INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# MC-SpEx: Towards Effective Speaker Extraction with Multi-Scale Interfusion and Conditional Speaker Modulation

*Jun Chen*[1,2,†], *Wei Rao*[2], *Zilin Wang*[1], *Jiuxin Lin*[1], *Yukai Ju*[2], *Shulin He*[2]
*Yannan Wang*[2], *Zhiyong Wu*[1,3,*]

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

y-chen21@mails.tsinghua.edu.cn, ellenwrao@tencent.com, zywu@sz.tsinghua.edu.cn

## Abstract

The previous SpEx+ has yielded outstanding performance in speaker extraction and attracted much attention. However, it still encounters inadequate utilization of multi-scale information and speaker embedding. To this end, this paper proposes a new effective speaker extraction system with multi-scale interfusion and conditional speaker modulation (ConSM), which is called MC-SpEx. First of all, we design the weight-share multi-scale fusers (ScaleFusers) for efficiently leveraging multi-scale information as well as ensuring consistency of the model's feature space. Then, to consider different scale information while generating masks, the multi-scale interactive mask generator (ScaleInterMG) is presented. Moreover, we introduce ConSM module to fully exploit speaker embedding in the speech extractor. Experimental results on the Libri2Mix dataset demonstrate the effectiveness of our improvements and the state-of-the-art performance of our proposed MC-SpEx.

**Index Terms**: speaker extraction, multi-scale interfusion, conditional speaker modulation

## 1. Introduction

Speech separation, commonly known as the cocktail-party problem, is a fundamental challenge in the field of speech processing that intends to separate each source signal from the mixed speech of multiple speakers. Most studies on speech separation are limited by the requirement for prior knowledge of the number of speakers, and additionally they entail addressing the global permutation ambiguity challenge [1] to channel the correct speaker to the correct output voice stream. In order to avoid these constraints, speaker extraction is proposed as a strategy that only extracts target speaker's speech from the mixture according to the reference speech from target speaker. It can be used in a variety of downstream applications, including automatic speech recognition (ASR), real-time communication (RTC) and speaker diarization, just to name a few.

Motivated by humans' top-down auditory attention to the target speaker [2, 3], deep learning-based speaker extraction methods [4–12] primarily adopt a two-subnet architecture consisting of a speaker encoder and a speech extractor, in which the speaker encoder models the speaker representation of the target speaker, and then directs the speech extractor to extract the speech signal belonging to the target speaker. While following the previous practice, SpEx+ [13] further introduces a twin speech encoder with shared parameters to capture multi-scale speech features directly from the speech waveform for the speaker encoder and speech extractor, and reverts their processed multi-scale features to the waveform via the speech decoder. By doing so, a uniform latent feature space containing multi-scale information is introduced for both the subnets of speaker encoder and the speech extractor, resulting in outstanding results for SpEx+.

Despite the impressive performance, SpEx+ is still not perfect. The remarkable achievements of Conformer in source separation [14, 15] and speaker verification [16] demonstrate that the effective fusion of multi-scale speech information enables neural networks to leverage more comprehensive acoustic features, which is conducive to the model performance. Nevertheless, in SpEx+, the fusion of the multi-scale information extracted by speech encoder could be further boosted. And, while the twin speech encoder of SpEx+ considers the consistency of the two subnets' feature space [13], the original independent fusion modules within each subnet do not take this into account so far. Besides, when decoding the mask, SpEx+ merely employs three separate branches to generate three masks for different scales, which fails to adequately combine and utilize information in multiple scales. Furthermore, it has been pointed out that the speaker embedding, which is concatenated with the frame-level speech features, cannot be sufficiently exploited by the stacked temporal convolutional network (TCN) blocks [17]. We argue that this underutilization of speaker embedding arisen in SpEx+ may compromise the model's discrimination for target speaker in mixed speech, consequently leading to a limited capability of the speaker extraction system.

In this paper, to tackle the insufficient utilization of multi-scale information as well as speaker embedding in SpEx+, we propose an efficient speaker extraction system called MC-SpEx. We design the ScaleFuser to more effectively leverage multi-scale information extracted from the speech waveform. Afterwards, for the consistency of two subnets' feature space, we incorporate the newly designed ScaleFusers with shared parameters into the twin speech encoders to obtain twin multi-scale fusion speech encoders. Correspondingly, the ScaleInterMG is also presented to substitute the original three independent branches, so as to take the different scale information into account when generating masks. Together with the original speech decoder, the ScaleInterMG constitutes the multi-scale interactive speech decoder. Furthermore, we introduce ConSM module to fully blend speaker embedding into the speech extractor. Experimental results show that MC-SpEx significantly outperforms the performance of our baseline SpEx+ in extracting the speech signals of the target speaker, which confirms the effectiveness of our improvements. Moreover, our proposed MC-SpEx also yields state-of-the-art results for the speaker extraction task on the Libri2Mix dataset [18].

## 2. Methodology

As shown in Figure 1, MC-SpEx is composed of four main parts: multi-scale fusion speech encoders, speaker encoder,

[†] Work conducted when the first author was intern at Tencent.
[*] Corresponding author.


10.21437/Interspeech.2023-1130

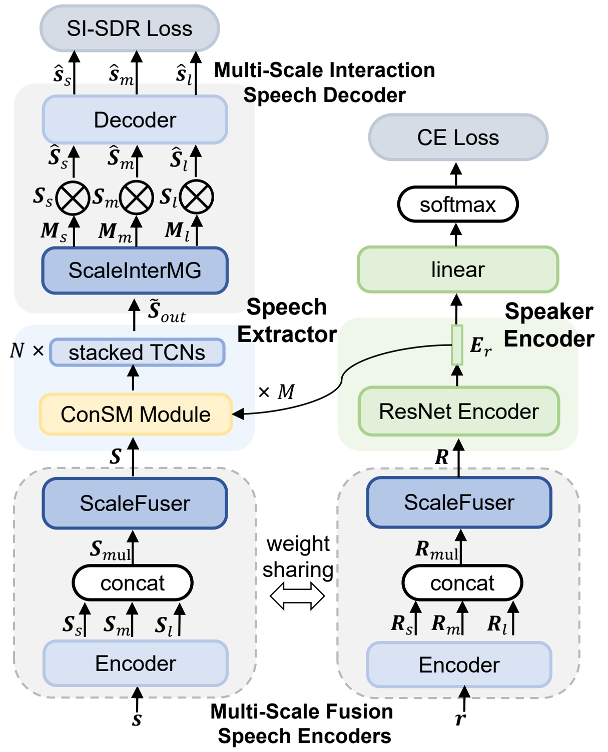
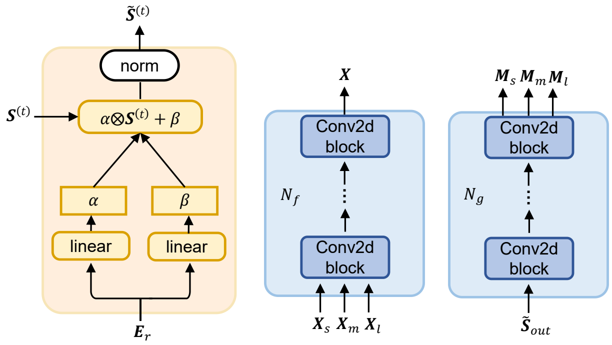



Figure 1: *The overall diagram of the MC-SpEx. The dotted border represents the module with shared weights. The "⊗" means element-wise multiplication. The "SI-SDR Loss" and "CE Loss" refer to the scale-invariant signal-to-distortion ratio loss and cross-entropy loss in multi-task learning [1].*

speech extractor, and multi-scale interaction speech decoder. One of the subnets, the speaker encoder, uses a residual network (ResNet) encoder [19] comprising of stacked ResNet blocks and a pooling layer. As the other subnet, the speech extractor consists of $M$ groups of speaker-guided stacked TCN blocks. Inside each group, there are the ConSM module in the front end and $N$ stacked TCNs with exponentially growing dilation factors $\{2^n\}(n \in \{0, \cdots, N-1\})$.

The proposed model takes the mixed speech waveform $\mathbf{s}$ and reference speech waveform $\mathbf{r}$ as inputs. In the weight-share multi-scale fusion speech encoders, the encoder respectively extracts multi-scale speech features $\mathbf{S}_{mul} = [\mathbf{S}_s, \mathbf{S}_m, \mathbf{S}_l]$ and $\mathbf{R}_{mul} = [\mathbf{R}_s, \mathbf{R}_m, \mathbf{R}_l]$ from waveforms $\mathbf{s}$ and $\mathbf{r}$, where "$[,]$" denotes the concatenation operation and the subscripts $s$, $m$ and $l$ refer to small, middle, and large scales respectively as in [13]. And then, the ScaleFusers are responsible for fusing the multi-scale information, in which $\mathbf{S}_{mul}$ and $\mathbf{R}_{mul}$ are fused into $\mathbf{S}$ and $\mathbf{R}$, respectively. $\mathbf{R}$ is fed to speaker encoder and then we get the speaker embedding $\mathbf{E}_r$ from it to represent the characteristics of the target speaker. The speech extractor, under the direction of $\mathbf{E}_r$, outputs processed speech features $\widetilde{\mathbf{S}}_{out}$ based on another input $\mathbf{S}$. Eventually, in multi-scale interaction speech decoder, the ScaleInterMG interactively generates multi-scale receptive masks $\mathbf{M}_s$, $\mathbf{M}_m$ and $\mathbf{M}_l$ from $\widetilde{\mathbf{S}}_{out}$. After the element-wise multiplication between multi-scale masks and the corresponding multi-scale features in $\mathbf{S}_{mul}$, we obtain the speech features $\widehat{\mathbf{S}}_s$, $\widehat{\mathbf{S}}_m$ and $\widehat{\mathbf{S}}_l$ predicted by the model, and these features are transformed by decoder to gain the estimated speech waveform $\hat{\mathbf{s}}_s$, $\hat{\mathbf{s}}_m$ and $\hat{\mathbf{s}}_l$. These predicted speech waveforms and $\mathbf{E}_r$ are eventually used to calculate the multi-task




(a) ConSM Module (b) ScaleFuser (c) ScaleInterMG

Figure 2: *(a) The details of the ConSM, where "⊗" indicates the element-wise product. (b) The details of the ScaleFuser. (c) The details of the ScaleInterMG.*

learning loss for training the model as described in [1]. We will elaborate on our improvements in the following subsections.

### 2.1. Multi-Scale Fuser

The notable successes of Conformer in source separation [14, 15] and speaker verification [16] evince that the effective fusion of temporal information of speech on multiple time scales [20, 21] enables neural networks to exploit richer acoustic features, which is conducive to the performance of the model. However, the fusion of the multi-scale information in SpEx+ is quite rough. It simply concatenates $\mathbf{S}_l$, $\mathbf{S}_m$ and $\mathbf{S}_s$ in the feature dimension, and fuses them using a 1-D convolution with kernel size of 1. This approach fails to consider the local information of the feature dimension as well as the information in neighboring frames, and is not enough to sufficiently fuse multi-scale features.

As a consequence, we propose the more efficient multi-scale feature fusion module ScaleFuser as an alternative to the original coarse method. The structure of ScaleFuser is shown in Figure 2(b), which consists of $N_f$ stacked Conv2d blocks, each of them contains 2-D convolution, an activation function ELU [22]. The module takes the multi-scale features $\mathbf{X}_s$, $\mathbf{X}_m$ and $\mathbf{X}_l$ as input, where $\mathbf{X}$ stands for $\mathbf{S}$ or $\mathbf{R}$. We consider $\mathbf{X}_s$, $\mathbf{X}_m$ and $\mathbf{X}_l$ as different channels, and then interactively fuse the features of different channels through the 2-D convolution kernels in Conv2d blocks. Eventually the feature $\mathbf{X}$ that adequately combines information from different scales is obtained:

$$\mathbf{X} = \mathcal{F}([\mathbf{X}_s, \mathbf{X}_m, \mathbf{X}_l]). \tag{1}$$

where $\mathcal{F}(\cdot)$ represents the mapping function defined by the ScaleFuser. Compared with the fusion in SpEx+, ScaleFuser takes the information of local feature dimensions and adjacent frames into account through 2-D convolution, which enables a more effective fusion of multi-scale features. Additionally, for the consistency of two subnets' feature space, we share the parameters of ScaleFusers in the multi-scale fusion speech encoders.

### 2.2. Conditional Speaker Modulation Module

It has been pointed out that the speaker embedding, which is concatenated with the frame-level speech features, cannot be sufficiently exploited by the stacked TCN blocks in the SpEx+ [17]. Meanwhile, in text-to-speech (TTS) works [23, 24], a small condition network called conditional layer normalization is implemented to modulate the hidden representations for synthesizing the target speaker's speech. Inspired by this and accounting for the differences between TTS and speaker extraction tasks, we present the ConSM for addressing the above speaker embedding under-utilization problem based on conditional layer normalization.

Specifically, the conditional layer normalization [24] adopts speaker-conditional affine transformation to modulate the hidden features that are performed layer normalization [25] at first. In contrast to TTS, in speaker extraction, the speech features extracted from the mixture contain the target speaker's speech information, in which case we believe that applying layer normalization first may result in the loss of such information. According to this, we tune the locations of layer normalization and speaker-conditional affine transformation, and apply it to the speaker extraction task, which is what we call ConSM. As in Figure 2(a), the ConSM takes speaker embedding $\mathbf{E}_r$ and the frame-level speech feature $\mathbf{S}^{(t)}$ in speech extractor as the inputs, where $t = 1, ..., T$ denotes the frame indices. We feed $\mathbf{E}_r$ to a linear layer with mapping function $\mathcal{D}(\cdot)$ and another linear layer with mapping function $\mathcal{R}(\cdot)$ to gain the adaptive scale vector $\alpha$ and bias vector $\beta$ respectively:

$$\alpha = \mathcal{D}(\mathbf{E}_r), \quad \beta = \mathcal{R}(\mathbf{E}_r). \tag{2}$$

We perform an affine transformation of $\mathbf{S}^{(t)}$ with $\alpha$ and $\beta$, then conduct a layer normalization on the result of it:

$$\widetilde{\mathbf{S}}^{(t)} = Norm(\alpha \otimes \mathbf{S}^{(t)} + \beta). \tag{3}$$

where $Norm$ is the layer normalization and $\otimes$ represents the element-wise product. Through this, depending on the condition of the given speaker embedding, we are able to scale up or down the speech features, negate them, and selectively set thresholds on them. Consequently, we can reinforce the information in the speech features that are relevant for extracting the target speaker in the ConSM module, instead of relying on the subsequent stacked TCNs to handle this, which resolves the speaker embedding under-utilization issue of the stacked TCNs in SpEx+. The conditional affine transformation in ConSM may be relatively similar to FiLM [26–28], but as described above, the problem it seeks to solve is different from FiLM. Furthermore, compared to FiLM, ConSM has an extra layer normalization operation after the affine transformation, which we argue is more appropriate for the speaker extraction task.

### 2.3. Multi-Scale Interactive Mask Generator

Previous works [1, 13, 17] all apply three independent branches to generate three scales of receptive masks. This leads to the issue that, when generating the mask of corresponding scale, each branch solely utilizes the information of one scale. To this end, we introduce ScaleInterMG. Through the interaction of different scales, it allows the model to make reasonable use of the valid information of other scales when generating the mask at a certain scale.

As shown in the Figure 2(c), the ScaleInterMG is composed of $N_g$ stacked Conv2d blocks, with each block containing a 2-D convolution, an ELU [22] and a layer normalization. The module takes $\widetilde{\mathbf{S}}_{out}$ as input, and then interactively fuses the features of different channels through the 2-D convolution kernel in the Conv2d block, ultimately treating the obtained features of three channels as multi-scale masks $\mathbf{M}_s$, $\mathbf{M}_m$ and $\mathbf{M}_l$:

$$\mathbf{M}_s, \mathbf{M}_m, \mathbf{M}_l = \mathcal{H}(\widetilde{\mathbf{S}}_{out}). \tag{4}$$

where $\mathcal{H}(\cdot)$ is the mapping function that is intended to describe the ScaleInterMG. By means of the above, ScaleInterMG can generate the mask at a certain scale with valid information from multiple scales, while reducing the number of parameters required by SpEx+ that employed three branches.

## 3. Experiments

### 3.1. Datasets

We conduct our experiments on the popular Libri2Mix dataset [18]. The *train-100* subset, which is used for training, contains a total of 58 hours of utterances from 291 speakers. The *dev* subset and *test* subset consist of 40 unseen speakers respectively, with an overall audio duration of 11 hours in each subset. The *dev* subset is served as the validation set during model training, while the *test* subset is for the evaluation of the model's final performance. For all speech audio, the sampling rate is 8 kHz. Moreover, all of mixtures are in the 'minimum' mode.

### 3.2. Training Setup and Baselines

We employ Adam optimizer with an initial learning rate of 1e-3. The learning rate decays by 0.5 once the performance on the validation set is not improved in 3 consecutive epochs. The training of the model will be stopped when the best model is not found in the validation set after 8 consecutive epochs. During the training procedure, both the mixed speech and the reference speech of target speaker are sliced into 3-second segments, while the full-length audio is applied at inference.

To testify the effectiveness of our improvement, the following models were compared. Aiming at a fair comparison, the identical experimental setup is implemented for each model. (1) **SpEx+:** Following [13], the convolutional filtering lengths of the speech encoder and decoder in SpEx+ are {2.5, 10, 20} ms respectively. The number of ResNet blocks [19] in ResNet speaker encoder is set to 3, and the dimension of speaker embedding is 256. The hyperparameters are $M = 4$ and $N = 8$ for speaker-guided stacked TCNs in speech extractor. The SpEx+ has a total of 11.78 M parameters. (2) **MC-SpEx:** There are 4 Conv2d blocks in ScaleFuser with 2-D convolution channels of {3, 32, 32, 1} and kernal sizes of {3, 3, 3, 3}. The ScaleInterMG contains 4 Conv2d blocks having kernal sizes of {3, 3, 3, 3}, and their 2-D convolution channels are {1, 32, 32, 3}. The other configurations are the same as described above. The number of parameters for MC-SpEx is 10.77 M.

In our experiments, we mainly evaluate the model performance through three objective metrics, SI-SDR, PESQ, and ESTOI. Among them, SI-SDR is measured from the signal perspective, while PESQ and ESTOI are considered from the perceptual quality perspective, and the higher values of them are all positively correlated with better outcomes.

Table 1: *The performance in terms of SI-SDR [dB], PESQ [MOS] and ESTOI [%] on the Libri2Mix test set.*

| Methods | SI-SDR | PESQ | ESTOI |
|---|---|---|---|
| Mixture | 0.001 | 1.603 | 53.8 |
| TD-SpeakerBeam [9] | 12.86 | 2.750 | - |
| sDPCCN [29] | 11.65 | 2.738 | 78.9 |
| TD-SpeakerBeam + $\text{PL}_2$+ $\text{PF}^{lin}$ [30] | 13.88 | 2.860 | - |
| SpEx+ [13] | 13.41 | 2.936 | 82.4 |
| MC-SpEx | **14.61** | **3.195** | **84.9** |

Table 2: *Performance of SI-SDR and PESQ in section 3.4.1 using the Libri2Mix test set. "SF" and "SIMG" mean ScaleFuser and ScaleInterMG respectively.*

| Ablation Modules | ID | Settings | SI-SDR | PESQ |
|---|---|---|---|---|
| | #1 | SpEx+ | 13.41 | 2.936 |
| Weight-share ScaleFusers | #2 | #1 + weight-share SFs | 14.05 | 3.075 |
| ScaleInterMG | #3 | #1 + SIMG | 13.72 | 3.059 |
| | #4 | #2 + SIMG | 14.32 | 3.153 |
| ConSM module | #5 | #1 + ConSM | 13.69 | 2.969 |
| | #6 | #4 + ConSM | **14.61** | **3.195** |

Table 3: *Performance of SI-SDR and PESQ in the investigation of weight-share ScaleFusers with the Libri2Mix test set. The "$Enc_{Ext}$" and "$Enc_{Spk}$" respectively denote the multi-scale fusion speech encoder in the subnet of speech extractor and speaker encoder.*

| Methods | SI-SDR | PESQ |
|---|---|---|
| SpEx+ | 13.41 | 2.936 |
| + ScaleFuser in $Enc_{Ext}$ | 13.55 | 2.955 |
| + ScaleFuser in $Enc_{Spk}$ | 13.70 | 3.031 |
| + shared weights | **14.05** | **3.075** |

### 3.3. Comparison with Baseline and State-of-the-art Methods

Table 1 shows the performance of different speaker extraction methods on the Libri2Mix test dataset. In the last two rows of the Table, the performances of SpEx+ and the proposed MC-SpEx are compared, and the results show that MC-SpEx outperforms the SpEx+ in all evaluation metrics. This indicates that our improvements have indeed improved the model's ability to extract the speech of target speakers.

We further compare the proposed MC-SpEx to some other top-ranked methods [9,29,30] on the Libri2Mix dataset in Table 1. Among them, the sDPCCN is a frequency-domain method with powerful capability [29]. While the TD-SpeakerBeam + $\mathrm{PL}_2$ + $\mathrm{PF}^{lin}$ has resolved the target confusion using the strategies of prototypical loss (PL) and post-filtering (PF) [30], which is the previous best speech extraction system on the Libri2Mix dataset. It can be concluded that comparing with these latest methods, MC-SpEx achieves the state-of-the-art results on the Libri2Mix dataset.

### 3.4. Ablation Studies

In this section, taking SpEx+ as backbone, we perform a series of ablation studies about our improvements.

#### 3.4.1. Overall Analysis on Proposed Modules

To start with, we analyze the contribution of our proposed modules to the model and the compatibility among them. From Table 2, we can observe that appending weight-share ScaleFusers, ScaleInterMG and ConSM individually to SpEx+ shows certain effectiveness (#2, #3 and #5). It follows that the interfusion of multi-scale information in ScaleFuser and ScaleInterMG, as well as using the ConSM to modulate, are beneficial to the model's performance. Furthermore, all of the weight-share ScaleFusers, ScaleInterMG and ConSM do not conflict with each other (#2, #4 and #6) and the combination of the three, the MC-SpEx, achieves the best results (#6).

Table 4: *The performance in terms of SI-SDR and PESQ in the investigation of ConSM module using the Libri2Mix test set. The "Conditional LN" refers to the conditional layer normalization.*

| Methods | SI-SDR | PESQ |
|---|---|---|
| SpEx+ | 13.41 | 2.936 |
| SpEx+ with Conditional LN [24] | 13.42 | 2.935 |
| SpEx+ with FiLM [27] | 13.55 | 2.955 |
| SpEx+ with ConSM module | **13.69** | **2.969** |

#### 3.4.2. Investigation of Weight-share ScaleFusers

We then further explore the role of the weight-share ScaleFusers. As can be seen from Table 3, the performance of the model has been improved after the additions of ScaleFuser to the multi-scale fusion speech encoder in the subnet of speech extractor and speaker encoder. This illustrates that our proposed ScaleFuser indeed enables more effective fusion of multi-scale features than the original method in SpEx+. Moreover, after sharing the weights of the ScaleFusers, the model performance is further boosted, which confirms effectiveness of our strategy to ensure the consistency of two subnets' feature space in the multi-scale feature fusion stage.

#### 3.4.3. Investigation of ConSM Module

In order to investigate the effects of the ConSM module, in table 4, we compare it with the approaches mentioned in section 2.2. One can see that, after applying a dedicated module to exploit speaker embedding, the usage of either the FiLM or ConSM module brings a performance gain respectively, which indicates that this strategy does improve the SpEx+ with insufficient utilization of speaker embedding. However, the addition of conditional layer normalization leads to almost no change in performance. This is probably because that, as analyzed in subsection 2.2, the normalization on the speech feature first operated in the conditional layer normalization introduces an information loss, which renders it inappropriate for the speaker extraction task. In addition, the SpEx+ with ConSM module exceeds the SpEx+ with FiLM in terms of both the signal (SI-SDR) and perceptual quality (PESQ). This also demonstrates that, compared to FiLM, our proposed ConSM is more suitable for speaker extraction.

## 4. Conclusions

In this paper, we propose a framework with multi-scale interfusion and conditional speaker modulation named MC-SpEx. It adopts the weight-sharing ScaleFusers to effectively exploit the multi-scale information and ensure the consistency of the two subnets' feature space. And then, the ScaleInterMG is presented to take the different scale information into account while generating masks. Furthermore, we introduce ConSM module to fully blend the speaker embedding in the speech extractor. Experimental results[1] on the Libri2Mix dataset show that MC-SpEx achieves a impressive performance and achieves state-of-the-art results for the speaker extraction task, which demonstrate the effectiveness of our improvements.

**Acknowledgement**: This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004), Tencent AI Lab Rhino-Bird Focused Research Program (RBFR2022005) and Tsinghua University - Tencent Joint Laboratory.

[1] Demo page: https://rookiejunchen.github.io/MC-SpEx_demo

# 5. References


[1] 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.

[2] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[3] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160101, 2017.

[4] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures." in *Interspeech*, 2017, pp. 2655–2659.

[5] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 327–334.

[6] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "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, 2019.

[7] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6965–6969.

[8] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika *et al.*, "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," *Proc. Interspeech 2020*, pp. 2677–2681, 2020.

[9] 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)*. IEEE, 2020, pp. 691–695.

[10] S. He, H. Li, and X. Zhang, "Speakerfilter: Deep learning-based target speaker extraction using anchor speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 376–380.

[11] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, Y. Wang, L. Xie, and S. Shang, "TEA-PSE: Tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS CHALLENGE," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9291–9295.

[12] Y. Ju, S. Zhang, W. Rao, Y. Wang, T. Yu, L. Xie, and S. Shang, "TEA-PSE 2.0: Sub-Band Network for Real-Time Personalized Speech Enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 472–479.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," *Proc. Interspeech 2020*, pp. 1406–1410, 2020.

[14] C. Li, Y. Wang, F. Deng, Z. Zhang, X. Wang, and Z. Wang, "EAD-Conformer: a Conformer-Based Encoder-Attention-Decoder-Network for Multi-Task Audio Source Separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 521–525.

[15] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5749–5753.

[16] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," *Proc. Interspeech 2022*, pp. 306–310, 2022.

[17] W. Wang, C. Xu, M. Ge, and H. Li, "Neural Speaker Extraction with Speaker-Speech Cross-Attention Network." in *Interspeech*, 2021, pp. 3535–3539.

[18] 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.

[19] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[20] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT," *PloS one*, vol. 13, no. 10, p. e0205355, 2018.

[21] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullSubNet+: Channel Attention FullSubNet with Complex Spectrograms for Speech Enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[22] D. Kim, J. Kim, and J. Kim, "Elastic exponential linear units for convolutional neural networks," *Neurocomputing*, vol. 406, pp. 253–266, 2020.

[23] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[24] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," *arXiv preprint arXiv:2204.00436*, 2022.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[27] R. Rikhye, Q. Wang, Q. Liang, Y. He, and I. McGraw, "Closing the Gap between Single-User and Multi-User VoiceFilter-Lite," *arXiv preprint arXiv:2202.12169*, 2022.

[28] Y. Lei, S. Yang, J. Cong, L. Xie, and D. Su, "Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion," *Proc. Interspeech 2022*, pp. 2563–2567, 2022.

[29] J. Han, Y. Long, L. Burget, and J. Černockỳ, "DPCCN: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation And Extraction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7292–7296.

[30] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target Confusion in End-to-end Speaker Extraction: Analysis and Approaches," *Proc. Interspeech 2022*, pp. 5333–5337, 2022.