What to Refer and How? - Exploring Handling of Auxiliary Information in Target Speaker Extraction

Tomohiro Hayashi*, Riku Ogino*, Kohei Saijo* and Tetsuji Ogawa*
* Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
E-mail: thayashi@pcl.cs.waseda.ac.jp

***Abstract*—This study explored the appropriate auxiliary information for target speaker extraction (TSE). In TSE, the conventional method involves using pre-recorded speech as an auxiliary signal. However, in multi-channel environments where spatial information is available, alternative approaches that utilize a rough pre-extraction of the target speech as an auxiliary signal are being explored. Pre-recorded speech is typically clean but contains different content, whereas pre-extracted speech has content aligned with the target but includes various types of distortion depending on the extraction method used. These two approaches differ in the characteristics of the auxiliary signals and the conditioning methods applied. Previous research has not thoroughly examined the relationship between the nature of auxiliary signals and TSE model performance. This study aims to explore the effective use of auxiliary signals in TSE models and to provide a comprehensive analysis of their impact. Through TSE experiments conducted in a multi-channel environment, we investigated the effects of auxiliary signals from three key perspectives: the trade-off between content consistency and distortion, the level of referencing, and the type of distortion.**

I. Introduction

Speech separation has long served as a crucial front-end component for speech applications like automatic speech recognition (ASR) [1], [2]. Recently, efforts have focused on isolating the target source from mixtures containing various sources by incorporating auxiliary information about the target source into separation networks. This approach, known as target speaker extraction (TSE) [3], has seen significant advancements due to the adoption of neural networks. Alongside improvements in network architecture, considerable research has also been dedicated to enhancing the auxiliary information used to identify the target speaker [4], [5].

In TSE, an enrollment utterance, which is a pre-recorded sample of the target speaker's voice, is typically used as auxiliary information. The speaker embedding derived from this enrollment utterance via a speaker encoder is employed to condition the extraction network. Since the enrollment utterance and the target speech in the mixture differ, the speaker embedding is often transformed into a global feature, such as by applying mean pooling across time frames.

On the other hand, in some scenarios, it is possible to estimate the target speech without enrollment utterances, for instance, by leveraging the spatial position of the target speaker. For example, if multi-channel microphones are available and the direction of the target speaker is known, blind source separation (BSS) [6] can be used to roughly extract the target speech. In these situations, the pre-extracted signal can be employed as auxiliary information in place of enrollment utterances, which may be more advantageous for target speech extraction (TSE) because both the speech content and the recording environment are identical. However, this approach introduces a trade-off between content consistency and processing distortion, as the pre-extracted speech is likely to include various distortions. The nature and extent of these distortions, such as nonlinear artifacts and residual noise, depend on the specific signal processing methods used [7], [8].

In such scenarios, where the pre-extracted speech and the target speech in the mixture share the same content, it is possible to condition the model on a frame-by-frame basis, rather than relying solely on global features as in traditional TSE systems. In speech separation, various methods like Sequential Neural Beamforming [9] and Beam-guided TasNet [10] have been developed to use pre-separated speech to guide an additional separation network, thereby enhancing performance. This approach has also been extended to TSE, known as Beamformer-guided TSE [11], where the TSE network utilizes the pre-extracted beamforming output as frame-by-frame auxiliary information. However, as previously mentioned, the output from the pre-extraction stage often includes distortions, and it remains uncertain whether frame-by-frame conditioning is superior to systems employing global speaker features. Although Beamformer-guided TSE has demonstrated that frame-by-frame conditioning can surpass systems using global speaker features, the influence of distortions in pre-extracted speech on TSE has not been thoroughly investigated.

This raises several critical questions regarding the optimal auxiliary signals for TSE, particularly concerning the trade-offs between distortions and content alignment with the target source, the nature of distortions in the auxiliary signals, and the level at which these signals should be referenced:

- **Question 1**: When referencing signals at the statistical level, which is more effective: undistorted but unaligned enrollment utterances, or distorted but content-aligned pre-extracted signals?
- **Question 2**: When referencing distorted but content-aligned pre-extracted signals, which is more advantageous: statistical-level conditioning or frame-level conditioning?
- **Question 3**: When referencing distorted but content-aligned pre-extracted signals, which type of distortion is preferable: signals containing artifacts or those with

residual noise?

In this paper, we address these questions by exploring various aspects of providing auxiliary information to TSE models. Through experiments conducted in a multi-microphone environment, we aim to elucidate how different types of auxiliary signals and conditioning methods affect TSE performance. The findings from this study are expected to contribute valuable insights for the effective utilization of auxiliary information in TSE.

The remainder of the paper is organized as follows. Section II discusses the key technologies utilized in this study. Section III presents the experiments on target speaker extraction and summarizes insights related to the use of auxiliary signals. Finally, Section IV concludes with a summary of the findings.

II. Key Technologies

In this study, we address the three questions posed in the introduction by conducting experiments that compare five TSE models. These models utilize three types of auxiliary signals and two conditioning methods. The auxiliary signals include pre-recorded speech, which is commonly used in conventional TSE, as well as pre-extracted speech obtained using Independent Vector Analysis (IVA) [12] and Ideal Binary Masking (IBM) [13] in a multi-channel environment. The conditioning methods involve referencing the auxiliary signals either at the statistical level or at the frame level. This section provides an overview of the key technologies employed in our experiments.

A. Ideal Binary Masking (IBM)

IBM is a widely used technique for speech separation, where an ideal binary mask is generated to isolate target signals from interference. In this approach, each time-frequency bin is classified as binary: bins where the target signal dominates the interference are retained, while others are masked. Specifically, the spectral amplitude of the target signal $\mathbf{T}(t, f)$ is compared to that of the interference signal $\mathbf{I}(t, f)$, generating a mask that assigns a value of one to bins dominated by the target signal and zero otherwise. Applying this mask to the spectrogram of the mixture signal enhances the target signal while suppressing interference. The time-domain target signal is then reconstructed using the inverse Short-Time Fourier Transform (ISTFT) [14].

In this study, we estimated IBM using a Deep Neural Network (DNN) composed of convolutional layers and Long Short-Term Memory (LSTM) [15] layers. Multichannel mixture signals and corresponding clean target speech were used as ground truth. The DNN inputs the spectrogram of the mixture signal and outputs a binary mask that indicates whether each time-frequency bin corresponds to the target speech. During training, network parameters were optimized to minimize the error between the estimated and ideal masks using the binary cross-entropy loss function. Due to the non-linear nature of IBM processing, the output audio from DNN estimation contains artifacts.

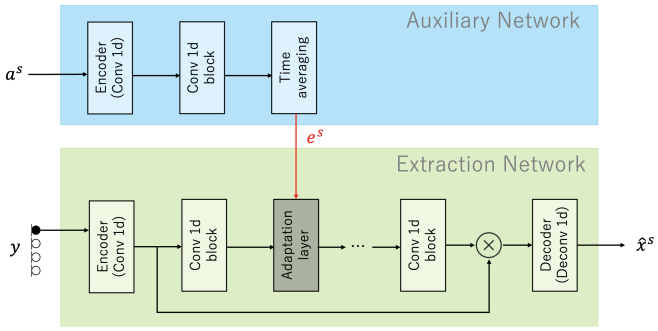

Fig. 1. Model structure for statistics-level referencing of auxiliary information. a^s can be either clean enrollment utterance or pre-extracted signal of speaker s.

B. Independent Vector Analysis (IVA)

IVA is a sophisticated Blind Source Separation (BSS) technique that leverages the statistical independence of multi-dimensional signals. Building upon Independent Component Analysis (ICA) [16], IVA excels in separating multiple sources captured by multiple microphones.

The core principle of IVA is based on a model where the observed signal vector $\mathbf{y}(t)$ is a linear mixture of unknown, independent source vectors $\mathbf{x}(t)$, indexed by time t. This relationship is mathematically represented as:

$$\mathbf{y}(t) = \mathbf{A}\mathbf{x}(t), \tag{1}$$

where $\mathbf{A}$ is the mixing matrix, with each column representing the propagation of signals from different sources to the microphones. IVA estimates this mixing matrix $\mathbf{A}$ from the observed signals, enabling the recovery of the original source signals $\mathbf{x}(t)$.

In IVA, the number of separated output signals typically matches the number of input channels. In this experiment, we assumed that the direction of arrival (DOA) of the target speaker was known in advance and used this information to perform the separation. From the separated signals, we identified the one corresponding to the target speaker by calculating the steering vector based on the speaker's DOA. We then assessed the similarity by computing the average absolute value of the inner product between the steering vectors of the separated signals and that of the target speaker. The signal with the highest similarity was selected as the target speech. However, due to the separation process, the pre-extracted speech obtained using IVA still contains residual noise.

C. Statistics-Level Referencing of Target Speaker Information

Although any existing TSE architectures could be employed for this investigation, we selected the time-domain implementation of SpeakerBeam (TD-SpeakerBeam) [17]–[21] due to its foundational role in many TSE models. Figure 1 provides an overview of TD-SpeakerBeam. This architecture consists of two networks: an auxiliary network [22] and an extraction network. The auxiliary network is designed to extract the

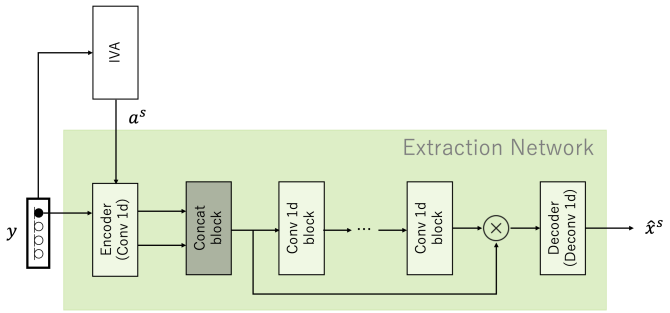

Fig. 2. Model structure of frame-level referencing of auxiliary information.

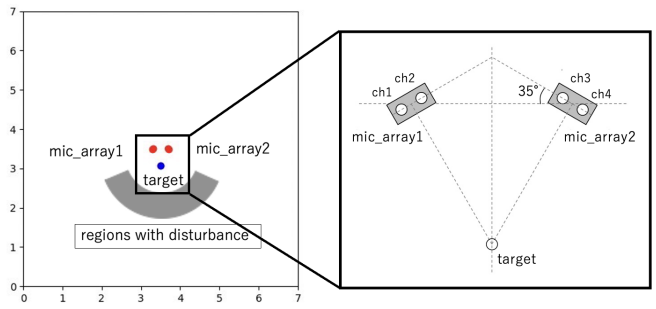

Fig. 3. Simulational acoustic field used in experiments.

global speaker embedding of the target speaker from either the enrollment utterance or the pre-extracted signal a^s. Meanwhile, the extraction network processes the mixture y and uses the speaker embedding to condition its internal features for extracting the target speech.

The auxiliary network includes an encoder layer and a single convolutional block to process the time-domain auxiliary signal. The output from the convolutional block is averaged over time frames to produce the global speaker embedding e^s.

The extraction network is configured similarly to Conv-TasNet [23]. The time-domain mixture y is input into the encoder layer, passes through several convolutional blocks, and generates a mask in the latent space. An adaptation layer, placed between the first and second convolutional blocks, incorporates the speaker embedding vector e^s from the auxiliary network to guide the mask estimation through element-wise multiplication. This approach ensures that only the target speech is extracted by tailoring the mask estimation based on the speaker embedding. The final mask, applied to the mixed signal and processed through the decoder layer, yields the time-domain extracted signal $\hat{x}^s$. Since this model uses the global speaker embedding averaged over time frames as statistical values, it exemplifies "referencing auxiliary information at the statistical level."

D. Frame-Level Referencing of Target Speaker Information

We investigate a framework in which the extraction network refers to the auxiliary information of the target speaker on a frame-by-frame basis, similar to the Beamformer-guided TSE approach. In scenarios where it is feasible to isolate and extract the target source from multi-channel observation signals, the extraction network can use the pre-extracted target source signal as auxiliary information. In this study, we employ IVA and IBM for pre-extraction to obtain this auxiliary information.

Figure 2 illustrates the structure of this model. Initially, signals recorded by multi-channel microphones are separated into source signals using IVA or IBM, and the extracted signal corresponding to the target source is referenced by the extraction network. The core extraction network retains the architecture of the SpeakerBeam. In the frame-level referencing approach, both the mixture and auxiliary signals are input into the encoder, and the outputs are concatenated along the feature dimension. Despite containing distortions, the auxiliary signals are aligned with the target source in the mixture. This allows the model to utilize the pre-extracted signal on a frame-by-frame basis, potentially leading to improved performance compared to statistical-level referencing due to the richer information available. However, when the pre-extracted signal contains significant distortion, global speaker embedding might offer greater stability. The trade-off between stability against distortions and content alignment remains an area for further investigation.

III. Target Speaker Extraction Experiment

To address the key questions concerning the trade-off between content consistency and processing distortion (**Q1**), the referencing level of auxiliary signals (**Q2**), and the type of distortion in auxiliary signals (**Q3**), we conducted target speaker extraction (TSE) experiments in a multi-channel input setting.

To achieve this, we compared the following models:

1) A model that refers to undistorted pre-recorded speech at the statistical level,
2) A model that refers to speech containing artifacts pre-extracted by IBM at the statistical level,
3) A model that refers to speech containing residual noise pre-extracted by IVA at the statistical level,
4) A model that refers to speech containing artifacts pre-extracted by IBM at the frame level, and
5) A model that refers to speech containing residual noise pre-extracted by IVA at the frame level.

Comparing Models 1, 2, and 3 addresses Question 1. To investigate Question 2, we compare Models 2 and 4, as well as Models 3 and 5. Finally, to explore Question 3, we compare Models 2 and 3, as well as Models 4 and 5.

A. Experimental Setups

Figure 3 illustrates the experimental simulation environment. The room dimensions are $7 \times 7 \times 3$ meters. Two microphone arrays, each with two channels, are arranged as shown in the figure. The microphone arrays are spaced 40 cm apart, with a 3 cm distance between microphones within each array. The

TABLE I
RESULTS OF TARGET SPEAKER EXTRACTION EXPERIMENTS IN TWO-SPEAKER MIXED ENVIRONMENT.

System	Target Speaker Extraction Models		Evaluation Results	
	Auxiliary Signal	Referencing Level	SISDR [dB]	STOI [%]
Observation	-	-	-0.5	70.3
IBM	-	-	7.3	86.5
IVA	-	-	7.8	88.4
TSE-1	Pre-recorded clean signal	Statistics	12.1	90.6
TSE-2	Pre-extracted signal by IBM	Statistics	12.6	91.3
TSE-3	Pre-extracted signal by IVA	Statistics	12.4	91.1
TSE-4	Pre-extracted signal by IBM	Frame-by-frame	15.2	94.5
TSE-5	Pre-extracted signal by IVA	Frame-by-frame	18.0	96.3

target source is placed in front of the microphone arrays, while the interfering sources are located one to two meters behind. This setup simulates the recording of mixed speech from two speakers [24]. The position of the target source remains fixed, whereas the interfering sources are positioned at angles ranging from 15° to 165° in 10° increments, allowing for interference from various directions. The number of interfering sources remains constant for each direction.

Both the target and interfering speeches are selected from the LibriSpeech corpus [25], using speaker combinations consistent with those in the Libri2Mix corpus [26]. The sampling frequency is set to 8 kHz, and the reverberation time is randomly chosen between 0.2 and 0.5 seconds. The dataset comprises 27000 samples for training, 6000 samples for validation, and 6000 samples for evaluation.

The model is trained for 200 epochs with a batch size of five, using the negative SI-SDR [27] as the loss function to measure the difference between the ground truth and the extracted speech. The Adam optimizer is employed with a learning rate of 0.001, and no weight decay is applied. To prevent convergence to local optima, the learning rate is halved if there is no improvement in validation performance for 20 epochs. If performance does not improve after 120 epochs, training is halted early.

B. Experimental Results

Table I presents the results from the target speaker extraction experiments. Based on these results, insights are organized to address the following three key questions: *i)* the trade-off between content consistency and distortion, *ii)* the approach to referencing auxiliary signals, and *iii)* the impact of different types of distortion in auxiliary signals.

1) Q1: Trade-off between Content Consistency and Distortion: We compared the use of undistorted pre-recorded speech (TSE-1), which differs in content from the target source, with pre-extracted speech containing distortions but aligned with the speech content (TSE-2 and TSE-3), referencing them at the statistical level. The results indicate that, in the trade-off between content consistency and distortion, the latter has a more favorable impact on model performance. This is likely because, when conditioning at the statistical level, the time-averaging process in the auxiliary network compresses the signal, thereby mitigating the adverse effects of distortion in the auxiliary signal.

2) Q2: Referecing Level of Auxiliary Singals: The results from using TSE-2 through TSE-5 indicate that models conditioned at the frame level outperform those conditioned at the statistical level when referencing pre-extracted speech, whether by IVA or IBM. This suggests that when the auxiliary signal is aligned with the target speech, learning local features that capture the temporal variations within the utterance is more effective than aggregating information to learn global features.

3) Q3: Type of Distortion in Auxiliary Singals: In a multi-channel environment, using IBM for source separation introduces specific artifacts due to nonlinear processing, whereas using IVA results in residual noise. The artifacts typically manifest as artificial, harsh sounds caused by missing spectral components of the target signal, while residual noise comprises leftover interference and environmental noise from the recording setup.

When comparing TSE-2 with TSE-3 and TSE-4 with TSE-5, it can be observed that at the statistical level, residual noise has a more detrimental impact on performance than artifacts, whereas at the frame level, artifacts are more problematic than residual noise. This suggests that for statistical-level conditioning, which captures global features of the auxiliary signal, it is crucial to ensure that no extraneous elements are present, leaving only the target speech. In contrast, for frame-level conditioning, which captures local features, it is more important that, even if extraneous elements are present, the target speech remains intact and clear.

IV. CONCLUSION

This study examined the effects of various types of auxiliary signals and conditioning methods on the performance of TSE models in a multi-channel environment. Our experiments revealed that pre-extracted signals with distortions, when aligned with the target speech, enhance TSE performance more effectively than undistorted but unaligned pre-recorded utterances. Frame-level conditioning proved particularly advantageous, as it captures local features and utilizes richer information compared to statistical-level conditioning. Furthermore, the type of distortion in the auxiliary signal was found to influence model performance: residual noise has a more adverse impact than artifacts in statistical-level conditioning, while the reverse holds

true for frame-level conditioning. These findings underscore the importance of maintaining the purity of auxiliary signals for global feature extraction and preserving the integrity of target speech information for local feature extraction, offering valuable insights for the design and optimization of TSE systems.

REFERENCES

[1] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[2] N. Kanda, C. Boeddeker, J. Heitkaemper, *et al.*, "Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr," in *CHiME-5 Workshop*, 2019.

[3] T. Nakatani, T. Yoshioka, M. Delcroix, and K. Kinoshita, "Neural target speech extraction: An overview," *arXiv preprint arXiv:2301.13341*, 2020.

[4] A. Ephrat, I. Mosseri, S. Lang, *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 11 792–11 802.

[5] J. Heitkaemper, T. Feher, M. Freitag, and R. Haeb-Umbach, "A study on online source extraction in the presence of changing speaker positions," in *International Conference on Statistical Language and Speech Processing 2019, Ljubljana, Slovenia*, 2019.

[6] P. Comon and C. Jutten, "Blind source separation: The state of the art," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 80–88, 2010.

[7] P. Bofill, J. M. Amigó, and E. Parra, "Residual crosstalk in blind source separation: Analysis and solutions," *Neurocomputing*, vol. 72, no. 1-3, pp. 1–13, 2008.

[8] K. Nishikawa and T. Yamasaki, "Nonlinear distortions in blind source separation algorithms: An experimental study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 244–254, 2013.

[9] Z.-Q. Wang, H. Erdogan, S. Wisdom, *et al.*, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021.

[10] H. Chen, Y. Yi, D. Feng, and P. Zhang, "Beam-guided tasnet: An iterative speech separation framework with multi-channel output," *arXiv preprint arXiv:2102.02998*, 2021.

[11] M. Elminshawi, S. R. Chetupalli, and E. A. P. Habets, "Beamformer-guided target speaker extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[12] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.

[13] D. Wang, "Ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[14] J. B. Allen and L. R. Rabiner, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural computation*, MIT Press, vol. 9, 1997, pp. 1735–1780.

[16] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[17] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. ICASSP*, 2018, pp. 5554–5558.

[18] K. Žmolíková, M. Delcroix, K. Kinoshita, *et al.*, "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.

[19] M. Delcroix, S. Watanabe, T. Ochiai, *et al.*, "End-to-end speakerbeam for single channel target speech recognition.," in *Proc. Interspeech*, 2019, pp. 451–455.

[20] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *Proc. ICASSP*, 2019, pp. 6965–6969.

[21] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. ICASSP*, 2020, pp. 691–695.

[22] K. Vesely, S. Watanabe, K. Žmolíková, M. Karafiat, and L. Burget, "Sequence summarizing neural network for speaker adaptation," in *In Proc. ICASSP*, 2016.

[23] Y. L. *et al.*, "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.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[26] 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.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *Proc. ICASSP*, 2019, pp. 626–630.