INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speaker Extraction with Detection of Presence and Absence of Target Speakers

Ke Zhang[1,2], *Marvin Borsdorf*[3], *Zexu Pan*[2], *Haizhou Li*[4,3,2], *Yangjie Wei*[1], *Yi Wang*[1]

[1]Key Laboratory of Intelligent Computing in Medical Image, Northeastern University, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Machine Listening Lab (MLL), University of Bremen, Germany
[4]SDS, SRIBD, The Chinese University of Hong Kong, Shenzhen, China

1910621@stu.neu.edu.cn, marvin.borsdorf@uni-bremen.de

Abstract

Target speaker extraction extracts a target voice from a given cocktail party mixture signal. Most studies are restricted to conditions in which the target speaker is present in the mixture (PT), which often fail when the target speaker is absent (AT). Training on both PT and AT situations helps, but degrades the PT performance as the model intrinsically tries to detect the target presence. We propose a new model, called TSEJoint, that jointly performs target speaker detection and extraction. Both tasks share the low-level modules, allowing the detection branch to use a pre-separated signal and keeping the overall processing pipeline length similar, while at the high-level they have different branches to ensure the performance of each task. We evaluate our proposed methods under PT and AT conditions comprising one and two talkers. The TSEJoint model shows better extraction performance under the PT condition and better detection performance on all conditions compared with the baseline.

Index Terms: cocktail party problem, target speaker extraction, speaker detection, selective auditory attention, absent speaker

1. Introduction

Speech constitutes an important role in human-computer interactions. However, speech processing algorithms [1–6] are adversely affected by overlapping speakers. Blind source separation is a method that separates the overlapping speech signal into the individual clean source streams [7–10]. This method faces the global permutation ambiguity, especially in continuous separation, which makes it hard to maintain the output permutation of speech signals.

Target speaker extraction (TSE) is an alternative technique, that only extracts the target speaker's voice from a cocktail party mixture signal by utilizing a given reference signal of the target speaker. In addition to the pre-recorded speech as reference signal [11–16], other modalities such as face [17,18], gesture [19], or text [20] information could also be used.

In natural speech communication, interlocutors typically take turns to speak. Thus, the target speaker may or may not be present in a speech mixture signal to be processed by TSE algorithms. When the mixture speech signal contains the speech signal of the target speaker, it is referred to as present target (PT) condition. When the target speaker says nothing in the mixture, it is called the absent target (AT) condition [21]. For real-world applications, a TSE algorithm has to produce a reliable output, independent of the number of speakers in the mixture and regardless of whether the target speaker is present or not. However, most TSE algorithms have been studied only under the PT condition. Applying those methods under the AT condition results in extracting wrong signals from the mixture as the model has been trained to extract a speech signal that follows the given reference signal [21].

There are mainly two strategies for addressing the TSE problem under AT conditions. The first approach directly trains the TSE model on both PT and AT conditions [21, 22], to output the target speech for the PT condition and a zero (silent) signal for the AT condition. Borsdorf et al. [21] proposed a silence-evaluating scale-invariant signal-to-distortion ratio (SE-SI-SDR) loss function to cater to the AT condition. It maximizes the scale-invariant similarity between the target speech and the extracted speech on the PT condition, while minimizing the energy of the extracted speech on the AT condition. Therefore, it is hard to determine whether the TSE model trained by SE-SI-SDR has the ability to distinguish the absence and presence situations, or only adjust the energy of the outputs. Delcroix et al. [22] proposed an approach named TSE-IS, for addressing the TSE problem on AT conditions, but only considering the two-talker situation. TSE-IS is similar to Borsdorf et al. [21], but selects two scale-dependent loss functions for PT and AT samples, respectively. A disadvantage of this strategy is that the performance under the PT condition usually degrades, as the model not only has to extract the target speech, but also has to verify whether the speech belongs to the target speaker intrinsically.

The second strategy trains the TSE model on the PT condition only, and applies an additional speaker verification module on the extracted speech to determine the target presence [22,23]. However, the additional verification module increases the processing time. In addition, there are also some works about speaker verification under multi-talker environments [24–26].

In this work, we study the TSE model under four conditions, simulating the presence and absence of target speakers for speech of one and two talkers. Following the definitions in [21], we refer to the described conditions as 2T-PT, 1T-PT, 2T-AT, and 1T-AT in this paper. 2T and 1T describe the two-talker and the one-talker conditions and with PT and AT we indicate present target and absent speaker conditions. On 2T-PT and 1T-PT conditions, we expect the TSE model to output the speech of the target speaker while on 2T-AT and 1T-AT conditions, the model is expected to output a zero (silent) signal.

To balance the training for different conditions, we intro-

This research is supported by Huawei Noah's Ark Lab, National Natural Science Foundation of China (Grant No. 62271432), Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen (Grant No. B10120210117-KP02), Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046), and German Research Foundation (DFG) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

10.21437/Interspeech.2023-655

duce a modified loss function based on TSE-IS [22] with an adjustable weight for the AT loss. The weight controls the learning direction of the TSE model towards different conditions, thus enabling the study of the relationship between extraction and detection as opposed to [21, 22].

In addition, we believe that extraction and detection are closely related tasks that could benefit from each other. We propose a new model, called TSEJoint, that jointly performs target speaker detection and extraction. Both tasks share the same low-level modules, allowing the detection branch to use a pre-separated input signal and keeping the overall processing pipeline length similar, while at the high-level they have different branches to ensure the performance of each task. The TSEJoint model achieves better extraction results on the PT condition compared to the baseline TSE model, and a better detection rate on all conditions compared to the approach of only training the TSE model on PT and AT conditions together.

2. Proposed Methods

First, we introduce a modified loss function for directly training the TSE model on PT and AT conditions. Secondly, we propose a new model, called TSEJoint, which adopts the SpEx+ [13] architecture combined with an additional detection module.

2.1. Weighted loss function for baseline TSE model

The key of adapting TSE models to the AT condition is given by the loss function, which should balance the learning progress on both PT and AT conditions, avoiding dominating the gradient by one of these conditions. As discussed in Section 1, the loss function should be scale-dependent for all conditions. As in [22], we select the negative threshold SNR (tSNR) loss [27, 28] and the logtmse loss [29] for present and absent target samples, respectively. Additionally, we add a weight α for the AT loss to control the learning direction as follows:

$$\mathcal{L}_{mix} = \begin{cases} \mathcal{L}_{tSNR}, & s \neq 0 \\ \alpha \mathcal{L}_{logtmse}, & s = 0 \end{cases} \tag{1}$$

$$\mathcal{L}_{tSNR} = -10 \log_{10} \frac{||s||^2}{||s-x||^2 + \tau_1 ||s||^2} \tag{2}$$

$$\mathcal{L}_{logtmse} = 10 \log_{10}(||x||^2 + \tau_2 ||y||^2) \tag{3}$$

where $s \in \mathcal{R}^T$ is the target signal, $x \in \mathcal{R}^T$ is the extracted speech, $y \in \mathcal{R}^T$ is the mixture signal, and T is the duration of the speech signal. We set τ_1 as 10^{-3} and τ_2 as 10^{-2} to create two soft thresholds for preventing the samples with nearly perfect extraction output from dominating the gradients. Compared with the vanilla negative signal-to-distortion ratio (SDR) loss:

$$\mathcal{L}_{SDR} = -10 \log_{10} \frac{||s||^2}{||s-x||^2} \tag{4}$$

the value of $\mathcal{L}_{tSNR}$ is limited by τ_1 to -30 dB (see Section 4).

2.2. TSEJoint model

By analyzing the approach introduced in Section 2.1, we find the performance of the TSE model on the AT condition is achieved by sacrificing its performance on the PT condition. Inspired by this, we split the task and the output of the TSE model, and propose a family of models, called TSEJoint (TSEJoint2, TSEJoint3, and TSEJoint4), which is based on SpEx+ with an additional detection module. Figure 1 shows the structure of TSEJoint2, in which the detection module branches out after the second stacked temporal convolutional networks (TCNs) block of the SpEx+ model.

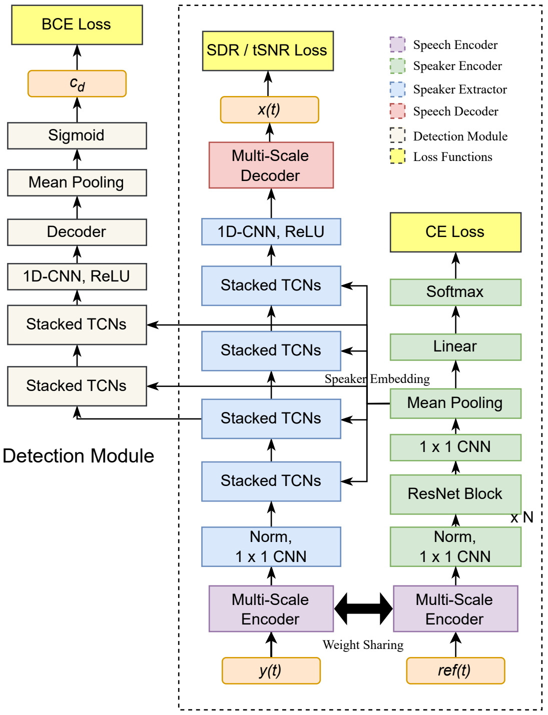

Fig. 1: *The model structure of TSEJoint2. The dotted box indicates the original SpEx+ architecture and the left part shows the detection module. In TSEJoint3, the detection module starts at the 3rd stacked TCNs block of the Speaker Extractor and comprises one stacked TCNs. In TSEJoint4, the detection module starts at the 4th shared TCNs block without any own stacked TCNs.*

The detection module in TSEJoint is designed for the detection of the target speaker. For simplicity, we use the same TCNs block, 1-D CNN layer, and mean-pooling layer as in SpEx+ for building the detection module. Binary cross entropy (BCE) loss is applied for training the detection module. For maintaining the extraction performance of TSE on PT conditions, we only train the extraction part in TSEJoint by using $\mathcal{L}_{tSNR}$ or the negative SDR loss with PT samples. The remaining part of TSEJoint is trained on both PT and AT conditions.

The TSEJoint models determine the final output on all conditions by analyzing the output of the detection module as follows:

$$x_d = \begin{cases} x, & c_d \geq c_{threshold} \\ 0, & c_d < c_{threshold} \end{cases} \tag{5}$$

where x_d is the output with detection, x is the extracted speech, c_d is the output of the detection module, and $c_{threshold}$ is a threshold commonly determined by calculating the Equal Error Rate (EER). We refer to this step as detection, which is essential for applying the TSEJoint model on AT conditions.

We create different model instances, namely TSEjoint2, TSEJoint3, and TSEjoint4. The number in the name represents on how many stacked TCNs the detection module runs on. In order to not extend the processing pipeline length of the TSE model, TSEJoint2 and TSEJoint3 have two and one stacked TCNs in the detection module, respectively. The TSEJoint4 model has no additional stacked TCNs.

Table 1: *Speaker extraction with/without detection on two-talker conditions. Atten. refers to the attenuation from the mixture. The numbers before $\mathcal{L}_{logtmse}$ are the value of α referring to the weight. The loss of the TSEJoint contains two parts for training both the extraction module and the detection module. We set the loss for extraction to zero for the AT condition. The cross entropy (CE) loss for the speaker encoder module in SpEx+ and TSEJoint has been omitted for simplicity. The input SDR is 0 dB.*

		Training				Without detection		With detection			
Sys	Models	Data sets		Loss Functions		SDR (dB)	Atten. (dB)	SDR (dB)	Atten. (dB)	EER	# Params
		2T-PT	2T-AT	PT	AT	2T-PT	2T-AT	2T-PT	2T-AT	(%)	
1	SpEx+	✓	✗	$\mathcal{L}_{SDR}$	-	17.0	-5.1	-	-	37.6	11.14 M
2				$\mathcal{L}_{tSNR}$	-	16.9	-5.1	-	-	38.5	
3	SpEx+	✓	✓	$\mathcal{L}_{tSNR}$	$0.001 \cdot \mathcal{L}_{logtmse}$	16.8	-29.2	13.3	-147.4	26.7	11.14 M
4					$0.01 \cdot \mathcal{L}_{logtmse}$	16.4	-130.9	14.8	-164.7	17.9	
5					$0.05 \cdot \mathcal{L}_{logtmse}$	16.1	**-167.1**	15.7	-176.4	12.0	
6					$0.1 \cdot \mathcal{L}_{logtmse}$	15.4	-167.0	15.2	-176.8	11.8	
7	SpEx+_V	✓	✗	$\mathcal{L}_{SDR}$	-	17.0	-5.1	12.3	-144.7	28.3	11.14 M
8	TSEJoint2	✓	✓	$\mathcal{L}_{SDR}$	0	17.2	-5.0	15.5	-178.9	10.8	15.70 M
9	TSEJoint3			+	+	**17.4**	-5.1	**16.1**	**-179.2**	**10.6**	13.48 M
10	TSEJoint4			$\mathcal{L}_{BCE}$	$\mathcal{L}_{BCE}$	16.5	-5.2	15.1	-178.3	11.1	11.21 M

3. Experimental Setup

3.1. Data set

As in [21], we evaluate our proposed methods on an extended version of the WSJ0-2mix-extr data set [30] to cover four different conditions: 2T-PT, 1T-PT, 2T-AT, and 1T-AT. The 2T-PT data set is the same as the WSJ0-2mix-extr data set, which contains 20,000, 5,000, and 3,000 utterances for training, validation, and testing, respectively. The training and validation sets share the same 101 speakers, whereas the 18 speakers in the test set are different. We utilize the utterances twice by switching the target speaker in the mixture. The 1T-PT data set is created by replacing the input mixtures with a clean speech signal of the target speakers. For the absent target conditions, the 2T-AT and 1T-AT data sets use the same mixtures of the corresponding PT data sets, but the reference utterances are different from the speakers given in the input mixture. The target signal data is replaced with a silent audio file with the same length as the respective input mixture. The details can be found in [21]. The data is sampled with 8 kHz. For the two-talker conditions, we use the min version and an overlap ratio of 100%.

3.2. Training details

We select the SpEx+ [13] as the baseline TSE model for evaluating the approach proposed in Section 2.1. First of all, we train the SpEx+ and the TSEJoint models on two-talker condition (2T-PT and 2T-AT) only, since we think the two-talker condition is more difficult and representative. Secondly, we study the conditions of PT (2T-PT and 1T-PT) and one-talker (1T-PT and 1T-AT) separately. Finally, we train and test the models on all four conditions.

We train all models for 100 epochs on segments with 4 seconds. The learning rate is set to $1e^{-4}$ and decays by 0.5 if the development loss does not improve within two consecutive epochs. The configuration of the SpEx+ is the same as in [13].

3.3. Evaluation metrics

To evaluate the extraction performance for PT conditions, we apply the SDR:

$$SDR = 10 \log_{10} \frac{||s||^2}{||s-x||^2} \tag{6}$$

For AT conditions, we apply the attenuation from mixtures [22]:

$$\mathcal{A} = 20 \log_{10}(\frac{||x||}{||y||} + 1e^{-10}) \tag{7}$$

We use the EER to evaluate the detection performance. For SpEx+, the EER is calculated based on the value of attenuation. For TSEJoint, the EER is calculated on c_d which is the output of the detection module. Detection (Equation 5) is essential for applying the TSEJoint model on AT conditions, and optional for SpEx+ by analyzing the value of attenuation:

$$x_d = \begin{cases} x, & \mathcal{A} \geq \mathcal{A}_{threshold} \\ 0, & \mathcal{A} < \mathcal{A}_{threshold} \end{cases} \tag{8}$$

where $\mathcal{A}_{threshold}$ is a threshold of the attenuation determined by the EER. We present the results without detection and with detection in Section 4 for the models that have been trained on both PT and AT conditions.

4. Results

4.1. Two-talker conditions

We first train and test SpEx+ as well as TSEJoint on the two-talker conditions. Table 1 shows the results. Sys 1 with $\mathcal{L}_{SDR}$ and Sys 2 with $\mathcal{L}_{tSNR}$ trained only on 2T-PT provide the baseline results and show similar performances. Both scale-dependent loss functions lead to an EER around 38%.

Sys 3-6 are the results of SpEx+ trained with $\mathcal{L}_{mix}$ and with different values of α under PT and AT conditions. With smaller α, the performance on 2T-PT is closer to the baseline results, but the EER degrades. The outputs' energy of Sys 3-6 are naturally lower than Sys 1-2 on 2T-AT. When we gradually increase the weight of $\mathcal{L}_{logtmse}$ for the samples with absent target speaker, the ability of SpEx+ in terms of speaker detection improves. However, this improvement is not unlimited. When we set α to one, SpEx+ always outputs zero signals on both conditions, and the model loses both the ability of speaker extraction and detection. We balance Sys 5 with α set to 0.05, to obtain a good performance on both 2T-PT and 2T-AT conditions.

For Sys 3-6 with detection by using Equation 8, the attenuation on AT conditions reaches a relatively good level, while the SDR on 2T-PT becomes worse because some samples on 2T-PT are wrongly detected as AT conditions, resulting in zero output instead of extracting the speech signal.

Table 2: *Speaker extraction results with/without detection on different test conditions when being trained on different training data compositions.*

Line	Models	Training Data sets	Testset	Loss Functions		Without detection		With detection		EER (%)	
						SDR (dB)	Atten. (dB)	SDR (dB)	Atten. (dB)		
				PT	AT	PT	AT	PT	AT	Testset	All
1	SpEx+	2T-PT & 1T-PT	2T-PT	$\mathcal{L}_{SDR}$	-	14.2	-	-	-	-	-
2			1T-PT			49.3	-	-	-	-	-
3			2T-PT	$\mathcal{L}_{tSNR}$	-	16.7	-	-	-	-	-
4			1T-PT			**55.0**	-	-	-	-	-
5	SpEx+	All 4 conditions	2T	$\mathcal{L}_{tSNR}$	$0.05 \cdot \mathcal{L}_{logtmse}$	15.9	**-161.4**	14.9	-176.8	11.8	9.0
6			1T			50.0	**-183.3**	48.7	-187.8	6.1	
7	TSEJoint3	All 4 conditions	2T	$\mathcal{L}_{tSNR} + \mathcal{L}_{BCE}$	$0 + \mathcal{L}_{BCE}$	**16.8**	-5.1	**15.7**	**-180.4**	**10.0**	**7.9**
8			1T			54.8	-24.0	**51.5**	**-187.9**	**6.0**	

Sys 7 uses the same approach as TSE-V [22]. Considering the speaker embedding extraction module has not been trained on the extracted speech which contains some residual noise and interference, we think a high EER is reasonable. In addition to Sys 7, we also test this approach on SpEx+ trained with the SI-SDR [31] loss, leading to an EER of 43.2%.

Among the TSEJoint models, Sys 9 shows the best performance and can also outperform the baseline models on the 2T-PT condition, even if those are experts on this condition. The results in terms of SDR are notable, since adding AT samples seems to not harm the performance under PT conditions but rather improves it. The SDR of Sys 10 is lower compared to Sys 8-9 and Sys 1. We think this could be attributed to the missing TCN block in the structure of the detection module in TSEJoint4. The tasks of extraction and detection may interact with each other for the dominance of the shared TCN blocks in the extractor module. This could be a possible explanation for the trade-off performance on PT and AT conditions for Sys 3-6.

For applying Sys 8-10 on AT conditions, the detection is necessary. With detection, the SDR on 2T-PT has a certain degree of decline, while the attenuation becomes a good level. Comparing Sys 9 (with detection) with Sys 5 (without detection), Sys 9 shows a similar SDR on 2T-PT and a better EER, resulting in a better ability of target speaker detection. Sys 9 has substantial advantages when being only applied on PT conditions, and it also has a better applicability on all conditions, because the performance on PT and AT conditions can be easily controlled by adjusting the threshold of detection.

In addition to the above models, we also train the detection branch in the TSEJoint individually without the extraction. The EER is 12.8% which is not as good as the performance of the TSEJoint models that perform extraction and detection at the same time. This result validates our assumption that detection and extraction can benefit from each other.

4.2. One-talker conditions and PT conditions

We test the detection ability of SpEx+ and TSEJoint3 on one-talker conditions (1T-PT and 1T-AT). The TSEJoint3 achieves an EER of 7.9%. However, for the SpEx+ trained by $\mathcal{L}_{mix}$, there are only two kinds of outputs. With a small α in $\mathcal{L}_{mix}$, the SpEx+ always outputs the extracted speech on both PT and AT conditions. With a relatively larger α, the output becomes a zero signal, and the attenuation is close to the lower limit at -200 dB on both conditions. In either case, the EER is around 50%. We don't find a suitable value of α, since the intervals of α values in the two cases overlap. Training the SpEx+ by $\mathcal{L}_{mix}$ is influenced by different factors, such as the learning difficulty of the model on PT and AT conditions, and the gradient change caused by the convergence.

The results on one- and two-talker PT conditions (2T-PT and 1T-PT) are shown in Line 1-4 of Table 2. We use $\mathcal{L}_{SDR}$ and $\mathcal{L}_{tSNR}$ to train the SpEx+ on PT conditions, and show the results on 2T and 1T conditions separately. Compared with the SpEx+ trained on 2T-PT only in Table 1, the SDR on 2T-PT decreases by 0.2 dB when using $\mathcal{L}_{tSNR}$, which is acceptable. The SDR on 1T-PT shows a good result with 55.0 dB. However, by using $\mathcal{L}_{SDR}$, the SDR on 2T-PT decreases by 2.8 dB. We think the reason is the imbalance of extraction difficulty on 2T-PT and 1T-PT conditions, when using vanilla negative SDR as training loss and training from scratch.

4.3. All four conditions

From Table 1, we select the SpEx+ (Sys 5) and the TSEJoint3 (Sys 9) and train each model on all four conditions. We replace the training loss $\mathcal{L}_{SDR}$ by $\mathcal{L}_{tSNR}$ based on the results on one- and two-talker PT conditions. The results are shown in Table 2.

Line 5-6 show the results of the SpEx+ trained on all four conditions. Compared with line 3-4, the SDR (without detection) decreases by 0.8 dB and 5.0 dB on 2T-PT and 1T-PT, respectively. The EER on the two-talker conditions (2T-PT and 1T-PT) is 11.8%, which is slightly better compared to Sys 5 (Table 1) that is trained on the two-talker conditions only. The problem when only being trained on the one-talker conditions, depicted in Section 4.2, does not arise, and the EER on one-talker conditions is 6.1%.

Line 7-8 show the results for the TSEJoint3 trained on all 4 conditions together. Without detection, the model achieves similar extraction results compared to the SpEx+ when being trained on PT conditions only with $\mathcal{L}_{tSNR}$. For applying the TSEJoint3 on all conditions, the detection is essential. The TSEJoint3 with detection achieves similar extraction performance compared to the SpEx+ trained with $\mathcal{L}_{mix}$ without detection. However, the EER improves to 10.0% and 6.0% on two-talker and one-talker conditions, respectively.

5. Conclusion and Future Work

In this work, we studied different strategies to train TSE models on present and absent target speaker conditions. We proposed a new model based on the combination of a SpEx+ architecture and an additional target speaker detection module. Our model achieves superior performance on both the speaker extraction task and the speaker detection task without extending the processing pipeline length of the whole system. In our future work, we will focus on the optimization of the detection module to increase the accuracy of the target speaker detection.

6. References

[1] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target doa estimation with an audio-visual fusion mechanism," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4280–4284.

[2] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *Proc. INTERSPEECH*, 2020, pp. 364–368.

[3] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proc. of the 29th ACM Int. Conf. on Multimedia*, 2021, pp. 3927–3935.

[4] J. Wang, X. Qian, and H. Li, "Predict-and-update network: Audio-visual speech recognition inspired by human speech perception," *arXiv preprint arXiv:2209.01768*, 2022.

[5] T. Liu, R. K. Das, K. Aik Lee, and H. Li, "Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7517–7521.

[6] Y. Ma, K. A. Lee, V. Hautamäki, and H. Li, "Pl-eesr: Perceptual loss based end-to-end robust speaker representation extraction," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 106–113.

[7] 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.

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

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.

[10] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[11] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech 2019*, 2019, pp. 2728–2732. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1101

[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.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1397

[14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6109–6113.

[15] 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.

[16] 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)*, 2020, pp. 691–695.

[17] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[18] Z. Pan, R. Tao, C. Xu, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.

[19] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Processing Letters*, vol. 29, pp. 1467–1471, 2022.

[20] J. Li, M. Ge, Z. Pan, L. Wang, and J. Dang, "VCSE: Time-Domain Visual-Contextual Speaker Extraction Network," in *Proc. Interspeech 2022*, 2022, pp. 906–910.

[21] M. Borsdorf, C. Xu, H. Li, and T. Schultz, "Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers," in *Proc. Interspeech 2021*, 2021, pp. 1469–1473.

[22] M. Delcroix, K. Kinoshita, T. Ochiai, K. Zmolikova, H. Sato, and T. Nakatani, "Listen only to me! How well can target speech extraction handle false alarms?" in *Proc. Interspeech 2022*, 2022, pp. 216–220.

[23] C. Xu, W. Rao, J. Wu, and H. Li, "Target speaker verification with selective auditory attention for single and multi-talker speech," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 29, pp. 2696–2709, 2021.

[24] L. Zhang, Z. Chen, and Y. Qian, "Enroll-Aware Attentive Statistics Pooling for Target Speaker Verification," in *Proc. Interspeech 2022*, 2022, pp. 311–315.

[25] Y. Shi and T. Hain, "Supervised speaker embedding de-mixing in two-speaker environment," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 758–765.

[26] A. Aloradi, W. Mack, M. Elminshawi, and E. A. Habets, "Speaker verification in multi-speaker environments using temporal feature fusion," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 354–358.

[27] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised speech separation using mixtures of mixtures," in *Advances in Neural Information Processing System*, vol. 33, 2020, pp. 3846–3857.

[28] Y. Luo and N. Mesgarani, "Separating Varying Numbers of Sources with Auxiliary Autoencoding Loss," in *Proc. Interspeech 2020*, 2020, pp. 2622–2626.

[29] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 186–190.

[30] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6990–6994.

[31] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.