TYPE Original Research
PUBLISHED 24 December 2024
DOI 10.3389/frsip.2024.1440401

OPEN ACCESS

EDITED BY
Tsubasa Ochiai,
Nippon Telegraph and Telephone, Japan

REVIEWED BY
Hangting Chen,
Tencent Holdings Limited, China
Kateřina Žmolíková,
Meta AI, United Kingdom
Takanori Ashihara,
Nippon Telegraph and Telephone, Japan

*CORRESPONDENCE
Mohamed Elminshawi,
✉ mohamed.elminshawi@audiolabs-erlangen.de

RECEIVED 29 May 2024
ACCEPTED 21 October 2024
PUBLISHED 24 December 2024

CITATION
Elminshawi M, Mack W, Chetupalli SR, Chakrabarty S and Habets EAP (2024) New insights on the role of auxiliary information in target speaker extraction.
Front. Sig. Proc. 4:1440401.
doi: 10.3389/frsip.2024.1440401

New insights on the role of auxiliary information in target speaker extraction

Mohamed Elminshawi*, Wolfgang Mack, Srikanth Raj Chetupalli, Soumitro Chakrabarty and Emanuël A. P. Habets

International Audio Laboratories Erlangen, Erlangen, Germany

Speaker extraction (SE) aims to isolate the speech of a target speaker from a mixture of interfering speakers with the help of auxiliary information. Several forms of auxiliary information have been employed in single-channel SE, such as a speech snippet enrolled from the target speaker or visual information corresponding to the spoken utterance. The effectiveness of the auxiliary information in SE is typically evaluated by comparing the extraction performance of SE with uninformed speaker separation (SS) methods. Following this evaluation procedure, many SE studies have reported performance improvement compared to SS, attributing this to the auxiliary information. However, recent advancements in deep neural network architectures, which have shown remarkable performance for SS, suggest an opportunity to revisit this conclusion. In this paper, we examine the role of auxiliary information in SE across multiple datasets and various input conditions. Specifically, we compare the performance of two SE systems (audio-based and video-based) with SS using a unified framework that utilizes the commonly used dual-path recurrent neural network architecture. Experimental evaluation on various datasets demonstrates that the use of auxiliary information in the considered SE systems does not always lead to better extraction performance compared to the uninformed SS system. Furthermore, we offer new insights into how SE systems select the target speaker by analyzing their behavior when provided with different and distorted auxiliary information given the same mixture input.

KEYWORDS

speaker extraction, speaker separation, deep learning, auxiliary information, single-channel

1 Introduction

Speech is the primary means through which humans communicate. In typical acoustic scenes, a recording of a speaker of interest is often degraded by other acoustic sources, such as background noise and interfering speakers. Remarkably, human brains have the ability to focus on a specific acoustic source in a noisy environment while ignoring others, a phenomenon commonly referred to as the *cocktail party effect* (Cherry, 1953). In contrast, speech corrupted with concurrent interfering speakers has been shown to severely deteriorate the performance of several speech processing algorithms, including automatic speech recognition (Cooke et al., 2010) and speaker verification (SV) (Martin and Przybocki, 2001). Over the past several decades, a considerable amount of research has been devoted to dealing with overlapped speech as a speaker separation (SS) task, i.e., separating all speakers from the observed mixture signal (Wang and Chen, 2018). In particular, deep learning has considerably advanced the performance of single-channel SS methods (Hershey et al., 2016; Yu et al., 2017; Chen et al., 2017; Luo and Mesgarani, 2019; Luo et al., 2020; Chen et al., 2020; Zeghidour and Grangier, 2021; Subakan et al., 2021; Byun and Shin, 2021). One fundamental issue associated with SS is the permutation problem, i.e., the correspondence between the separated output signals and the speakers is arbitrary. This ambiguity poses a challenge when training deep neural networks (DNNs) for separation since the loss function needs to be computed between each output signal and the ground-truth speech of its corresponding speaker. To address this challenge, permutation invariant training (PIT) (Yu et al., 2017; Kolbæk et al., 2017) has been proposed, which enables optimizing DNNs that directly separate the speech signals by finding the permutation of the ground-truth signals that best matches the output signals.

In many scenarios, it may not be necessary to reconstruct all speakers from the mixture; instead, it suffices to extract a single target speaker. This task has been given numerous names in the literature, among which are *target speaker extraction* (Delcroix et al., 2018; 2021), *informed speaker extraction* (Ochiai et al., 2019b), or simply speaker extraction (SE) (Xu et al., 2020; Zmolikova et al., 2019). In contrast to SS, SE systems do not suffer from the permutation ambiguity since only a single output exists. Early works on SE (Du et al., 2014; 2016) were target-dependent, i.e., systems designed to extract speech from only a particular speaker and cannot generalize to other speakers. Such systems require abundant training data from the target speaker, which is infeasible in many applications. To realize speaker-independent SE systems, prior knowledge or auxiliary information must be provided to specify the target signal. SE approaches can be categorized based on the modality of the auxiliary information. Audio-based SE (SE-A) methods rely on a speech snippet from the target speaker that guides the system towards that speaker (Zmolikova et al., 2019; Wang et al., 2019; Wang et al., 2018; Delcroix et al., 2020). Video-based SE (SE-V) methods[1] have also been proposed that leverage visual information from the target speaker, such as lip movements (Gabbay et al., 2018; Hou et al., 2018; Afouras et al., 2018a; Wu et al., 2019) or cropped facial frames (Ephrat et al., 2018; Afouras et al., 2019). Other methods have exploited multi-modal information, for example, by utilizing both visual features of the target speaker as well as an enrollment utterance (Afouras et al., 2019; Ochiai et al., 2019a; Sato et al., 2021). Finally, brain signals (Ceolini et al., 2020) and speaker activity (Delcroix et al., 2021) have also been utilized as auxiliary signals for SE.

Clearly, SS and SE are related problems in the sense that both deal with overlapped speech. In fact, SE can be realized by using a SS system followed by a SV module, where all speakers are first separated, and then SV is applied on all outputs to select the target speaker. However, SS and SE exhibit notable distinctions in terms of their underlying assumptions and the nature of errors that could arise. In SS, all speakers in the mixture are to be recovered, whereas only a unique speaker is assumed to be the target in SE. In addition, knowledge about the number of speakers in the mixture is often assumed in SS (Hershey et al., 2016; Isik et al., 2016; Yu et al., 2017; Kolbæk et al., 2017), while such an assumption is not necessary in SE. Furthermore, SE necessitates prior knowledge about the target speaker in the form of an auxiliary signal, while SS blindly isolates the speech signals. With respect to evaluation, any permutation of the outputs of SS is a valid solution and leads to the same objective metrics. In contrast, SE systems are prone to *speaker confusion*, i.e., recovering an interfering speaker instead of the target (Zhao et al., 2022). The above points should not be overlooked when evaluating and, especially, comparing the performance of SS and SE.

The utility of the auxiliary information in SE is generally assessed by comparing the extraction performance of SE to that of SS with target speaker selection (e.g., in an oracle fashion) (Zmolikova et al., 2019; Delcroix et al., 2020). Using this evaluation procedure, the majority of SE works often report performance improvement over SS, attributing this to the use of auxiliary information. In particular, it has been argued that the use of auxiliary information improves the performance in scenarios involving mixtures having similar voice characteristics (e.g., same-gender mixtures) (Gabbay et al., 2018), long mixtures with complicated overlapping patterns (Zmolikova et al., 2019), or adverse acoustic conditions, e.g., very low signal-to-noise ratios (SNRs) or more interfering speakers (Chuang et al., 2020; Michelsanti et al., 2021). Another work in SE-V has demonstrated that the auxiliary visual information improves the extraction performance compared to SS, especially for visually distinguishable sounds (Aldeneh et al., 2021).

However, ongoing advancements in DNN architectures, which have demonstrated significant performance improvements in SS (Luo et al., 2020; Chen et al., 2020; Subakan et al., 2021), suggest an opportunity to revisit these findings. Having a clear understanding of the contribution of the auxiliary information in SE would not only give us more insights into how such systems function, but it could also allow us to develop robust SE systems against unreliable auxiliary information, e.g., noisy enrollment utterances, occluded or temporally misaligned visual features.

In this work, we conduct an empirical study to objectively examine the role of auxiliary information in SE from two aspects. Firstly, through a comprehensive analysis over multiple datasets and various mixing conditions, the utility of the auxiliary information in improving the extraction performance of SE-A and SE-V in comparison with uninformed SS is revisited. To ensure a fair comparison, all SE and SS systems are implemented within a unified framework employing the commonly used dual-path recurrent neural network (DPRNN) architecture (Luo et al., 2020). Secondly, inspired by previous works (Afouras et al., 2019; Sato et al., 2021) that address corrupted auxiliary signals in SE, we offer new insights into how SE systems select the target speaker by inspecting their behavior for various samples from the embedding space of the auxiliary information. This analysis also highlights the difference in how SE systems select the target speaker when trained on 2-speaker mixtures versus those trained on 3-speaker mixtures, potentially addressing the issue of speaker confusion, which typically occurs in scenarios involving inactive target speakers (Borsdorf et al., 2021; Delcroix et al., 2022). The remainder of the paper is structured as follows. Section 2 describes the tasks of SS and SE as well as the systems employed in this study. In Section 3, the experimental setup

[1] Also known in the literature as audio-visual speaker enhancement/separation methods.

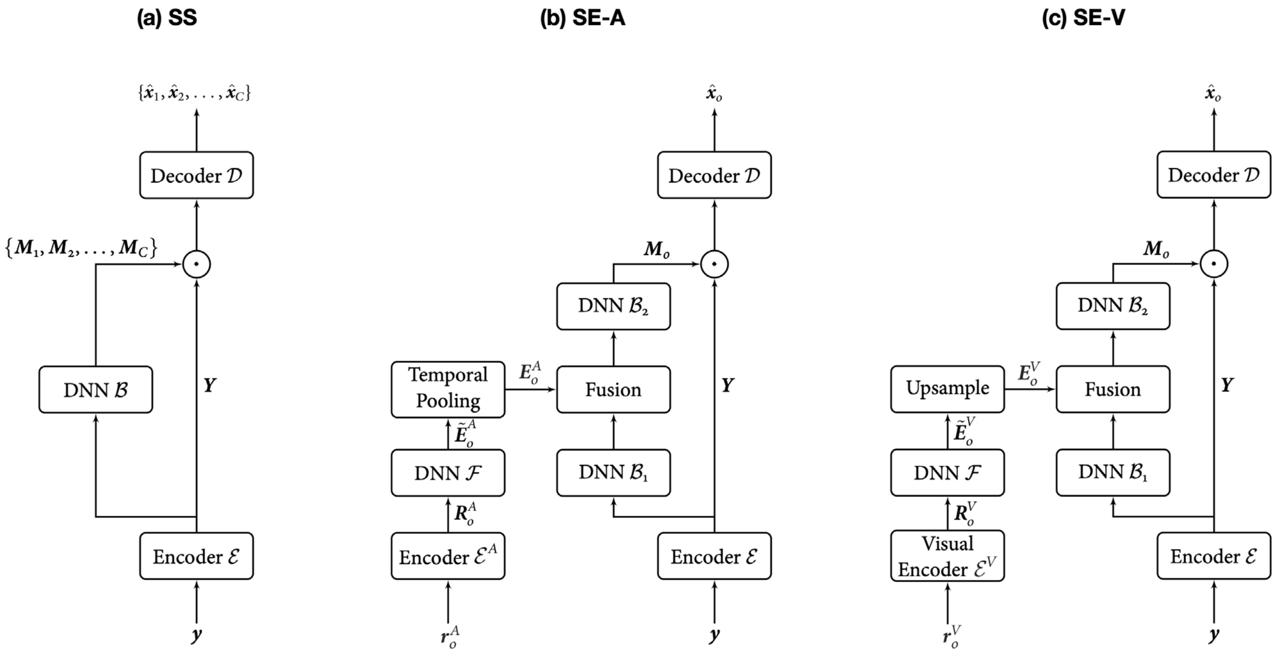

FIGURE 1
Block diagrams of the speaker separation (SS) and speaker extraction (SE) systems: **(A)** SS trained with uPIT, **(B)** audio-based SE (SE-A), and **(C)** video-based SE (SE-V).

is discussed, and Section 4 presents the experimental results. Finally, the discussion is provided in Section 5.

2 Speaker separation and extraction systems

In this section, we formally define the problems of SS and SE, and provide a detailed description of the different systems used in this study, which resemble, to a great extent, recently developed methods in the literature on SS and SE (Luo and Mesgarani, 2019; Luo et al., 2020; Delcroix et al., 2020; Wu et al., 2019; Ochiai et al., 2019a; Ge et al., 2020; Pan et al., 2022). Figure 1 shows the block diagrams of the different systems. Note that a common backbone is employed for all systems to ensure a fair comparison. Further details about the systems' configurations are described in Section 3.2.

Let $y \in \mathbb{R}^S$ be S samples of an observed single-channel time-domain mixture signal consisting of speech from C speakers, denoted by $x_1, \ldots, x_C \in \mathbb{R}^S$, i.e.,

$$y = \sum_{i=1}^{C} x_i. \tag{1}$$

2.1 Speaker separation (SS)

The objective of SS is to reconstruct all the constituent speech signals in the mixture, i.e.,

$$\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_C\} = f(y), \tag{2}$$

where $\hat{x}_i \in \mathbb{R}^S$ denotes the estimated speech signal at the i-th output, and f represents the transformation applied by the separation system on the mixture signal y. Note that the order of the output signals is arbitrary, and a mapping between each output and its corresponding speaker identity is typically required. Following prior SS works (Luo and Mesgarani, 2018; 2019; Chen et al., 2020; Subakan et al., 2021), we adopt an encoder-masker-decoder structure for f, as shown in Figure 1A. In particular, f comprises three main blocks: an encoder, a mask estimator using a DNN, and a decoder, represented by $\mathcal{E}$, $\mathcal{B}$, and $\mathcal{D}$, respectively. The encoder $\mathcal{E}$ transforms the time-domain waveform of the mixture into frame-wise features $Y \in \mathbb{R}^{N\times T}$, where N is the dimensionality of the encoded features of each time frame, and T denotes the number of time frames, i.e.,

$$Y = \mathcal{E}(y). \tag{3}$$

The mask estimator $\mathcal{B}$ is a DNN that maps the encoded features Y to a mask for each speaker in the mixture, i.e.,

$$\{M_1, M_2, \ldots, M_C\} = \mathcal{B}(Y), \tag{4}$$

where $M_i \in \mathbb{R}^{N\times T}$ denotes the mask for the i-th output. Each mask is then applied to the encoded features Y and subsequently fed to the decoder $\mathcal{D}$ to reconstruct the time-domain waveform of the corresponding speaker, denoted by $\hat{x}_i \in \mathbb{R}^S$, as

$$\hat{x}_i = \mathcal{D}(Y \odot M_i), \tag{5}$$

where $\odot$ denotes the Hadamard product. To optimize the parameters of the separation system, utterance-level PIT (uPIT) (Kolbæk et al., 2017) is used first to find the bijective mapping between each output signal $\hat{x}_i$ and its corresponding speaker, as

$$\phi^{*} = \underset{\phi\in\mathcal{P}}{\operatorname{argmin}} \sum_{i=1}^{C} \boldsymbol{\ell}\left(\boldsymbol{x}_{\phi(i)}, \hat{\boldsymbol{x}}_i\right), \tag{6}$$

where $\boldsymbol{\ell}$ is a loss function defined between two time-domain signals, $\mathcal{P}$ represents the set of all possible permutations, and ϕ^{*} is the optimum permutation that provides the minimum loss. The total loss $\mathcal{L}$ is then computed, as

$$\mathcal{L} = \sum_{i=1}^{C} \boldsymbol{\ell}\left(\boldsymbol{x}_{\phi^{*}(i)}, \hat{\boldsymbol{x}}_i\right). \tag{7}$$

2.2 Speaker extraction (SE)

In contrast to SS, SE refers to the task of reconstructing a single target speaker from the mixture given auxiliary information about the target speaker. We refer to the auxiliary information as the reference signal and denote it by $\boldsymbol{r}_o$, where o is the index of the desired speaker. Note that the dimensionality of $\boldsymbol{r}_o$ depends on the type of information provided to the system. SE can be formulated as

$$\hat{\boldsymbol{x}}_o = g\left(\boldsymbol{y}, \boldsymbol{r}_o\right), \tag{8}$$

where g represents the transformation carried out by the SE system, and $\hat{\boldsymbol{x}}_o \in \mathbb{R}^S$ denotes the estimated speech of the target speaker. Existing SE methods typically decompose g into two stages (Ephrat et al., 2018; Wu et al., 2019; Ochiai et al., 2019a; Wang et al., 2019; Delcroix et al., 2020; Ge et al., 2020; Pan et al., 2022): an auxiliary network and an extraction network, represented by h and $\tilde{f}$, respectively. The auxiliary network h extracts informative features, denoted by $\boldsymbol{E}_o$, from the reference signal, which help to specify the target speaker, i.e.,

$$\boldsymbol{E}_o = h\left(\boldsymbol{r}_o\right). \tag{9}$$

The second stage is to condition the extraction network $\tilde{f}$ on the features $\boldsymbol{E}_o$ such that an estimate of the target speaker can be obtained as

$$\hat{\boldsymbol{x}}_o = \tilde{f}\left(\boldsymbol{y}, \boldsymbol{E}_o\right). \tag{10}$$

Following prior works (Wu et al., 2019; Delcroix et al., 2020), $\tilde{f}$ has a similar encoder-masker-decoder structure to the separation system f described in Section 2.1, except that the mask estimator consists of two DNN blocks $\mathcal{B}_1$ and $\mathcal{B}_2$, as well as a fusion layer inserted in between, such that the informative features $\boldsymbol{E}_o$ can be included. Many fusion techniques have been proposed, e.g., concatenation-based (Ephrat et al., 2018; Wu et al., 2019), and product-based (Ochiai et al., 2019a; Zmolikova et al., 2019; Delcroix et al., 2020). In this work, we adopt the product-based technique as the fusion mechanism, based on empirical results showing that it provided better performance compared to the concatenation-based mechanism. The transformations carried out by $\tilde{f}$ are represented by

$$\boldsymbol{M}_o = \mathcal{B}_2\left(\mathcal{B}_1\left(\boldsymbol{Y}\right) \odot \boldsymbol{E}_o\right), \tag{11}$$

$$\hat{\boldsymbol{x}}_o = \mathcal{D}\left(\boldsymbol{Y} \odot \boldsymbol{M}_o\right). \tag{12}$$

The loss function, in this case, is computed with respect to the target speaker only, and hence uPIT is not required, i.e.,

$$\mathcal{L} = \boldsymbol{\ell}\left(\boldsymbol{x}_o, \hat{\boldsymbol{x}}_o\right). \tag{13}$$

The design of the auxiliary network h in (Equation 9) depends on the modality of the reference signal $\boldsymbol{r}_o$. In this work, we focus on audio-based SE (SE-A) and video-based SE (SE-V). Both systems are illustrated in Figures 1B, C, and described in the following.

2.2.1 Audio-based SE (SE-A)

In SE-A, a reference speech signal from the target speaker is used as auxiliary information to guide the extraction system. Typical SE-A methods realize this process by mapping the reference speech signal to an embedding vector that encodes the voice characteristics of the target speaker. The well-known speaker representations developed for speaker recognition, such as i-vector (Dehak et al., 2011) and d-vector (Wan et al., 2018), have been employed in SE-A in (Zmolikova et al., 2019; Wang et al., 2019). Alternatively, speaker representations can also be learned in an end-to-end fashion via an auxiliary DNN that is jointly optimized with the extraction network (Zmolikova et al., 2019; Delcroix et al., 2020). The end-to-end approach was adopted in this study, as we empirically found it to be better than using a pre-trained speaker recognition model (Desplanques et al., 2020). The first block in the auxiliary network is an audio encoder $\mathcal{E}^A$ (similar to $\mathcal{E}$ in (Equation 3)), i.e.,

$$\boldsymbol{R}_o^A = \mathcal{E}^A\left(\boldsymbol{r}_o^A\right), \tag{14}$$

where $\boldsymbol{r}_o^A \in \mathbb{R}^{S_r}$ denotes the speech reference signal from the target speaker having a length of S_r samples, and $\boldsymbol{R}_o^A \in \mathbb{R}^{N\times T_a}$ represents the encoded frame-wise features, where T_a represents the number of time frames. The features $\boldsymbol{R}_o^A$ are further processed with a DNN, denoted by $\mathcal{F}$, which produces an embedding vector for each time frame, arranged in the matrix $\tilde{\boldsymbol{E}}_o^A \in \mathbb{R}^{N\times T_a}$, as

$$\tilde{\boldsymbol{E}}_o^A = \mathcal{F}\left(\boldsymbol{R}_o^A\right). \tag{15}$$

Temporal average pooling is then applied to the frame-wise features $\tilde{\boldsymbol{E}}_o^A$, such that an utterance-wise embedding vector, represented by $\boldsymbol{E}_o^A \in \mathbb{R}^{N\times 1}$, is obtained. Note that, in this case, the embedding vector $\boldsymbol{E}_o^A$ is *time-invariant*, and it is broadcasted over the different time frames in the fusion layer in (Equation 11).

2.2.2 Video-based SE (SE-V)

In an attempt to mimic the multimodality of human perception (Golumbic et al., 2013; Partan and Marler, 1999), video-based SE leverages visual cues, such as lip movements or facial expressions, as auxiliary information. The first component of the auxiliary network in the SE-V system is a visual encoder, denoted by $\mathcal{E}^V$. This encoder extracts visual features $\boldsymbol{R}_o^V \in \mathbb{R}^{N_v\times T_v}$, where N_v represents the dimensionality of the features, and T_v represents the number of time frames, from the given visual reference signal $\boldsymbol{r}_o^V \in \mathbb{R}^{D\times H\times W\times T_v}$, where D, H, and W denote the depth, height, and width, respectively, i.e.,

$$\boldsymbol{R}_o^V = \mathcal{E}^V\left(\boldsymbol{r}_o^V\right). \tag{16}$$

The design of the visual encoder $\mathcal{E}^V$ depends on the type of visual information used. When facial frames are utilized as a reference signal, typically $\mathcal{E}^V$ is a pre-trained face recognition network, e.g., FaceNet (Cole et al., 2017), from which an embedding vector for each facial frame is extracted (Ephrat et al.,

TABLE 1 Statistics of the training, validation, test splits of the different datasets.

Corpus	No. Speakers	Total duration [h]	No. Utterances
TCD-TIMIT (Harte and Gillen, 2015)	47/6/6	6.7/0.9/0.8	30k/5k/3k
LRS3 (Afouras et al., 2018b)	906/49/142	25.0/1.2/3.7	30k/5k/3k
WSJ0 (Garofolo et al., 1993)	101/(8/10)	24.9/(1.5/2.2)	WSJ0-2&3Mix (Hershey et al., 2016): 20k/5k/3k
LibriSpeech (Panayotov et al., 2015)	921/40/40	362.4/5.4/5.4	Libri-2Mix(3Mix) (Cosentino et al., 2020): 51k (34k)/3k/3k

2018; Ochiai et al., 2019a). In the case of lip frames as a reference signal, $\mathcal{E}^V$ typically consists of a spatio-temporal convolutional layer, i.e., 3-D ConvLayer, followed by ResNet-18 (Stafylakis and Tzimiropoulos, 2017) that outputs a lip embedding for each frame (Afouras et al., 2018a; Wu et al., 2019; Pan et al., 2022). It has been shown that using lip features as visual information in SE generally provides better extraction performance than facial features (Inan et al., 2019; Shetu et al., 2021).

In this study, we adopt lip frames as visual information and use the 3-D ConvLayer + ResNet-18 structure for the visual encoder $\mathcal{E}^V$. The visual embeddings $\boldsymbol{R}_o$ are further processed with a DNN $\mathcal{F}$, resulting in more task-specific features, represented by $\tilde{\boldsymbol{E}}_o^V \in \mathbb{R}^{N\times T_v}$, as

$$\tilde{\boldsymbol{E}}_o^V = \mathcal{F}(\boldsymbol{R}_o^V). \tag{17}$$

Finally, to match the sampling rates of the visual and audio streams, the learned features $\tilde{\boldsymbol{E}}_o^V$ are upsampled using linear interpolation along the temporal dimension, similar to (Owens and Efros, 2018; Pan et al., 2022), resulting in the frame-wise features $\boldsymbol{E}_o^V \in \mathbb{R}^{N\times T}$.

3 Experimental design

In this section, we introduce the datasets used in this study and describe the specific configurations of the systems presented in Section 2 and two baselines for SE. We then provide details of the training setup. Finally, we describe the evaluation procedure for SS systems within a SE setup.

3.1 Datasets

For experimentation, we consider the four most commonly used datasets in the literature on SE. These datasets differ in terms of the number of examples, the number of speaker identities, the vocabulary size, and the recording conditions. The details of these datasets are presented in Table 1. The audio files in all datasets have a sampling frequency of 16 kHz.

3.1.1 TCD-TIMIT

TCD-TIMIT (Harte and Gillen, 2015) consists of synchronized audio-visual recordings of 59 speakers reading sentences from the TIMIT corpus. TCD-TIMIT is collected in a controlled environment, thus comprising high-quality audio and video clips of speech. The video recordings are sampled at 25 frames per second.

Since there are no official SS/SE dataset splits for TCD-TIMIT, we created training, validation, and test splits based on speaker identities, i.e., the splits are ensured to form disjoint sets in terms of the speaker identities. We then simulated 2-speaker and 3-speaker mixtures by randomly sampling utterances from different speakers in each split. For consistency among the datasets and following prior works (Ochiai et al., 2019a; Sato et al., 2021), the utterances were mixed with a signal-to-interference ratio (SIR) sampled from −5 dB–5 dB. The mixtures have a duration of 3 s. In the case of 3-speaker mixtures, two SIRs were sampled, and each interferer was scaled by its corresponding SIR with respect to the target, and then all signals were superimposed.

3.1.2 LRS3

LRS3 (Afouras et al., 2018b) is a large-scale audio-visual corpus obtained from TED and TEDx talks. Unlike TCD-TIMIT, LRS3 is collected *in the wild*, resulting in a lower quality of samples compared to TCD-TIMIT. However, LRS3 has a tremendous variability in terms of the spoken sentences, visual appearances, and speaking styles, which allows developing robust DNN models that generalize to real-world conditions. Similar to TCD-TIMIT, the video recordings are sampled at 25 frames per second. In addition, we followed the same procedure as in TCD-TIMIT to create dataset splits suitable for SS/SE, since there are no official dataset splits for LRS3.

3.1.3 WSJ0Mix

Derived from the WSJ0 corpus (Garofolo et al., 1993), the WSJ0-2Mix and WSJ0-3Mix datasets (Hershey et al., 2016) for single-channel 2-speaker and 3-speaker mixtures, respectively, have become the standard benchmark for SS. The utterances are mixed with a SIR randomly sampled from −5 dB–5 dB. We used the min version of the datasets. Unlike TCD-TIMIT and LRS3, WSJ0Mix comprises only audio signals, with no corresponding visual recordings of the speakers.

3.1.4 LibriMix

LibriMix (Cosentino et al., 2020) is an audio-only dataset that comprises 2-speaker and 3-speaker mixtures created from the LibriSpeech corpus (Panayotov et al., 2015). For our experiments, we used the *clean* subset of the dataset, which comprises speech mixtures without noise or reverberation. The SIR of the mixtures in the dataset follows a normal distribution with a mean of 0 dB and a standard deviation of 4.1 dB. Similar to WSJ0Mix, we used the min version of the dataset.

For all the datasets mentioned above, each reference speech signal for the SE-A system was an utterance spoken by the target speaker that was different from the one in the mixture. For the SE-V system, each reference signal was derived by cropping the lip region of the target speaker from the video clip of the corresponding spoken utterance.

3.2 Model configurations

As mentioned in Section 2, we adopted an encoder-masker-decoder structure for the different SS and SE systems. In particular, we followed a TasNet-like structure (Luo and Mesgarani, 2018; 2019) for the encoder and decoder, which utilizes learnable kernels instead of the traditional pre-defined Fourier bases. Both audio encoders $\mathcal{E}$ and $\mathcal{E}^A$, shown in Figure 1, consisted of a 1-D convolutional layer followed by a rectified linear unit (ReLU) non-linearity. We set the parameters of this layer as follows: number of kernels $N = 256$, kernel size $L = 32$ (2 ms), and hop size $R = 16$ (1 ms). The decoder $\mathcal{D}$ was a 1-D transposed convolutional layer, having the same kernel and hop sizes as the encoder.

For the SE-V system, the visual encoder $\mathcal{E}^V$ was pre-trained[2] on a speech recognition task, and we kept its parameters fixed during training, similar to (Wu et al., 2019; Pan et al., 2022). The parameter count of the visual encoder is 11.2M. A linear layer was used to match the feature dimensionality of the visual encoder's output (i.e., $N_v = 512$) with the input of the DNN block $\mathcal{F}$ (i.e., $N = 256$).

For the DNN blocks in Figure 1, i.e., $\mathcal{B}$, $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{F}$, we employed the DPRNN architecture (Luo et al., 2020). We used the DPRNN implementation provided by SpeechBrain (Ravanelli et al., 2021) with the following hyperparameters. For the intra- and inter-chunk recurrent neural networks (RNNs), bi-directional long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) were used with 128 hidden units in each direction. The bottleneck size was set to 64. We used a chunk size of 90, except for $\mathcal{F}$ in SE-V, which was set to 12. This choice ensures a comparable sequence length for the intra- and inter-chunk RNNs and was calculated for a 3-s input. Global Layer Normalization (gln) (Luo and Mesgarani, 2019) was used, and ReLU non-linearity was applied at the output. For the SE-A and SE-V systems, the DNN blocks $\mathcal{B}_1$ and $\mathcal{B}_2$ have the same size, each consisting of 3 DPRNN blocks. In contrast, for the SS system, the DNN block $\mathcal{B}$ comprised 6 DPRNN blocks. This ensures that the separation network f and the extraction network $\tilde{f}$ are similar. In the auxiliary network of SE-A and SE-V, the DNN block $\mathcal{F}$ consists of only one DPRNN block. The number of trainable parameters of the SS, SE-A, and SE-V models is 2.6M, 3.2M, and 4.2M, respectively.

As visual information for the SE-V system, lip regions were extracted using facial landmark detection implemented in (King, 2009). The lip regions were transformed into grayscale (i.e., $D = 1$) and resized to 100×50 pixels corresponding to the width W and height H, respectively. For visual frames where the lip detection algorithm failed to detect the lip region, e.g., due to occlusion, a patch of zeros was used instead.

3.3 SE baselines

To validate the design choices of the adopted SE systems, we compare their performance with two SE baselines: SpEx+ (Ge et al., 2020) and USEV (Pan et al., 2022). These baselines were chosen because they follow a time-domain approach similar to the SE systems in this study, allowing for a fair comparison, and due to the availability of their implementation. SpEx+[3] is a complete time-domain SE-A method that comprises a multi-scale speech encoder and decoder. Following the notations in (Ge et al., 2020), the hyperparameters of SpEx+ were set as follows: $L_1 = 40$ (2.5 ms), $L_2 = 160$ (10 ms), $L_3 = 320$ (20 ms), $N = 256$, $B = 8$, $R = 4$, $O = 256$, $P = 512$, $Q = 3$, $N_R = 3$, $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0.5$, and the speaker embedding dimension was set to 256. The number of trainable parameters of the SpEx + model is 11.3M (for TCD-TIMIT and WSJ0Mix) and 11.5M (for LRS3 and LibriMix)[4].

As a baseline for SE-V, the USEV[5] method (Pan et al., 2022) was used. In general, both the USEV baseline and the SE-V system used in this study share a similar structure. One key difference lies in the type of architecture used for the DNN block $\mathcal{F}$ in the auxiliary network, where USEV employs a temporal convolutional network (TCN) instead of a DPRNN. Following the notations in (Pan et al., 2022), the hyperparameters of USEV were set as follows: $L = 40$ (2.5 ms), $B = 64$, $N = 256$, $R = 6$, $H = 128$, $K = 100$, and 5 repeated TCN blocks were used in the auxiliary network. The USEV model has a total of 4.1M trainable parameters.

3.4 Training setup

For each of the SS, SE-A, and SE-V systems, as well as the baselines, we trained two independent models per dataset: one on 2-speaker mixtures and another on 3-speaker mixtures. Note that the video-based extraction systems (i.e., SE-V & USEV (Pan et al., 2022)) were not trained on WSJ0Mix and LibriMix due to the lack of visual recordings in such datasets. Adam (Kingma and Ba, 2015) optimizer was used with an initial learning rate of 10^{-3} and a weight decay of 10^{-5}. The batch size was set to 24, and gradients were clipped if their L_2 norm exceeded a value of 5. The maximum number of epochs was set to 300. A scheduler was utilized to reduce the learning rate by a factor of two if no reduction in the validation loss occurred in 10 consecutive epochs, and early stopping was used with patience of 20 epochs. A common seed was set for the generator of each dataset to ensure that identical training examples were provided to all systems during training.

For the TCD-TIMIT and LRS3 datasets, we trained on 3-s segments, whereas 4-s segments were used for WSJ0Mix and

[2] The weights of the visual encoder is obtained from: https://github.com/smeetrs/deep_avsr

[3] Official implementation provided Online: https://github.com/xuchenglin28/speaker_extraction_SpEx

[4] The difference in the number of trainable parameters across the different datasets is due to the linear layer used for the speaker identification loss, which depends on the number of speakers in the respective training split.

[5] Official implementation provided Online: https://github.com/zexupan/USEV

LibriMix. This also holds for the length of the reference signal for the SE-A systems. Dynamic mixing was not applied during training. However, the enrollment utterances for SE-A were sampled randomly across the different epochs. As the loss function ℓ, the negative scale-invariant source-to-distortion ratio (SI-SDR) (Le Roux et al., 2019) was used.

3.5 SS evaluation in SE setup

Evaluating a SS system in a SE setup requires an identification step for selecting the target speaker. Following prior works (Wang et al., 2019; Zmolikova et al., 2019; Xu et al., 2020; Delcroix et al., 2020), we compare two variants for target speaker selection in SS: oracle selection (SS + Oracle) and speaker verification selection (SS + SV).

For oracle selection, the target speaker is selected by computing the SI-SDR between each output of the SS system and the ground-truth target signal. The signal that yields the maximum value is selected as the target estimate. Oracle selection excludes identification errors in SS, and thus provide an upper bound on the performance of SS when applied in a SE setup.

In contrast, selecting that target speaker in SS using a speaker verification (SV) module does not exclude identification errors, which gives some perspective on the practical performance of the SS system when applied in a SE setup. The SV module accepts two utterances as input and computes a similarity score between them. It consists of a speaker embedding network that extracts an embedding vector for each utterance. The similarity is then determined by the cosine distance between the two embeddings. Evaluating target speaker selection for SS using SV is carried out by computing a similarity score between the enrollment utterance and each output of the SS system. The output signal that yields the maximum similarity score is selected as the estimated target speaker. For the speaker embedding network, we employed the ECAPA-TDNN (Desplanques et al., 2020), which is pre-trained on the VoxCeleb 1 + 2 datasets (Nagrani et al., 2020) comprising utterances from over seven thousand speakers. We used the implementation provided in SpeechBrain[6] (Ravanelli et al., 2021).

To validate the generalizability of the pre-trained ECAPA-TDNN model to the datasets considered in this study, we evaluated its SV performance on each dataset using 3,000 positive and negative pairs from the clean signals in the respective test split. The resulting equal error rates are as follows: 0.4%, 2.1%, 0.5%, and 1.3%, for TCD-TIMIT, LRS3, WSJ0Mix, and LibriMix, respectively. This evaluation confirms that the pre-trained ECAPA-TDNN model generalizes well to the datasets used in this study.

4 Experimental results

The goal of this study is to gain a better understanding of the role of auxiliary information in SE systems. In the first set of experiments, we investigate whether and in which scenarios the auxiliary information improves the extraction performance compared to uninformed SS systems. The second set of experiments explores the behavior of SE systems when provided with different or distorted auxiliary information for a given mixture signal. As evaluation metrics, we use the SI-SDR (Le Roux et al., 2019) to assess speech quality and the extended short-time objective intelligibility (ESTOI) (Taal et al., 2010) to measure speech intelligibility. Audio examples are available online.[7]

4.1 Performance on fully overlapped mixtures

In this experiment, we evaluate the extraction performance of the different systems on fully overlapped 2-speaker and 3-speaker mixtures. We first compare the SE-A and SE-V systems with the baselines described in Section 3.3. Subsequently, a comparison between the SE systems and SS is provided. The mean results in terms of the SI-SDR improvement (ΔSI-SDR) are presented in Table 2. We also report the ESTOI scores, which generally follow the trends observed in the SI-SDR scores.

4.1.1 Comparison with SE baselines

By comparing the SE-A system to SpEx+ (Ge et al., 2020), we observe no clear trend with respect to the superiority of either system across the different datasets. Furthermore, the performance of both SE-V and USEV (Pan et al., 2022) is generally comparable, except in the TCD-TIMIT datasets for 3-speaker mixtures, where SE-V clearly outperforms USEV. These results affirm that the adopted SE-A and SE-V systems are, to a certain extent, competitive with existing SE methods in the literature.

4.1.2 Comparison with SS

We further compare the performance of the SE systems with SS + SV and SS + Oracle. We first observe that SS + SV generally yields comparable results to SS + Oracle for 2-speaker mixtures. Conversely, for 3-speaker mixtures, SS + SV generally exhibits worse mean performance than SS + Oracle. This can be attributed to the lower separation scores for 3-speaker mixtures and the fact that the SV module is pre-trained on clean speech signals rather than on separated signals that might have processing artifacts or residuals from the interfering speakers.

Comparing SE-A with SS + Oracle, it can be seen that SS + Oracle achieves comparable or better scores across the different datasets. This also holds for SS + SV, except for LRS3 on 3-speaker mixtures, where a 1.4 dB drop in performance can be seen compared to SE-A. By examining the results of SE-V and SS + Oracle, we observe comparable performance, except for LRS3 on 3-speaker mixtures, where SE-V achieves better mean scores than SS + Oracle. However, the gap in performance increases when SV is used instead of oracle selection for SS. The results also clearly demonstrate that the SE-V system exhibits better mean performance than SE-A. This could be due to the prominent correlation between the visual cues and the target

[6] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[7] https://www.audiolabs-erlangen.de/resources/2024-New-Insights-on-Target-Speaker-Extraction

TABLE 2 Extraction performance for fully-overlapped 2-speaker and 3-speaker mixtures.

		TCD-TIMIT		LRS3		WSJ0Mix		LibriMix	
		2	3	2	3	2	3	2	3
SpEx+ (Ge et al., 2020)	ΔSI-SDR	15.9	10.6	12.8	11.2	16.1	12.2	14.4	11.7
	ESTOI	81.6	59.4	77.3	62.2	91.3	78.3	85.4	70.3
USEV (Pan et al., 2022)	ΔSI-SDR	18.4	14.4	14.7	13.5	—	—	—	—
	ESTOI	86.2	69.0	80.8	69.5	—	—	—	—
SE-A	ΔSI-SDR	15.5	10.0	13.7	12.0	15.2	13.8	14.7	12.7
	ESTOI	81.3	60.1	79.1	65.0	89.1	81.4	85.2	73.0
SE-V	ΔSI-SDR	18.4	16.5	14.4	13.2	—	—	—	—
	ESTOI	86.1	74.7	80.4	68.5	—	—	—	—
SS + SV	ΔSI-SDR	17.9	15.4	13.7	10.6	17.0	13.8	16.4	12.9
	ESTOI	85.0	70.2	79.2	61.6	92.8	80.9	88.6	72.6
SS + Oracle	ΔSI-SDR	18.0	16.1	14.2	11.9	17.0	14.0	16.5	13.7
	ESTOI	85.2	70.9	79.8	63.4	92.8	81.2	88.8	74.0

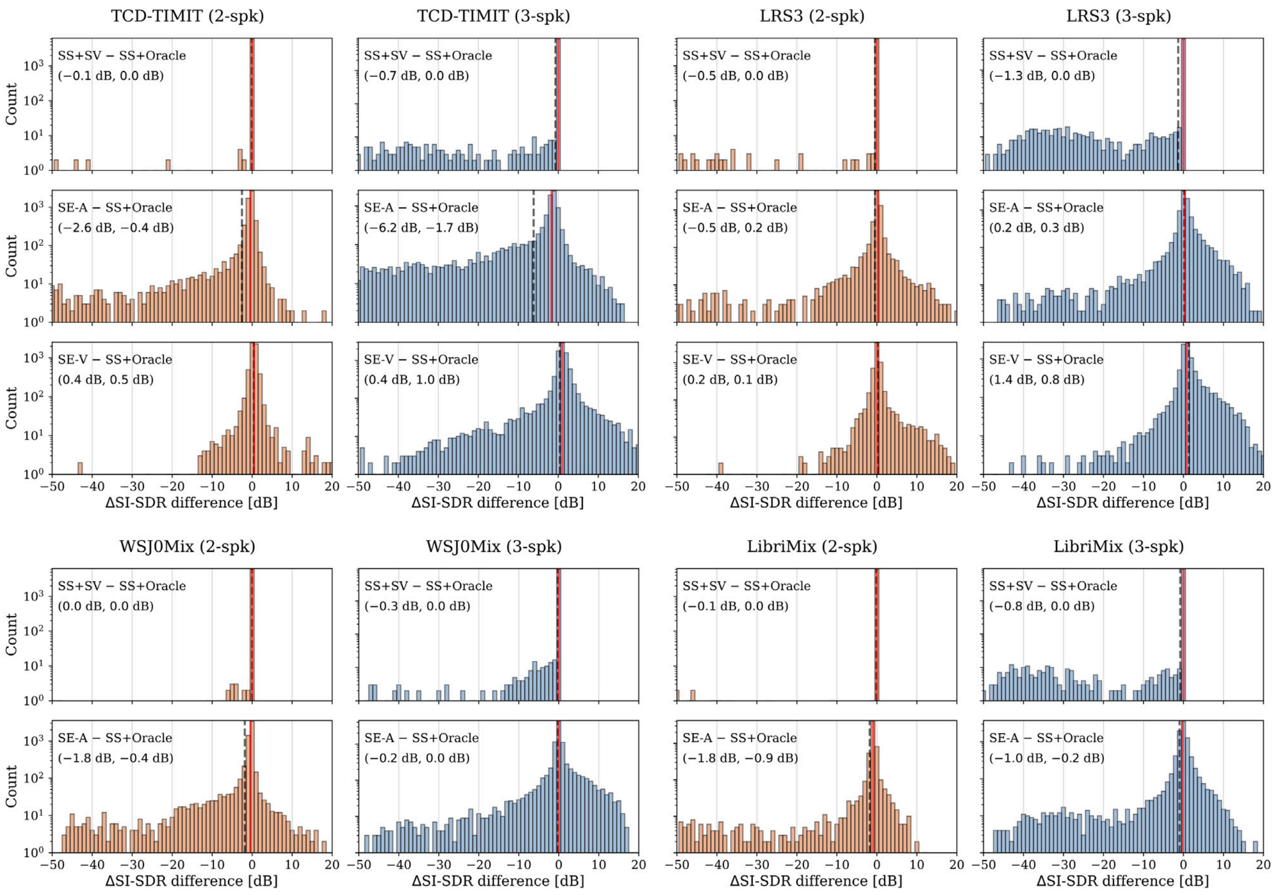

FIGURE 2
Histogram of the SI-SDR [dB] score difference between the different systems and SS + Oracle for 2-speaker (orange) and 3-speaker (blue) mixtures. The mean (dashed line, black) and median (solid line, red) are also visualized. The tuples provide the values of the mean and median, respectively.

signal in the mixture, as well as the use of time-varying embeddings in SE-V.

4.2 Per-sample performance analysis

The comparison in Section 4.1 provides a holistic view of the systems' performance, quantified by the arithmetic mean of the evaluation metrics over all samples in the test set. To gain more insights, we perform a per-sample analysis by inspecting the difference in the ΔSI-SDR performance between the different systems and SS + Oracle. A negative difference indicates that the SS + Oracle system is better than the respective system, and *vice versa*. It is important to note that this analysis differs from those in previous studies [e.g., (Ochiai et al., 2019a)], which compared histograms of performance scores computed for each system individually, rather than examining the performance differences between a pair of systems. The histograms of the ΔSI-SDR difference are shown in Figure 2. The mean and median values of the ΔSI-SDR difference are also provided. Interestingly, in many cases, the mean and median scores are quite different, which clearly shows that solely reporting the mean performance does not provide a full picture when comparing SE and SS systems.

Comparing SS + SV with SS + Oracle, it can be observed that the median is always centered around 0 dB. In contrast, the mean deviates to the negative side, especially for 3-speaker mixtures, where the SV module sometimes selects an interferer speaker instead of the target. When comparing SE-A with SS + Oracle, we observe that, in most cases, the median is close to 0 dB, except for LibriMix on 2-speaker mixtures and TCD-TIMIT on 3-speaker mixtures, where clearly the SE-A system performs poorly compared to SS + Oracle. It is also important to note how susceptible the mean performance of SE-A is to outliers, reflected by the gap between the mean and median values. The distributions of the performance difference between SE-V and SS + Oracle for 2-speaker mixtures exhibit a slight shift towards the positive side for TCD-TIMIT, whereas it is centered around zero for LRS3. However, for 3-speaker mixtures, the shift towards the positive side is more prominent, indicating an overall advantage of the visual information in this case.

The analysis here highlights the inadequacy of reporting only the mean scores over the samples when attempting to compare SS and SE systems due to the presence of outliers, e.g., caused by the incorrect selection of the target speaker. By excluding such outliers and considering the median values, the following conclusions can be drawn. The auxiliary information in SE-A does not consistently improve the quality of the extracted signals compared to SS, neither for 2-speaker mixtures nor for 3-speaker mixtures. To some extent, this is also the case for the auxiliary information in SE-V for 2-speaker mixtures. However, for 3-speaker mixtures, the visual information provides an overall improvement compared to SS, indicated by the shift of the distributions towards the positive side in Figure 2.

4.3 Effect of input SIR

In this experiment, we specifically study the impact of the input SIR on the performance of the different systems for 2-speaker and 3-speaker mixtures. This examines whether the auxiliary information in SE-A and SE-V improves the extraction performance compared to SS for different powers of the interfering signal(s), especially at low SIRs. For evaluation, 1,000 examples (target + interferer(s)) were selected from the test split of each dataset and mixed with a SIR swept from −10 dB to 20 dB with a step size of 10 dB. Figure 3 shows the results of this experiment, where we report the SI-SDR instead of ΔSI-SDR to better reflect the reconstruction quality of the extracted signals. As expected, the SI-SDR scores generally drop as the SIR decreases. Furthermore, it can be seen that SS + SV is comparable to SS + Oracle for higher SIRs (i.e., ≥ 10 dB). However, as the SIR decreases, the SS + SV system generally exhibits worse mean performance, especially for 3-speaker mixtures. Nonetheless, it is important to note that the median values of SS + SV and SS + Oracle are still close to each other, highlighting again the influence of the outliers on the mean values.

For 2-speaker mixtures, it can be seen that SE-A generally performs worse than SS + SV and SS + Oracle in terms of both mean and median values for all SIRs. Comparing SE-V with the SS systems, we observe comparable mean and median values, except for LRS3, where the mean of the SS + SV system is lower than SE-V and SS + Oracle. The results for 3-speaker mixtures follow different trends than the 2-speaker case. With the exception of TCD-TIMIT, it can be observed that SE-A yields better mean and median scores than SS + SV (and SS + Oracle for LRS3 and WSJ0Mix) for SIR = −10 dB. However, for higher SIRs, both SS + SV and SS + Oracle generally outperform SE-A across the datasets. We can also observe that the SE-A system exhibits poor generalization to unseen SIRs for TCD-TIMIT and WSJ0Mix. It is clear that SE-V generally exhibits the best performance in terms of mean and median scores for SIR ≤ 0 dB.

The results of this experiment show that the use of auxiliary information in SE does not always lead to better performance compared to SS for different SIRs. The only exception is for 3-speaker mixtures at low SIRs, where SE-A and SE-V generally exhibit better performance than the SS systems.

4.4 Long sequences

Thus far, only fully overlapped mixture signals were considered. In this experiment, we evaluate the effectiveness of the auxiliary information in SE for 2-speaker mixtures with different overlapping patterns and longer durations (up to 9 s), which mimic what typically occurs in natural conversations. Figure 4 depicts the different overlapping patterns considered in this experiment. For each dataset and each pattern, we generated 1,000 mixture signals with a SIR of 0 dB. For the SE-V system, the duration of the reference signal was always equal to the duration of the mixture signal. In segments where the target speaker was absent, we used static visual frames by repeating the starting frame of the silent segment. For SS + Oracle and SS + SV, target speaker selection was performed on the whole output waveforms rather than over the individual segments. Figure 5 shows the results of this experiment in terms of SI-SDR.

Consistent with the observations from the previous results on 2-speaker mixtures, both SS + Oracle and SS + SV attain comparable mean and median performance across the different patterns and datasets. Interestingly, the high scores of the SS systems indicate that they are able to track the speakers over time, even though there could be

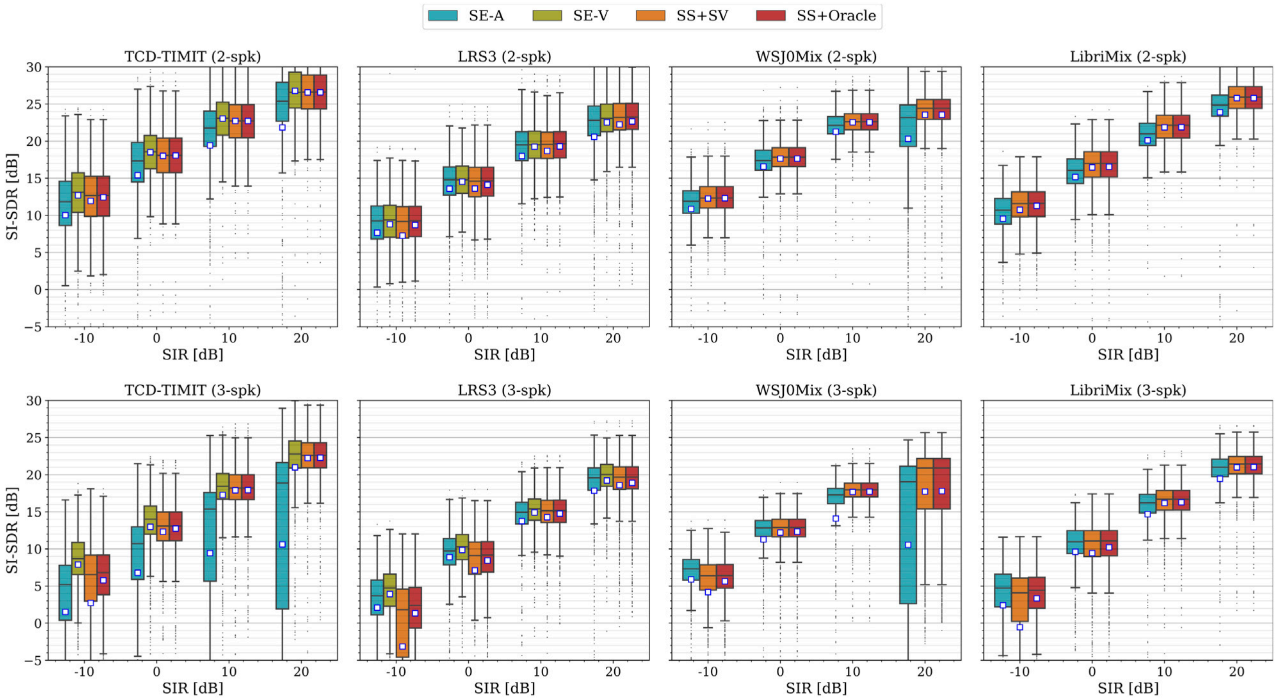

FIGURE 3
Extraction performance in terms of SI-SDR [dB] for different signal-to-interference ratios (SIRs) for 2-speaker (top) and 3-speaker (bottom) mixtures. The mean values are visualized by the square symbol.

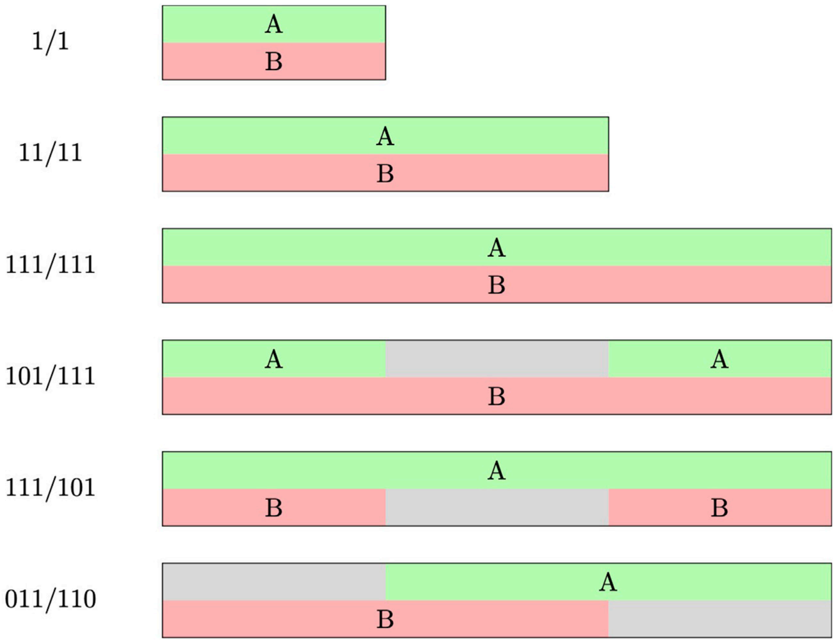

FIGURE 4
Different overlapping patterns for 2-speaker mixtures. The green and red regions represent the target speaker A and the interferer speaker B (red), respectively, while gray regions indicate silence. Each digit (1 or 0) within a pattern represents a 3-s segment, where 1 denotes a speech segment, whereas 0 represents silence. The last three patterns {101/111, 111/101, 011/110} also include their cyclic shifts by 3 s.

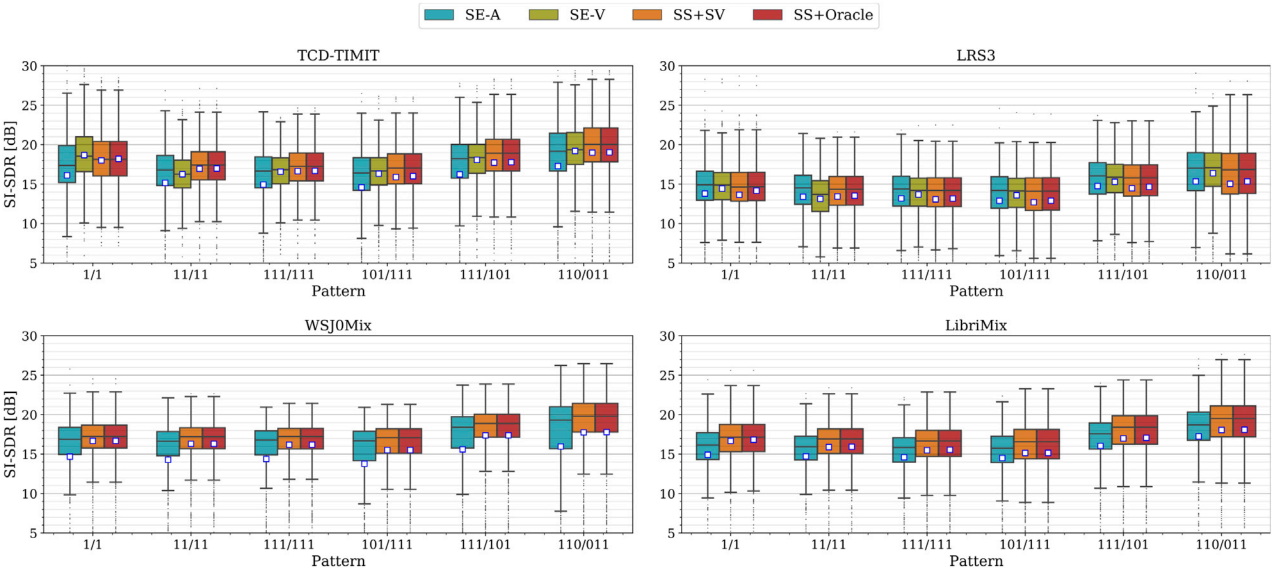

FIGURE 5
Extraction performance in terms of SI-SDR [dB] for the different overlapping patterns for 2-speaker mixtures. The mean values are visualized by the square symbol.

pauses in the streams of either the target or interferer, e.g., patterns 101/111, 111/101, and 110/011. This behavior could be attributed to the recurrent structure of the DPRNN architecture as well as the non-causality of the SS systems. From the results, it can be observed that SE-A achieves similar performance to SS + SV and SS + Oracle for LRS3, whereas it is generally worse for the other datasets. Note again how different the median and mean values are for SE-A, indicating the presence of many outliers where the SE-A system performs poorly. A comparison between SE-V and SS + SV/Oracle shows that the SE-V system generally achieves slightly higher mean values. However, the median scores of SE-V are mostly close to those of the SS systems, or in some cases, even lower, especially for TCD-TIMIT. This again emphasizes the importance of reporting additional statistical measures alongside the mean values when comparing SE and SS systems, to avoid drawing misleading conclusions. The findings presented here show that the use of auxiliary information in SE does not consistently lead to better performance than SS for the considered overlapping patterns for 2-speaker mixtures.

4.5 Performance for different reference signals

We further investigate the role of auxiliary information in SE by evaluating the extraction performance for different reference signals from the target speaker, given the same input mixture. This experiment addresses the question of whether different reference signals yield equivalent performance, or if some reference signals lead to better outcomes than others. The experiment was conducted only for SE-A, where multiple enrollment utterances from the target speaker are available. For each dataset, evaluation was performed on 1,000 2-speaker mixtures with a SIR of 0 dB. For each mixture, we evaluated the extraction performance of SE-A using five different randomly selected reference signals and then computed the standard deviation of the resulting ΔSI-SDR scores. In Figure 6, the histogram of the standard deviations is depicted for each dataset. Interestingly, across the various datasets, the standard deviation is less than 0.5 dB in more than 80% of the cases, demonstrating that the extraction performance for the five different reference signals is generally similar. These results suggest that the performance of SE-A for a given mixture is primarily determined by the capability of its extraction network, rather than by the specific enrollment utterance used, assuming that such utterances possess ‘sufficient’ discriminative information.

4.6 Distorted auxiliary information

Inpired by prior studies (Afouras et al., 2019; Sato et al., 2021) that address corrupted auxiliary information in SE, we analyze the behavior of the SE-A and SE-V systems when provided with distorted auxiliary information. Although no distortions were introduced during training, this experiment offers interesting insights into how SE systems leverage the auxilairy information in selecting the target speaker. We explore distortions in the embedding space of the auxiliary signals by linearly interpolating between two embeddings belonging to different speakers (indexed by A and B), i.e., $\boldsymbol{E}_{\text{interpolated}} = \alpha \quad \boldsymbol{E}_A + (1-\alpha) \quad \boldsymbol{E}_B$, for $\alpha \in [0, 1]$. In one case, we interpolate between embeddings from two in-mixture speakers, i.e., $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$ belong to two different speakers in the mixture. In the other case, we consider $\boldsymbol{E}_A$ to be from an in-mixture speaker, while $\boldsymbol{E}_B$ belongs to an out-of-mixture speaker. Note that when $\alpha = 0$ for the latter case, it is equivalent to the challenging inactive target speaker scenario, studied in (Borsdorf et al., 2021; Delcroix et al., 2022).

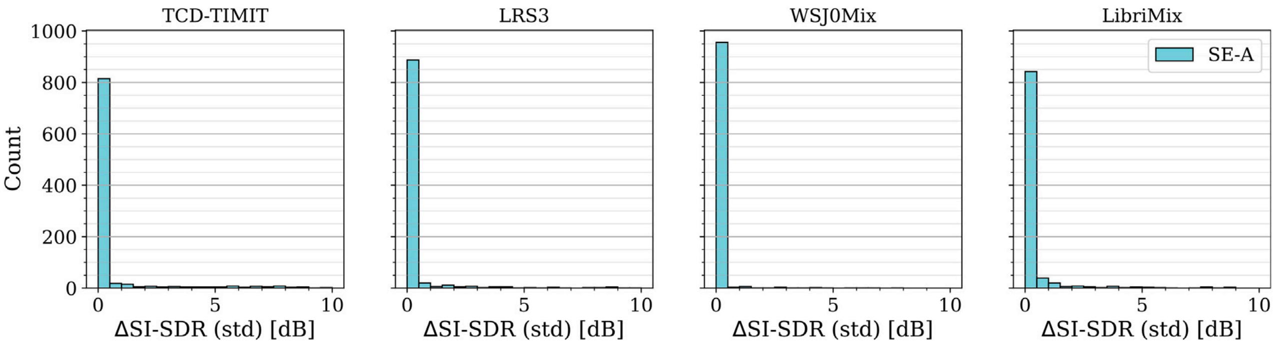

FIGURE 6
Histogram of the standard deviations in terms of ΔSI-SDR [dB] for the SE-A system. For each input mixture, the standard deviation of the extraction performance is computed for five different randomly selected reference signals belonging to the target speaker.

For evaluation, we considered two subsets, one consisting of 2-speaker mixtures and the other of 3-speaker mixtures, each containing 1,000 examples with a SIR of 0 dB. For both SE-A and SE-V systems, evaluation was performed in a matched condition utilizing the models trained on mixtures having the same number of speakers. We evaluate each output signal in terms of ΔSI-SDR with respect to the ground-truth signals of all in-mixture speakers to determine which speaker was extracted. In addition, we report the difference in system performance when provided with the interpolated embedding versus the undistorted embedding of each in-mixture speaker. A performance difference close to zero indicates that the interpolated embedding enables extracting an in-mixture speaker with a reconstruction quality close to the respective undistorted embedding. The histograms of the ΔSI-SDR scores are provided in Figure 7. For brevity, we only show the results for the LRS3 dataset, as the outcomes for the other datasets exhibit a similar pattern.

4.6.1 Two-speaker mixtures

Figure 7A shows the results for interpolating between two in-mixture speakers for mixtures with two speakers. It can be seen that the closer the interpolated embedding is to either speaker embedding, the more likely the corresponding speaker is extracted. Except for $\alpha = 0.5$, it is interesting to observe from the histogram of differences that the performance of both interpolated and undistorted embeddings is, in most cases, close to each other. For $\alpha = 0.5$, in at least 50% of the cases, both SE-A and SE-V systems still equally likely extract one of the speakers in the mixture, maintaining the same quality as when the undistorted embedding is provided.

When an embedding from an out-of-mixture speaker is interpolated with an in-mixture speaker (i.e., spk-1), it can be seen in Figure 7B that for higher values of α, the considered in-mixture speaker is more likely to be extracted compared to the other speaker in the mixture. Interestingly, when an utterance from an out-of-mixture speaker is used as a reference signal for SE-A (i.e., $\alpha = 0$), the system tends to extract one of the speakers in the mixture. This behavior has also been observed in previous studies (Borsdorf et al., 2021; Delcroix et al., 2022). However, we also demonstrate that this behavior holds for the SE-V system at $\alpha = 0$, although the reference signal (i.e., lip frames) corresponds to a different utterance spoken by a speaker not present in the mixture.

4.6.2 Three-speaker mixtures

For 3-speaker mixtures, the results for interpolating between in-mixture speakers are omitted for brevity, as they are consistent with those shown in Figure 7A. However, when interpolation is performed between an in-mixture speaker and an out-of-mixture speaker, it can be seen in Figure 7C that the SE systems generally do not extract any speaker in the mixture as α decreases. This behavior contrasts with systems trained on 2-speaker mixtures (as shown in Figure 7B), which tend to extract one of the speakers in the mixture, irrespective of the provided auxiliary information. These findings suggest that the training data composition, particularly the number of speakers in the mixtures, significantly influences the speaker selection mechanism in SE systems. Additionally, the results imply that training a SE system on both 2-speaker and 3-speaker mixtures may offer a potential strategy to mitigate the issue of speaker confusion.

5 Discussion

In this study, we examined the role of auxiliary information in two reference-based SE systems, namely audio-based SE (SE-A) and video-based SE (SE-V). In the first set of experiments (i.e., Sections 4.1–4.4), we compared the extraction performance of both SE systems to uninformed speaker separation (SS) systems evaluated in a SE setup. The comparison was carried out for 2-speaker and 3-speaker mixtures across various datasets and multiple input conditions. We demonstrated that the use of auxiliary information in the SE systems does not always result in better extraction performance than SS. However, as an exception, we found that the auxiliary information in SE generally leads to a performance improvement compared to SS for 3-speaker mixtures under low SIRs.

In the second set of experiments (i.e., Sections 4.5 and 4.6), we inspected the behavior of SE for different samples from the embedding space of the auxiliary information. We first showed that the performance of SE-A for a given mixture does not generally vary when provided with different enrollment signals from the target speaker with sufficient discriminative information. This indicates that the SE-A performance is mainly determined by how capable its

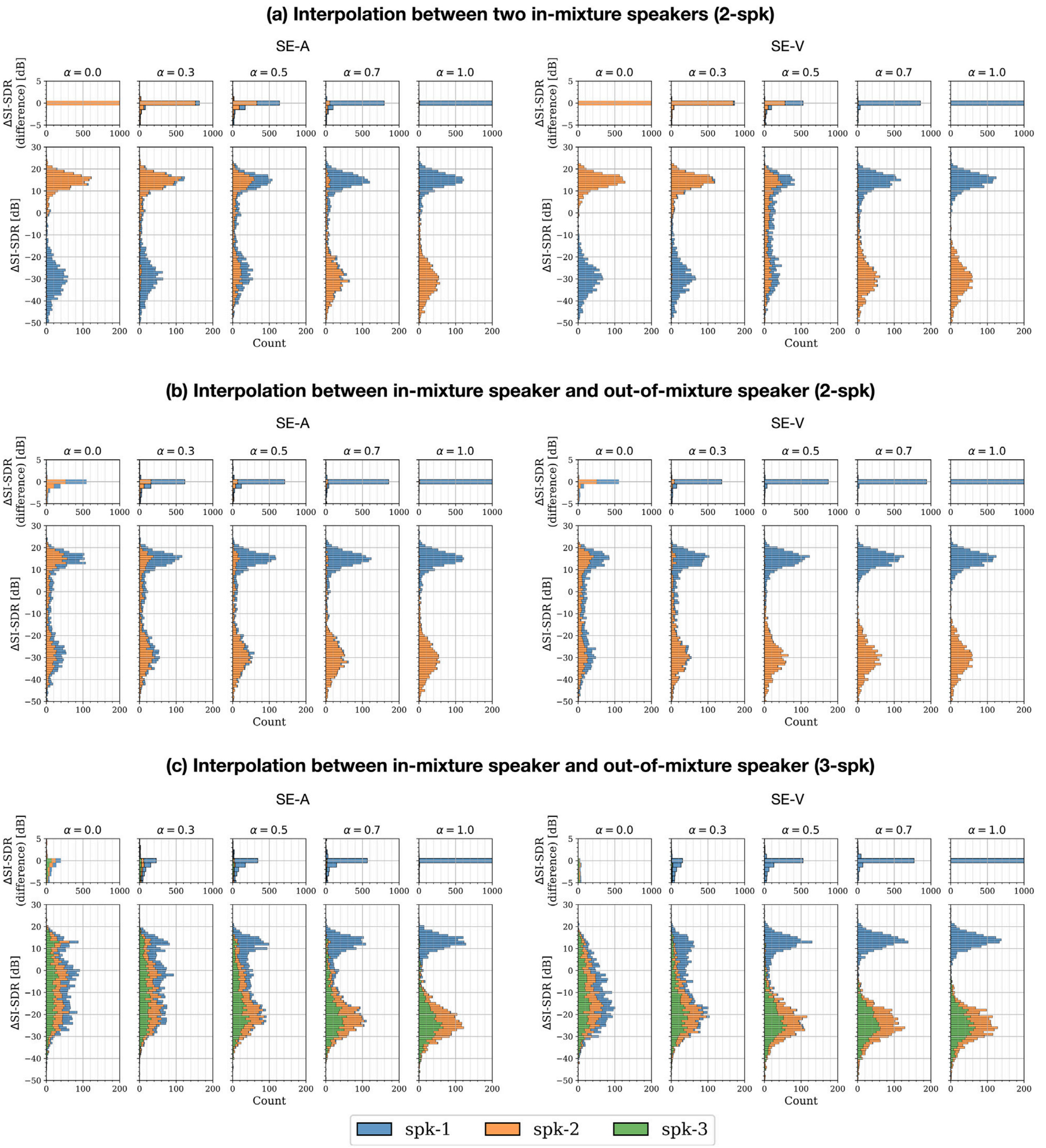

FIGURE 7
Performance for distorted auxiliary information via interpolation in the embedding space. **(A)** Interpolation between embeddings from two in-mixture speakers for 2-speaker mixtures; **(B)** and **(C)** interpolation between embeddings from an in-mixture and out-of-mixture speakers for 2-speaker and 3-speaker mixtures, respectively. For each value of the interpolation coefficient α, stacked histograms are provided: (bottom) histogram of ΔSI-SDR scores [dB] for interpolated embeddings, (top) histogram of the difference in performance between the interpolated embedding and the undistorted embedding. The output of the SE systems is evaluated with respect to all speakers in the mixture: spk-1 (blue), spk-2 (orange), and spk-3 (green).

extraction network is, rather than by the specific enrollment signal used. In addition, we evaluated the performance of SE systems when provided with distorted auxiliary information by interpolating embeddings either from two in-mixture speakers or from an in-mixture speaker and an out-of-mixture speaker. We showed that an interpolated embedding between two in-mixture speakers generally leads to extracting either one of the speakers, which is closest to the interpolated embedding.

Furthermore, the results for interpolating between an in-mixture and out-of-mixture speakers highlight the difference between SE systems trained on 2-speaker mixtures and those trained on 3-speaker mixtures in the way the auxiliary information is leveraged to select the target speaker. Particularly, when using an embedding from an out-of-mixture speaker, SE systems trained on 2-speaker mixtures tend to consistently extract one of the speaker in the mixture, whereas those trained on 3-speaker mixtures generally do not extract any in-mixture speaker. This suggests that training SE systems on both 2-speaker and 3-speaker mixtures could help mitigate the speaker confusion issue, typically occurring in scenarios with an inactive target speaker.

While this study provides valuable insights into the role of auxiliary information in SE, it is crucial to acknowledge its limitations to avoid over-generalization of the results. On the system level, we considered the DPRNN architecture as the main learning machine for all systems, although several newer architectures (Subakan et al., 2021; Wang et al., 2023) have been proposed and demonstrated better separation performance. Therefore, whether the conclusions reached in this study also extend to these architectures is yet to be validated. In addition, we considered only two forms of auxiliary information for SE, i.e., enrollment utterances and visual information. It would be interesting to extend this study to other forms of auxiliary information and possibly also their combinations, i.e., multi-modal SE. Finally, it is important to note that the list of input mixture scenarios covered in this study is by no means exhaustive. For example, neither non-speech interferers nor reverberation were considered. Exploring these and other more complex scenarios is left for future work.

This study highlights several key aspects that should be considered in future work on SE. Firstly, reporting only mean performance may not adequately capture the true behavior of SE systems due to the influence of outliers, mainly caused by the speaker confusion problem. We recommend including additional metrics, such as the median, and visualizing score distributions to provide a more comprehensive understanding of system performance. Furthermore, our findings suggest that existing SE systems do not fully exploit the potential of auxiliary information, which could intuitively lead to significantly improved performance compared to SS. Thus, future research should focus on developing new techniques to better leverage the auxiliary information in SE. Finally, further investigation is required to address the challenges of speaker confusion and inactive target speakers in order to improve the robustness of SE systems in practical scenarios.

Data availability statement

Publicly available datasets were analyzed in this study. More details can be found here: https://www.audiolabs-erlangen.de/resources/2024-New-Insights-on-Target-Speaker-Extraction/.

Author contributions

ME: Methodology, Writing-original draft. WM: Writing-review and editing. SRC: Supervision, Writing-review and editing. SoC: Writing-review and editing. EH: Supervision, Writing-review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Parts of this work have been funded by the Free State of Bavaria in the DSAI project.

Acknowledgments

The authors thank the Erlangen Regional Computing Center (RRZE) for providing compute resources and technical support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afouras, T., Chung, J. S., and Zisserman, A. (2018a). "The conversation: deep audio-visual speech enhancement," in *Proc. Interspeech conf.*, 3244–3248.

Afouras, T., Chung, J. S., and Zisserman, A. (2018b). LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv Prepr. arXiv:1809.00496*. doi:10.48550/arXiv.1809.00496

Afouras, T., Chung, J. S., and Zisserman, A. (2019). "My lips are concealed: audio-visual speech enhancement through obstructions," in *Proc. Interspeech conf.*, 4295–4299.

Aldeneh, Z., Kumar, A. P., Theobald, B.-J., Marchi, E., Kajarekar, S., Naik, D., et al. (2021). "On the role of visual cues in audiovisual speech enhancement," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 8423–8427.

Borsdorf, M., Xu, C., Li, H., and Schultz, T. (2021). "Universal speaker extraction in the presence and absence of target speakers for speech of one and two talkers," in *Proc. Interspeech conf.*, 1469–1473.

Byun, J., and Shin, J. W. (2021). Monaural speech separation using speaker embedding from preliminary separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 2753–2763. doi:10.1109/taslp.2021.3101617

Ceolini, E., Hjortkjær, J., Wong, D. D., O'Sullivan, J., Raghavan, V. S., Herrero, J., et al. (2020). Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception. *NeuroImage* 223, 117282. doi:10.1016/j.neuroimage.2020.117282

Chen, J., Mao, Q., and Liu, D. (2020). "Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech conf.*, 2642–2646.

Chen, Z., Luo, Y., and Mesgarani, N. (2017). "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 246–250.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi:10.1121/1.1907229

Chuang, S.-Y., Tsao, Y., Lo, C.-C., and Wang, H.-M. (2020). "Lite audio-visual speech enhancement," in *Proc. Interspeech conf.*, 1131–1135.

Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., and Freeman, W. T. (2017). "Synthesizing normalized faces from facial identity features," in *Proc. IEEE/CVF conf. on computer vision and pattern recognition (CVPR)*, 3386–3395.

Cooke, M., Hershey, J. R., and Rennie, S. J. (2010). Monaural speech separation and recognition challenge. *Comput. Speech & Lang.* 24, 1–15. doi:10.1016/j.csl.2009.02.006

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. (2020). LibriMix: an open-source dataset for generalizable speech separation. *arXiv*. doi:10.48550/arXiv.2005.11262

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 19, 788–798. doi:10.1109/tasl.2010.2064307

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

Delcroix, M., Ochiai, T., Zmolikova, K., Kinoshita, K., Tawara, N., Nakatani, T., et al. (2020). "Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 691–695.

Delcroix, M., Zmolikova, K., Kinoshita, K., Ogawa, A., and Nakatani, T. (2018). "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 5554–5558.

Delcroix, M., Zmolikova, K., Ochiai, T., Kinoshita, K., and Nakatani, T. (2021). "Speaker activity driven neural speech extraction," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 6099–6103.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech conf.*, 3830–3834.

Du, J., Tu, Y., Dai, L.-R., and Lee, C.-H. (2016). A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 24, 1424–1437. doi:10.1109/taslp.2016.2558822

Du, J., Tu, Y., Xu, Y., Dai, L., and Lee, C.-H. (2014). "Speech separation of a target speaker based on deep neural networks," in *Proc. Intl. conf. on signal processing*, 473–477.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.* 37, 1–11. doi:10.1145/3197517.3201357

Gabbay, A., Shamir, A., and Peleg, S. (2018). "Visual speech enhancement," in *Proc. Interspeech conf.*, 1170–1174.

Garofolo, J., Graff, D., Baker, J. M., Paul, D., and Pallett, D. (1993). *CSR-I (WSJ0) complete*. Philadelphia: Linguistic Data Consortium. doi:10.35111/ewkm-cg47

Ge, M., Xu, C., Wang, L., Chng, E. S., Dang, J., and Li, H. (2020). "SpEx+: a complete time domain speaker extraction network," in *Proc. Interspeech conf.*, 1406–1410.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* 33, 1417–1426. doi:10.1523/JNEUROSCI.3675-12.2013

Harte, N., and Gillen, E. (2015). TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans. Multimed.* 17, 603–615. doi:10.1109/tmm.2015.2407694

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 31–35.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neur. Comp.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., and Wang, H.-M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* 2, 117–128. doi:10.1109/tetci.2017.2784878

Inan, B., Cernak, M., Grabner, H., Tukuljac, H. P., Pena, R. C., and Ricaud, B. (2019). "Evaluating audiovisual source separation in the context of video conferencing," in *Proc. Interspeech conf.*, 4579–4583.

Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech conf.*, 545–549.

King, D. E. (2009). Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* 10, 1755–1758. doi:10.5555/1577069.1755843

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Proc. IEEE intl. conf. on learn. repr.* (ICLR), 1–15.

Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 25, 1901–1913. doi:10.1109/TASLP.2017.2726762

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). "SDR–half-baked or well done?," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 626–630.

Luo, Y., Chen, Z., and Yoshioka, T. (2020). "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 46–50.

Luo, Y., and Mesgarani, N. (2018). "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 696–700.

Luo, Y., and Mesgarani, N. (2019). Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 27, 1256–1266. doi:10.1109/taslp.2019.2915167

Martin, A. F., and Przybocki, M. A. (2001). "Speaker recognition in a multi-speaker environment," in *Proc. European conf. on speech communication and technology (Eurospeech)*, 787–790.

Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D., et al. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 1368–1396. doi:10.1109/taslp.2021.3066303

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020). Voxceleb: large-scale speaker verification in the wild. *Comput. Speech & Lang.* 60, 101027. doi:10.1016/j.csl.2019.101027

Ochiai, T., Delcroix, M., Kinoshita, K., Ogawa, A., and Nakatani, T. (2019a). "Multimodal SpeakerBeam: single channel target speech extraction with audio-visual speaker clues," in *Proc. Interspeech conf.*, 2718–2722.

Ochiai, T., Delcroix, M., Kinoshita, K., Ogawa, A., and Nakatani, T. (2019b). "A unified framework for neural speech separation and extraction," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 6975–6979.

Owens, A., and Efros, A. A. (2018). "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. European conf. on computer vision (ECCV)*, 631–648.

Pan, Z., Ge, M., and Li, H. (2022). USEV: universal speaker extraction with visual cue. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 30, 3032–3045. doi:10.1109/taslp.2022.3205759

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 5206–5210.

Partan, S., and Marler, P. (1999). Communication goes multimodal. *Science* 283, 1272–1273. doi:10.1126/science.283.5406.1272

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., et al. (2021). A general-purpose speech toolkit. *arXiv*. arXiv:2106.04624.

Sato, H., Ochiai, T., Kinoshita, K., Delcroix, M., Nakatani, T., and Araki, S. (2021). "Multimodal attention fusion for target speaker extraction," in *IEEE spoken language technology workshop (SLT)*, 778–784.

Shetu, S. S., Chakrabarty, S., and Habets, E. A. P. (2021). "An empirical study of visual features for DNN based audio-visual speech enhancement in multi-talker environments," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 8418–8422.

Stafylakis, T., and Tzimiropoulos, G. (2017). "Combining residual networks with LSTMs for lipreading," in *Proc. Interspeech conf.*, 3652–3656.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). "Attention is all you need in speech separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 21–25.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 4214–4217.

Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). "Generalized end-to-end loss for speaker verification," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 4879–4883.

Wang, D., and Chen, J. (2018). Supervised speech separation based on deep learning: an overview. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 26, 1702–1726. doi:10.1109/taslp.2018.2842159

Wang, J., Chen, J., Su, D., Chen, L., Yu, M., Qian, Y., et al. (2018). "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. Interspeech conf.*, 307–311.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., et al. (2019). Voicefilter: targeted voice separation by speaker-conditioned spectrogram masking. *Proc. Interspeech conf.*, 2728–2732.

Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y., and Watanabe, S. (2023). "TF-GridNet: making time-frequency domain models great again for monaural speaker separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 1–5.

Wu, J., Xu, Y., Zhang, S.-X., Chen, L.-W., Yu, M., Xie, L., et al. (2019). "Time domain audio visual speech separation," in *Proc. IEEE workshop on automatic speech recognition and understanding*, 667–673.

Xu, C., Rao, W., Chng, E. S., and Li, H. (2020). SpEx: multi-scale time domain speaker extraction network. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 28, 1370–1384. doi:10.1109/taslp.2020.2987429

Yu, D., Kolbæk, M., Tan, Z. H., and Jensen, J. (2017). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)*, 241–245.

Zeghidour, N., and Grangier, D. (2021). Wavesplit: end-to-end speech separation by speaker clustering. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 2840–2849. doi:10.1109/taslp.2021.3099291

Zhao, Z., Yang, D., Gu, R., Zhang, H., and Zou, Y. (2022). "Target confusion in end-to-end speaker extraction: analysis and approaches," in *Proc. Interspeech conf.*, 5333–5337.

Zmolikova, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., et al. (2019). SpeakerBeam: speaker aware neural network for target speaker extraction in speech mixtures. *IEEE J. sel. Top. Sig. Proc.* 13, 800–814. doi:10.1109/jstsp.2019.2922820