2024 International Conference on Electronic Engineering and Information Systems (EEISS) | 979-8-3503-5103-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/EEISS62553.2024.00048

Improved Speech Separation via Dual-Domain Joint Encoder in Time-Domain Networks

1st Lan Wang
Shantou Polytechnic
Shantou, China
lwang1216@stpt.edu.cn

2nd Haitao Zhang
Shantou Polytechnic
Shantou, China
htzhang@stpt.edu.cn

3rd Youli Qiu
School of Software
Liaoning Technical University
Huludao, China
superme32767@126.com

4th Yanji Jiang*
School of Software
Liaoning Technical University
Huludao, China
jyjvip@126.com

5th Hao Dong
Suzhou Automotive Research Institute, Tsinghua University
Suzhou, China
eason@utcer.com

6th Pengfei Guo
Shantou Polytechnic
Shantou, China
lntuguopf@163.com

***Abstract*—Time-domain algorithms have been validated to exhibit excellent performance in speech separation tasks. However, a single time-domain encoding feature is insufficient to comprehensively capture the necessary characteristics for effective speech separation. This study introduces an enhanced speech separation model that refines the time-domain encoder of the classical Conv-TasNet. To address the limitations of conventional time-domain encoders in feature extraction, which inherently impose an upper limit on separation accuracy, this study presents enhancements to the time-domain encoder. A novel dual-domain joint encoding module is devised to incorporate both time-domain and frequency-domain information, thereby bolstering the feature encoding capacity of the separation model. Experimental results demonstrate that, compared to the baseline model, the proposed model achieves improvements of 0.5896 dB and 0.5454 dB in SI-SNRi, and 0.5585 dB and 0.577 dB in SDRi on the WSJ0-2mix and Libri2Mix open-source datasets, respectively. Furthermore, the proposed model surpasses the baseline model in terms of both PESQ and STOI metrics, confirming the efficacy of the dual-domain joint encoder in enhancing separation accuracy and quality.**

Keywords-time-frequency domain fusion; dual-domain joint encoder; speech separation; monaural separation

I. INTRODUCTION

Speech separation arises from the 'cocktail party problem' [1], which entails extracting distinct speech signals from a blend of multiple speakers' voices. In recent years, substantial strides in research have been achieved with deep learning-based speech separation algorithms, harnessing the resilience and enhanced adaptability of abstract features extracted through these techniques. These deep learning approaches have showcased superior performance in comparison to conventional methods.

In the past, separation algorithms have generally fallen into two categories: frequency-domain algorithms, which take spectrograms as input data. For instance, in 2016, Hershey et al. (2016) introduced the Deep Clustering algorithm (DPCL) [2], which used ideal binary masks as training objectives and addressed permutation issues by exploiting the permutation invariance of affine transformations. While DPCL yielded state-of-the-art performance by mitigating permutation and output dimension mismatch challenges, it couldn't utilize reconstruction error as an optimization target. In the same year, Isik et al. (2016) extended the Deep Clustering framework (DPCL++) to multi-speaker speech separation tasks [3]. Their algorithm assumed dominant time-frequency units for each source signal and employed an expanded soft clustering sub-system to minimize separation errors, directly generating masks. They further leveraged an additional mask-enhancement network to enhance performance, achieving notable results in speaker-independent speech separation. In 2017, Yu et al. (2017) introduced Permutation Invariant Training (PIT) for speaker-independent multi-speaker speech separation, effectively resolving long-standing label permutation issues in speech separation techniques [4]. However, PIT is a frame-level permutation invariance technique and still requires permutation resolution during inference. In the same year, Kolbæk et al. (2017) proposed an utterance-level permutation invariant training technique (uPIT) [5], which extended PIT using a sentence-level cost function, eliminating additional permutation challenges encountered during inference. In 2018, Luo et al. (2018) presented the Deep Attractor Network (DANet) and its extended form, Anchored DANet (ADANet) [6]. DANet expanded the Deep Clustering framework by introducing attractor points in the embedding space, addressing the limitation of optimizing with reconstruction error. ADANet, meanwhile, tackled DANet's training and testing phase mismatch. Similarly, Li et al. (2018) in the same year introduced the Convolutional, Bidirectional Long Short-Term Memory, Deep Feedforward Neural Network (CBLDNN) combined with Generative Adversarial Training (GAT) for independent multi-speaker single-channel speech separation (CBLDNN-GAT) [7].

979-8-3503-5103-3/24/$31.00 ©2024 IEEE
DOI 10.1109/EEISS62553.2024.00048

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

Experimental results demonstrated new achievements in Signal-to-Noise Ratio (SDR). The advantages of the aforementioned frequency-domain methods lie in their ability to better integrate traditional signal processing techniques, yielding sparser and more structured acoustic features. In 2018, Luo et al. (2018) proposed the Time-domain Audio Separation Network (TasNet) [8], which directly modeled signals in the time domain using an encoder-decoder framework and separated sources from non-negative encoder outputs. TasNet eliminated the need for frequency decomposition, framing the separation task as estimating source masks from the encoder outputs for subsequent synthesis by the decoder. Although TasNet outperformed prior time-frequency speech separation methods, its use of Deep Long Short-Term Memory (LSTM) networks as separation modules restricted its applicability to low-resource, low-power platforms. In 2019, Luo et al. (2019) introduced Convolutional Time-domain Audio Separation Network (Conv-TasNet) [9], a deep learning framework for end-to-end time-domain speech separation. The most recent advancements in research predominantly leverage the Transformer architecture in proposing models for speech separation. For instance, SepFormer pioneers the adoption of a dual-path Transformer structure, supplanting conventional CNN and RNN architectures [10]. Conversely, DPTNet introduces an improved dual-path Transformer network, facilitating direct interaction among elements in the speech sequence [11]. MossFormer employs a gated single-headed Transformer architecture, thereby attaining an upper bound [12]. Conv-TasNet exhibited smaller model sizes and shorter minimum latency, rendering it suitable for offline and real-time speech separation applications. While frequency-domain methods constructing masks for each source signal from a mixed signal's time-frequency representation hold advantages in handling complex signals, efficient feature extraction, and parallel computation, they may lose temporal information and encounter challenges in blind source separation. Conversely, individual time-domain methods can retain temporal information, adapt to dynamic properties, and feature simple model structures. However, they too confront issues with amplitude and phase errors in time-domain encoding features and might struggle with complex signal handling. Therefore, a sole-domain feature encoder fails to capture the requisite features adequately. In order to explore methods for improving speech separation performance, this work combines frequency-domain and time-domain encoders and proposes an improved end-to-end speech separation model based on the classical Conv-TasNet framework. By incorporating frequency-domain features that integrate temporal relationship information as auxiliary information for the separation network, an innovative dual-domain joint encoder is introduced, effectively enhancing the separation performance and quality of the baseline model. Finally, experimental comparative validation is conducted in this study.

II. METHODS

After conducting research [13], it has been discovered that both phase and amplitude information play a crucial role in effectively separating speech. However, traditional time-domain model encoders are limited to extracting temporal information from signals, leading to amplitude and phase errors. To address these issues, this study introduces frequency-domain information as additional input features. A dual-domain joint encoder is designed to extract more useful features from both the time-domain and frequency-domain, thereby equipping it with the ability to integrate temporal and spectral inference. Due to the disparate time durations of frequency-domain and time-domain features, the simple incorporation of frequency-domain information presents challenges in accurately expressing semantic features. To tackle this, a domain alignment module is devised. Prior to extracting frequency-domain features, a temporal relationship modeling of mixed speech is performed to align the lengths of the two features and leverage the temporal characteristics to further align the semantic information of the frequency-domain features. The detailed design is described as follows.

A. Comprehensive Framework of the Separation Network

Conv-TasNet stands as a noteworthy time-domain speech separation model, acclaimed for its compact design and remarkable separation performance [9]. As a result, the model developed in this study represents an enhanced iteration built upon the Conv-TasNet foundation. The comprehensive architecture of the separation network is illustrated in Fig. 1, encompassing a dual-domain joint encoder (TFEncoder), a separation network, and a time-domain decoder (Decoder). The structural integrity of the separation network and time-domain decoder remains consistent with the Conv-TasNet model. The overarching framework is encapsulated by (1), (2), and (3):

$$Encoded\ feat = \text{TFEncoder}(mix) \tag{1}$$

$$Mask_{1,2} = \text{Separation network}(Encoded\ feat) \tag{2}$$

$$Speaker_{1,2} = \text{Decoder}(Encoded\ feat * Mask_{1,2}) \tag{3}$$

Here, $mix \in \mathrm{R}^{1\times T}$ represents the input mixed audio signal, where 1 denotes the number of signal channels, and T signifies the signal's temporal duration.

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

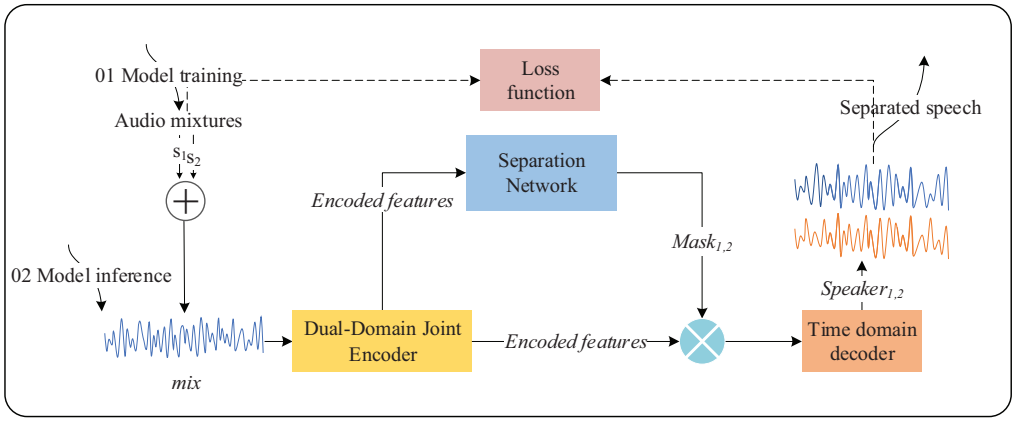

Figure 1. Speech Separation Model based on Dual-domain Joint Encoder.

B. *Dual-domain Joint Encoder*

This work introduces a dual-domain joint encoder (TFEncoder) to replace the conventional time-domain encoder, thereby providing improved encoding features to the separation network to enhance speech separation performance. In comparison to previous time-domain encoders, the TFEncoder module incorporates a frequency-domain feature extractor that integrates temporal relationships and a time-frequency domain fusion module. This results in a more effective extraction of both temporal and spectral information from the mixed audio, followed by their fused output, thereby enriching the input features for the separation network. The internal structure of the TFEncoder, as depicted in Fig. 2, comprises a time-domain encoder, a frequency domain feature extractor, and a time-frequency domain fusion module. The internal architecture of TFEncoder can be mathematically represented as (4), (5), and (6).

$$Time\_feat = \text{Time Domain Encoder}\ (mix) \tag{4}$$

$$Freq\_feat = \text{Frequency Domain Extractor}\ (mix) \tag{5}$$

$$Encoded\_feat = \text{TFFusion Module}(Time\_feat,\ Freq\_feat) \tag{6}$$

The time-domain encoder is constructed using 1D convolutions. Extraction of temporal relationships involves partitioning the input speech segment into distinct cyclic fragments based on the magnitude of its amplitude. These fragments are then arranged in parallel and subjected to 2D convolutions to extract their temporal relationships. The frequency domain feature extractor utilizes an STFT (Short-Time Fourier Transform) built with 1D convolutions. The time-frequency domain fusion module encompasses a domain alignment module, which aligns the length and semantic features of frequency domain features, and a feature fusion module. The latter merges time-domain features with the aligned frequency domain features. This module is crafted with a combination of components, including multiple 1D convolutional layers, pooling layers, activation function layers, normalization layers, and a dual-stage self-attention mechanism.

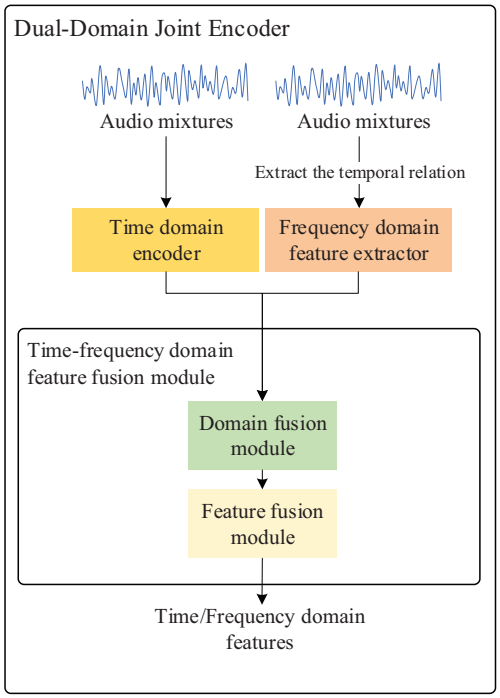

Figure 2. Dual-Domain Joint Encoder.

C. *Objective function and evaluation metrics*

The training objective of this study is scale-invariant signal-to-noise ratio (SI-SNR), with scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal distortion ratio improvement (SDRi) serving as objective evaluation metrics for separation accuracy. SI-SNR is defined by (7), (8), and (9), where $\hat{S} \in \mathrm{R}^{1\times T}$ and $S \in \mathrm{R}^{1\times T}$ represent the model's output and the original clean source, respectively.

$$S_T = (\hat{S}\cdot S)/\|S\|^2 \cdot S \tag{7}$$

$$S_E = \hat{S} - S_T \tag{8}$$

$$SI\text{-}SNR = 10\log_{10}(\|S_T\|^2/\|S_E\|^2) \tag{9}$$

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

III. Results and Discussion

A. Dataset

In the experiment, this study utilizes the publicly available WSJ0-2mix and Libri2Mix datasets as the foundation for its experimental research [14,15]. The WSJ0-2mix dataset stands as a widely employed resource within speech separation studies. It involves a deliberate selection of speech samples from various speakers found within the WSJ0 dataset. The process entails a controlled mixing of signal-to-noise ratios (SNR) ranging from -5 to 5 dB, resulting in the creation of a diverse corpus of multi-speaker mixed speech data. The WSJ0-2mix dataset encompasses an extensive training set, amounting to 30 hours and 22 minutes of content, a validation set of 7 hours and 40 minutes, and a further 4 hours and 49 minutes allocated to the test set.

On a parallel note, the Libri2mix dataset is fashioned from the well-regarded LibriSpeech dataset. However, it distinguishes itself from the WSJ0-2mix by the absence of added noise and a lower sampling rate of 8kHz. The Libri2mix dataset is structured with meticulous care, offering a training set, a development set, and a test set, all aimed at nurturing and appraising the efficacy of speech separation models. Notably, the training set boasts a substantial collection of approximately 42 hours of speech, spanning a total of 13,900 segments, and occupying an approximate space of 7 GB. Meanwhile, the development set is characterized by its encompassment of around 4 hours and 30 minutes of speech, distilled into 3,000 segments, with an associated size of roughly 744 MB. Echoing this pattern, the test set for each dataset mirrors the 4 hours and 11 minutes duration, containing 3,000 speech segments, and possessing a size of approximately 700 MB.

B. Model parameter configuration

The training process involves the network being trained on 1s speech segments, with an initial learning rate set at 10^{-3}, and a speech sampling rate of 8 kHz. Convergence of the model is determined if there is no enhancement in the SI-SDR (Scale-Invariant Signal-to-Distortion Ratio) of the validation set over a span of 10 consecutive training iterations. The optimization of the model is achieved using the AdamW optimizer, with the loss function being SI-SDR. Further details regarding the configuration of the remaining network parameters can be found in Table 1.

TABLE I. Model hyperparameter Settings

Baseline model	Identifies	Parameter interpretation	Set value
ConvTasNet	N	The number of output channels of the time domain encoder	256
	K	The convolution kernel size of the time domain encoder	16
	S	Stride size of the time domain encoder	8
	Num_spks	The number of separated speakers	2
	N	The number of output channels of the time domain encoder	256
	H	The number of hidden layers separating the network	512
	Layer	Number of layers inside the TCN	8
	Stack	Number of TCN stacks	3
	Kernel	Convolution kernel size in TCN	3
	Sr	Rate of sampling	8000
	Chunk_size	The length of the input signal	8000
	Norm	The normalized way	GroupNorm
	B	Batch size	36

C. Analysis of results

To ensure an equitable evaluation of the effectiveness of the proposed method, this study conducted separate training sessions for the ConvTasNet-KS16 and TFE-ConvTasNet models under identical experimental conditions. Both models were configured with the same settings and were rigorously evaluated using the WSJ0-2mix dataset. Notably, TFE-ConvTasNet, the model proposed in this work, leverages the fusion of time-domain and frequency-domain features for speech separation. As indicated in Table 2, a direct comparison reveals that, when compared to the ConvTasNet-KS16 model, the TFE-ConvTasNet model introduced in this study exhibits substantial enhancements on the WSJ0-2mix test set. Specifically, it achieves an impressive improvement of 0.5896 dB in SI-SNRi and 0.5585 dB in SDRi.

TABLE II. A Comparison of the Model Proposed and the Baseline Model on the WSJ0-2mix Dataset

Model	SI-SDR		Test	
	Train	*Validation*	*SI-SNRi*	*SDRi*
ConvTasNet -KS16	13.7493	12.4853	10.3575	11.1742
TFE-ConvTasNet (Ours)	14.5567	12.9590	10.9471	11.7327

Table 2 provides an overview of the model's comprehensive performance scores on the WSJ0-2mix test set. Below, a comparison is made using 10 randomly selected speech samples from the test set, illustrating the overall trend of differences between the two models in terms of SDRi and SI-SNRi, as depicted in Fig. 3. In the Figure, the solid line represents SDRi scores, while the dashed line corresponds to SI-SNRi scores. The visual representation highlights that the enhanced TFE-ConvTasNet model proposed in this study achieves an impressive peak SI-SNRi of 15.72 dB and a notable maximum SDRi of 15.4663 dB for individual speech samples. In comparison to the baseline model, the novel model introduced in this study demonstrates a remarkable

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

improvement of 1.7624 dB in SI-SNRi and 1.7324 dB in SDRi.

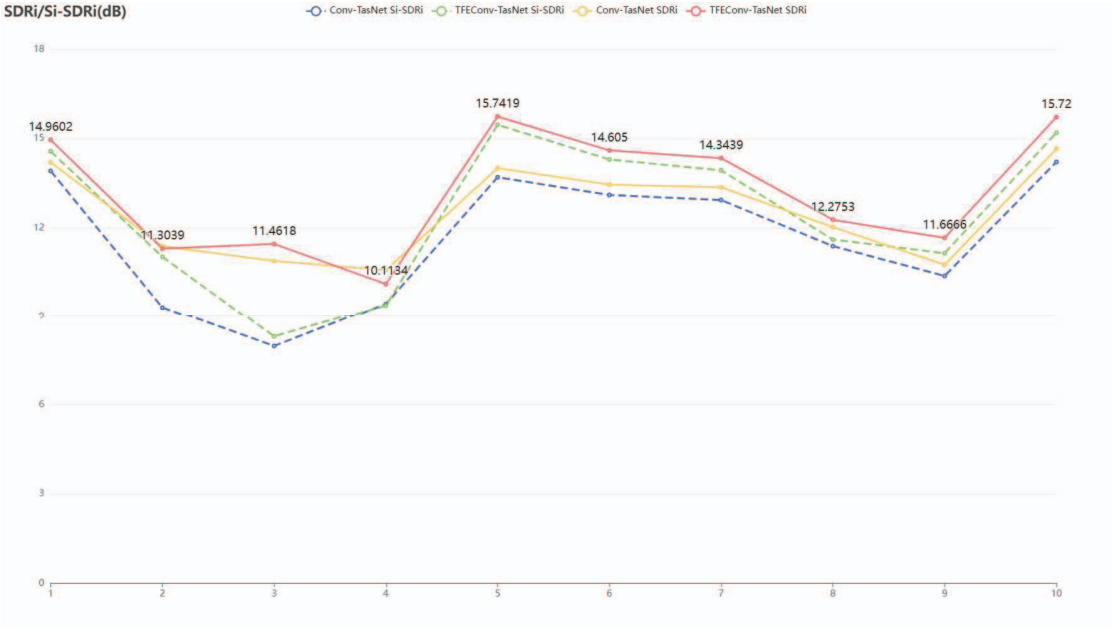

Figure 3. A Comparison of SDRi and SI-SNRi Scores on the WSJ0-2mix Test Set.

In addition to evaluating SDRi and SI-SNRi, this study also assesses the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) of the separated speech, comparing the results with those of the baseline model. Illustrated in Fig. 4, the graph presents an overview of the overall PESQ and STOI scores for both the model proposed in this study and the baseline model on the WSJ0-2mix test set. The data indicates that the TFE-ConvTasNet model achieves a commendable average PESQ score of 2.7857 for both speakers, exhibiting a notable improvement of 0.1054 over the ConvTasNet model's overall score of 2.6803. Furthermore, the TFE-ConvTasNet model attains an average STOI score of 0.8706 for both speakers, outperforming the ConvTasNet model's average STOI score of 0.8637.

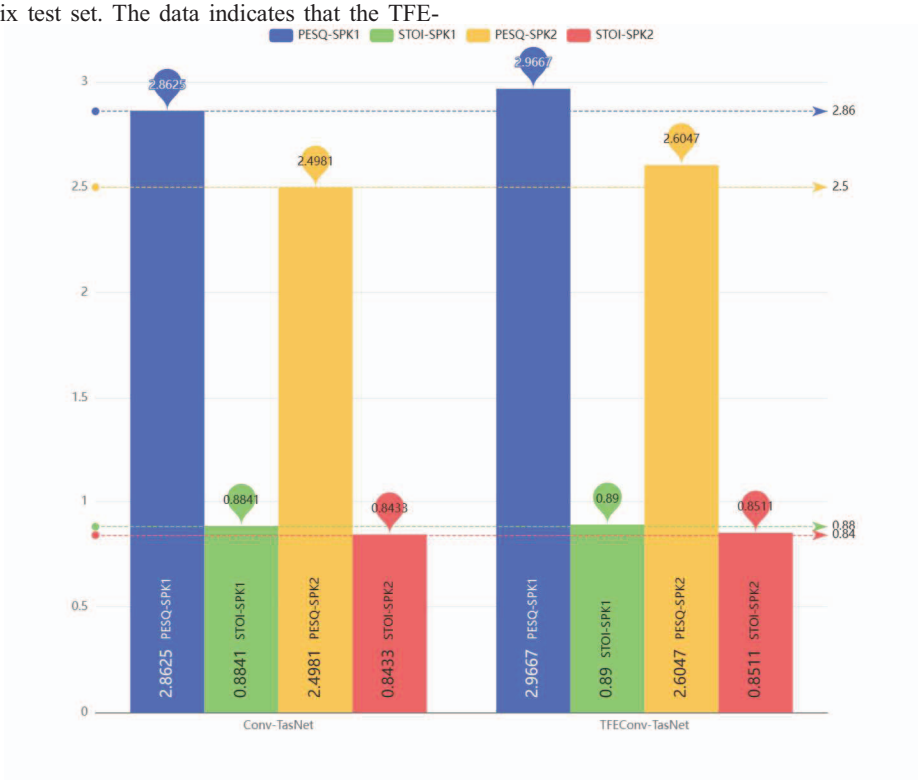

Figure 4. A Comparison of PESQ and STOI Scores on the WSJ0-2mix Test Set.

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

This study undertakes a comprehensive comparison of the TFE-ConvTasNet model, considering its SI-SNRi, SDRi scores, and model parameter sizes. Table 3 offers a detailed overview of how the model proposed in this study stacks up against other models in terms of SI-SNRi, SDRi, and Params, utilizing the WSJ0-2mix dataset. Please note that some values in the table are absent due to the lack of reported data in the referenced research. The table unmistakably highlights that the model introduced in this study exhibits a notable edge, excelling in both separation accuracy and model parameter count when contrasted with counterparts like DPCL++, uPIT-BLSTM, DANet, ADANet, cuPIT-Grid-RD, and CBLDNN-GAT. Additionally, the dual-domain encoder proposed in this study is also suitable for certain Transformer-based separation models. Subsequent validation of the primary contributions of this study will be conducted on these models, such as SepFormer and MossFormer. It is noteworthy that these models exhibit a substantial parameter count, yet they achieve the current optimal performance, as illustrated in Table 3. To substantiate the value of this work, this work replicated the experimental results of ConvTasNet-KS16. In comparison with the baseline model ConvTasNet-KS16, despite a modest increase of 0.8 million parameters, the separation accuracy demonstrates a commendable growth of approximately 0.6 dB.

TABLE III. A COMPARISON OF SI-SNRI, SDRI, AND PARAMS BETWEEN THE MODEL PROPOSED AND OTHER MODELS

Model	WSJ0-2mix		Param (Mb)
	SI-SNRi	*SDRi*	
DPCL++	10.8	-	13.6Mb
uPIT-BLSTM	-	10.0	92.7Mb
DANet	10.5	-	9.1Mb
ADANet	10.4	10.8	9.1Mb
uPIT-Grid-RD	-	10.2	47.2Mb
CBLDNN-GAT	-	11.0	39.5Mb
Conv-TasNet	15.3	15.6	5.1Mb
SepFormer	20.4	20.5	26Mb
MossFormer	22.8	-	42.1Mb
ConvTasNet-KS16	10.3575	11.1742	4.9Mb
TFE-ConvTasNet(Ours)	10.9471	11.7327	5.8Mb

To provide additional validation for the proposed methodology, experimental verification was carried out utilizing the Libri2mix dataset in this study. As indicated in Table 4, in comparison to the ConvTasNet-KS16 model, the newly introduced TFE-ConvTasNet exhibited notable enhancements of 0.5454 dB in SI-SNRi and 0.5770 dB in SDRi on the Libri2mix test set.

TABLE IV. A COMPARISON BETWEEN THE MODEL PROPOSED AND THE BASELINE MODEL ON THE LIBRI2MIX DATASET.

Model	SI-SDR		Test	
	Train	*Validation*	*SI-SNRi*	*SDRi*
ConvTasNet -KS16	16.0774	13.9032	13.4763	13.8839
TFE-ConvTasNet (Ours)	16.6074	14.5610	14.0217	14.4609

IV. CONCLUSION

Given the dynamic nature of speech signals, influenced by both temporal changes and frequency domain attributes such as amplitude and phase, conventional time-domain models that exclusively extract temporal variations from waveform signals frequently encounter challenges in accurately representing amplitude and phase. To tackle this issue, this study introduces an innovative speech separation model based on dual-domain fusion encoding. This architectural design harmonizes time-domain encoding features derived from the original mixed speech waveform with frequency domain characteristics, enabling accurate prediction of the target speech waveform. By adopting Conv-TasNet as the baseline model and introducing a groundbreaking dual-domain joint encoding approach, the performance of traditional time-domain separation models experiences a substantial boost. Experimental findings demonstrate that the proposed method outperforms the baseline model when applied to separation tasks. Specifically, on the testing sets of the WSJ0-2mix and Libri2Mix datasets, the proposed model yields remarkable improvements of 0.5896 dB and 0.5454 dB in SI-SNRi, along with 0.5585 dB and 0.5770 dB enhancements in SDRi, respectively. Furthermore, the experimental results illustrate that the newly proposed approach surpasses the baseline system across evaluation metrics such as PESQ and STOI. This methodology effectively addresses the limitations of traditional time-domain separation algorithms, specifically the inadequate extraction of sequence information by the time-domain encoder. The approach not only holds promise in enhancing the precision and quality of speech separation but also provides a robust framework for future advancements in the field.

ACKNOWLEDGMENT

This research was supported by the Fund of Liaoning Provincial Department of Education, Project No. LJKZ0338; Huludao Science and Technology Plan Project, Project No.2023JH (1)4/02b; Guangdong Province Science and Technology Innovation Strategy special city and county science and Technology innovation Support project, Project No. STKJ2023071.

REFERENCES

[1] Ya-Ting, H., Jing, S., Jia-Ming, X. U., & Bo, X. U. (2019). Research advances and perspectives on the cocktail party problem and related auditory models. Acta Automatica Sinica, 45(02), 234-251.

[2] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 31-35).

[3] Isik, Y.Z., Le Roux, J., Chen, Z., Watanabe, S., & Hershey, J.R. (2016). Single-Channel Multi-Speaker Separation Using Deep Clustering. arXiv preprint.

[4] Yu, D., Kolbæk, M., Tan, Z. H., & Jensen, J. (2017, March). Permutation invariant training of deep models for speaker-

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.

independent multi-talker speech separation. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 241-245).

[5] Kolbæk, M., Yu, D., Tan, Z. H., & Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(10), 1901-1913.

[6] Luo, Y., Chen, Z., & Mesgarani, N. (2018). Speaker-independent speech separation with deep attractor network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(4), 787-796.

[7] Li, C., Zhu, L., Xu, S., Gao, P., & Xu, B. (2018, April). CBLDNN-based speaker-independent speech separation via generative adversarial training. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 711-715).

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

[9] Luo, Y., & Mesgarani, N. (2019). Conv-Tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, 27(8), 1256-1266.

[10] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2020). Attention is All You Need in Speech Separation. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Canada, 21-25.

[11] Chen, J., Mao, Q., & Liu, D. (2020). Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation.

[12] Zhao, S., & Ma, B. MossFormer: Pushing the Performance Limit of Monaural Speech Separation Using Gated Single-Head Transformer with Convolution-Augmented Joint Self-Attentions. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 1-5.

[13] Wang, Z. Q., Wichern, G., & Le Roux, J. (2021). On the compensation between magnitude and phase in speech separation. IEEE Signal Processing Letters, 28, 2018-2022.

[14] Garofolo, J., Graff, D., Paul, D., & Pallett, D. (1993). CSR-I (WSJ0) Complete LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium.

[15] Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., & Vincent, E. (2020). Librimix: An open-source dataset for generalizable speech separation. arXiv preprint.

Authorized licensed use limited to: University of Illinois. Downloaded on March 08,2025 at 04:00:25 UTC from IEEE Xplore. Restrictions apply.