Interspeech 2022
18-22 September 2022, Incheon, Korea

Listen only to me! How well can target speech extraction handle false alarms?

Marc Delcroix[1], *Keisuke Kinoshita*[1], *Tsubasa Ochiai*[1], *Katerina Zmolikova*[2], *Hiroshi Sato*[1], *Tomohiro Nakatani*[1]

[1]NTT Corporation, Japan, [2]Brno University of Technology, Speech@FIT, Czechia,
marc.delcroix@ieee.org

Abstract

Target speech extraction (TSE) extracts the speech of a target speaker in a mixture given auxiliary clues characterizing the speaker, such as an enrollment utterance. TSE addresses thus the challenging problem of simultaneously performing separation and speaker identification. There has been much progress in extraction performance following the recent development of neural networks for speech enhancement and separation. Most studies have focused on processing mixtures where the target speaker is actively speaking. However, the target speaker is sometimes silent in practice, i.e., inactive speaker (IS). A typical TSE system will tend to output a signal in IS cases, causing false alarms. This is a severe problem for the practical deployment of TSE systems. This paper aims at understanding better how well TSE systems can handle IS cases. We consider two approaches to deal with IS, (1) training a system to directly output zero signals or (2) detecting IS with an extra speaker verification module. We perform an extensive experimental comparison of these schemes in terms of extraction performance and IS detection using the LibriMix dataset and reveal their pros and cons.

Index Terms: Speech enhancement, Target speech extraction, Inactive speaker

1. Introduction

Enhancing a speech signal corrupted by interfering speakers has been one of the major challenges of speech signal processing. One way to tackle this problem is to use speech separation [1], which separates a speech mixture into all its sources. Research in speech separation has progressed rapidly with the advent of deep learning [2–4]. However, there are two fundamental limitations with most separation techniques. First, separation requires knowing or estimating the number of sources in the mixture. Then, there is a global permutation ambiguity; the mapping between outputs speakers is arbitrary.

Target speech extraction (TSE) [5] has been proposed as an alternative to enhance speech in a mixture. TSE focuses on extracting only the speech signal of a target speaker instead of separating all sources by exploiting a speaker clue to identify which speaker to extract [5–15]. For example, we can use an enrollment utterance, which consists of a short recording containing only the voice of the target speaker [6, 7, 10]. Because TSE estimates only the speech of the target speaker, it naturally alleviates the issues of separation systems, i.e., the processing is independent of the number of sources in the mixtures, and there is no speaker ambiguity at the output.

We can realize TSE using a neural network (NN) conditioned on the target speaker clue, which directly estimates the target speech from the mixture [6, 10, 11, 16]. Such a TSE system must perform thus both *separation* and *speaker identification* internally. Most studies about TSE have assumed the target speaker was always actively speaking in the mixtures, i.e., *active speaker (AS)* case. However, we argue that measuring TSE performance in such conditions does not fully represent the *speaker identification* capabilities of TSE systems. Indeed, in practice, a target speaker may be silent, i.e., *inactive speaker (IS)* case. In such a case, a TSE should output nothing or a zero signal. However, a TSE system trained only on AS conditions would always try to output a speech-like signal, which would cause *false alarms* or false positive. It is thus essential to consider IS conditions in the design and evaluation of TSE systems.

There have been only a few works dealing with the IS issue of TSE [17–19]. These works offer two different strategies to address the problem. The TSE with internal IS detection (TSE-IS) scheme trains a TSE system to directly output zero signals for IS cases by including IS samples during training [17, 18]. The TSE+Verification (TSE-V) scheme combines TSE with speaker verification and detects IS samples when the extracted signals do not match the target speaker characteristics of the enrollment [19][1]. TSE-IS is a simpler system than TSE-V, but it is potentially easier to control false alarm and *miss detection*[2] with TSE-V. However, these schemes have not been compared, and their impact on TSE performance has not been fully revealed. In this paper, we address this shortcoming and perform a comprehensive comparison in terms of the detection of IS and extraction performance in order to answer the following question: *How well can TSE systems handle IS samples?*

The contribution of this paper is as follows: (1) We propose two simple implementations of the TSE-IS and TSE-V schemes based on the SpeakerBeam TSE framework [16], and perform an comprehensive experimental comparison in terms of extraction and AS/IS detection performance. (2) We reveal that a TSE-IS system trained with a modified signal-to-noise ratio (SNR) loss can predict IS in about 90% of the cases but also significantly increases the number of extraction failures for AS cases. (3) We show that we can build a TSE-V system from a TSE system trained only with AS samples. Such a simple TSE-V system can detect AS/IS better than a TSE-IS, while maintaining high extraction performance. (4) Finally, we reveal that the enrollment duration impacts moderately extraction performance but greatly affects AS/IS detection errors of TSE-V. With enrollment of 15 sec or more, we can achieve AS/IS detection with a Equal error rate (EER) of about 5 %. The results of this study demonstrate the potential of current capabilities of TSE systems to detect and extract a target speaker.

2. Related works

Prior works [17, 18] considered TSE with IS cases. They introduced a modified scale-invariant SNR (SI-SNR) [20] loss to allow training a TSE-IS system with IS samples. However, using a scale-invariant loss makes the output scale arbitrary. It

[1]Note that in [19] the goal is not TSE but speaker verification.

[2]Here miss detection or false negative means that the TSE systems wrongly predicted an AS as IS.

10.21437/Interspeech.2022-11252

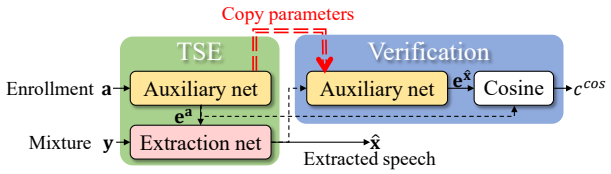

Figure 1: *Overview of a TSE system and its extension to TSE-V.*

may thus be challenging to determine if the output can actually be considered an AS or not. Besides, it is thus unclear how well the system proposed in [18] can detect IS cases since the approach was only evaluated with signal extraction measures and not in terms of AS/IS detection. In contrast, we propose using a modified SNR loss to train our TSE-IS system, which preserves the scale at the output of the system and allows thus performing the AS/IS detection based on the attenuation from the mixture.

There have been two prior studies combining TSE with speaker verification [19, 21], which are related to TSE-V. Both works aimed at improving speaker verification for speech in a mixture and used a TSE system as a pre-processing. However, as their goal was speaker verification and not TSE, they did not evaluate their systems in terms of extraction performance although, e.g., miss detection errors caused by zeroing out the output when AS cases are detected as IS can have a severe impact on extraction performance.

3. TSE problem and baseline system

3.1. Problem formulation

TSE aims at extracting speech of a target speaker, $\mathbf{x}^s \in \mathcal{R}^T$ from a mixture $\mathbf{y} \in \mathcal{R}^T$ defined as,

$$\mathbf{y} = \mathbf{x}^s + \sum_{i \neq s} \mathbf{x}^i + \mathbf{n}, \tag{1}$$

where $\mathbf{x}^i$ and $\mathbf{n}$ represent the interference speech and the background noise signals, respectively. T is the duration of the signal. We assume having an enrollment utterance of the target speaker, $\mathbf{a}^s \in \mathcal{R}^{T^a}$, of duration T^a. Note that when the target speaker is active, $\mathbf{x}^s$ is a speech signal, while when it is inactive $\mathbf{x}^s = \mathbf{0}$, where $\mathbf{0}$ denotes a vector of all zeros.

3.2. SpeakerBeam

We use time-domain SpeakerBeam [16] as a basis for our study as it represents a typical enrollment-based neural TSE system [10, 11]. The left part of Fig. 1 shows a diagram of the system. It consists of two modules. (1) An auxiliary network that computes a target speaker embedding, $\mathbf{e}^s \in \mathcal{R}^N$, from the enrollment, $\mathbf{a}^s$. (2) An extraction network that estimates the target speech from the mixture given the speaker embedding. The operation of the network is summarized as follows,

$$\mathbf{e}^s = \mathrm{NN}^{\mathrm{Aux}}(\mathbf{a}^s), \tag{2}$$

$$\hat{\mathbf{x}}^s = \mathrm{NN}^{\mathrm{Ext}}(\mathbf{y}, \mathbf{e}^s), \tag{3}$$

where $\hat{\mathbf{x}}^s \in \mathcal{R}^T$ is the estimated target speech signal, $\mathrm{NN}^{\mathrm{Aux}}(\cdot)$ and $\mathrm{NN}^{\mathrm{Ext}}(\cdot)$ represent the auxiliary and extraction NN, respectively.

With time-domain SpeakerBeam both $\mathrm{NN}^{\mathrm{Aux}}(\cdot)$ and $\mathrm{NN}^{\mathrm{Ext}}(\cdot)$ are implemented with 1-D convolutional blocks proposed for the fully-convolutional time-domain audio separation network (Conv-TasNet) [22]. The extraction network uses an element-wise multiplication [23, 24] to combine the embedding vector with the hidden representation obtained after the first convolutional block of the extraction network.

3.3. Training objective for active speaker cases

With SpeakerBeam, we train jointly both the auxiliary and extraction networks, which enables learning speaker embeddings optimal for the TSE task. Speech separation and TSE systems are usually trained using a time-domain criterion such as SNR or SI-SNR [20, 22]. We chose to use a scale-dependent loss, to ensure that the system preserves the scale of the signals as it may be important to detect AS/IS samples. In particular, we use the negative thresholded SNR [25] loss defined as,

$$\mathcal{L}^{\mathrm{active}}(\hat{\mathbf{x}}^s, \mathbf{x}^s) = -10 \log_{10} \left(\frac{\|\mathbf{x}^s\|^2}{\|\mathbf{x}^s - \hat{\mathbf{x}}^s\|^2 + \tau \|\mathbf{x}^s\|^2} \right), \tag{4}$$

where τ is a threshold that we set at $\tau = 10^{-3}$. It avoids that the low distortion training samples dominate the gradient. We train our baseline TSE model with only AS training samples.

4. Handling inactive speakers

4.1. TSE-IS: Learning direct IS detection with inactive loss

The first approach for handling IS, TSE-IS, consists of training a system to output zero signals for IS cases. The loss functions derived from the SNR such as Eq. (4) are ill-defined when the reference signal is zero. Thus, we cannot use such losses directly with IS samples. This problem was first revealed for the training of separation systems that can accommodate a varying number of sources in the mixture [26], i.e., the number of sources can be less than the number of outputs of the separation system. In this case, a separation system needs thus to be able to output zero signals, which is similar to the IS problem of TSE.

We propose to use the modified SNR loss introduced in [26] to train our TSE-IS system. The loss is defined as,

$$\mathcal{L}(\hat{\mathbf{x}}^s, \mathbf{x}^s, \mathbf{y}) = \begin{cases} \mathcal{L}^{\mathrm{active}}(\hat{\mathbf{x}}^s, \mathbf{x}^s), & \text{if } \mathbf{x}^s \neq \mathbf{0}, \\ \mathcal{L}^{\mathrm{inactive}}(\hat{\mathbf{x}}^s, \mathbf{y}), & \text{if } \mathbf{x}^s = \mathbf{0}, \end{cases} \tag{5}$$

where the inactive loss is given by

$$\mathcal{L}^{\mathrm{inactive}}(\hat{\mathbf{x}}^s, \mathbf{y}) = 10 \log_{10} \left(\|\hat{\mathbf{x}}^s\|^2 + \tau^{\mathrm{inactive}} \|\mathbf{y}\|^2 \right), \tag{6}$$

and τ^{inactive} is a soft threshold set at $\tau^{\mathrm{inactive}} = 10^{-2}$. $\mathcal{L}^{\mathrm{inactive}}$ consists of the denominator term of Eq. (4) with a different setting for the soft threshold (i.e., $\mathbf{x}^s$ replaced by $\mathbf{y}$).

We opt here for a scale-dependent SNR loss, unlike [18], because we believe that the scale of the output signal may matter in practical applications to detect IS cases. For example, we can evaluate how well the system could internally detect AS/IS cases by looking at the output scale and, e.g., measuring the attenuation from the mixtures, $\mathcal{A}^{\mathrm{mixture}} = 10 \log_{10} \left(\frac{\|\hat{\mathbf{x}}^s\|^2}{\|\mathbf{y}\|^2} \right)$. We can thus define a classifier based on the attenuation as,

$$c^{\mathrm{Att}} = \begin{cases} 1, & \text{if } \mathcal{A}^{\mathrm{mixture}} > \eta^{\mathrm{Att}}, \\ 0, & \text{if } \mathcal{A}^{\mathrm{mixture}} \leq \eta^{\mathrm{Att}}, \end{cases} \tag{7}$$

where η^{Att} is a threshold. The target speaker is considered active when $c^{\mathrm{Att}} = 1$ and inactive when $c^{\mathrm{Att}} = 0$.

We introduced the above classifier to measure the AS/IS detection capability of the system, but in practice, we do not need it as TSE-IS performs the AS/IS detection internally and directly output a speech signal or a zero signal. There is no increase in computational complexity compared to an existing TSE system. However, it allows little control to, e.g., balance the false alarms or miss detection errors of the system at test time. Besides, adding IS cases during training may hurt the extraction performance for the AS cases.

Table 1: *Description of the dataset*

	Train-100k	Train-360k	Val	Test
Nb. of mixtures	13900	50800	3000	3000
Nb. of Speakers	251	921	40	40

4.2. TSE-V: Post AS/IS detection with speaker verification

Another approach to handle IS cases consists of using a TSE system trained on AS cases, which always try to output a speech-like signal, and then perform post verification to check that the speech characteristics of the extracted speech, $\mathbf{x}^s$, correspond to those of the enrollment. Figure 1 shows a schematic diagram of such a system.

In this work, we propose using the auxiliary network to compute a speaker embedding for the extracted speech, $\mathbf{e}^{\hat{\mathbf{x}}^s} = \mathrm{NN}^{\mathrm{Aux}}(\hat{\mathbf{x}}^s)$, since we showed in prior works that it could extract discriminative speaker embeddings [5]. We then make the AS/IS decision by looking at the cosine similarity between the embeddings computed from the enrollment and from the extracted speech as,

$$c^{\mathrm{Cos}} = \begin{cases} 1, & \text{if } \mathcal{C}(\mathbf{e}^{\hat{\mathbf{x}}^s}, \mathbf{e}^s) > \eta^{\mathrm{Cos}}, \\ 0, & \text{if } \mathcal{C}(\mathbf{e}^{\hat{\mathbf{x}}^s}, \mathbf{e}^s) \leq \eta^{\mathrm{Cos}}, \end{cases} \tag{8}$$

where $\mathcal{C}(\mathbf{e}^{\hat{\mathbf{x}}^s}, \mathbf{e}^s)$ is the cosine similarity between $\mathbf{e}^{\hat{\mathbf{x}}^s}$ and $\mathbf{e}^s$, and η^{Cos} is a threshold. We can then define an extracted signal after detection as $\bar{\mathbf{x}}^s = c^{\mathrm{Cos}}\hat{\mathbf{x}}^s$, which simply zeros out the samples detected as IS.

Note that this approach checks whether the extracted speech matches the enrollment characteristics. It can thus possibly detect not only IS but also extraction failures. Such failures occur when the TSE system wrongly outputs the mixture or the interference speakers instead of the target speech.

Compared to TSE-IS, TSE-V increases the computational complexity slightly as it requires an additional pass through the auxiliary network. However, since the AS/IS detection is performed independently of the TSE process, it allows better control at test time and also does not require the training of the TSE module with IS samples. Note that in contrast to our proposed TSE-V, the system proposed in [19] used a pre-trained speaker embedding extractor and retrained it on extracted speech. We can view our TSE-V system as a simplified version of [19].

5. Experiments

We performed experiments using the LibriMix dataset [27], which consists of noisy two-speaker mixtures derived from the LibriSpeech dataset [28]. We used the open implementation of SpeakerBeam [29] based on the asteroid toolkit [30].

5.1. Dataset

We performed experiments using the full-overlap (i.e., min version) two-speaker noisy mixtures of the LibriMix dataset. Table 1 provides more details about the dataset. For each mixture, we randomly sampled enrollment utterances from the speakers in the mixture for AS cases and from a different speaker for the IS cases. In both cases, the enrollment differed from the utterances used in the mixture. At test time, we considered enrollment utterances from three speakers for each test mixture, i.e., two from the ASs in the mixture and one from another speaker, i.e., IS.

5.2. Experimental settings

We used the same network architecture for all experiments, which consists of the SpeakerBeam system provided in [29], except that we used the training loss of Eq. (5). We followed a similar configuration as Conv-TasNet [22]. We used blocks of eight stacked 1-D convolution layers for the auxiliary and extraction networks, repeated three times for the extraction network. We used an element-wise multiplication to combine the embedding vector with the hidden representation at the output of the first convolution block. We trained the systems for 200 epochs with the Adam optimizer [31].

We compared the following four TSE systems.

Baseline TSE corresponds to the baseline system of Section 3, which was trained with the train-100k set with only AS samples. It does not perform neither internal nor post AS/IS detection.

TSE-IS corresponds to the system described in Section 4.1, which was trained with the train-100k training set including 10% of IS cases, i.e., we used an enrollment from a speaker not present in the mixture and a zero signal as target for 10% of the training samples.

TSE-V corresponds to the system described in Section 4.2. The TSE module corresponds to the above baseline TSE system trained with only AS samples. At test time, we re-used the auxiliary network to compute the embedding vector for the extracted speech and performed AS/IS detection with Eq. (8).

TSE-V(360) consists of the TSE module of the above TSE-V system retrained on AS samples of the train-360k dataset for 100 epochs. It is used to measure the impact of using a larger training set with more speakers.

5.3. Evaluation metrics

We evaluated the systems in terms of the following evaluation metrics: (1) **EER** measures the AS/IS detection errors using the detection error tradeoff (DET) curves shown in Fig. 3 obtained with the classifiers of Eq. (7) and Eq. (8) for TSE-IS and TSE-V, respectively. (2) **Signal-to-distortion ratio improvement (SDRi)** measures the extraction performance for the AS cases using the BSS eval toolkit [32]. We report the SDRi before and after AS/IS detection, i.e. using $\hat{\mathbf{x}}^s$ or $\bar{\mathbf{x}}^s = c\hat{\mathbf{x}}^s$, respectively, where c is given by either Eq. (7) or Eq.(8) using the threshold that gives the EER. We do not need to compute $\bar{\mathbf{x}}^s$ for TSE-IS, but we perform it anyway to provide a fair comparison with TSE-V. SDRi-after accounts for the impact of miss detection errors on the extraction performance. Note that samples detected as IS are replaced by a zero signal, thus resulting in a SDR of 0 dB. (3) **Failure rate (Fail)** is defined as $Fail = \frac{NF^{\mathrm{AS}}}{N^{\mathrm{AS}}}$, where NF^{AS} is number of AS samples with SDRi below 1dB and N^{AS} is the total number of AS samples. Failures happen when, e.g., the TSE system extracts the wrong speaker, output the mixture or a zero signal when using TSE-IS. (4) **Failure and miss detection rate (Fail&Miss)** is defined as $Fail\&Miss = \frac{NFM^{\mathrm{AS}}}{N^{\mathrm{AS}}}$, where NFM^{AS} is the number of AS samples that result in extraction or detection errors, i.e. SDRi below 1dB, miss detection or both. It measures the total error rate for the AS cases. For example, even if a sample is correctly detected as AS its extraction performance may be low, and it should thus be considered as an error. (5) **Attenuation (Att.)** measures the attenuation from the mixture, $\mathcal{A}^{\mathrm{mixture}}$, defined in Section 4.1. It shows how well a TSE system can output zero signals for IS cases.

Note that we use only AS samples to compute SDRi, Fail and Fail&Miss, but both AS and IS for EER and Attenuation.

5.4. Experimental results

Table 2 shows the extraction and AS/IS detection results for the different systems using enrollment of average duration of 10

Table 2: *Extraction and detection performance with enrollment of average duration of 10 sec. The input SDR is -1.8 dB.*

	SDRi before(after) detection [dB] ↑	Fail [%] ↓	EER [%] ↓	Fail&Miss [%] ↓
Baseline TSE	12.4 (na)	3.4	-	-
TSE-IS	10.8 (11.4)	8.6	11.6	13.4
TSE-V	12.4 (11.9)	3.4	8.9	10.5
TSE-V(360)	13.6 (13.1)	1.7	6.3	7.1

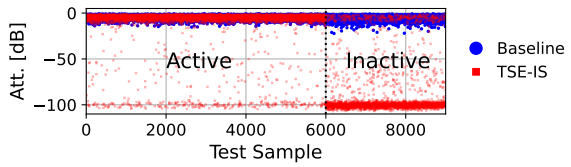

Figure 2: *Attenuation for each test sample. The first 6000 samples correspond to AS and the last 3000 to IS samples.*

sec. Figure 2 shows the attenuation with respect to the mixture, $\mathcal{A}^{mixture}$, as a function of the samples in the test set.

The baseline TSE system, which was trained only with AS samples, achieves 12.4 dB SDRi (the input SDR is -1.8 dB) and only 3.4 % of failures. However, as seen in Fig. 2, the attenuation values remain in a similar range for both AS and IS samples, meaning that it always outputs some signal even for IS cases, which would cause many false alarms.

The TSE-IS system, which we trained with IS samples, can output zero signals. We observe in Fig.2 that the attenuation is around -100 dB for most IS cases while it remains close to 0dB for most AS cases. This confirms that TSE-IS can internally perform AS/IS detection. However, we also observe that learning with IS has an impact on extraction performance for AS cases. Indeed, around 10% of the AS test samples have attenuation around -100 dB (i.e. miss detection). Consequently, the failure rate is high, i.e., close to 9%, and the average SDRi lower than the baseline. The impact on SDRi may be exaggerated as it includes miss detection errors, i.e., samples where the system wrongly outputs a signal close to zero. The SDRi after detection is slightly better but remains lower than the baseline. We can also evaluate the AS/IS detection capability of the TSE-IS by looking at the detection performance of a classifier based on the attenuation as introduced in Eq. (7). Figure 3 shows the DET curve and EER of such a classifier.

The TSE module in TSE-V corresponds to the above baseline system, and thus the performance before detection is the same for the AS cases. The proposed verification based on the cosine distance of the embeddings computed with the auxiliary network is simple yet effective. Indeed, it can detect relatively well AS/IS, with an EER of less than 9%. The SDRi after detection is 0.5 dB lower because it includes miss detection errors. The total error rate on AS cases, i.e., Fail&Miss rate, is 10.5%, which is better than the TSE-IS system by about 3%. Overall TSE-V achieves higher extraction and detection performance than TSE-IS.

We also explore the impact of training with a larger training set which includes more speakers with the TSE-V(360) system. Retraining on the larger training set improves SDRi by about 1.2 dB, but mainly it can greatly reduce the failure rate, the EER and the combined Fail&Miss.

Figure 3 plots the DET curves for the AS/IS detection with TSE-V and TSE-IS. We observe that the miss rate rapidly increases for the TSE-IS, while the curve for TSE-V is much smoother. Consequently, it is more challenging to tune at test time the false alarm or miss rate of TSE-IS then TSE-V.

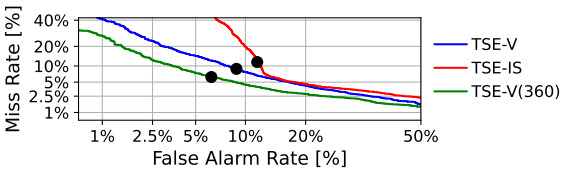

Figure 3: *DET curves for AS/IS detection with TSE-V and TSE-IS. The black circles indicate the EER.*

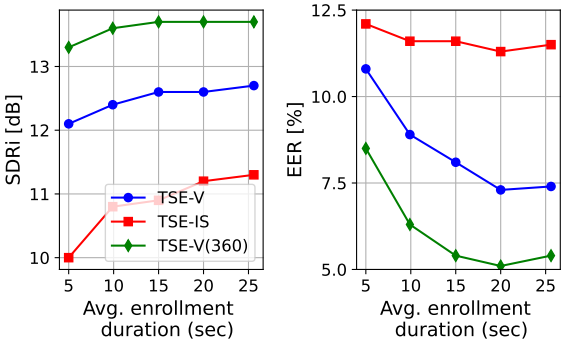

Figure 4: *Extraction and AS/IS detection performance as a function of the enrollment duration.*

Finally, Figure 4 plots the SDRi before detection and EER as a function of the average enrollment duration. Here we varied the enrollment duration by concatenating from 1 to 5 enrollment utterances for each test sample, which resulted in the average utterance length varying from 5 to 25 seconds. Increasing the enrollment duration improves extraction performance moderately but greatly reduces EER for TSE-V and TSE-V(360). For example, we can approach a EER of 5% with TSE-V(360) when using an enrollment utterance of 15 to 20 sec. We do not observe a similar trend for TSE-IS.

The results of our experiments demonstrate that with slight modifications, a TSE system can handle IS cases relatively well. The TSE-V approach provides better overall extraction and AS/IS detection performance than TSE-IS. It also allows more control to tune the miss detection and false alarm rates at test time. However, TSE-V requires an additional verification step. TSE-IS can learn to detect internally IS cases and output directly zeros signals without increasing the computational complexity. Although TSE-IS performs worse than TSE-V, it could still be advantageous for, e.g., low-latency systems where the batch verification step would not be allowed.

6. Conclusion

A TSE system must perform speech extraction and speaker identification. Most studies have focused on evaluating TSE systems in terms of extraction performance and have mostly ignored the impact of false alarms when the target speaker is inactive. In this paper, we carried out a systematic comparison of two possible schemes to handle IS. Our experiments revealed that we could exploit the auxiliary network of a TSE system to perform speaker verification at the output and detect AS/IS cases. We can detect AS/IS cases with a EER of around 5%, using TSE-V trained with a relatively large amount of speaker, and using enrollment utterances of more than 15 sec. This positive finding confirms the potential of TSE systems.

Our TSE-V system outperforms a TSE-IS system that can internally detect IS and output zero signals. However, the TSE-IS system may remain attractive for, e.g., low-latency systems, which we plan to explore in our future works.

7. References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP'16*, 2016, pp. 31–35.

[3] Y. Dong, K. Morten, T. Zheng-Hua, and J. Jesper, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of ICASSP'17*, 2017, pp. 241–245.

[4] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," in *Proc. of ICASSP'18*, 2018.

[5] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE JSTSP*, vol. 13, no. 4, pp. 800–814, 2019.

[6] 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 *Proc. of Interspeech'17*, 2017, pp. 2655–2659.

[7] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with SpeakerBeam," in *Proc. of ICASSP'18*, 2018, pp. 5554–5558.

[8] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics (TOG)*, vol. 37, no. 4, pp. 1–11, 2018.

[9] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. of Interspeech'18*, 2018, pp. 3244–3248.

[10] 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. of Interspeech'19*, 2019, pp. 2728–2732.

[11] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. of ASRU'19*. IEEE, 2019, pp. 327–334.

[12] J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. of Interspeech'18*, 2018, pp. 307–311.

[13] 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 *Proc. of ICASSP*, 2019, pp. 86–90.

[14] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žd'ánský, "Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors," in *Proc. of ICASSP'20*, 2020, pp. 676–680.

[15] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. of Interspeech'19*, 2019, pp. 4290–4294.

[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 *Proc. of ICASSP'20*, 2020, pp. 691–695.

[17] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and Accurate Time-Domain Speaker Extraction Network," in *Proc. of Interspeech'20*, 2020, pp. 1421–1425.

[18] 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. of Interspeech'21*, 2021, pp. 1469–1473.

[19] C. Zhang, M. Yu, C. Weng, and D. Yu, "Towards robust speaker verification with target speaker enhancement," in *Proc. of ICASSP'21*, 2021, pp. 6693–6697.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR-half-baked or well done?" in *Proc. of ICASSP'19*, 2019, pp. 626–630.

[21] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target Speaker Extraction for Multi-Talker Speaker Verification," in *Proc. of Interspeech'19*, 2019, pp. 1273–1277.

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] L. Samarakoon and K. C. Sim, "Subspace LHUC for fast adaptation of deep neural network acoustic models." in *Proc. of Interspeech'16*, 2016, pp. 1593–1597.

[24] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for SpeakerBeam target speaker extraction," in *Proc. of ICASSP'19*, 2019, pp. 6965–6969.

[25] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixtures of mixtures," *Proc. of NeurIPS'20*, 2020.

[26] 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 *Proc. of ICASSP*, 2021, pp. 186–190.

[27] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of ICASSP'15*, 2015, pp. 5206–5210.

[29] "https://github.com/BUTSpeechFIT/speakerbeam."

[30] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proc. of Interspeech'20*, 2020, pp. 2637–2641.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR'15*, 2015.

[32] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.