# Boosting the Performance of SpEx+ by Attention and Contextual Mechanism

*Chenyi Li*[1,2,†] *, Zhiyong Wu*[1,3,*]*, Wei Rao*[2,*]*, Yannan Wang*[2]*, Helen Meng*[1,3]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Tencent Ethereal Audio Lab, Tencent Corporation, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong SAR, China

licy20@mails.tsinghua.edu.cn, zywu@se.cuhk.edu.hk, {ellenwrao, yannanwang}@tencent.com, hmmeng@se.cuhk.edu.hk

## Abstract

Target speaker extraction (TSE) aims to mimic human selective attention to extracting our interested voice from the multi-talker environment. Time-domain methods represented by SpEx+ [1] have promoted the process of TSE tasks while residual noise, squeaks, and over-suppression still exist in the extracted speech. In this paper, we explore three ways to improve the performance of SpEx+, referring to two attention-based weight learning mechanisms on disparate dimensions to generate typical features and the context mechanism to refine the extracted masks. Experiments on both single-channel and multi-channel signals preliminarily demonstrated the effectiveness of our explored methods on SpEx+, especially on speech quality and alleviating squeaks, unexpected noises, and over-suppression.

**Index Terms**: Target speaker extraction, SpEx+, attention, contextual mechanism

## 1. Introduction

Real-world speech communication is invariably polluted by background noise and corrupted by speaker interference. Deep learning based speech separation methods such as deep clustering (DC) [2], permutation invariant training (PIT) [3] and deep attractor network (DANet) [4] are proposed to handle the permutation problem [2] of speaker separation. However, mismatch problems of estimated magnitude and original phase [5] exist in the above methods since they only estimate the magnitude of signals, leaving the phase unchanged. Phase-sensitive mask [6], complex ratio mask [7] and time-domain methods represented by TaSNet [8, 9] solve these issues by using phase information for extracted masks or directly modeling the mixture waveform by an encoder-decoder framework. However, they require knowing or estimating the number of speakers.

TSE, from a different perspective, aims at extracting target speaker speech from the mixture. It mimics the selective auditory attention of human brains to focus on the voice that appeals to us, which is rewarding to the scene of online meetings, discussion rooms, call centers, and so on. Simply capturing attention to the target speaker can avoid the problem of an unknown number of speakers. Early speaker extraction algorithms like SpeakerBeam [10] and VoiceFilter [11] are implemented in the frequency-domain as well, they inherently suffer from phase estimation problems like PIT et al [2, 3, 4]. Conv-TasNet [9] based time-domain approaches like SpEx [12], SpEx+ [1] and TD-SpeakerBeam [13] are proposed to solve the problem. However, a common problem with time-domain methods is the unpleasant residual in high frequencies, making squeaks, residual noise and over-suppression still exist in the extracted speech.

† Work done during internship at Tencent.
* Corresponding authors.

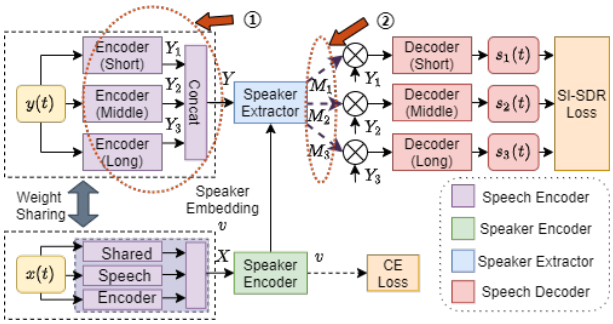


Figure 1: *Framework of SpEx+ [1]. Two orange dotted elliptic lines circled the location of our investigated mechanisms depicted in Figure 2 (①) and Figure 3 (②). The Shared Speech Encoder for the reference speech $x(t)$ is the twin speech encoder same as the speech encoder for the mixture speech $y(t)$.*

Studies on attention mechanism [14, 15] show that learning the attention weight based on feature interaction rather than equally weighting or averaging can improve the typicality of embedding features. [16] proved that speaker characteristics of time-varying, context-dependent bias vectors are beneficial to TSE task. Moreover, DeepFilter [17] adopts a complex filter concerning the context information of time and frequency domain to reduce the interference, especially at high frequencies.

Inspired by the above work, in this study, we mainly focus on attention-based weight learning (Figure 1①) and context-dependent masks (Figure 1②) for SpEx+ [1] to reduce squeaks, residual noise, and over-suppression of the extracted speech. First, as circled in Figure 1①, to generate representative features for the extractor network, we explore two attention-based weight learning methods upon the channel and scale dimension, respectively. Then, in Figure 1②, a context mechanism is applied to the masks extracted by speaker extractor network to refine the speaker masks, which can obviously improve the voice quality proven by objective evaluations. Experiments on both single-channel and multi-channel signals have demonstrated the effectiveness of our explored methods.

## 2. SpEx+

As depicted in Figure 1, SpEx+ is a complete time-domain speaker extraction network consisting of speech encoder, speaker encoder, speaker extractor, and speech decoder [1]. It solves the mismatch problem of latent feature space between time-domain speech encoder and frequency-domain speaker encoder in SpEx [12] by proposing time-domain twin speech encoders. The twin speech encoders obtain embedding coeffi-

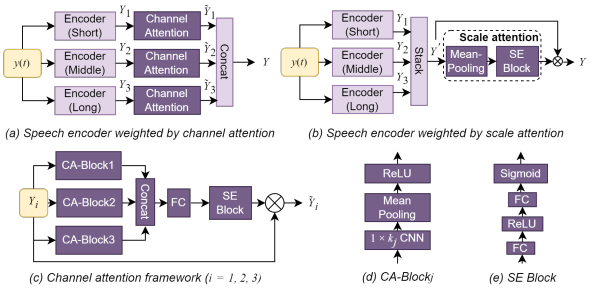


Figure 2: *Framework of the weighted speech encoder. We explore two ways for encoder weight learning: one depicted in (a) is channel attention for encoder in each scale, the other is scale attention in (b) for encoder stacked from each scale to learn weights of different scales. (c) and (d) are the detailed structures of channel attention and (e) depicts the SE block [20].*

cients in various temporal resolutions through several parallel 1-D CNNs with different filter lengths. Multi-scale speech encoders implemented in twin speech encoders can enhance the performance since multiple temporal scales represent detailed information in phonemic, prosodic, and other acoustic information in temporal structure [18]. Moreover, TCN [19] in the speaker extractor can replace deep LSTM networks for separation step in a more efficient and lighter-weight way [12]. And the speech decoder restores the signals of three scales from extracted speech features.

Nevertheless, Figure 4 shows that squeaks, residual noise, and over-suppression still exist in the speech extracted by SpEx+. To improve the mentioned shortcomings, we explore three methods on weighted speech features and refined masks. The first two are for the speech encoder to learn weights with attention to disparate dimensions, and the third is for the speaker extractor to extract more precise masks.

## 3. Explored Methods to Improve SpEx+

As circled in Figure 1, this paper mainly focuses on improving the speech encoder (①) and mask layer of speaker extractor (②). Speech encoder is considered since representative features provide rewarding information to the following extractor [14]. Thus, we explore the attention mechanism to learn weighted features with feature interaction. Since for time-domain feature embedding, each channel dimension has its own characteristic, we explore attention weights on channel dimension. As manifested in SpEx+ [1], multi-scale [18] speech encoder boosts the performance of the system, we use attention-based weight learning on scale dimension to highlight the contribution of multi-scale as well. As the mask layer of speaker extractor is vital to the extracted speech, the refined mask generated by the context mechanism is our strategy to reduce squeaks, residual noise, and over-suppression.

### 3.1. Multi-kernel channel attention for channel weight learning

We set the channel attention function after the three speech encoders as shown in Figure 2(a). Since time-domain features $Y_i \in R^{N\times T}$ $(i = 1, 2, 3)$ in each scale are extracted directly from the waveform, coefficients in the channel dimension $N$ of different scales in dimension $T$ contain plenty speech information. Attentionally weighting the coefficients rather than purely

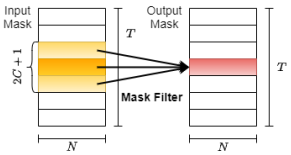


Figure 3: *Context mechanism on masks. The deep yellow block of the input mask is the one to be filtered, and all the yellow blocks are considered regions to be mapped to the one red block of the output mask. In this figure, C is set to 1.*

concatenating them would emphasize their conspicuous feature information. Considering this, multi-kernel is used for computing weights concerning multi-scale speech features for each scale in time domain.

As depicted in Figure 2(c), enlightened by the theory of multi-scale [18], we first configure three parallel Channel Attention Blocks (CA-Blocks, Figure 2(d)) consisting of 1-D CNN with different sizes of kernel size $k_j$ (disparate from the number of scale $i$, $j$ is the number of blocks for each scale, $j = 1, 2, 3$), followed by their corresponding 1-D mean-pooling layers and ReLU activation units. Then, the three parallel features are concatenated into a fully-connected layer before the general squeeze-and-excitation block [20] (SE block, shown in Figure 2(e)). Each scale has their own channel attention block $f_i(\cdot)$ for learning the weight coefficients $W_{CA} \in R^{N\times 1}$ to transform $\tilde{Y}_i \in R^{N\times T}$ from $Y_i \in R^{N\times T}$( $i = 1, 2, 3$) as follows:

$$\tilde{Y}_i = f_i(Y_i) = W_{CA} \cdot Y_i, i = 1, 2, 3 \quad (1)$$

$$Y = [\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3] \quad (2)$$

The output of channel attention blocks $\tilde{Y}_i$ are concatenated as $Y \in R^{3N\times T}$ before putting into the speaker extractor.

### 3.2. Scale attention for multi-scale weight learning

As depicted in Figure 2(b), different from channel attention, scale attention (abbreviated to SA in Table 1) for multi-scale weight learning is configured after the stack of the output of the original SpEx+ multi-scale speech encoder. For multi-scale features which retain multiple temporal resolutions, it is necessary for the system to attach different importance to different resolutions. We investigate attention-based weight learning since importing attention weights for different scale features will stick out the importance of different scales to the system. In detail, we stack the three encoder features to $Y' \in R^{3\times N\times T}$ and put it to a 2-D mean pooling layer for dimension reduction followed by the SE block [20] (shown in Figure 2(e)) to obtain a proper weight $W \in R^{3\times 1\times 1}$ of each scale as the factor of $Y'$ to calculate the weighted speech feature $\tilde{Y} \in R^{3\times N\times T}$ as follows:

$$\tilde{Y} =< W, Y' > \quad (3)$$

Here, $< \cdot, \cdot >$ means the dot product. Then, we chunk $\tilde{Y}$ into three latent features with dimension of $R^{N\times T}$ and concatenate them in the first dimension to obtain $Y \in R^{3N\times T}$.

### 3.3. Context mechanism for extracted masks

As shown in Figure 4, in the evaluation on the waveform extracted by original SpEx+, we found that there were squeaks, unexpected noises, and over-suppression in the signals. Our experiments demonstrate that one of the reasons is that the masks extracted by the original speaker extractor are not precise

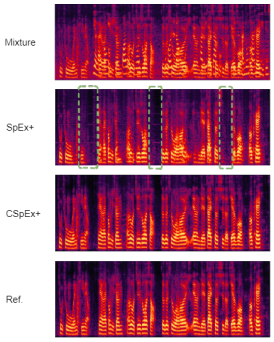


Figure 4: *Spectrogram comparison on speech extracted by SpEx+, CSpEx+ (context mechanism based SpEx+), Ref. (the reference speech), and Mixture (the mixture speech). Three green dotted boxes circled the main squeaks, along with over-suppression in the speech extracted by SpEx+. From other parts of the spectrogram in SpEx+, we can also find over-suppression and residual noise. However, they are obviously reduced by CSpEx+ as represented in this Figure.*

enough. As the context of the feature contains beneficial information, we explore a context mechanism inspired by DeepFilter [17] for the original masks by taking the context information of the masks into account. Considering time-domain embedding features only have continuity in the time domain, we mainly use the context information from time domain and regard the $N$ channels as an entirety. Figure 3 shows the scheme of context mechanism where $2C + 1$ is the context length. We use a Conv1D layer to implement the context mechanism where the different kernel size indicates the different context length. Spectrogram comparison in Figure 4 demonstrates that context mechanism could effectively reduce squeaks, noises, and over-suppression well.

# 4. Experiments

## 4.1. Dataset

Our experiments are performed on the multi-channel reverberant WSJ0-2mix corpus (MC-WSJ0-2mix) [13, 21]. The WSJ0-2mix dataset has been widely used in plenty of single-channel speech separation studies [1, 2, 3, 4, 8, 9, 10, 12] since the debut of DC [2]. In this work, we also choose MC-WSJ0-2mix in order to explore the performance of improved SpEx+ on the reverberant dataset and even dual-channel spatialized dataset.

The multi-channel recordings are generated by RIR generator using image method for reverberation time of up to 600ms [13]. The dataset containing 8 channel recordings, and we use the first channel for the single-channel TSE task and the first two channels for the multi-channel task's spatial feature extraction. There are respectively 20k (∼30h), 5k (∼10h), and 3k (∼5h) utterances in the training, validation, and test set resampled to 8kHz from 16kHz sampling rate. Note that the test speakers are unseen in the training and validation set and the reference recordings for evaluation are the clean reverberant speech of each source at the first channel.

## 4.2. Implementation details

Our experimental setup follows the same configuration in SpEx+ [1]. The filter lengths of convolutions in multi-scale speech encoder and decoder are $L_1 = 2.5ms, L_2 = 10ms, L_3 = 20ms$ for speech of 8kHz sample rate. For attention mechanism, the three kernel sizes we use in multi-kernel channel attention are $k_1 = 3, k_2 = 5, k_3 = 9$, and the three parallel Conv1D layers are convoluted with parameter *groups* set to the number of channels $N = 256$ to interrupt the relationship between each channel and reduce the cost. For scale attention, the number of hidden channels between two fully-connected layers is set to 6. As for the contextual mechanism, select a context length $C$, the kernel size of the Conv1D layer will be $2 \times C + 1$.

Since our experiments are conducted on the multi-channel corpus, each of our methods are evaluated on both single-channel tasks and multi-channel tasks. For multi-channel tasks, we import the original Channel Decorrelation (CD) [22] to extract the multi-channel spatial information from the first two channels and add it up with the mixture speech encoder from the first channel as the updated mixture speech encoder. The second channel features are calculated the same way as the first but with separated parameters. Note that the attention mechanism is performed on the updated speech encoder on the multi-channel task here.

Table 1: *Evaluations on single-channel attention-based weight learning of different methods. CA and SA mean channel attention depicted in section 3.1 and scale attention depicted in section 3.2. "Ours" means our explored attention mechanism.*

| Method | | Para./M | SI-SDR | PESQ | ESTOI/% |
|---|---|---|---|---|---|
| Mixture | | - | 2.490 | 2.323 | 60.9 |
| SpEx+ | | 11.27 | 10.911 | 3.016 | 78.9 |
| CA | +spk | 11.70 | 11.094 | 3.051 | 79.5 |
| | +SC | 11.48 | 10.876 | 3.003 | 78.7 |
| | dec. | 11.48 | 10.893 | 3.018 | 79.4 |
| | time | 19.04 | 10.878 | 2.998 | 78.8 |
| | Ours | 11.48 | **11.105** | **3.054** | **79.6** |
| SA | S-MaxPool | 11.14 | 10.885 | 3.022 | 79.1 |
| | sum. | 11.14 | 10.836 | 3.014 | 78.8 |
| | +spk | 11.27 | 11.117 | **3.059** | 79.6 |
| | +SC | 11.27 | 11.068 | 3.046 | 79.5 |
| | dec. | 11.27 | 11.000 | 3.024 | 79.2 |
| | Ours | 11.27 | **11.190** | 3.057 | **79.7** |

## 4.3. Comparative study on attention mechanism

We do comparative studies to explore the potential of channel attention (*CA*) and scale attention (*SA*), the two investigated attention mechanisms stated in Table 1. In Table 1, *Mixture* is the original mixture signal to be extracted and *SpEx+* is the evaluation result of our baseline on the first channel of MC-WSJ0-2mix. Comparative studies are done in the following ways:

- *+spk*: simultaneously performing the same but standalone corresponding attention mechanism on speaker embedding encoded by twin speech encoder;
- *+SC*: using skip connection on the learned weight;
- *dec.*: performing the attention mechanism in the speech decoder rather than the speech encoder;
- *time*: performing the weight learning on the time dimension $T$ rather than the channel dimension $N$;
- *sum.*: adding the three weighted features to feature at the size of $R^{N\times T}$ at the end of the scale attention mechanism, while other ways concatenate them to the size of $R^{3N\times T}$;
- *S-MaxPool*: replacing the concatenation of speech encoder of three scales without attention with a max-pooling operation.

Table 2: *Comparative results of context mechanism on single-channel signals. Context (past only) means only considering the past information while Context (ours) means considering the context information. For Context (past only) and Context (ours), $C$ is the past information or context length in the time dimension. For Context&Channel, we apply context mechanism simultaneously on time and channel dimension where $C$ is the context length in the channel dimension with the length in time domain fixed at $\{8,4,1\}$.*

| Method | $C$ | Para./M | SI-SDR | PESQ | ESTOI/% |
|---|---|---|---|---|---|
| SpEx+ | | 11.27 | 10.911 | 3.016 | 78.9 |
| Context&Channel | 0 | 11.27 | 11.015 | 3.028 | 78.8 |
| | 1 | 11.28 | 10.682 | 2.980 | 77.7 |
| | 4 | 11.28 | 10.730 | 2.986 | 78.1 |
| Context (past only) | 1 | 11.67 | 11.094 | 3.046 | 79.5 |
| | 2 | 11.86 | 10.994 | 3.033 | 79.1 |
| Context (ours) | 0 | 11.47 | 10.954 | 3.032 | 79.2 |
| | 1 | 11.86 | **11.221** | **3.070** | **79.9** |
| | 2 | 12.26 | 10.889 | 3.025 | 78.9 |

For *CA*, the result of *time* shows learning weights on time dimension achieves worse performance than *Ours* with computational cost proliferates. *+SC* reduces the performance since it undermines the advantage of weights learned here. Compared with weights learned in the decoder, learning weights in the encoder has an improvement of 0.212 in SI-SDR and 0.036 in PESQ since speech features in the encoder, the start of the system, have more fruitful information.

Compared *S-MaxPool* and *sum.* in *SA* with other configurations with concatenated weighted features, we find retaining the weight in each scale is crucial to the system, which indicates the advantage of multi-scale as well. Moreover, performances of *+spk* in *CA* and *SA* are slightly worse than *Ours*, which manifests learning weights from mixture speech is enough to boost the performance of SpEx+.

### 4.4. Comparative study on context mechanism

Table 2 shows our comparative studies on *Context&Channel* referencing DeepFilter [17] and mechanism with past information only (*past only*). For *Context&Channel*, we select the kernel size of $[3, 5]$ for the mask filter. The context length on time dimension for each scale aims to complementary to the proportion of filter length $L_1, L_2, L_3$ in the speech encoder. For *Context (past only)*, given the context size of $C$, kernel size will be $C + 1$ and zero-padding will only be configured at the start of the embedding at the size of $C$.

Results in Table 2 show that *Context&Channel* performs poor. Since in time-domain methods, coefficients in channel dimension have little continuity and interpretability, it is improper to pay contextual attention. Comparisons on *past only* and *ours* also demonstrate the effectiveness of our context mechanism in time-domain signals, especially in speech quality and intelligibility. Moreover, evaluations upon different context lengths of *Context* indicate that proper context length is also crucial to the performance of the context mechanism.

### 4.5. Extended experiments on multi-channel signals

To further verify our investigated methods, we do experiments on multi-channel signals as well. Table 3 shows the main experiments of our explored methods. Experimental rules on single-channel signals are also applicable to multi-channel signals. SpEx+ with context mechanism outperforms all of the other methods and achieves 0.051 and 0.8% improvements in PESQ and ESTOI respectively, compared with SpEx+ with CD. We compare our explored methods with the improved CD algorithm [23] with experiments on speaker adaptation in Table 4. Results show that our explored method with context mechanism w/o *CDSA* gets the best speech quality. Table 3 and Figure 4 demonstrate that our explored attention-based weight learning and context mechanism can improve speech perceptual quality by reducing squeaks, residual noise, and over-suppression in the extracted speech.

Last but not least, experiments on systems combined with attention and context mechanism or two attention mechanisms show that the performance is worse than that only with context mechanism or with one attention mechanism. However, a system with either outperforms the baseline in both single-channel and multi-channel tasks, which manifests the effectiveness of our explored methods. Compared all the results we can also find that importing spatial features to the system can get a greater effect gain than optimizing the framework of network.

Table 3: *Evaluations of multi-channel signals on our studied methods with CD. As illustrated in Table 1, CA and SA mean channel attention and scale attention. Context length $C$ of context mechanism is set to 1.*

| Method | | Para./M | SI-SDR | PESQ | ESTOI/% |
|---|---|---|---|---|---|
| SpEx+ | | 11.41 | 12.010 | 3.149 | 81.6 |
| CA | dec. | 11.62 | 11.908 | 3.124 | 81.2 |
| | Ours | 11.62 | 12.078 | 3.166 | 81.8 |
| SA | dec. | 11.41 | 12.080 | 3.161 | 81.7 |
| | Ours | 11.41 | 12.041 | 3.150 | 81.5 |
| Context | past only | 12.00 | 12.031 | 3.155 | 81.8 |
| | Ours | 12.00 | **12.302** | **3.200** | **82.4** |

Table 4: *PESQ performance of SpEx+ and CSpEx+ (SpEx+ with context mechanism) compared with the best two improved CD algorithm [23] on multi-channel signals. +CDSA here means CD with speaker adaptation stated in [23].*

| Method | CD-Unrolled | | CD-Para-a | SpEx+ | | CSpEx+ | |
|---|---|---|---|---|---|---|---|
| | *+CDSA* | - | *+CDSA* | *+CDSA* | - | *+CDSA* | - |
| PESQ | 3.143 | 3.107 | 3.142 | 3.123 | 3.149 | 3.059 | **3.200** |

## 5. Conclusion

In this paper, we mainly explore three approaches to reduce the squeak, residual noise, and over-suppression of the extracted speech based on SpEx+ [1]. First is the weight-learning mechanism based on attention on channel dimension, second is on scale dimension, and the rest is the context mechanism on masks for purification. Experiments on single-channel and multi-channel speech both demonstrate the effectiveness of our investigations, especially in speech quality. Future work will continue to improve performance of SpEx+ by importing an extended post-filter to the system, applying context mechanism to different parts in SpEx+, and do additional evaluations for different tasks (e.g., speech denoising).

## 6. Acknowledgements

This work is supported by Shenzhen Science and Technology Innovation Committee (WDZC20200818121348001), National Natural Science Foundation of China (62076144).

# 7. References


[1] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network." *conference of the international speech communication association*, 2020.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35, iSSN: 2379-190X.

[3] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, 2017.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 246–250, iSSN: 2379-190X.

[5] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," *Learning*, 2019.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 708–712, iSSN: 2379-190X.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[8] Y. Luo and N. Mesgarani, "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)*, Apr. 2018, pp. 696–700, iSSN: 2379-190X.

[9] ——, "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, Aug. 2019, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[10] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, Aug. 2019, conference Name: IEEE Journal of Selected Topics in Signal Processing.

[11] H. Muckenhirn, I. L. Moreno, J. R. Hershey, K. W. Wilson, P. Sridhar, Q. Wang, R. A. Saurous, R. Weiss, Y. Jia, and Z. Wu, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *conference of the international speech communication association*, 2019.

[12] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[13] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving Speaker Discrimination of Target Speech Extraction With Time-Domain Speakerbeam," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 691–695, iSSN: 2379-190X.

[14] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-net: Target speaker separation via attention-based neural network," *conference of the international speech communication association*, 2020.

[15] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 171–178.

[16] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 86–90.

[17] W. Mack and E. A. P. Habets, "Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2020, conference Name: IEEE Signal Processing Letters.

[18] Z. Zhu, J. Engel, and A. Hannun, "Learning multiscale features directly from waveforms," *conference of the international speech communication association*, 2016.

[19] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[21] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5.

[22] J. Han, X. Zhou, Y. Long, and Y. Li, "Multi-channel target speech extraction with channel decorrelation and target speaker adaptation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6094–6098.

[23] J. Han, W. Rao, Y. Wang, and Y. Long, "Improving channel decorrelation for multi-channel target speech extraction," *conference of the international speech communication association*, 2021.