[19] C. Quan and X. Li, "Multi-channel Narrow-band Deep Speech Separation with Full-band Permutation Invariant Training," in ICASSP, May 2022.

[20] G. Li, S. Liang, S. Nie, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Direction-aware speaker beam for multi-channel speaker extraction," in Proc. Interspeech, 2019, pp. 2713–2717.

[21] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, "Sceneagnostic multi-microphone speech dereverberation," in Proc. Interspeech, 2021, pp. 1129–1133.

[22] S. Zhang and X. Li, "Microphone array generalization for multichannel narrowband deep speech enhancement," arXiv:2107.12601, 2021.

[23] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," arXiv:2110.09625, 2022.

[24] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, “Personalized percepnet: Real-time, low-complexity target voice separation and enhancement,” arXiv:2106.04129, 2021.

[25] L. Drude, J. Heitkaemper, C. Boeddeker, R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition." arXiv, 2019. DOI: 10.48550/arXiv.1910.13934

[26] Kolbaek M, Yu D, Tan Z, et al. Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1901-1913.

[27] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462–1469, July 2006.

[28] Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1.

[29] C. Quan and X. LI, “SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol.32, 2024, pp. 1310-1323.

[30] Benesty J, Chen J, Huang Y. Microphone array signal processing[M]. Springer Science & Business Media, 2008.

[31] Taherian H, Eskimez S E, Yoshioka T, et al. “One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement” IEEE International Conference on Acoustics, Speech and Signal Processing (*ICASSP*). IEEE, 2022: 271-275.

[32] Quan, Changsheng, and Xiaofei Li. "NBC2: Multichannel speech separation with revised narrow-band conformer." arXiv preprint arXiv:2212.02076 (2022).

[33] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *ICASSP*, 2001, pp. 749–752.

[34] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, N. Kanda, "VarArray: Array-Geometry-Agnostic Continuous Speech Separation", in *ICASSP* 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Singapore, Singapore 2022, pp. 6027-6031.

[35] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, “SDR – Half-baked or Well Done?,” in *ICASSP*, Brighton, United Kingdom, May 2019, pp. 626–630.

ACKNOWLEGEMENT

Source of the selected topic, research background:
The topic of this project comes from the combination of my personal interest and the current research trend in artificial intelligence.

The work and contribution of each team member(s):
The project is completed independently under the guidance of the supervisors.

Xupeng Bai is responsible for preparing and transforming WSJ0 clean speech dataset, design and implementation of iterative joint full-band and sub-band learning, as well as setup and evaluation of methods for comparison. The most challenging part is to design and implement the models to process non-intuitive signal data especially in frequency domain, comparing to other intuitive tasks like computer vision or natural language process. As well, set-up and evaluation of the models of both our method and comparison methods are time-consuming and exhausted, taking dozens of hour to train models even on a server with dual GPU over and over.

Deyan Ye is responsible for generating RIR datasets with different configurations of microphone, design and implementation of self-attention based adaptive multi-channel encoding, and development of components for model testing with speech related evaluation metrics. The most challenging task is getting familiar with and apply complex self-attention mechanism in signal domain. Understand the complex principles of acoustic modeling and generating RIR dataset with volume of 14.7G are also very difficult.

The relationship between the supervisors and the student, the role supervisors played in the process of writing thesis, and whether the tutoring is paid:
Dr. Xiangyu Kong is an off-campus supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Acadamy, Beijing No. 101 High School. Dr. Yu Chen Zhou is the supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Academy, Beijing No. 101 High School. Xupeng Bai and Deyan Ye are project team members of the laboratory.

Research completed with the assistance of others:
The project is completed independently under the guidance of the supervisors.

Team Profile：

Xupeng Bai: International Accelerated Class, grade 11, Beijing No. 101 High School. Highest score winner in the high school entrance examination, International Department of the school.

He has a strong interest in AI and mathematics. By himself, he learnt linear algebra, elective calculus, mathematical modeling, as well as the textbooks of "Machine Learning", "Neural Networks and Machine Learning", "Modern Speech Signal Processing", etc. He is familiar with machine learning development platforms such as Pytorch and Matlab, and proficient in 3D modeling by using Fusion360 and CAD. He participated several AI related projects and won awards in the competitions:

- Harvard China Think Big Top 5% - Research on Optimization Strategies for Elderly Fitness Activities Based on Sun Tzu's Art of War - Taking Dart Sports as an Example
- Conrad Challenge China 2023-2024 Best Innovation Award: Gravity driven water-saving toilets
- Mathematical Modeling Competition - Dandelion Diffusion and Invasive Assessment Mathematical Modeling
- Technical report: Dandelion Spread Dynamics and Invasion Assessment
- The National Scientificand Technological Innovation Talent Training Program-Computer Science: a kind of foldable amphibious drone,

Deyan Ye: class monitor, Qian Xuesen Experimental Class, Beijing 101 High School. He has a strong interest in mathematics and programming, familiar with information science related knowledge such as dynamic programming, graph theory, number theory, data structures, etc. He self studied textbooks of "Machine Learning," "Neural Networks and Deep Learning," and "Modern Speech Signal Processing". He is familiar with C/C++, Python, Pytorch, CNN, RNN and other deep learning models. He won several awards in computer and information related competitions:

- Second Prize, 2023 National Youth Informatics Olympiad, China Computer Association Non-Professional Software Capability Certification CSP-S
- Second Prize, 2022 China Computer Association Non-Professional Software Capability Certification CSP-J
- B Rating, Beijing Youth Program Exhibition Activity, Haidian District Middle School Group

Supervisor Profile：

Xiangyu Kong: lecturer, Computer School, Beijing Information Science and Technology University. He has been serving as a researcher at media computing group, Microsoft Research Asia (2019-2023) and an adjunct researcher at multi-agent group, Beijing Institute of General Artificial Intelligence (BIGAI) from 2023. He received a doctor degree from Peking University under the supervision of Prof. Yizhou Wang. From 2014-2015, he visited vision group of Queen Mary, University of London funded by China Scholarship Council (CSC). His current research interests includes speech processing ,computer vision and multi-agent embodied AI.

Yuchen Zhou: Ph.D., certificated research fellow, supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. He is a senior member of the ACM and IEEE, and former member of Technical Committee, Embedded System Society, China Computer Federation. With 20 years of technical innovation experience in IBM, he served as senior research manager of AI perception in IBM Research China, a member of IBM Academy of Science and Technology, IBM Master Inventor, chair of technical committee and patent review committee of the center, etc. He had won 3 outstanding technical achievement awards, published 1 book, participated and contributed to 2 international standards, obtained around 50 international patents, and published more than 30 papers.