Neural Speech Separation with Parallel Amplitude and Phase Spectrum Estimation

Fei Liu, Yang Ai*, Zhen-Hua Ling
National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
E-mail: fliu215@mail.ustc.edu.cn, yangai@ustc.edu.cn, zhling@ustc.edu.cn

***Abstract*—This paper proposes APSS, a novel neural speech separation model with parallel amplitude and phase spectrum estimation. Unlike most existing speech separation methods, the APSS distinguishes itself by explicitly estimating the phase spectrum for more complete and accurate separation. Specifically, APSS first extracts the amplitude and phase spectra from the mixed speech signal. Subsequently, the extracted amplitude and phase spectra are fused by a feature combiner into joint representations, which are then further processed by a deep processor with time-frequency Transformers to capture temporal and spectral dependencies. Finally, leveraging parallel amplitude and phase separators, the APSS estimates the respective spectra for each speaker from the resulting features, which are then combined via inverse short-time Fourier transform (iSTFT) to reconstruct the separated speech signals. Experimental results indicate that APSS surpasses both time-domain separation methods and implicit-phase-estimation-based time-frequency approaches. Also, APSS achieves stable and competitive results on multiple datasets, highlighting its strong generalization capability and practical applicability.**

I. Introduction

In noisy indoor environments such as cocktail parties, multiple people often speak simultaneously, accompanied by background noise. Yet, people always seem able to effortlessly separate a target speaker's voice from the mixture of sound sources and focus solely on that speaker. This phenomenon is known as the "cocktail party problem" [1]. Speech separation arises from this problem. In real-world scenarios, audio recordings typically contain not only the voice of the primary speaker but also interference from other speakers and background noise. Therefore, the goal of speech separation is to extract the useful speech signal from the corrupted mixture. The separated speech can then be used in downstream tasks such as automatic speech recognition [2], improving the accuracy and robustness of these systems. This paper primarily focuses on monaural two-speaker speech separation, which aims to separate the voices of two speakers from a mixture recorded using a single microphone.

In the early stages, researchers use signal-processing-based methods to separate speech [3]. Under the assumption of known prior distributions for speech and interference, these methods infer the spectral coefficients of speech from the mixed signal to achieve separation. For example, Wiener filtering is an optimal filter that separates speech in the sense of minimizing mean square error [3]. Signal processing methods may perform well under ideal conditions where their assumptions are met; however, in real-world scenarios, these assumptions rarely hold, resulting in a substantial degradation in performance. Later, researchers introduce decomposition-based methods, which assume that the sound spectrogram has a low-rank structure and can therefore be represented by a small set of basis vectors [4]. However, these methods suffer from poor generalization and degraded performance when the input speech does not match the training conditions. For example, the non-negative matrix factorization (NMF) method [5] factorizes any non-negative matrix into the product of two non-negative matrices, extracting local basis representations. Nonetheless, it struggles to capture deep nonlinear features and involves time-consuming inference, making it impractical for real-time applications. To overcome these limitations, Wang *et al.* propose computational auditory scene analysis (CASA) method [6], which simulates the auditory masking mechanism of the human ear by estimating an ideal binary mask to achieve speech separation. This approach does not rely on strict prior assumptions and thus offers better generalization. However, it heavily depends on the accuracy of pitch estimation, which is challenging in complex environments, limiting its practical applicability.

In recent years, with the advancement of deep learning, data-driven neural networks have been widely applied to speech separation. Initially, most researchers adopt approaches based on the time-frequency representation of the mixed speech. They apply the short-time Fourier transform (STFT) to the mixed speech waveform to obtain its spectra, and then estimate either the amplitude spectrum of each individual source or a corresponding mask [7]. For example, early deep clustering methods use ideal binary masks (IBM) for separation [8]. However, whether estimating the amplitude directly or using a mask, the waveform of each source is ultimately reconstructed by combining the estimated amplitude spectrum with the phase spectrum of the original mixture using the inverse STFT (iSTFT). Even with an ideal amplitude spectrum, errors in the phase spectrum impose an upper limit on the accuracy of the reconstructed speech. Although phase reconstruction techniques can partially mitigate this issue, accurate phase

* Corresponding author. This work was funded by the Anhui Province Major Science and Technology Research Project under Grant S2023Z20004, the National Nature Science Foundation of China under Grant 62301521 and the Anhui Provincial Natural Science Foundation under Grant 2308085QF200.

979-8-3315-7206-8/25/$31.00

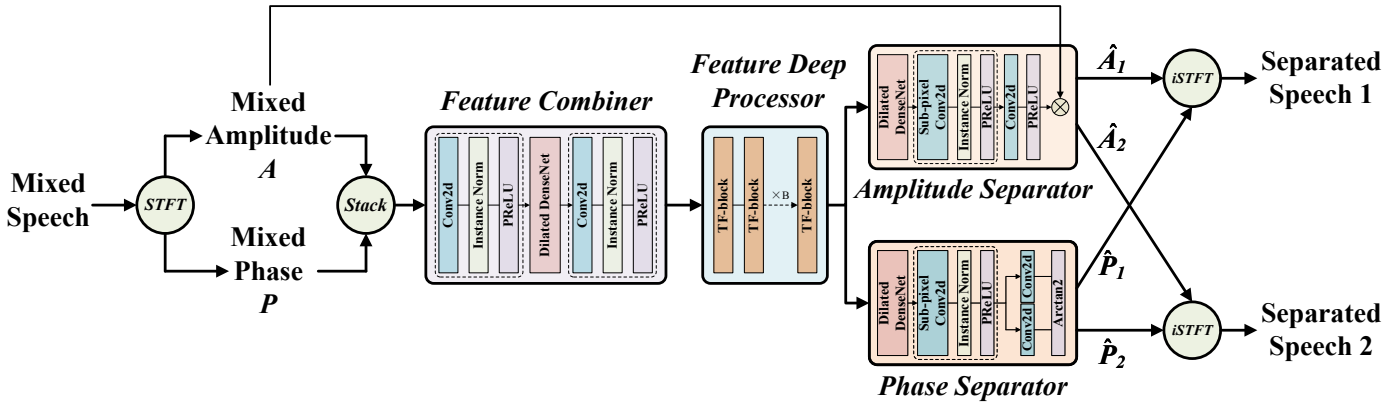

Fig. 1. Overview of the proposed APSS. The "Stack" denotes the spectral concatenation operation and the "Arctan2" denotes the two-argument arc-tangent function.

estimation remains a complex challenge, and current methods still fall short of achieving satisfactory performance. Some researchers adopt a real-and-imaginary component modeling approach in the time-frequency domain to implicitly model phase [9], [10], achieving promising speech separation results. Alternatively, some approaches model speech separation directly in the time domain to circumvent the decoupling of amplitude and phase [11]–[22]. For instance, TasNet [19] introduces learnable encoder and decoder modules and estimates a mask from the encoder output to achieve source separation. However, the improvement brought by this method is limited, and modeling long sequences in the time domain introduces greater challenges and significantly increases computational cost. Consequently, the lack of direct phase prediction still inevitably impacts speech separation performance.

Therefore, this paper proposes a novel Amplitude-Phase-estimation-based neural Speech Separation model called APSS. The APSS adopts a time-frequency domain modeling approach and is primarily composed of a feature combiner, a feature deep processor, an amplitude separator and a phase separator. First, the model extracts the amplitude and phase spectra of the mixed speech waveform using STFT. These spectra are then jointly fed into the feature combiner to produce high-dimensional fusion features. The feature deep processor leverages both time-domain and frequency-domain Transformers to capture temporal and spectral dependencies in the fusion features. Next, parallel amplitude and phase separators independently estimate the original amplitude and phase spectrum for each speaker. Finally, the iSTFT is applied to reconstruct each separated speech waveform. By explicitly modeling phase information, APSS effectively mitigates the compensation effect between amplitude and phase [23]. It also avoids the challenge of modeling long sequences in the time domain. The experimental results validate that our proposed APSS outperforms both time-domain speech separation methods and time-frequency domain approaches relying on implicit phase modeling.

This paper is organized as follows. In Section II, we provide details of the proposed APSS. In Section III, we present our experimental results. Finally, we give conclusions in Section IV.

II. Proposed Method

A. Overview

The overall architecture of the proposed APSS model is illustrated in Figure 1. APSS primarily consists of a feature combiner ϕ_{FC}, a feature deep processor ϕ_{FDP}, an amplitude separator ϕ_{AS} and a phase separator ϕ_{PS}, working together to separate monaural speech signals from different speakers in the time-frequency domain. We denote the mixture of two speech signals ($\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^L$) as $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2 \in \mathbb{R}^L$, where L is the length of the waveform. The mixed speech $\boldsymbol{x}$ is first transformed using STFT to extract the mixed amplitude spectrum $\boldsymbol{A} \in \mathbb{R}^{T\times F}$ and the mixed phase spectrum $\boldsymbol{P} \in \mathbb{R}^{T\times F}$, where T is the number of frames and F is the number of frequency bins. The amplitude and phase spectra are stacked together to form the feature combiner's input $\boldsymbol{X} \in \mathbb{R}^{2\times T\times F}$, which is then processed by the feature combiner to produce a high-dimensional fused feature $\boldsymbol{E} \in \mathbb{R}^{C\times T\times F}$, i.e.,

$$\boldsymbol{X} = Stack(\boldsymbol{A}, \boldsymbol{P}), \tag{1}$$

$$\boldsymbol{E} = \phi_{FC}(\boldsymbol{X}), \tag{2}$$

where C denotes the number of feature channels. The feature combiner effectively exploits the coupling and correlation between amplitude and phase, enhancing the performance of the subsequent separation process. Then, the fused feature $\boldsymbol{E}$ is passed through the feature deep processor to produce deep feature $\boldsymbol{S} \in \mathbb{R}^{C\times T\times F}$, i.e.,

$$\boldsymbol{S} = \phi_{FDP}(\boldsymbol{E}). \tag{3}$$

The feature deep processor effectively captures temporal and spectral dependencies, thereby reducing the complexity of the subsequent separation process. The deep feature is then fed into two parallel separators: an amplitude separator and a phase separator, which estimate the separated amplitude

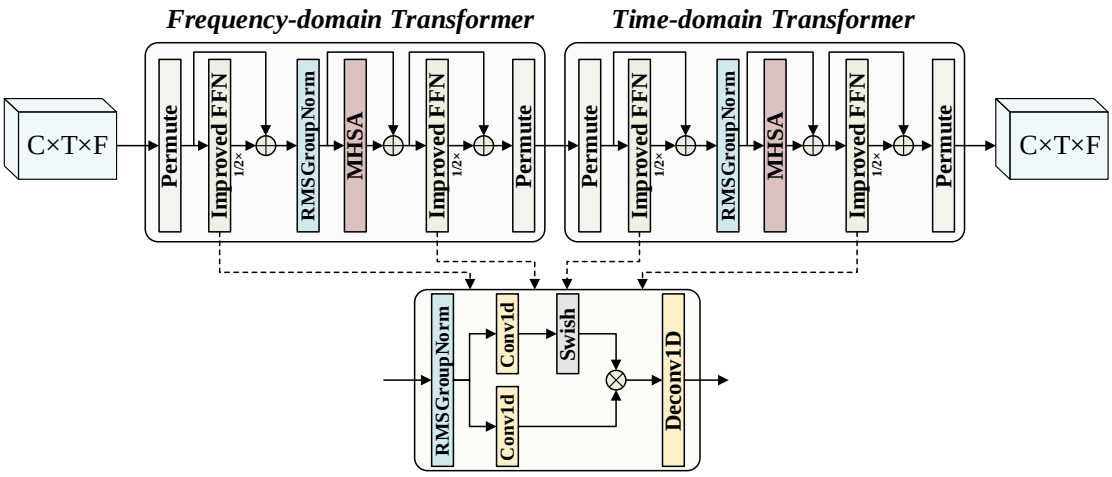

Fig. 2. The details of the TF-block employed in APSS.

spectra $\hat{\boldsymbol{A}}_1, \hat{\boldsymbol{A}}_2 \in \mathbb{R}^{T\times F}$ and the separated phase spectra $\hat{\boldsymbol{P}}_1, \hat{\boldsymbol{P}}_2 \in \mathbb{R}^{T\times F}$, respectively, i.e.,

$$\hat{\boldsymbol{A}}_1, \hat{\boldsymbol{A}}_2 = \phi_{AS}(\boldsymbol{S}), \tag{4}$$

$$\hat{\boldsymbol{P}}_1, \hat{\boldsymbol{P}}_2 = \phi_{PS}(\boldsymbol{S}). \tag{5}$$

Although the two separators estimate amplitude and phase independently, their inputs incorporate both amplitude and phase information. This is due to the intrinsic correlation between amplitude and phase, where each can facilitate the prediction of the other. Finally, the separated speech signals ($\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2 \in \mathbb{R}^L$) are respectively reconstructed via iSTFT, i.e.,

$$\hat{\boldsymbol{x}}_1 = iSTFT(\hat{\boldsymbol{A}}_1 e^{j\hat{\boldsymbol{P}}_1}), \tag{6}$$

$$\hat{\boldsymbol{x}}_2 = iSTFT(\hat{\boldsymbol{A}}_2 e^{j\hat{\boldsymbol{P}}_2}). \tag{7}$$

B. Model Details

1) Feature Combiner: As shown in Figure 1, the feature combiner consists of two convolutional blocks and a dilated DenseNet [24]. It actually applies dimensional transformation to the input feature $\boldsymbol{X}$, producing the fused feature $\boldsymbol{E}$ with C channels. Specifically, each convolutional block is composed of a 2D convolution layer, an instance normalization layer, and a parametric rectified linear unit (PReLU) activation function. The first convolutional block increases the number of channels from 2 to C, fully integrating amplitude and phase information. The following dilated DenseNet expands the receptive field along the time axis and concatenates the output of each convolutional layer with all preceding layers, enhancing feature reuse and temporal modeling. Finally, the second convolutional block performs preliminary processing along the frequency axis to produce the final fused feature $\boldsymbol{E}$.

2) Feature Deep Processor: The dual-path attention structure has been shown to deliver excellent perfermance in speech separation tasks [14]. As illustrated in Figure 1, the feature deep processor consists of B TF-blocks, each of which sequentially captures temporal and spectral dependencies, making it easier to learn the distinctions between different speakers. As shown in Figure 2, each TF-block includes a frequency-domain Transformer and a time-domain Transformer, both of which share the same architecture. Compared with the standard Transformer, we adopt an improved feed-forward network (FFN), replacing the two linear layers with a one-dimensional convolution layer and a one-dimensional transposed convolution layer. Additionally, we substitute the original ReLU activation function with the SwiGLU [25] activation function to improve the Transformer's performance. SwiGLU combines the smoothness of Swish with the gating mechanism of GLU, enabling the model to more effectively learn diverse features from the input data. This activation function has been widely adopted in various Transformer architectures [26], [27]. We also introduce an improved normalization method: instead of computing the mean of the input, we apply a root mean square (RMS)-like normalization across grouped features. This approach not only simplifies computation but also avoids instability caused by mean estimation, resulting in faster and more stable convergence during training. Specifically, each TF-block takes the fused feature $\boldsymbol{E}$ with a shape $C \times T \times F$ as input. It first applies a dimension transformation to produce an intermediate representation of shape $T \times F \times C$. This is passed through an FFN and then into a multi-head self-attention layer to capture frequency dependencies, followed by another FFN. After processing through the frequency-domain Transformer, the output is reshaped to $F \times T \times C$ and then passed through the time-domain Transformer to capture temporal dependencies. Finally, the output is reshaped back to $C \times T \times F$, forming the output of the TF-block.

3) Amplitude Separator: As shown in Figure 1, the amplitude separator consists of a dilated DenseNet, a deconvolutional block, and a mask estimation module. The deconvolutional block is structurally similar to the convolutional blocks in the feature combiner, but it replaces the 2D convolution with a 2D sub-pixel convolution [28]. The mask estimation module is composed of a 2D convolutional layer followed by a PReLU activation function. The deep feature $\boldsymbol{S}$ produced by the feature

deep processor are passed through the amplitude separator to generate the amplitude masks $\hat{\boldsymbol{M}}_1, \hat{\boldsymbol{M}}_2 \in \mathbb{R}^{T\times F}$ for separated speech signals. These masks are then element-wise multiplied with the mixed amplitude spectrum $\boldsymbol{A}$ to obtain the separated amplitude spectrum $\hat{\boldsymbol{A}}_1$, $\hat{\boldsymbol{A}}_2$, i.e.,

$$\hat{\boldsymbol{A}}_1 = \hat{\boldsymbol{M}}_1 \odot \boldsymbol{A}, \tag{8}$$

$$\hat{\boldsymbol{A}}_2 = \hat{\boldsymbol{M}}_2 \odot \boldsymbol{A}, \tag{9}$$

where $\odot$ represents the element-wise multiplication.

4) Phase Separator: As shown in Figure 1, the phase separator shares a similar structure with the amplitude separator. However, due to the inherent phase wrapping problem, we do not directly reuse the same architecture for both separators. Instead, inspired by [29], we adopt a parallel estimation architecture (PEA) to predict the wrapped phase. Specifically, two parallel 2D convolutional branches are used to estimate the pseudo-real and pseudo-imaginary components, which are then activated using the two-argument arctangent function to produce the separated phase spectra $\hat{\boldsymbol{P}}_1, \hat{\boldsymbol{P}}_2$.

C. Training Criteria

During training, we use scale-invariant signal-to-noise ratio (SI-SNR) as the loss function for optimization, which is defined as follows:

$$\mathcal{L} = -20\log_{10}\frac{\|\boldsymbol{e}_1\|_2\,\|\boldsymbol{e}_2\|_2}{\|\hat{\boldsymbol{x}}_1-\boldsymbol{e}_1\|_2\,\|\hat{\boldsymbol{x}}_2-\boldsymbol{e}_2\|_2}, \tag{10}$$

where

$$\boldsymbol{e}_1 = \frac{\hat{\boldsymbol{x}}_1^\top \boldsymbol{x}_1}{\|\boldsymbol{x}_1\|_2^2}\boldsymbol{x}_1, \quad \boldsymbol{e}_2 = \frac{\hat{\boldsymbol{x}}_2^\top \boldsymbol{x}_2}{\|\boldsymbol{x}_2\|_2^2}\boldsymbol{x}_2, \tag{11}$$

and $\|\cdot\|_2$ denotes L2-norm. Since the output order of the separation model is inherently ambiguous, we adopt permutation invariant train (PIT) [30]. PIT computes the prediction error for all possible output-target permutations and selects the permutation with the lowest error to guide network training. This process continues until the model converges.

III. Experiments

A. Dataset

We evaluated the performance of the proposed APSS model on two commonly used monaural speech separation datasets, i.e., WSJ0-2Mix [8] and Libri2Mix [31]. For all datasets, we used the fully overlapping minimum versions and set the sampling rate to 8 kHz.

- **WSJ0-2Mix**: WSJ0-2Mix consists of two-speaker mixed speech, where each mixture is generated by randomly selecting utterances from the corresponding sets and mixing them at a randomly chosen signal-to-noise ratio (SNR) between 0 dB and 5 dB. The dataset contains approximately 30 hours of training data, 10 hours of validation data, and 5 hours of test data.
- **Libri2Mix**: The Libri2Mix dataset was constructed by randomly selecting speech from two different speakers in the train-100 subset of LibriSpeech [32], and mixing them with uniformly sampled loudness units relative to full scale (LUFS) to get a mixture at an SNR between -25 and -33 dB. The dataset contains approximately 58 hours of training data, 11 hours of validation data, and 11 hours of test data. Following previous work [14], [20], we used the clean version of the dataset for our experiments.

B. Experimental Setup

When extracting amplitude and phase spectra from the mixed speech, we set the window length to 16 ms, the hop size to 8 ms, and used 128 FFT points, resulting in 65 frequency bins (i.e. $F = 65$). In the APSS model, all 1D convolutions and 1D transposed convolutions used a kernel size of 4, while all 2D convolutions used a kernel size of 1×3. And the number of intermediate channels is 128 (i.e. $C = 128$). The model employed 6 (i.e. $B = 6$) TF-blocks, each with 8 attention heads. We trained the APSS using the AdamW optimizer on a single Nvidia GeForce RTX 4090 GPU, with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.01 for 200 epochs. The initial learning rate was set to 0.001, and a warm-up strategy was applied during the first 4,000 steps. If the validation loss did not improve for two consecutive epochs, the learning rate was reduced by a factor of 0.5.

C. Evaluation Metrics

To evaluate the performance of the APSS model, we followed common practice in monaural speech separation tasks and adopted widely used objective evaluation metrics, including scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-distortion ratio improvement (SDRi). SI-SNRi measured the improvement in SI-SNR ratio relative to the mixed speech signal. It eliminated the influence of volume differences and focused solely on the quality of the target speech reconstruction, providing a fairer assessment of separation performance. SDRi, on the other hand, measured the improvement in SDR ratio relative to the mixture. It accounted for both target speech distortion (e.g., waveform deformation) and residual interference, reflecting the method's ability to preserve the fidelity of the target speech while suppressing interference.

D. Experimental Results

Table I reports the performance of our proposed APSS model compared to several baseline speech separation models on the WSJ0-2Mix dataset. First, compared with time-domain speech separation models such as TasNet [19] and Conv-TasNet [12], our proposed APSS demonstrated superior performance. This provides evidence for the effectiveness of time-frequency-domain approaches. Performing deep feature extraction on the amplitude and phase spectra obtained from the mixture may be more beneficial for speech separation than learning directly from the raw waveform. For two-stage separation methods such as Two-Step CTN [21] and DeepCASA [10], while they delivered certain performance gains, they introduced greater operational complexity. In contrast, APSS not only achieved better performance but also offered a more streamlined and practical solution. Although TFPSNet

TABLE I
EXPERIMENTAL RESULTS OF APSS AND BASELINE SPEECH SEPARATION MODELS ON WSJ0-2MIX. "-" DENOTES UNAVAILABLE RESULT IN ORIGINAL WORK. "T" DENOTES TIME-DOMAIN-BASED MODEL AND "TF" DENOTES TIME-FREQUENCY-DOMAIN-BASED MODEL.

Model	Domain	**SI-SNRi (dB)**	**SDRi (dB)**
TasNet [19]	T	10.8	11.1
Conv-TasNet [12]	T	15.3	15.6
DeepCASA [10]	TF	17.7	18.0
Two-Step CTN [21]	T	16.1	-
SuDoRM-RF [18]	T	18.9	-
DPRNN [13]	T	18.8	19.0
DPTNet [14]	T	20.2	20.3
WaveSplit [22]	T	21.0	21.2
A-FRCNN [11]	T	18.3	18.6
SepFormer [17]	T	20.4	20.5
Sandglasset [16]	T	20.8	21.0
TFPSNet [9]	TF	21.1	21.3
TDANet [20]	T	18.5	18.7
S4M [15]	T	20.5	20.7
APSS	TF	**21.3**	**21.5**

TABLE II
EXPERIMENTAL RESULTS OF APSS AND BASELINE SPEECH SEPARATION MODELS ON LIBRI2MIX. "T" DENOTES TIME-DOMAIN-BASED MODEL AND "TF" DENOTES TIME-FREQUENCY-DOMAIN-BASED MODEL.

Model	Domain	**SI-SNRi (dB)**	**SDRi (dB)**
Conv-TasNet [12]	T	12.2	12.7
SuDoRM-RF [18]	T	14.0	14.4
DPRNN [13]	T	16.1	16.6
DPTNet [14]	T	16.7	17.1
WaveSplit [22]	T	16.6	17.2
A-FRCNN [11]	T	16.7	17.2
SepFormer [17]	T	16.5	17.0
S4M [15]	T	16.9	17.4
APSS	TF	**17.1**	**17.6**

[9] operates in the time-frequency domain, it lacks explicit phase modeling capability and instead processes the real and imaginary parts of the mixture directly. As shown in Table I, our APSS model achieved a 0.2 dB improvement in both SI-SNRi and SDRi over TFPSNet, suggesting that explicit phase modeling in APSS is indeed effective for speech separation. Overall, APSS outperformed all other models across the board, further validating that joint amplitude and phase modeling in the time-frequency domain is an effective approach for monaural speech separation.

To validate the generalization ability of our proposed APSS model, we also conducted comparisons with existing separation models on the Libri2Mix dataset. We selected the models from Table I that were also evaluated on the Libri2Mix dataset reported in their papers as our baselines. The results are shown in Table II. As we can see, the proposed APSS consistently outperformed all compared baseline speech separation models. This demonstrates that APSS has strong generalization capability and delivers robust performance across different datasets.

E. Ablation Studies

Next, we conducted three ablation experiments to evaluate the contributions of different modules in APSS, as shown in Table III. First, we replaced the feature combiner with a simple 2D convolution for feature extraction (i.e., APSS w/o FC). The

TABLE III
EXPERIMENTAL RESULTS OF ABLATION STUDIES ON WJS0-2MIX.

Model	**SI-SNRi (dB)**	**SDRi (dB)**
APSS	**21.3**	**21.5**
APSS w/o FC	17.3	17.5
APSS w/o PEA	19.9	20.0
APSS w/o AM	20.0	20.1

results show a noticeable drop in SI-SNRi and SDRi, indicating that effective feature combination helps extract richer and deeper information, which benefits the subsequent separation. Inspired by [29], we used the PEA which included two parallel 2D convolutions combined with a two-argument arctangent activation in the phase separator of APSS to address the phase wrapping issue. In the ablation study, we replaced the PEA with a single 2D convolution (i.e., APSS w/o PEA). The results show that both SI-SNRi and SDRi dropped by more than 1 dB. This suggests that imprecise phase estimation has a substantial negative impact on the separation performance of APSS. In addition, we ablated the use of an amplitude mask to examine its impact on separation performance (i.e., APSS w/o AM). The results show that directly predicting the amplitude spectrum harms performance by reducing estimation accuracy, which agrees with findings in some speech enhancement studies [23].

IV. CONCLUSION

In this paper, we propose a novel time-frequency-domain neural speech separation model based on amplitude and phase estimation, named APSS. The APSS explicitly models both amplitude and phase through a feature combiner, a feature deep processor, and parallel amplitude-phase separators, effectively leveraging the coupling and correlation between amplitude and phase. This approach introduces a new paradigm for speech separation tasks. Experimental results demonstrate that APSS outperforms existing time-domain and implicit-phase-estimation-based time-frequency-domain speech separation baseline models. In future work, we plan to further enhance separation performance of APSS and extend the amplitude-phase estimation framework to multi-speaker speech separation tasks.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, pp. 975–979, 1953.

[2] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. ICASSP*, 2020, pp. 6134–6138.

[3] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[4] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms and applications," *Acoustical Society of America Journal*, vol. 124, no. 1, p. 13, 2008.

[7] S. Liang, W. Liu, W. Jiang, and W. Xue, "The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense," *Speech Communication*, vol. 59, pp. 22–30, 2014.

[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[9] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-frequency domain path scanning network for speech separation," in *Proc. ICASSP*, 2022, pp. 6842–6846.

[10] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2092–2102, 2019.

[11] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network," in *Proc. NeurIPS*, vol. 34, 2021, pp. 22 509–22 522.

[12] 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, 2019.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[14] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.

[15] C. Chen, C.-H. H. Yang, K. Li, Y. Hu, P.-J. Ku, and E. S. Chng, "A neural state-space modeling approach to efficient speech separation," in *Proc. Interspeech*, 2023, pp. 3784–3788.

[16] M. W. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *Proc. ICASSP*, 2021, pp. 5759–5763.

[17] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[18] E. Tzinis, Z. Wang, and P. Smaragdis, "SuDo RM-RF: Efficient networks for universal audio source separation," in *Proc. MLSP*, IEEE, 2020, pp. 1–6.

[19] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018, pp. 696–700.

[20] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," in *Proc. ICLR*, 2023.

[21] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. ICASSP*, 2020, pp. 31–35.

[22] N. Zeghidour and D. Grangier, "WaveSplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[23] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement," *Neural Networks*, vol. 189, p. 107 562, 2025.

[24] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *Proc. ICASSP)*, 2020, pp. 6629–6633.

[25] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[26] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[27] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[28] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proc. CVPR*, 2016, pp. 1874–1883.

[29] Y. Ai and Z.-H. Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," in *Proc. ICASSP*, 2023, pp. 1–5.

[30] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[31] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.