TFPSNET: TIME-FREQUENCY DOMAIN PATH SCANNING NETWORK FOR SPEECH SEPARATION

Lei Yang Wei Liu Weiqin Wang

Samsung Research China – Beijing (SRC-B)

{lei81.yang, wei.liu, wq88.wang}@samsung.com

ABSTRACT

Speech separation has been very successful with deep learning techniques. In this paper, we propose time-frequency (T-F) domain path scanning network (TFPSNet) for speech separation task. The connections between frequency bins in frequency path, time path, and T-F path are modeled by transformer. We also introduce T-F path loss function to improve the performance further. The proposed TFPSNet could learn more details of frequency structure and separate the feature in T-F domain. Experiments show that proposed model achieves state-of-the-art (SOTA) performance on public WSJ0-2mix datasets. It reaches 21.1dB SI-SDRi on WSJ0-2mix, and 19.7dB SI-SDRi on Libri-2mix. Furthermore, our approach has good generalizability. The model trained on WSJ0-2mix dataset achieves 18.6dB SI-SDRi on Libri-2mix test set without any fine-tuning work. This result is even 0.4dB higher than DPTNet trained on Libri-2mix dataset.

Index Terms— speech separation, source separation, transformer, deep learning, T-F domain

1. INTRODUCTION

Speech separation is a fundamental task in signal processing [1, 2]. Substantial efforts have been made on this problem, it can be considered as a pre-processing step for many downstream tasks such as speech recognition [3], speaker identification [4], and speaker diarization [5, 6]. In this paper, we focus on the single channel speech separation task.

Deep learning techniques have accomplished a big step forward on speech separation task. The current leading methods are based on the time-domain audio separation network (TasNet) [7]. TasNet uses a learnable encoder and decoder to replace the fixed T-F domain transformation. It takes waveform inputs and directly reconstructs sources, and computes time-domain loss with utterance-level permutation invariant training (uPIT) [8]. Several approaches are proposed based on TasNet framework, such as the Conv-TasNet [9, 10], the dual-path recurrent neural network (DPRNN) [11], the dual-path Transformer network (DPTNet) [12], RNN-free transformer-based neural network (SepFormer) [13] . Moreover, a self-attentive network with a novel sandglass-shape, namely Sandglasset [14] advances the SOTA speech separation performance.

Time domain separation methods achieved impressive results. However, the specific space generated in time domain methods lacks interpretability and the performance is unstable in extreme conditions [15]. In the meantime, T-F domain separation methods are more robust, and highly correlated to the phonetic structure of speech. But STFT is not learnable, and is a generic signal transformation that is not necessarily optimal for speech separation [7]. To overcome this problem, auxiliary encoder after STFT and separation approach designed for T-F domain are necessary. A complex T-F domain method is proposed in [16], but the performance is not as good as time domain dual path methods. In this paper, we propose a T-F domain path scanning network (TFPSNet) for end-to-end monaural speech separation, which leads to superior separation performance. We incorporate STFT within the encoder, the separator works in the T-F domain, the decoder converts the T-F signal to waveform with iSTFT.

The contributions of this work are summarized as follows:

1) Our separation network is designed for T-F domain, and uses transformer to scan three kinds of paths and separate mixed T-F feature. The new design could learn more details of frequency structure, and improve the generalizability as well.
2) We design T-F path loss function in T-F domain to make the model reconstruct frequency structure.
3) Experiments show that:
   - On WSJ0-2mix dataset, our approach achieves SOTA performance of 21.1dB SI-SDRi.
   - On Libri-2mix dataset, our approach achieves 19.7dB SI-SDRi, outperforms DPTNet and DPRNN.
   - Our approach has good generalizability. The separation model trained on WSJ0-2mix dataset achieves 18.6dB SI-SDRi on Libri-2mix test set without any fine-tuning work. It is even 0.4dB higher than DPTNet trained on Libri-2mix data corpus.
   - Our approach has consistency to different networks. We change our transformer-based network to RNN-based network, and use DPRNN as the baseline, the result is 0.9dB higher than DPRNN.

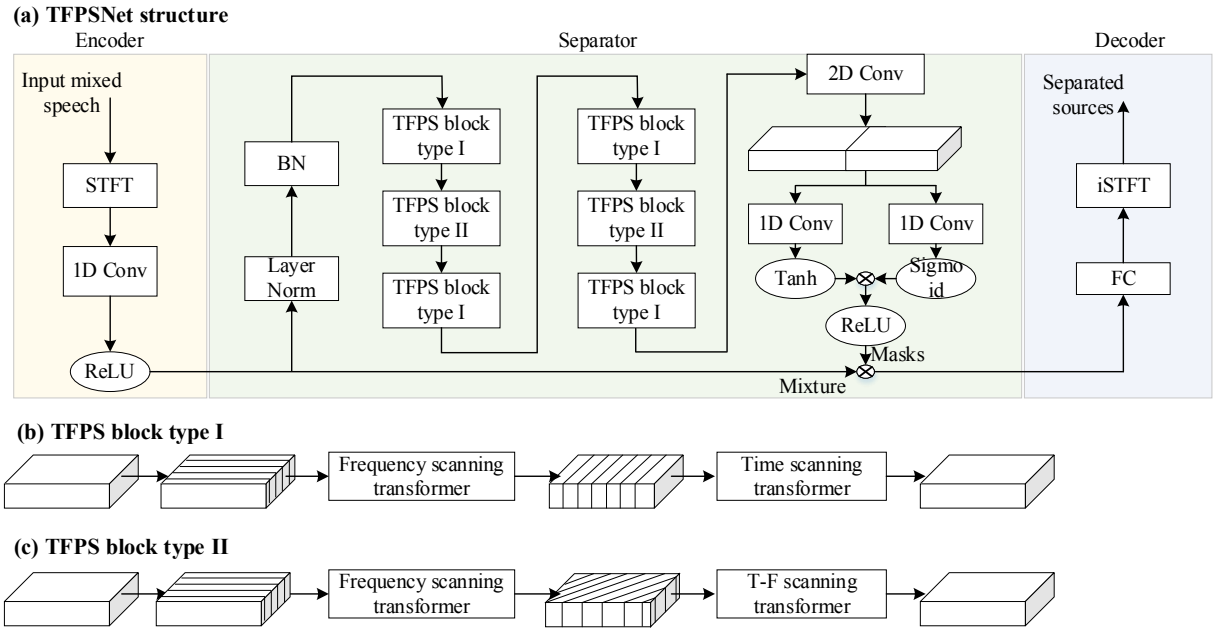

Fig. 1: The overall architecture of the proposed TFPSNet

2. TFPSNET

2.1. Overall structure

As depicted in Fig. 1, our speech separation system consists of three stages: encoder, separator and decoder, which is similar to that of Conv-TasNet. First, an encoder is used to convert the mixture waveform into T-F features. Then the features are fed to the separation layer to predict mask vector for each source. Finally, the decoder reconstructs the source waveforms by iSTFT. In the following, we describe the encoder, separator, and decoder in details.

2.2. Encoder

In this module, encoder segments the T samples input signal $x \in \mathbb{R}^{1\times T}$ into L segments (frames) of length J_w $x \in \mathbb{R}^{L\times J_w}$ with overlap. Then, the signal x converts to STFT domain.

$$X = \mathrm{STFT}(x) \tag{1}$$

$X \in \mathbb{R}^{L\times K\times 2}$, where and K is the number of frequency bins. The last dimension is the real part and imaginary part of the frequency bins.

The spectrogram is highly correlated to the phonetic structure of speech. So, the spectrum structure and resolution are very important for the following separator. 1D Conv is used here as an auxiliary encoder after STFT. We regard the complex signal as a 2 channels vector. The purpose of designing the auxiliary encoder is to encode the 2 channels vector to a high dimension non-negative vector.

The encoder encodes $X \in \mathbb{R}^{L\times K\times 2}$ to high dimension mixture feature $U \in \mathbb{R}^{L\times K\times H}$ as follows

$$U = \mathrm{ReLU}(\mathrm{Conv1D}(X)) \tag{2}$$

2.3. Separator

Fig. 1(a) shows the detailed architecture of the separator. The separator is fed by the encoded representations U and estimates a group of masks $M_s \in \mathbb{R}^{L\times K\times H}$ in the mixture U. The masked encoder features for the s-th source $Z_s \in \mathbb{R}^{L\times K\times H}$ are obtained by the element-wise multiplication between U and M_s:

$$Z_s = U \cdot M_s \tag{3}$$

The mask M_s is estimated by several T-F domain path scanning (TFPS) blocks. There are 2 kinds of TFPS blocks – TFPS block type I (Fig. 1(b)), and TFPS block type II (Fig. 1(c)). Inspired by dual path network, our model consists three kinds of T-F path scanning layers as follows to model the T-F feature.

- **Frequency path**. It is applied to model the transitions from frequency bin 0 to frequency bin $K-1$ in one frame. It processes the T-F feature in each frame independently.
- **Time path**. It is applied to model the transitions of same frequency along time axis. It processes the T-F feature in each frequency bin independently. It has more obvious physical meaning than time domain dual path network.
- **Time-frequency path**. Frequency path and time path connect frequency bins in one frame, and same frequency along time axis directly. All the frequency bins in the utterance have implicit connections by stacking frequency and time paths. But the connection is not strong and direct. Here we model the transitions of adjacent frequency bins of adjacent frames directly by T-F path. The connection is important and meaningful for speech, for example, speech pitch and formant usually change frame by frame. The time-frequency path could trace the changing of adjacent frequency bins.

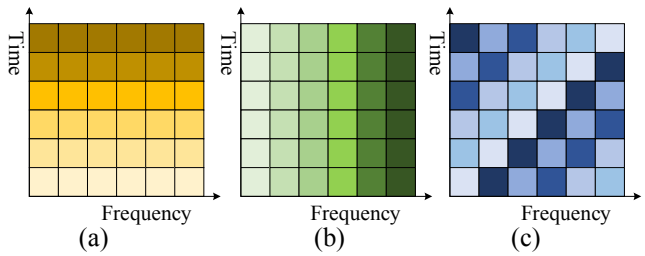

Fig. 2: Illustration of path scanning. (a) Frequency path scanning (b) Time path scanning (c) T-F path scanning

The path scanning illustration is shown in Fig. 2. The cells represent frequency bins. The cells with same colors mean the frequency bins are connected in one path. Fig. 2 (c) shows the T-F path scanning along the diagonal direction.

Transformer has been shown impressive performance in dual path network, such as DPTNet, Sepformer, and Transmask [17]. We use transformer to scan these 3 kinds of paths. It is comprised of three core modules: scaled dot-product attention, multi-head attention and position-wise feed-forward network. And recurrent neural network is used to learn the order information of the speech sequences without positional encodings. The transformer structure is as same as DPTNet.

2.4. Decoder

In decoder, a fully connected layer $V \in \mathbb{R}^{H\times 2}$ is used to reconstruct separated speech T-F signals $D_s \in \mathbb{R}^{L\times K\times 2}$ for the *s*-th source:

$$D_s = Z_s * V \tag{4}$$

It converts the H-dimension feature to 2-dimension. Then iSTFT is applied to obtain the final waveforms $y_s \in \mathbb{R}^{1\times T}$

$$y_s = \text{iSTFT}(D_s) \tag{5}$$

2.5. Training objective: T-F path loss function

We train the proposed model with uPIT to maximize scale-invariant source-to-distortion ratio (SI-SDR) [7, 18]. SI-SDR is defined as:

$$s_{target} = \frac{\langle \tilde{z}, z\rangle z}{\|z\|^2} \tag{6}$$

$$e_{noise} = \tilde{z} - s_{target} \tag{7}$$

$$\text{SI-SDR}(\tilde{z}, z) = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \tag{8}$$

where z and $\tilde{z}$ are clean and estimated sources, respectively. Both are normalized to zero-mean before the calculation.

Instead of using waveform SI-SDR directly, we calculate SI-SDR along frequency path and time path. By the T-F path loss method, the network learns more details of frequency structure. The proposed loss function consists of three parts:

- **Frequency path SI-SDR**. The SI-SDR is designed for real value, but the frequency bin is complex value. In order to use SI-SDR along frequency path, we interleave the real and imaginary parts per frequency bin, and reshape $D_s \in \mathbb{R}^{L\times K\times 2}$ to $C_s \in \mathbb{R}^{L\times 2K}$. c_s^l and $\tilde{c}_s^l$ are $2K$-dimension vectors, representing the idea signal and estimated signal of the l-th frame, respectively. The frequency path SI-SDR is

$$F_f = \frac{1}{L}\sum_l \text{SI-SDR}(\tilde{c}_s^l, c_s^l) \tag{9}$$

- **Time path SI-SDR**. In order to use SI-SDR along frequency path for each frequency bin, we transpose and reshape $D_s \in \mathbb{R}^{L\times K\times 2}$ to $B_s \in \mathbb{R}^{K\times 2L}$, b_s^k are $\tilde{b}_s^k$ $2L$-dimension vectors, representing the signal of the k-th frequency bin, respectively. The time path SI-SDR is

$$F_t = \frac{1}{K}\sum_k \text{SI-SDR}(\tilde{b}_s^k, b_s^k) \tag{10}$$

- **Waveform SI-SDR**. Waveform SI-SDR is the same as end-to-end separation training objective, such as TasNet. $\tilde{y}_s$ represents the estimated waveform signal of the *s*-th source

$$F_w = \text{SI-SDR}(\tilde{y}_s, y_s) \tag{11}$$

Loss function is defined as below

$$\mathcal{L}_{TFW} = \alpha \cdot F_f + \alpha \cdot F_t + \beta \cdot F_w \tag{12}$$

The α and β value don't have significant impact on performance according to our experiment. Here we set $\alpha = 0.25$, $\beta = 0.5$ in Section 4.

3. EXPERIMENTAL PROCEDURES

3.1. Dataset

We evaluate our proposed model on two-speaker speech separation task using the WSJ0-2mix dataset. The dataset is derived from WSJ0 data corpus [19]. The 30 hours of training data and 10 hours of validation data contain two-speaker mixtures generated by randomly selecting utterances from different speakers in the WSJ0 training set.

WSJ0-2mix is widely used to evaluate the speech separation performance. But WSJ0 dataset only contains 101 different speakers, and 25 hours of training data. The number of speakers is too small to evaluate model's generalizability [17]. In order to test generalizability of our model, we use the model trained by WSJ0-2mix dataset to test Libri-2mix test set [20], which is created based on the Librispeech dataset [21]. The Librispeech dataset contains 1,172 speakers, and 465 hours of training data. We also use sub train set train-100 as the training dataset to evaluate our approach performance in Libri-2mix. All the datasets are 8kHz sampling rate.

3.2. Experiment configurations

In our model, the window length and hop size are 32.5ms and 16.25ms, respectively. The DFT length is 256, we use 129 frequency bins for each frame. The 1D Conv layer in the encoder converts 129 frequency bin signals to 256-dimension feature, the kernel size is 6. In separator, the 256-dimension feature is converted to 64-dimension feature by bottleneck (BN) layer. The dimension size of transformer is $d = 64$. We stack 6 TFPS blocks, the stacking details are shown in Fig. 1. The models with proposed T-F path loss function and only

waveform SI-SDR loss function are called "TFPSNet" and "TFPSNet-WaveLoss" in the experiment, respectively.

We train the model for 140 epochs on 4-second long segments. The learning rate is initialized to 5e-4. We increase the learning rate linearly for the first 4000 training steps:

$$lrate = 0.2 \cdot d^{0.5} \cdot n \cdot 4000^{-1.5} \quad \text{when } n < 4000 \quad (13)$$

and then decay it by 0.98 for every two epochs. Adam [22] is used as the optimizer. All models are trained with uPIT.

In order to test the consistency of our method on different networks, we change our transformer-based network to BLSTM-based network, the output dimension of BLSTM is 64 with 128 hidden units, which is the same as DPRNN. The BLSTM-based TFPSNet with T-F path loss function is called "TFPSNet-RNN" in the experiment.

We train BLSTM-based TFPSNet without warmup step, the learning rate is initialized to 5e-4.

4. RESULTS AND DISCUSSIONS

4.1. Results on WSJ0-2mix

Table 1 compares the performance achieved by the proposed TFPSNet with the best results reported in the literature on the WSJ0-2mix dataset. The TFPSNet achieves 21.1dB SI-SDRi on the test set. The proposed model outperforms current SOTA approach "Sandglasset (MG)" 0.3dB. It also outperforms DPTNet and DPRNN by a large margin. Please notice that "SepFormer+DM" is better than our TFPSNet, and "Sandglasset (MG)+PT" achieves competitive performance with us, but both use dynamic mixing (DM) data augmentation, which doesn't relate to network design. To be fair, we don't compare with "SepFormer+DM", and "Sandglasset (MG)+PT" in this paper. The TFPSNet with T-F path loss function is 0.1dB better than with waveform SI-SDR loss.

4.2. Results of RNN and transformer path scanning

To test that our method is consistent across different networks, we change our transformer-based network to RNN (BLSTM) -based network. We use DPTNet and DPRNN as the baseline of transformer-based network, and RNN-based network, respectively. Table 2 compares the performances of the proposed models and the baseline models. We can see that the proposed methods outperform baseline systems.

4.3. Results on Libri-2mix

To prove our approach could work on other dataset, we conduct related experiments on the Libri-2mix dataset. Table 3 compares the performance achieved by TFPSNet, DPTNet and DPRNN. We reproduce these 3 models on Libri-2mix dataset. The proposed approach achieves 19.7dB SI-SDRi on the test set. It is 1.5dB higher than DPTNet, and 3.2dB higher than DPRNN.

4.4. Results of generalizability test

We also test the generalizability of our approach in Table 4. We reproduce TFPSNet, DPTNet, and DPRNN models on WSJ0-2mix dataset, and then use these models to test the Libri-2mix test set without fine-tuning. Benefiting from the path scanning modeling, the TFPSNet's SI-SDRi degradation is smaller than others. It achieves 18.6dB SI-SDRi on Libri-2mix test set. It is even 0.4dB higher than DPTNet trained on Libri-2mix dataset. This presents the generalizability of our method and further demonstrates the effectiveness of it.

Table 1: Comparison of performances on WSJ0-2mix

Model	Params.	SI-SDRi	SDRi
BLSTM-TasNet [7]	23.6M	13.2	13.6
Conv-TasNet [9]	8.8M	15.3	15.6
DC-Conv-TasNet-MSO [16]	5.8M	17.5	-
FurcaNeXt [23]	51.4M	18.4	-
DPRNN [11]	2.6M	18.8	19.1
DPTNet [12]	2.7M	20.2	20.6
Sepformer [13]	26M	20.4	20.5
Sandglasset (MG) [14]	2.3M	20.8	21.0
Sandglasset (MG) + PT*	2.3M	21.0	21.2
SepFormer + DM*	26M	22.3	22.4
TFPSNet – WaveLoss	2.7M	21.0	21.2
TFPSNet	**2.7M**	**21.1**	**21.3**

* These methods use dynamic mixing (DM) data augmentation. To be fair, we don't compare with them in this paper

Table 2: Comparison of performances for transformer-based and RNN-based network

Model	Params.	Si-SDRi	SDRi
DPTNet	2.7M	20.2	20.6
TFPSNet	**2.7M**	**21.1**	**21.3**
DPRNN	2.6M	18.8	19.1
TFPSNet – RNN	**2.6M**	**19.7**	**19.9**

Table 3: Comparison of performances on Libri-2mix

Model	Params.	SI-SDRi	SDRi
DPRNN	2.6M	16.5	16.8
DPTNet	2.7M	18.2	18.4
TFPSNet	**2.7M**	**19.7**	**19.9**

Table 4: Comparison of generalizability

Model	SI-SDRi test on WSJ0-2mix	SI-SDRi test on Libri-2mix	SI-SDRi degradation
DPRNN	18.8	13.0	5.8
DPTNet	20.2	15.7	4.5
TFPSNet	**21.1**	**18.6**	**2.5**

5. CONCLUSIONS

In this paper, we propose a novel neural model called TFPSNet for speech separation. The TFPSNet could learn more details of frequency structure by T-F path scanning transformer. Our experiment results demonstrate that TFPSNet achieves new SOTA performance on WSJ0-2mix data corpus. Moreover, our model has good generalizability. The model achieves 18.6dB SI-SDRi on unmatched test set Libri-2mix. As a future work, we would like to explore speech separation problem in noisy and reverberant environments.

6. REFERENCES

[1] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] W. Xiong, L.Wu, F. Alleva, et al., "The microsoft 2017 conversational speech recognition system," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018: 5934–5938.

[4] A. Alenin, A. Okhotnikov, R. Makarov, et al., "The ID R&D System Description for Short-Duration Speaker Verification Challenge 2021," in *Proc. Interspeech 2021*, 2021: 2297-2301.

[5] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, et al., "Front-end processing for the CHiME-5 dinner party scenario," *in Proceedings 5th Intl. Workshop on Speech Processing in Everyday Environments (CHiME)*, Hyderabad, 2018, pp. 35–40.

[6] N. Kanda, R. Ikeshita, S. Horiguchi, et al. "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *Proceedings 5th Intl. Workshop on Speech Processing in Everyday Environments (CHiME)*, Hyderabad, 2018, pp. 6–10.

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

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

[9] 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.

[10] S.J. Bai, J.Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

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

[12] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. of Interspeech 2020*, 2020.

[13] C. Subakan, M. Ravanelli, S. Cornell, et al. "Attention is all you need in speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021: 21-25.

[14] M.W.Y. Lam, J. Wang, D. Su, et al., "Sandglasset: A Light Multi-Granularity Self-Attentive Network for Time-Domain Speech Separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021: 5759-5763.

[15] J. Heitkaemper, D. Jakobeit, C. Boeddeker, et al., "Demystifying tasnet: A dissecting approach," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6359–6363.

[16] K. Wang, H. Huang, Y. Hu, et al., "End-to-End Speech Separation Using Orthogonal Representation in Complex and Real Time-Frequency Domain," in *Proc. of Interspeech 2021*, 2021.

[17] Z. Zhang, B. He, and Z. Zhang, "TransMask: A Compact and Fast Speech Separation Model Based on Transformer," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021: 5764-5768.

[18] R.J. Le, S. Wisdom, H. Erdogan, et al., "SDR-half-baked or well done?," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019: 626-630.

[19] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[20] 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

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints, p. arXiv:1412.6980*, 2014

[23] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 653–665.