INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Improved Speech Separation with Time-and-Frequency Cross-Domain Feature Selection

Tian Lan, Yuxin Qian, Yilan Lyu, Refuoe Mokhosi, Wenxin Tai, Qiao Liu

School of Information and Software Engineering,
University of Electronic Science and Technology of China, Chengdu, Sichuan, China
lantian1029@uestc.edu.cn, yxqian@std.uestc.edu.cn, lylan@std.uestc.edu.cn,
refuoemokhosi@yahoo.com, wxtai@std.uestc.edu.cn, qliu@uestc.edu.cn

Abstract

Most deep learning-based monaural speech separation models only use either spectrograms or time domain speech signal as the input feature. The recently proposed cross-domain network (CDNet) demonstrates that concatenated frequency domain and time domain features helps to reach better performance. Although concatenation is a widely used feature fusion method, it has been proved that using frequency domain and time domain features to reconstruct signal makes minor difference compared with only using time domain feature in CDNet. To make better use of frequency domain feature in decoder, we propose using selection weights to select and fuse features from different domains and unify the features used in separator and decoder. In this paper, we propose using trainable weights or the global information calculated from the different domain features to generate selection weights. Given that our proposed models use element-wise fusing in the encoder, only one deconvolution layer in the decoder is needed to reconstruct signals. Experiments show that proposed methods achieve encouraging results on the large and challenging Libri2Mix dataset with a small increasing in parameters, which proves the frequency domain information is beneficial for signal reconstruction. Furthermore, proposed method has shown good generalizability on the unmatched VCTK2Mix dataset.

Index Terms: monaural speech separation, feature selection, selective kernel network, multi-talker

1. Introduction

Monaural multi-speaker speech separation is the task of extracting speech signals from multiple speakers in overlapped speech. Although humans can focus on one voice in overlapped speech[1], research approaches still face difficulty in achieving this. The speech separation method enables many speech processing algorithms such as Automatic Speech Recognition (ASR) realize better performance under multi-speaker conditions.

Various methods such as Computational Auditory Scene Analysis (CASA), Nonnegative Matrix Factorization (NMF) [2, 3] are proposed to solve this problem. With the development of deep learning, neural network based methods such as Deep Clustering (DPCL) [4] and Permutation Invariant Training (PIT) [5, 6] perform better than conventional methods. On the basis of DPCL and PIT, deep attractor network (DANet) [7, 8] achieves improved performance by using the attractor mechanism to estimate masks for each source. Different from the above methods which take a magnitude spectrogram as the input feature, TasNet and Conv-TasNet [9, 10] use the time domain signal as the input feature. The main idea behind TasNet is to replace the generic transformation like Short-Time Fourier Transformation (STFT) with a convolution encoder, which is used to directly extract features from a time domain signal. By training the network in an end-to-end manner, the convolutional encoder is trained to find best weights to encode the time domain signal. On the basic of TasNet, [11, 12] explore the different designs of encoder component and achieve better performance.

Based on methods directly model the time domain signal, cross-domain network (CDNet) [13] proposes jointly embedding and clustering the time and frequency domain features. The encoder calculates the convolution encoded time domain feature and STFT extracted magnitude spectrogram in parallel, and concatenates them along the channel dimension. Then a separator network projects the concatenated feature map into a high-dimension embedding space and uses the attractor mechanism to produce masks for each source. The decoder uses two domain feature maps to reconstruct separated speech signal by using inverse Short-Time Fourier Transform (iSTFT) and deconvolution respectively. Experiments in [13] show that jointly modeling time and frequency domain features outperforms modeling time domain feature only. However, compared with reconstructing from time domain, reconstructing signal from two domains only makes minor difference [13]. The reason behind this result is that the design of encoder and decoder are still immature in processing cross-domain features, especially the reconstructing signal by using cross-domain features respectively in decoder.

In this work, we propose a novel idea that uses selection weights to select different features and fuse them together to solve the problem exists in the encoder and decoder design of CDNet. Our proposed method uses cross-domain feature to reconstruct the signal jointly in decoder which is same as the way masks are generated in separator and unifies the feature used in separator and decoder. Following this idea, we propose two designs of cross-domain feature selection module: using trainable weights and using the global information generated from different domain features to calculate selection weights. Using the trainable weights enables the network to learn the importance between different domain features directly, whereas using the global information improves the generalizability. Both methods significantly improve the performance of cross-domain separation model.

The remainder of this paper is organized as follows. Section 2 introduces the network design and details the proposed feature selection modules. Section 3 introduces the experiment

http://dx.doi.org/10.21437/Interspeech.2021-2246

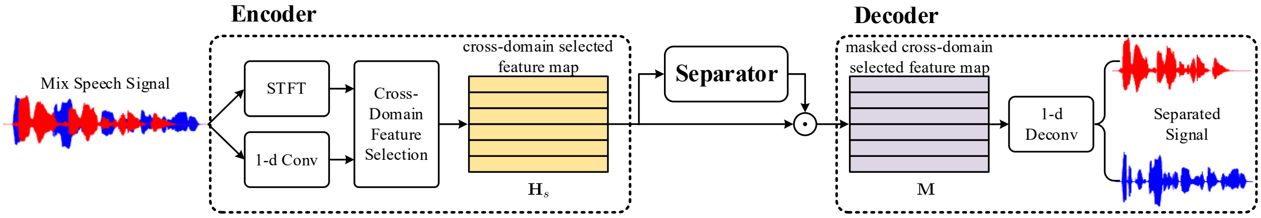

Figure 1: *Building blocks of cross-domain feature selection separation model* .

settings, results and discussions. We finally conclude this paper in Section 4.

2. Cross-Domain Feature Selection

In this section, we introduce the proposed cross-domain feature selection network in subsection 2.1 and describe the proposed encoder which uses trainable parameters to aggregate different domain features in subsection 2.2 and uses global information in subsection 2.3.

2.1. Cross-Domain Feature Selection Network

The entire network composes of three major processing modules: Encoder, Separator and Decoder. The Encoder is used to select cross-domain features, the Separator produces masks for each source and the Decoder reconstructs the speech signals from selected cross-domain features. Figure 1 shows the building blocks of network.

The Encoder encodes a mixed speech signal $\mathbf{x}$ into a selected cross-domain feature map $\mathbf{H}_s \in \mathbb{R}^{T\times C}$, which combines different domains information. Figure 2 depicts the architecture of the proposed cross-domain encoder.

$$\mathbf{H}_s = Encoder(\mathbf{x}) \tag{1}$$

Given mixed speech signal $\mathbf{x}$ consisting of N sources $\mathbf{s}_1$, $\mathbf{s}_2$,..., $\mathbf{s}_N$, time domain feature $\mathbf{F}_{conv}$ and frequency domain feature $\mathbf{F}_{spec}$ are derived in a parallel manner using convolution and STFT respectively, which can be defined as:

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{s}_i \tag{2}$$

$$\mathbf{F}_{spec} = \mathcal{S}(\mathbf{x}), \mathbf{F}_{conv} = \mathcal{F}(\mathbf{x}) \tag{3}$$

where $\mathcal{S}$ denotes the STFT operation and $\mathcal{F}$ denotes the 1-dimension convolution operation. Then a linear layer is used to transform $\mathbf{F}_{spec}$ into a feature map with the same feature dimensions as $\mathbf{F}_{conv}$, which is necessary for the feature selection operation. A 1-dimension convolution with kernel size 3 transforms $\mathbf{F}_{spec}$ into the same latent representation space as time domain feature. Two feature maps are selected and fused into one feature map by proposed cross-domain feature selection modules introduced in following subsections.

$\mathbf{H}_s$ is then sent to the separator to estimate masks for each source. The separator is consistent with Conv-TasNet, which is built with constructed TCN [14] blocks. Finally, masked feature maps are generated as:

$$\mathbf{M} = \hat{\mathbf{M}} \odot \mathbf{H}_s \tag{4}$$

The Decoder then reconstructs separated speech signal from the masked cross-domain selected feature map using a 1-dimension deconvolution.

We use negative Signal-to-Distortion Ratio (SDR) [15] as the objective function and PIT [6] to train the network, which can be expressed as:

$$\mathcal{L} = \min_{\pi\in\mathcal{P}} \left(-\frac{1}{N}\sum_{i}^{N} SDR(\mathbf{s}_i, \hat{\mathbf{s}}_{\pi(i)}) \right) \tag{5}$$

$$SDR(\mathbf{s}, \hat{\mathbf{s}}) = 10\log_{10} \frac{\langle \mathbf{s}, \hat{\mathbf{s}} \rangle^2}{\|\mathbf{s}\|^2 \|\hat{\mathbf{s}}\|^2 - \langle \mathbf{s} - \hat{\mathbf{s}} \rangle^2} \tag{6}$$

where $\mathcal{P}$ denotes the set of permutations, $\|\cdot\|^2$ denotes the signal power and $\langle\cdot\rangle^2$ denotes dot product, $\hat{\mathbf{s}}_{\pi(i)}$ denotes the estimated speech for one source and $\mathbf{s}_i$ denotes the target speech.

2.2. Feature Selection Using Trainable Weights

We propose the trainable weights cross-domain feature selection module (TCD) which uses trainable weights to weighted sum two domain feature maps together. Figure 3 depicts the architecture of the proposed encoder. The main idea of this method is using the trainable weight parameters to control what information should be propagated from different domains, which can be defined as:

$$\hat{\mathbf{a}}_c = \frac{e^{\mathbf{a}_c}}{e^{\mathbf{a}_c} + e^{\mathbf{b}_c}}, \hat{\mathbf{b}}_c = \frac{e^{\mathbf{b}_c}}{e^{\mathbf{a}_c} + e^{\mathbf{b}_c}} \tag{7}$$

$$\mathbf{H}_s = \hat{\mathbf{a}} \odot \mathbf{F}_{conv} + \hat{\mathbf{b}} \odot \hat{\mathbf{F}}_{spec} \tag{8}$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{1\times L}$ are trainable parameters and L is set to 1 or C. $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are selection weights calculated from trainable parameters with softmax function.

2.3. Feature Selection Using Global Information

Using trainable weights to control the different domain feature fusion may be sensitive to the mismatch between training and test conditions. To address this problem, we propose two methods that calculate selection weights from the global feature information generated from different domain features.

On the basis of the method proposed in subsection 2.2, we propose the global cross-domain feature selection module (GCD) which uses the concatenation of two vectors embedded by global average pooling from different domains feature maps to select different domain features, which is shown in Figure 4. We first embed the global information vector by using global average pooling. Then two global vectors are concatenated together along the channel dimension. Further, a fully-connected layer is used to generate weights $\mathbf{a}$, $\mathbf{b}$ from the global information. Finally, the softmax operation is applied on the generated weights to calculate the selection weight $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$ and the selected feature map $\mathbf{H}_s$ is calculated as equation (8).

Inspired by selective kernel network (SKNet) [16] in image classification which selects multiple convolution kernels with different receptive fields to adaptively adjust receptive

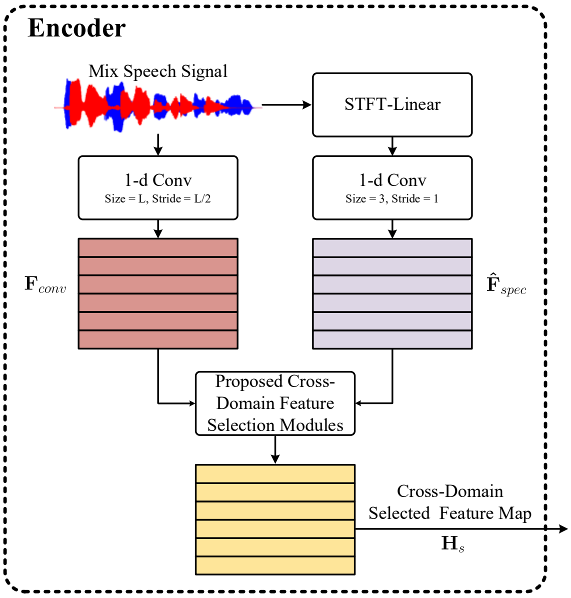

Figure 2: *Details of the proposed cross-domain encoder architecture.*

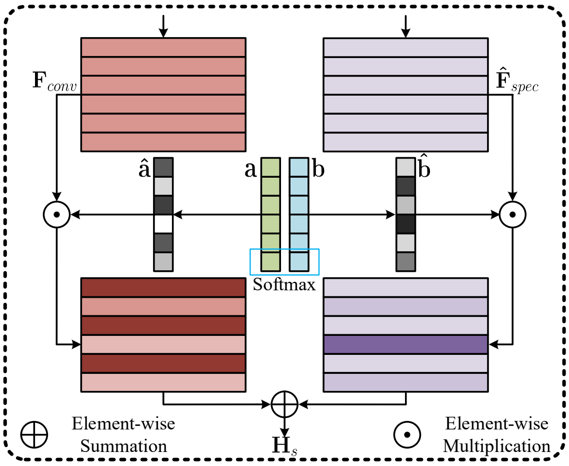

Figure 3: *Details of proposed trainable weights cross-domain feature selection module (TCD).*

field size by global information, we propose selective cross-domain feature selection module (SCD) which adopts the design of SKNet. Figure 5 depicts the architecture of the cross-domain feature selection encoder used to aggregate different domain features. Two feature maps $\mathbf{F}_{spec}$ and $\mathbf{F}_{conv}$ are then fused via an element-wise summation:

$$\mathbf{U} = \mathbf{F}_{conv} + \hat{\mathbf{F}}_{spec} \tag{9}$$

Then, the global information $\mathbf{s} \in \mathbb{R}^{1\times C}$ is calculated using the global average pooling along the time dimension. A fully connected layer generates a compact feature vector $\mathbf{z} \in \mathbb{R}^{1\times m}$ which guides the feature selection procedure, which can be defined as:

$$\mathbf{z} = \delta(\mathcal{N}(\mathbf{W}_c \mathbf{s})) \tag{10}$$

where δ denotes rectified linear unit (ReLU) [17] activation function, $\mathcal{N}$ denotes global Layer Normalization (gLN) [10] and $\mathbf{W}_c \in \mathbb{R}^{m\times C}$ denotes the weight of the fully connected layer. Furthermore, soft selection vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{1\times C}$ are generated to select different feature maps, which are guided by the impact feature vector $\mathbf{z}$. The similarity between proposed method and SKNet is that both methods select multiple feature maps, but proposed method selects different domain features instead of different receptive fields.

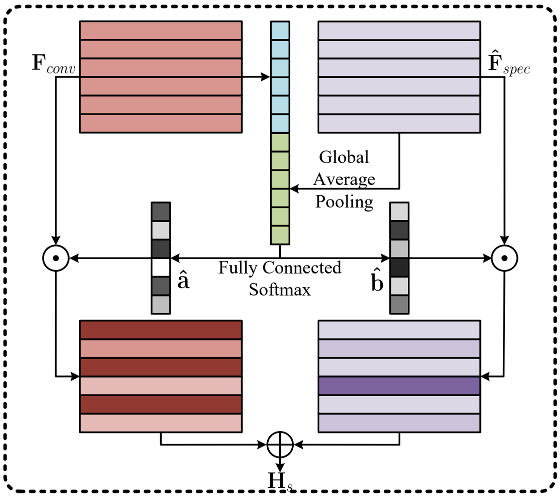

Figure 4: *Details of the proposed global cross-domain feature selection module (GCD).*

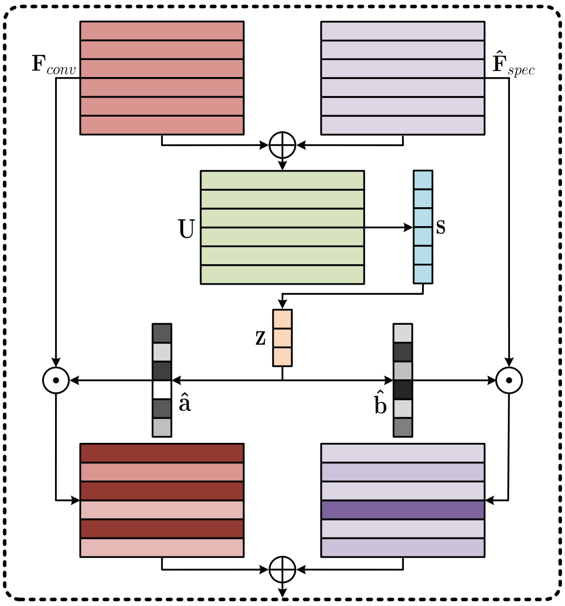

Figure 5: *Details of the proposed method using selective kernel module (SCD).*

3. Experiments

3.1. Datasets

Libri2Mix [18] is a recently released open-source dataset for single channel speech separation derived from Librispeech dataset [19]. We resample all speech data down to 8kHz to reduce computational and memory costs. We choose the sub train set train-100 as the training dataset, dev and test are chosen as validation and test set. VCTK2Mix [18] is used to evaluate the generalization ability. The VCTK-2Mix is an un-

Table 1. *Performance of different models achieved on Libri2Mix test set in clean condition (L: length of each selection weight).*

Model	L	SDR$_i$ (dB)	SI-SNR$_i$ (dB)
Conv-Tasnet	-	13.02	12.61
CD-TasNet	-	12.81	12.93
ACD-TasNet	-	13.49	13.10
SCD-TasNet	256	13.74	13.49
TCD-TasNet	1	13.53	13.61
TCD-TasNet	256	13.72	13.30
GCD-TasNet	1	**14.03**	**13.63**

matched test set derived from VCTK[20]. The mixing procedure for VCTK2Mix is identical to that for LibriMix [18].

3.2. Experimental Setup

For all of our experiments, all models are trained on 3s segments utterances. We use the ADAM [21] optimizer to train the network. The learning rate starts at 1×10^{-3}, and is then halved after 3 epochs with no reduction in validation loss. The train procedure stops if no best validation loss is found in 10 consecutive epochs. The maximum of training epochs is set to 100. The objective function SDR on the validation set is used to choose the best model, while SDR improvement (SDRi) and scale-invariant signal-to-noise ratio improvement (SI-SNRi) are taken as performance evaluation criterions.

We select Conv-TasNet, which is implemented with same configuration with [10] as a baseline model.

For all cross-domain models, the frequency domain log magnitude spectrogram is calculated using STFT with 20-points window length, 10-points window shift and the square root of the Hann window. The 256-dimensional time domain feature is calculated by a convolution layer with same window size and striding, and ReLU is chosen as activation function. The separator part is consistent with [10].

We build CD-TasNet, which is constructed using the concatenation based encoder proposed in [13] to replace the encoder used in Conv-TasNet as a baseline model. In decoder, the hyper parameter α which controls the signal reconstruction summation is set to 1 according to [13].

For the cross-domain model that uses trainable weights to select the cross-domain features, we build two models TCD-TasNet-1 and TCD-TasNet-256. The difference between two models is the shape of the trainable parameters. In TCD-TasNet-1, the trainable parameters $\mathbf{a}, \mathbf{b} \in \mathbb{R}^1$ are scalars. In TCD-TasNet-256, the trainable parameters $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{1\times 256}$ are vectors. The purpose of building these two models is to verify whether the fine control of each feature channel can improve the final separation performance. We also create a model called ACD-TasNet which directly sums two domains features together. The ACD-TasNet can be regarded as a special case of TCD-TasNet when the selection weights are set to 1.

For the cross-domain model that uses global information to select the cross-domain features, we build two models GCD-TasNet and SCD-TasNet. GCD-TasNet is built with the GCD encoder proposed in subsection 2.3 and the length of selection weight for each domain is set to 1. SCD-TasNet is built with the SCD encoder proposed in subsection 2.3. Hyper parameter m is set to 32.

Table 2. *Performance of different models achieved on VCTK2Mix test set in clean condition (L: length of each selection weight).*

Model	L	SDR$_i$ (dB)	SI-SNR$_i$ (dB)
Conv-Tasnet	-	11.72	11.05
CD-TasNet	-	11.30	11.07
ACD-TasNet	-	11.96	11.29
SCD-TasNet	256	11.89	11.45
TCD-TasNet	1	11.82	11.62
TCD-TasNet	256	12.26	11.54
GCD-TasNet	1	**12.69**	**11.97**

3.3. Results

The results of models evaluated on Libri2Mix *test* set are reported in Table 1. Compared with CD-TasNet and Conv-TasNet, all proposed method shows better performance. According to the comparison between ACD-TasNet and CD-TasNet, we can see the importance of fusing different domain features into one and using cross-domain feature to reconstruct the signal jointly in decoder. From the comparison between ACD-TasNet and other cross-domain feature selection methos, we can see that using a selection weights to select and fuse cross-domain feature achieves a significant improvement in SI-SNRi. The comparison between TCD-TasNet-1 and TCD-TasNet-256 indicates that increasing the number of trainable weights does not improve the performance, which may be due to the channel-wise impact has been learned by the previous convolutional layer of each domain, thus we only need to consider the impact of domain-wise. The TCD and GCD show similar performance on the matched test set.

The results of models evaluated on unmatched VCTK-2Mix test set are reported in Table 2. GCD-TasNet achieves an improvement of 0.35dB in SI-SNRi than TCD-TasNet. Considering two models has similar performance on matched test set, this implies that using trainable weights is sensitive to the mismatch between training and test conditions and using global information has better generalizability.

4. Conclusion

In this paper, we propose several feature selection encoders for time and frequency domain features in speech separation. The results on the matched dataset Libri2Mix confirm the validity of using frequency domain features to reconstruct separated signals in the decoder and the results on the unmatched dataset VCTK2Mix shows that using global information to generate selection weights is benefical to the generalizability of cross domain model. Our further work includes investigating more different cross domain feature selection module, the impact of selecting more different domain features such as complex spectrums, and using the proposed mechanism in other components of model.

5. Acknowledgment

This work was supported in part by the National Science Foundation of China (U19B2028, 61772117), National Science and Technology Commission Innovation Project (19-163-21-TS-001-042-01), CETC Big Data Research Institute (10-2018039), and the Fundamental Research Funds for the Central Universities (ZYGX2019J077).

6. References

[1] E Colin 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] Peng Li, Yong Guan, Bo Xu, and Wenju Liu, “Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2014-2023, 2006.

[3] Zi Wang, and Fei Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proc. ICASSP*, 2014, pp. 3749-3753.

[4] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31-35.

[5] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper 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.

[6] Dong Yu, Morten Kolbæk, Zheng Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241-245.

[7] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246-250.

[8] Yi Luo, Zhuo Chen, and Nima Mesgarani, “Speaker-independent speech separation with deep attractor network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787-796, 2018.

[9] Yi Luo, and Nima Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018, pp. 696-700.

[10] Yi Luo, and Nima 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.

[11] David Ditter, and Timo Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," in *Proc. ICASSP*, 2020, pp. 36-40.

[12] Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Filterbank design for end-to-end speech separation," in *Proc. ICASSP*, 2020, pp. 6364-6368.

[13] Gene-Ping Yang, Chao-I Tuan, Hung-Yi Lee, and Lin-shan Lee, "Improved Speech Separation with Time-and-Frequency Cross-domain Joint Embedding and Clustering," in *Proc. Interspeech*, 2019, pp. 1363-1367.

[14] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. ECCV*, 2016, pp. 47-54.

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.

[16] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang, "Selective kernel networks," in *Proc. CVPR*, 2019, pp. 510-519.

[17] Vinod Nair, and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, 2010.

[18] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, “LibriMix: An Open-Source Dataset for Generalizable Speech Separation,” *arXiv preprint arXiv:2005.11262*, 2020.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206-5210.

[20] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, “CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit,” *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[21] Diederik P Kingma, and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.