Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu

©SHUTTERSTOCK.COM/ARTEMISDIANA

Neural Target Speech Extraction

An overview

Humans can listen to a target speaker even in challenging acoustic conditions that have noise, reverberation, and interfering speakers. This phenomenon is known as the *cocktail party effect*. For decades, researchers have focused on approaching the listening ability of humans. One critical issue is handling interfering speakers because the target and nontarget speech signals share similar characteristics, complicating their discrimination. Target speech/speaker extraction (TSE) isolates the speech signal of a target speaker from a mixture of several speakers, with or without noises and reverberations, using clues that identify the speaker in the mixture. Such clues might be a spatial clue indicating the direction of the target speaker, a video of the speaker's lips, and a prerecorded enrollment utterance from which the speaker's voice characteristics can be derived. TSE is an emerging field of research that has received increased attention in recent years because it offers a practical approach to the cocktail party problem and involves such aspects of signal processing as audio, visual, and array processing as well as deep learning. This article focuses on recent neural-based approaches and presents an in-depth overview of TSE. We guide readers through the different major approaches, emphasizing the similarities among frameworks and discussing potential future directions.

Digital Object Identifier 10.1109/MSP.2023.3240008
Date of current version: 1 May 2023

Introduction

In everyday life, we are constantly immersed in complex acoustic scenes consisting of multiple sounds, such as a mixture of speech signals from multiple speakers and background noise from air conditioners and music. Humans naturally extract relevant information from such noisy signals as they enter our ears. The cocktail party problem is a typical example [1], where we can follow the conversation of a speaker of interest (the target speaker) in a noisy room with multiple interfering speakers. Humans can manage this complex task due to selective attention, or a selective hearing mechanism, that allows us to focus our attention on a target speaker's voice and ignore others. Although the mechanisms of human selective hearing are not fully understood yet, many studies have identified essential cues exploited by humans to attend to a target speaker in a speech mixture: spatial, spectral (audio), visual, and semantic cues [1]. One long-lasting goal of speech processing research is designing machines that can achieve similar listening abilities as humans, i.e., selectively extracting the speech of a desired speaker, based on auxiliary cues.

In this article, we present an overview of recent developments in TSE, which estimates the speech signal of a target speaker in a mixture of several speakers, given auxiliary cues to identify the target. Alternative terms in the literature for TSE include *informed source separation*, *personalized speech enhancement*, and *audiovisual speech separation*, depending on the context and the modalities involved. In the following, we call auxiliary cues *clues* since they represent hints for identifying the target speaker in the mixture. Figure 1 illustrates the TSE problem and shows that by exploiting the clues, TSE can focus on the voice of the target speaker while ignoring other speakers and noise. Inspired by psychoacoustic studies [1], several clues have been explored to tackle the TSE problem, such as spatial clues that provide the direction of the target speaker [2], [3], visual clues from video of the speaker's face [4], [5], [6], [7], [8], [9], and audio clues extracted from a prerecorded enrollment recording of the speaker's voice [10], [11], [12].

The TSE problem is directly related to human selective hearing, although we approach it from an engineering point of view and do not try to precisely mimic human mechanisms. TSE is related to other speech and audio processing tasks, such as noise reduction and blind source separation (BSS), that do not use clues about the target speaker. Although noise reduction does suppress the background noise, it cannot well handle interfering speakers. BSS estimates each speech source signal in a mixture, which usually requires estimating the number of sources, a step that is often challenging. Moreover, it estimates the source signals without identifying them, which leads to global permutation ambiguity at its output; it remains ambiguous which of the estimated source signals corresponds to the target speaker. In contrast, TSE focuses on the target speaker's speech by exploiting clues without assuming knowledge of the number of speakers in the mixture and avoids global permutation ambiguity. It thus offers a practical alternative to noise reduction and BSS when the use case requires focusing on a desired speaker's voice.

Solving the TSE problem promises real implications for the development of many applications: 1) robust voice user interfaces and voice-controlled smart devices that respond only to a specific user, 2) teleconferencing systems that can remove interfering speakers close by, and 3) hearing aids/hearables that can emphasize the voice of a desired interlocutor.

TSE ideas can be traced back to early works on beamformers [2]. Several works also extended BSS approaches to exploit clues about the target speaker [4], [5], [12]. Most of these approaches required a microphone array [5] or models trained on a relatively large amount of speech data from the target speaker [4]. The introduction of neural networks (NNs) enabled the building of powerful models that learn to perform complex conditioning on various clues by leveraging large amounts of speech data of various speakers. This evolution resulted in impressive extraction performance. Moreover, neural TSE systems can operate with a single microphone and with speakers unseen during the training of the models, allowing more flexibility.

This overview article covers recent TSE development and focuses on neural approaches. Its remaining sections are organized as follows. In the "Problem Definition" section, we formalize the TSE problem and its relation to noise reduction and BSS and introduce its historical context. We then present a taxonomy of TSE approaches and motivate the focus of this overview article in the "TSE Taxonomy" section. We describe a general neural TSE framework in the "General Framework for Neural TSE" section. The later sections ("Audio-Based TSE," "Visual/Multimodal Clue-Based TSE," and "Spatial Clue-Based TSE") introduce implementations of TSE with different clues. We discuss extensions to other tasks in the "Extension to Other Tasks" section. Finally, we conclude by describing the outlook on remaining issues in the "Remaining Issues and Outlook" section and provide pointers to available resources for experimenting with TSE in the "Resources" section.

Problem definition

Speech recorded with a distant microphone

Imagine recording a target speaker's voice in a living room by using a microphone placed on a table. This scenario represents a typical use case of a voice-controlled smart device or a video conferencing device in a remote work situation. Many sounds may co-occur while the speaker is speaking, e.g., a vacuum cleaner, music, children screaming, voices from another conversation, and a TV. The speech signal captured at a microphone thus consists of a mixture of the target speaker's speech and interference from the speech of other speakers and background noise. In this article, we do not explicitly consider the effect of reverberation caused by the reflection of sounds on the walls and surfaces in a room, which also corrupt the recorded signal. Some of the approaches we discuss implicitly handle reverberation.

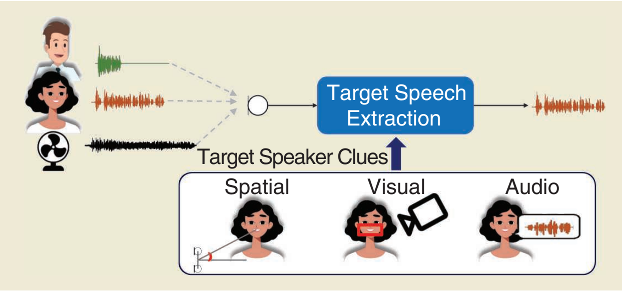


FIGURE 1. The TSE problem and examples of clues.

We can express the mixture signal recorded at a microphone as

$$\mathbf{y}^m = \mathbf{x}_s^m + \underbrace{\sum_{k\neq s} \mathbf{x}_k^m + \mathbf{v}^m}_{\triangleq \mathbf{i}^m} \tag{1}$$

where $\mathbf{y}^m = [y^m[0], \ldots, y^m[T]] \in \mathbb{R}^T$, $\mathbf{x}_s^m \in \mathbb{R}^T$, $\mathbf{x}_k^m \in \mathbb{R}^T$, and $\mathbf{v}^m \in \mathbb{R}^T$ are the time-domain signal of the mixture, the target speech, the interference speech, and noise signals, respectively. Variable T represents the duration (number of samples) of the signals, m is the index of the microphone in an array of microphones, s represents the index of the target speaker, and k is the index for the other speech sources. We drop microphone index m whenever we deal with single-microphone approaches. In the TSE problem, we are interested in recovering only the speech of the target speaker s, $\mathbf{x}_s^m$, and view all the other sources as undesired signals to be suppressed. We can thus define the interference signal as $\mathbf{i}^m \in \mathbb{R}^T$. Note that we make no explicit hypotheses about the number of interfering speakers.

TSE problem and its relation to BSS and noise reduction

The TSE problem is to estimate the target speech, given a clue, $\mathbf{C}_s$, as

$$\hat{\mathbf{x}}_s = \mathrm{TSE}(\mathbf{y}, \mathbf{C}_s; \theta^{\mathrm{TSE}}) \tag{2}$$

where $\hat{\mathbf{x}}_s$ is the estimate of the target speech and $\mathrm{TSE}(\cdot;\theta^{\mathrm{TSE}})$ represents a TSE system with parameters θ^{TSE}. The clue, $\mathbf{C}_s$, allows identifying the target speaker in the mixture. It can be of various types, such as a prerecorded enrollment utterance, $\mathbf{C}_s^{(a)}$; a video signal capturing the face and lip movements of the target speaker, $\mathbf{C}_s^{(v)}$; and such spatial information as the direction of arrival (DOA) of the speech of the target speaker, $\mathbf{C}_s^{(d)}$.

In the later sections, we expand on how to design TSE systems. Here, we first emphasize the key differences among TSE and BSS and noise reduction. Figure 2 compares these three problems.

BSS [13], [14] estimates all the source signals in a mixture without requiring clues:

$$\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_K\} = \mathrm{BSS}(\mathbf{y}; \theta^{\mathrm{BSS}}) \tag{3}$$

where $\mathrm{BSS}(\cdot;\theta^{\mathrm{BSS}})$ represents a separation system with parameters θ^{BSS}, $\hat{\mathbf{x}}_k$ are the estimates of the speech sources, and K is the number of sources in the mixture. As seen in (3), BSS does not and cannot differentiate the target speech from other speech sources. Therefore, we cannot know in advance which output corresponds to the target speech; i.e., there is a global permutation ambiguity problem between the outputs and the speakers. Besides, since the number of outputs is given by the number of sources, the number of sources K must be known or estimated. Comparing (2) and (3) emphasizes the fundamental difference between TSE and BSS: 1) TSE estimates only the target speech signal, while BSS estimates all the signals, and 2) TSE is conditioned on speaker clue $\mathbf{C}_s$, while BSS relies only on the observed mixture. Another setup sitting between TSE and BSS is a task that extracts multiple target speakers, e.g., extracting the speech of all the meeting attendees, given such information about them as enrollment and videos of all the speakers. Typical use cases for BSS include applications that require estimating speech signals of every speaker, such as automatic meeting transcription systems.

Noise reduction is another related problem. It assumes that the interference consists only of background noise, i.e., $\mathbf{i} = \mathbf{v}$, and can thus enhance the target speech without requiring clues:

$$\hat{\mathbf{x}}_s = \mathrm{Denoise}(\mathbf{y}; \theta^{\mathrm{Denoise}}) \tag{4}$$

where $\mathrm{Denoise}(\cdot;\theta^{\mathrm{Denoise}})$ represents a noise reduction system with parameters $\theta^{\mathrm{Denoise}}$. Unlike BSS, a noise reduction system's output consists only of target speech $\hat{\mathbf{x}}_s$, and there is thus no global permutation ambiguity. This is possible if the background noise and speech have distinct characteristics. For example, we can assume that ambient noise and speech signals exhibit different spectrotemporal characteristics that enable their discrimination. However, noise reduction cannot suppress interfering speakers because it cannot discriminate among different speakers in a mixture without clues. Some works propose to exploit clues for noise reduction and apply ideas similar to TSE to reduce background noise and, sometimes, interfering speakers. In the literature, this is called *personalized speech enhancement*, which, in this article, we view as a special case of the TSE problem, where only the target speaker is actively speaking [15]. Noise reduction is often used, e.g., in video conferencing systems and hearing aids.

TSE is an alternative to BSS and noise reduction, using a clue to simplify the problem. Like BSS, it can handle speech mixtures. Like noise reduction, it estimates only the target speaker, thus avoiding global permutation ambiguity and the need to estimate the number of sources. However, TSE requires access to clues, unlike BSS and noise reduction.

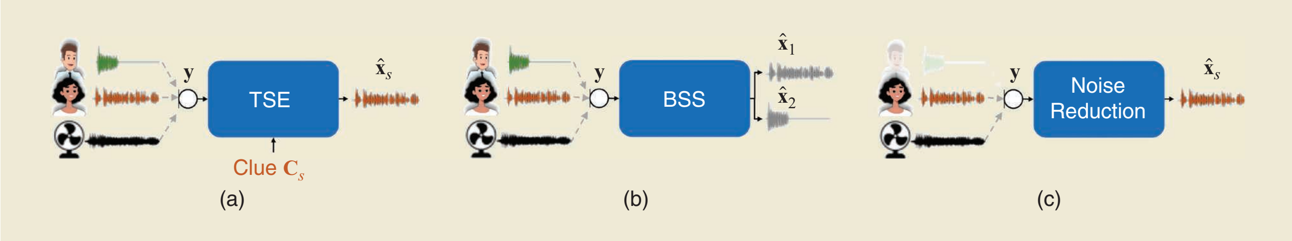


FIGURE 2. A comparison of (a) the TSE problem, (b) the BSS problem, and (c) the noise reduction problem.

Moreover, it must internally perform two subtasks: 1) *identifying the target speaker* and 2) *estimating the speech of that speaker in the mixture*. TSE is thus a challenging problem that introduces specific issues and requires dedicated solutions.

A straightforward way to achieve TSE using BSS methods is to first apply BSS and next select the target speaker among the estimated sources. Such a cascade system allows the separate development of BSS and speaker identification modules. However, this scheme is usually computationally more expensive and imports some disadvantages of BSS, such as the need to estimate the number of speakers in the mixture. Therefore, we focus on approaches that directly exploit the clues in the extraction process. Nevertheless, most TSE research is rooted in BSS, as argued in the following discussion on the historical context.

Historical context

The first studies related to TSE were performed in the 1980s. Flanagan et al. [2] explored enhancing a target speaker's voice in a speech mixture, assuming that the target speech originated from a fixed and known direction. They employed a microphone array to record speech and designed a fixed beamformer that enhanced the signals from the target direction [2], [16]. We consider that this work represents an early TSE system that relies on spatial clues.

In the mid-1990s, the BSS problem gained attention with pioneering works on independent component analysis (ICA). ICA estimates spatial filters that separate the sources by relying on the assumption of the independence of the sources in the mixture and the fact that speech signals are non-Gaussian [13]. A frequency-domain ICA suffers from a frequency permutation problem because it treats each frequency independently. In the mid-2000s, independent vector analysis (IVA) addressed the frequency permutation problem by working on vectors spanning all frequency bins, which allowed modeling dependency among frequencies [13]. Several works have extended ICA and IVA to perform TSE, which simplifies inference by focusing on a single target source. For example, in the late 2000s, TSE systems were designed by incorporating the voice activity information of the target speaker derived from video signals into the ICA criterion, allowing identification and extraction of only the target source [5]. In the late 2010s, independent vector extraction (IVE) extended IVA to extract a single source out of the mixture. In particular, IVE exploits clues to guide the extraction process, such as the enrollment of the target speaker, to achieve TSE [12]. All these approaches require a microphone array to capture speech.

In the first decade of the 2000s, single-channel approaches for BSS emerged, such as the factorial hidden Markov model (F-HMM) [17] and nonnegative matrix factorization (NMF) [18]. These approaches relied on pretrained spectral models of speech signals learned on clean speech data. An F-HMM is a model of speech mixtures, where the speech of each speaker in the mixture is explicitly modeled using a separate HMM. The parameters of each speaker HMM are learned on the clean speech data of that speaker. The separation process involves inferring the most likely HMM state sequence associated with each speaker HMM, which requires approximations to make inference tractable. This approach was the first to achieve superhuman performance using only single-channel speech [17]. In the early 2000s, the F-HMM was also among the first approaches to exploit visual clues [4]. This framework needs clues for all the speakers, a requirement that negates some of the advantages of TSE; e.g., the number of speakers must be known beforehand. Despite that, the method does not suffer from global permutation ambiguity since visual clues identify the target speaker, and we thus include this work in the broader view of TSE methods. In NMF, the spectrogram of each source is modeled as a multiplication of prelearned bases, representing the basic spectral patterns and their time-varying activations. NMF methods have also been extended to multichannel signals [13] and used to extract a target speaker [19] by using a flexible multisource model of the background. The main shortcoming of the F-HMM and NMF methods is that they require pretrained source models and thus struggle with unseen speakers. Furthermore, the inference employs a computationally expensive iterative optimization.

In the mid-2010s, deep NNs (DNNs) were first introduced to address the BSS problem. These approaches rapidly gained attention with the success of deep clustering and permutation invariant training [20], [21], which showed that single-channel speaker-open BSS was possible, i.e., separation of unseen speakers that are not present in the training data. In particular, the introduction of DNNs enabled more accurate and flexible spectrum modeling and computationally efficient inference. These advances were facilitated by supervised training methods that can exploit a large amount of data.

Neural BSS rapidly influenced TSE research. For example, Du et al. [22] trained a speaker-close NN to extract the speech of a target speaker by using training data with mixed various interfering speakers. This work is an initial neural TSE system using audio clues. However, using speaker-close models requires a significant amount of data from the target speaker and cannot be extended to speakers unseen during training. Subsequently, the introduction of TSE systems conditioned on speaker characteristics derived from an enrollment utterance significantly mitigated this requirement [10], [11], [23]. Enrollment consists of a recording of a target speaker's voice, which amounts to a few seconds of speech. With these approaches, audio clue-based TSE became possible for speakers unseen during training as long as an enrollment utterance was available. Furthermore, the flexibility of NNs to integrate different modalities combined with the high modeling capability of face recognition and lipreading systems offered new possibilities for speaker-open visual clue-based TSE [7], [8]. More recently, neural approaches have also been introduced for spatial clue-based TSE [3], [24].

TSE has gained increased attention. For example, dedicated tasks were part of such recent evaluation campaigns as the Deep Noise Suppression (DNS) (https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2022/) and Clarity (https://claritychallenge.github.io/clarity_CC_doc) challenges. Many works have extended

TSE to other tasks, such as a direct automatic speech recognition (ASR) of a target speaker from a mixture, which is called *target speaker ASR* (*TS-ASR*) [25], [26], and personalized voice activity detection (VAD)/diarization [27], [28]. Notably, target speaker VAD (TS-VAD)-based diarization [28] has been very successful in such evaluation campaigns as CHiME-6 (https://chimechallenge.github.io/chime6/results.html) and DIHARD-3 (https://dihardchallenge.github.io/dihard3/results), outperforming state-of-the-art diarization approaches in challenging conditions.

TSE taxonomy

TSE is a vast research area spanning a multitude of approaches. This section organizes them to emphasize their relations and differences. We categorize the techniques using four criteria: 1) the type of clues, 2) the number of channels, 3) speaker close versus open, and 4) generative versus discriminative. Table 1 summarizes the taxonomy; the works in the scope of this overview article are emphasized in red.

Type of clue

The type of clue used to determine the target speaker is an important factor in distinguishing among TSE approaches. The most prominent types are audio, visual, and spatial clues. This classification also defines the main organization of this article, which covers such approaches in the "Audio-Based TSE," "Visual/Multimodal Clue-Based TSE," and "Spatial Clue-Based TSE" sections. Other types have been and could be proposed, as we briefly discuss in the "Remaining Issues and Outlook" section.

An audio clue consists of a recording of a speech signal of the target speaker. Such a clue can be helpful, e.g., in the use case of personal devices, where the user can prerecord an example of his or her voice. Alternatively, for long recordings, such as meetings, clues can be obtained directly from part of the recording. The interest in audio clues sharply increased recently with the usage of neural models for TSE [10], [11], [12]. Audio clues are perhaps the most universal because they do not require using any additional devices, such as multiple microphones and a camera. However, the performance may be limited compared to other clues since discriminating speakers based only on their voice characteristics is prone to errors due to inter- and intraspeaker variability. For example, the voice characteristics of different speakers, such as family members, often closely resemble one another. On the other hand, the voice characteristics of one speaker may change depending on such factors as emotions, health, and age.

A visual clue consists of a video of the target speaker talking. This type is often constrained to the speaker's face, sometimes just to the lip area. Unlike audio clues, visual clues are typically synchronized with audio signals that are processed, i.e., not prerecorded. A few works also explored using just a photo of the speaker [37]. Visual clues have been employed to infer the activity pattern and location of the target speaker [5] and to jointly model audio and visual signals [4], [5]. Recent works usually use visual clues to guide discriminative models toward extracting the target speaker [7], [8], [9]. Visual clues are especially useful when speakers in the recording have

Table 1. A taxonomy of TSE works.

				Type of Clues			Number of Microphones		Speaker Close/ Open	
	Representative Approaches	References	Year	Audio	Visual	Spatial	Single	Multiple	Close	Open
	Fixed beamforming	[2], [16]*	1985	–	–	✓	–	✓	–	✓
Generative	Audiovisual F-HMM	[4]	2001	✓[†]	✓	–	✓	–	✓	–
	ICA with visual voice activity	[5]	2007	–	✓	–	–	✓	–	✓
	Multichannel NMF	[19]	2011	✓[†]	–	–	–	✓	✓	–
	IVE with x-vectors	[12]	2020	✓	–	–	–	✓	–	✓
	Audiovisual variational autoencoder	[29]	2020	–	✓	–	✓	–	–	✓
Discriminative	Speaker-specific network	[22]	2014	✓[†]	–	–	✓	–	✓	–
	Multichannel SpeakerBeam	[30], [10]	2017	✓	–	–	–	✓	–	✓
	SpeakerBeam	[10]	2019	✓	–	–	✓	–	–	✓
	VoiceFilter	[11]	2019	✓	–	–	✓	–	–	✓
	SpEx	[31]	2020	✓	–	–	✓	–	–	✓
	The conversation	[7]	2018	–	✓	–	✓	–	–	✓
	Looking to listen	[8]	2018	–	✓	–	✓	–	–	✓
	On/off-screen audiovisual separation	[9]	2018	–	✓	–	✓	–	–	✓
	Landmark-based audiovisual speech enhancement	[32]	2019	–	✓	–	✓	–	–	✓
	Multimodal SpeakerBeam	[33], [34]	2019	✓	✓	–	✓	–	–	✓
	Audiovisual speech enhancement through obstructions	[35]	2019	✓	✓	–	✓	–	–	✓
	Neural spatial filter	[3]	2019	✓	–	✓	–	✓	–	✓
	Spatial speaker extractor	[24]	2019	✓	–	✓	–	✓	–	✓
	Multichannel multimodal TSE	[36]	2020	✓	✓	✓	–	✓	–	✓

Approaches within the scope of this overview article are emphasized in red.

*Since the first works that proposed beamforming were not model-based, we consider them neither generative nor discriminative.

[†]In speaker-close cases, the models are trained on target speaker's audio. In this table, we consider this an audio clue.

similar voices [8]. However, they might be sensitive to physical obstructions of the speaker in the video.

A spatial clue refers to the target speaker's location, e.g., the angle from the recording devices. The location can be inferred, in practice, from a video of the room or a recording of a speaker in the same position. Extracting the speaker based on his or her location has been researched from the mid-1980s with beamforming techniques that pioneered this topic [2], [16]. More recent IVE models use location for initialization [12]. Finally, several works have shown that NNs informed by location can also achieve promising performance [3], [24]. Spatial clues are inherently applicable only when a recording from multiple microphones is available. However, they can identify the target speaker in the mixture rather reliably, especially when the speakers are stationary.

Different clues may work better in different situations. For example, the performance with audio clues might depend on the similarity of the voices of the present speakers, and obstructions in the video may influence visual clues. Hence, it is advantageous to use multiple clues simultaneously to combine their strengths. Many works have combined audio and visual clues [4], [33], and some have even added spatial clues [36].

Number of microphones

Another way to categorize the TSE approaches is based on the number of microphones (channels) they use. Multiple channels allow the spatial diversity of the sources to be exploited to help discriminate the target speaker from interference. Such an approach also closely follows human audition, where binaural signals are crucial for solving the cocktail party problem.

All approaches with spatial clues require using a microphone array to capture the direction information of the sources in the mixture [2], [3], [16], [24], [36]. Some TSE approaches that exploit audio and visual clues also assume multichannel recordings, such as the extensions of ICA/IVA approaches [5], [12].

Multichannel approaches generally generate extracted signals with better quality and are thus preferable when recordings from a microphone array are available. However, sometimes, they might fail when the sources are located in the same direction from the viewpoint of the recording device. Moreover, adopting a microphone array is not always an option when developing applications, due to cost restrictions. In such cases, single-channel approaches are requested. They rely on spectral models of speech mixture, using either the F-HMM or, recently, NNs, and exploit audio [10], [11] and visual clues [7], [8] to identify the target speech.

Recent single-channel neural TSE systems have achieved remarkable performance. Interestingly, such approaches can also be easily extended to multichannel processing by augmenting the input with spatial features [3] and combining the processing with beamforming [24], [30], as discussed in the "Integration With Microphone Array Processing" section. For example, using a beamformer usually extracts a higher-quality signal due to employing a spatial linear filter to perform extraction, which can benefit ASR applications [10].

Speaker-open versus speaker-close methods

We usually understand the clues used by TSE as short evidence about the target speaker obtained at the time of executing the method, e.g., one utterance spoken by the target speaker, a video of him/her speaking, and his/her current location. There are, however, also methods that use a more significant amount of data from the target speaker (e.g., several hours of his or her speech) to build a model specific to that person. These methods can also be seen as TSE except that the clues involve much more data.

We refer to these two categories as the *speaker-open method* and *speaker-close method*. Speaker-open and speaker-close categories are sometimes referred to as *speaker independent* and *speaker dependent*, respectively. We avoid this terminology, as in TSE, all systems are informed about the target speaker, and therefore, the term *speaker independent* might be misleading. In speaker-open methods, the data of the target speaker are available only during the test time; i.e., the model is trained on the data of different speakers. In contrast, the target speaker is part of the training data in speaker-close methods. Many methods in the past were speaker close, e.g., [4] and [19], where the models were trained on the clean utterances of the target speaker. Also, the first neural models for TSE used a speaker-specific network [22]. Most recent works on neural methods, which use a clue as an additional input, are speaker-open methods [3], [7], [8], [10], [11]. Recent IVE methods [12] are also speaker open; i.e., they guide the inference of IVE by using the embedding of a previously unseen speaker.

Generative versus discriminative

We can classify TSE into approaches using generative and discriminative models. Generative approaches model the joint distribution of the observations, target signals, and clues. The estimated target speech is obtained by maximizing the likelihood. In contrast, discriminative approaches directly estimate the target speech signal, given observations and clues.

In the TSE literature, generative models were the dominant choice in the pioneering works, including one [4] that used HMMs to jointly model audio and visual modalities. IVE [12] is also based on a generative model of the mixtures.

The popularity of discriminative models, in particular, NNs, has increased since the mid-2010s, and such models today are the choice for many problems, including TSE. With discriminative models, TSE is treated as a supervised problem, where the parameters of a TSE model are learned using artificially generated training data. The modeling power of NNs enables us to exploit large amounts of such data to build strong speech models. Moreover, the versatility of NNs enables complex dependencies to be learned among different types of observations (e.g., speech mixture and video/speaker embeddings), which allows the successful conditioning of the extraction process on various clues. However, NNs also bring new challenges, such as generalization to unseen conditions and high computational requirements [38]. Some recent works have also explored using generative NNs, such as variational autoencoders [29], which might represent a middle ground between the traditional generative approaches and those using discriminative NNs.

Scope of overview article

In the remainder of our article, we focus on the neural methods for TSE emphasized in Table 1. Recent neural TSE approaches opened the possibility of achieving high-performance extraction with various clues. They can be operated with a single microphone and applied for speaker-open conditions, which are very challenging constraints for other schemes. Consequently, these approaches have received increased attention from both academia and industry.

In the following section, we introduce a general framework to provide a uniformized view of the various NN-based TSE approaches for both single- and multichannel approaches and independent of the type of clues. We then respectively review the approaches relying on audio, visual, and spatial clues in the "Audio-Based TSE," "Visual/Multimodal Clue-Based TSE," and "Spatial Clue-Based TSE" sections.

General framework for neural TSE

In the previous section, we introduced a taxonomy that described the diversity of approaches to tackle the TSE problem. However, recent neural TSE systems have much in common. In this section, we introduce a general framework that provides a unified view of a neural TSE system, which shares the same processing flow independent of the type of clue used. By organizing the existing approaches into a common framework, we hope to illuminate their similarities and differences and establish a firm foundation for future research.

A neural TSE system consists of an NN that estimates the target speech conditioned on a clue. Figure 3 is a schematic diagram of a generic neural TSE system that consists of two main modules: a clue encoder and a speech extraction module, described in more detail in the following.

Clue encoder

The clue encoder pulls out (from the clue, $\mathbf{C}_s$) information that allows the speech extraction module to identify and extract the target speech in the mixture. We can express the processing as

$$\mathbf{E}_s = \mathrm{ClueEncoder}(\mathbf{C}_s; \theta^{\mathrm{Clue}}) \tag{5}$$

where $\mathrm{ClueEncoder}(\cdot; \theta^{\mathrm{Clue}})$ represents the clue encoder, which can be an NN with learnable parameters θ^{Clue}, and $\mathbf{E}_s$ are the clue embeddings. Naturally, the specific implementation of the clue encoder and the information carried within $\mathbf{E}_s$ largely depend on the type of clues. For example, when the clue is an enrollment utterance, $\mathbf{E}_s = \mathbf{E}_s^{(a)} \in \mathbb{R}^{D^{\mathrm{Emb}}}$ will be a speaker embedding vector of dimension D^{Emb} that represents the voice characteristics of the target speaker. When dealing with visual clues, $\mathbf{E}_s = \mathbf{E}_s^{(v)} \in \mathbb{R}^{D^{\mathrm{Emb}} \times N}$ can be a sequence of the embeddings of length N, representing, e.g., the lip movements of the target speaker. Here, N represents the number of time frames of the mixture signal.

Interestingly, the implementation of the speech extraction module does not depend on the types of clues used. To provide a description that is independent of the types of clues, hereafter, we consider that $\mathbf{E}_s \in \mathbb{R}^{D^{\mathrm{Emb}} \times N}$ consists of a sequence of embedding vectors of dimension D^{Emb} of length N. Note that we can generate a sequence of embedding vectors for audio clue-based TSE systems by repeating the speaker embedding vector for each time frame.

Speech extraction module

The speech extraction module estimates the target speech from the mixture, given the target speaker embeddings. We can use the same configuration independent of the type of clue. Its process can be decomposed into three main parts: a mixture encoder, a fusion layer, and a target extractor:

$$\mathbf{Z}_y = \mathrm{MixEncoder}(\mathbf{y}; \theta^{\mathrm{Mix}}) \tag{6}$$

$$\mathbf{Z}_s = \mathrm{Fusion}(\mathbf{Z}_y, \mathbf{E}_s; \theta^{\mathrm{Fusion}}) \tag{7}$$

$$\hat{\mathbf{x}}_s = \mathrm{TgtExtractor}(\mathbf{Z}_s, \mathbf{y}; \theta^{\mathrm{TgtExtractor}}) \tag{8}$$

where $\mathrm{MixEncoder}(\cdot; \theta^{\mathrm{Mix}})$, $\mathrm{Fusion}(\cdot; \theta^{\mathrm{Fusion}})$, and TgtExtractor $(\cdot; \theta^{\mathrm{TgtExtractor}})$ respectively represent the mixture encoder, the

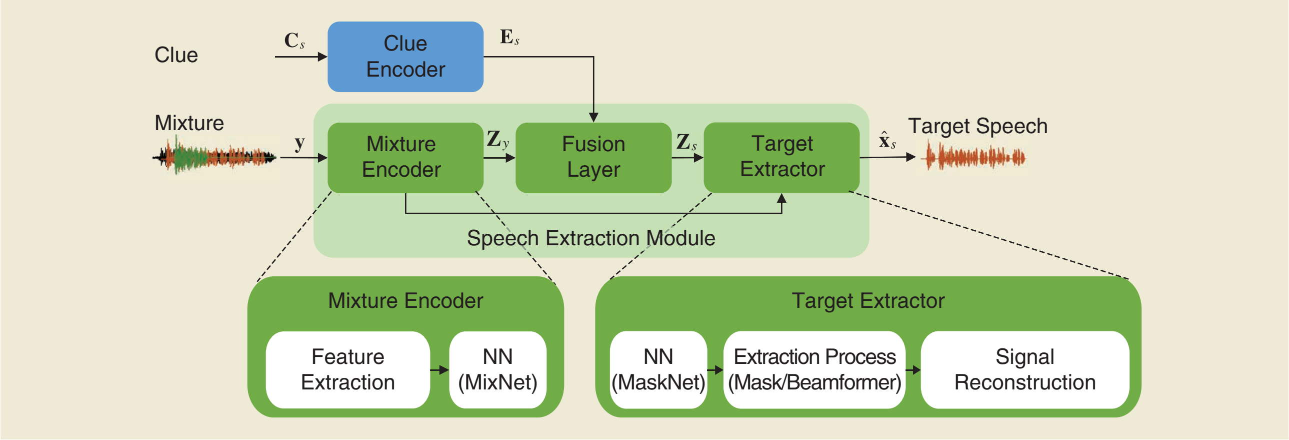


FIGURE 3. The general framework for neural TSE.

fusion layer, and the target extractor with parameters $\theta^{\mathrm{Mix}}, \theta^{\mathrm{Fusion}}$, and $\theta^{\mathrm{TgtExtractor}}$; $\mathbf{Z}_y \in \mathbb{R}^{D^y \times N}$ and $\mathbf{Z}_s \in \mathbb{R}^{D^s \times N}$ are the internal representations of the mixture before and after conditioning on embedding $\mathbf{E}_s$.

The mixture encoder performs the following:

$$\mathbf{Y} = \mathrm{FE}(\mathbf{y}; \theta^{\mathrm{FE}}) \tag{9}$$

$$\mathbf{Z}_y = \mathrm{MixNet}(\mathbf{Y}; \theta^{\mathrm{MixNet}}) \tag{10}$$

where FE(·) and MixNet(·) respectively represent the feature extraction process and an NN with parameters θ^{FE} and θ^{MixNet}. The feature extractor computes the features from the observed mixture signal, $\mathbf{Y} \in \mathbb{R}^{D \times N}$. These can be such spectral features as magnitude spectrum coefficients derived from the short-time Fourier transform (STFT) of the input mixture [7], [8], [10], [11]. When using a microphone array, spatial features, such as the interaural phase difference (IPD), defined in (21) in the "Spatial Clue-Based TSE" section, can also be appended. Alternatively, the feature extraction process can be implemented by an NN, such as a 1D convolutional layer, that operates directly on the raw input waveform of the microphone signal [23], [39]. This enables the learning of a feature representation optimized for TSE tasks.

The features are then processed with an NN, MixNet(·), which performs a nonlinear transformation and captures the time context, i.e., several past and future frames of the signal. The resulting representation, $\mathbf{Z}_y$, of the mixture is (at this point) agnostic of the target.

The fusion layer, sometimes denoted as an adaptation layer, is a key component of a TSE system and allows the conditioning of the process on the clue. It combines $\mathbf{Z}_y$ with the clue embeddings, $\mathbf{E}_s$. Conditioning an NN on auxiliary information is a general problem that has been studied for multimodal processing and the speaker adaptation of ASR systems. TSE systems have borrowed fusion layers from these fields. Table 2 lists several options for the fusion layer. Some widely used fusion layers include 1) the concatenation of $\mathbf{Z}_y$ with the clue embeddings $\mathbf{E}_s$ [7], [8], 2) addition after transforming the embeddings with linear transformation $\mathbf{L}$ to match the dimension of $\mathbf{Z}_y$, 3) multiplication [10], 4) a combination of addition and multiplication denoted as feature-wise linear modulation (FiLM), and 5) a factorized layer [10], [30], i.e., the combination of different transformations of the mixture representation weighted by the clue embedding values. Note that concatenation is similar to addition if a linear transformation follows it. Other alternatives have also been proposed, including attention-based fusion [40]. Note that the fusion operations described here assume just one clue. It is also possible to use multiple clues, as discussed in the "Audiovisual Clue-Based TSE" section. Some works also employ the fusion repeatedly at multiple positions in the model [31].

Table 2. The types of fusion layers.

Fusion Type	Equation	Parameters $(\theta^{\mathrm{Fusion}})$
Concatenation*	$\mathbf{Z}_s = [\mathbf{Z}_y, \mathbf{E}_s]$	–
Addition*	$\mathbf{Z}_s = \mathbf{Z}_y + \mathbf{L}\mathbf{E}_s$	$\mathbf{L} \in \mathbb{R}^{D^Z \times D^{\mathrm{Emb}}}$
Multiplication	$\mathbf{Z}_s = \mathbf{Z}_y \odot (\mathbf{L}\mathbf{E}_s)$	$\mathbf{L} \in \mathbb{R}^{D^Z \times D^{\mathrm{Emb}}}$
FiLM	$\mathbf{Z}_s = \mathbf{Z}_y \odot (\mathbf{L}_1\mathbf{E}_s) + \mathbf{L}_2\mathbf{E}_s$	$\mathbf{L}_1 \in \mathbb{R}^{D^Z \times D^{\mathrm{Emb}}}, \mathbf{L}_2 \in \mathbb{R}^{D^Z \times D^{\mathrm{Emb}}}$
Factorized layer	$\mathbf{Z}_s = \Sigma_{i=1}^{D^{\mathrm{Emb}}} \mathbf{L}_i \mathbf{Z}_y \mathrm{diag}(\mathbf{e}_i)$	$\mathbf{L}_i \in \mathbb{R}^{D^Z \times D^Z}$

FiLM: feature-wise linear modulation.

*Concatenation is similar to addition if a linear transformation follows it.

Here, $\mathbf{L}$, $\mathbf{L}_1$, and $\mathbf{L}_2$ are linear transformations for mapping the dimension of the clue embeddings, D^{Emb}, to the dimension of $\mathbf{Z}_y$ D^Z; $\odot$ represents the element-wise Hadamard multiplication operation of matrices; $\mathbf{e}_i$ is a vector containing the elements of the *i*-th row of $\mathbf{E}_s$; and diag(·) is an operator that converts a vector into a diagonal matrix.

The last part of the speech extraction module is the target extractor, which estimates the target signal. We explain in the following the time-frequency masking-based extractor, which has been widely used [3], [7], [8], [41]. Recent approaches also perform a similar masking operation in the learned feature domain [23], [39].

The time-frequency masking approach was inspired by early BSS studies that relied on the sparseness assumption of speech signals, an idea based on the observation that the energy of a speech signal is concentrated in a few time-frequency bins of a speech spectrum. Accordingly, the speech signals of different speakers rarely overlap in the time-frequency domain in a speech mixture. Thus, we can extract the target speech by applying a time-frequency mask on the observed speech mixture, where the mask indicates the time-frequency bins where the target speech is dominant over other signals. Figure 4 shows an example of an ideal binary mask for extracting a target

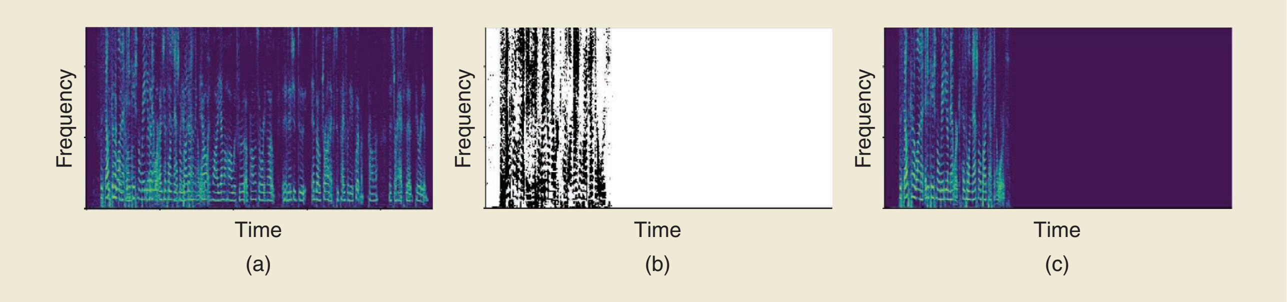


FIGURE 4. An example of a time-frequency mask processing for speech extraction. The figure shows (a) the spectrogram of the mixture, (b) the time-frequency mask, and (c) the spectrogram of the extracted signal. The time-frequency mask shows spectrogram regions where the target source is dominant. By applying this mask to the mixture, we obtain an extracted speech signal that estimates the target speech.

speech in a mixture of two speakers. Such an ideal binary mask assumes that all the energy in each time-frequency bin belongs to one speaker. In recent mask-based approaches that use real-valued (or complex) masks, this assumption, or observation, is not needed.

The processing of the masking-based extractor can be summarized as

$$\mathbf{M}_s = \text{MaskNet}\,(\mathbf{Z}_s; \theta^{\text{Mask}}) \tag{11}$$

$$\hat{\mathbf{X}}_s = \mathbf{M}_s \odot \mathbf{Y} \tag{12}$$

$$\hat{\mathbf{x}}_s = \text{Reconstruct}\,(\hat{\mathbf{X}}_s; \theta^{\text{Reconst}}) \tag{13}$$

where MaskNet($\cdot$) is an NN that estimates the time-frequency mask for the target speech, $\mathbf{M}_s \in \mathbb{R}^{D \times N}$ and θ^{Mask} are the network parameters, and $\odot$ denotes the element-wise Hadamard multiplication; $\mathbf{Y}$ and $\hat{\mathbf{X}}_s$ are the mixture and the estimated target speech signals in the feature domain. Equation (12) shows the actual extraction process. Here, Reconstruct($\cdot$) is an operation to reconstruct the time-domain signal by performing the inverse operation of the feature extraction of the mixture encoder, i.e., either the inverse STFT (iSTFT) or a transpose convolution if using a learnable feature extraction. In the latter case, the reconstruction layer has learnable parameters, θ^{Reconst}.

There are other possibilities to perform the extraction process. For example, we can modify the MaskNet($\cdot$)NN to directly infer the target speech signal in the feature domain. Alternatively, as discussed in the "Integration With Microphone Array Processing" section, we can replace the mask-based extraction process with beamforming when a microphone array is available.

Integration with microphone array processing

If we have access to a microphone array to record the speech mixture, we can exploit the spatial information to extract the target speech. One approach is to use spatial clues to identify the speaker in the mixture by informing the system about the target speaker's direction, as discussed in the "Spatial Clue-Based TSE" section. Another approach combines TSE with beamforming and uses the latter to perform the extraction process instead of (12). For example, we can use the output of a TSE system to estimate the spatial statistics needed to compute the coefficients of a beamformer steering in the direction of the target speaker. This approach can also be used with audio clue- and visual clue-based TSE systems and requires no explicit use of spatial clues to identify the target speaker in the mixture.

We briefly review the mask-based beamforming approach, which was introduced initially for noise reduction and BSS [42], [43]. A beamformer performs the linear spatial filtering of the observed microphone signals:

$$\hat{X}_s[n,f] = \mathbf{W}^{\text{H}}[f]\,\mathbf{Y}[n,f] \tag{14}$$

where $\hat{X}_s[n,f] \in \mathbb{C}$ is the STFT coefficient of the estimated target signal at time frame n and frequency bin f, $\mathbf{W}[f] \in \mathbb{C}^M$ is a vector of the beamformer coefficients, $\mathbf{Y}[n,f] = [Y^1[n,f], \ldots, Y^M[n,f]]^T \in \mathbb{C}^M$ is a vector of the STFT coefficients of the microphone signals, M is the number of microphones, and $^{\text{H}}$ is the conjugate transpose. We can derive the beamformer coefficients from the spatial correlation matrices of the target speech and the interference. These correlation matrices can be computed from the observed signal and the time-frequency mask estimated by the TSE system [30].

This way of combining a TSE system with beamforming replaces the time-frequency masking operation of (12) with the spatial linear filtering operation of (14). It allows distortionless extraction, which is often advantageous when using TSE as a front end for ASR [10].

Training a TSE system

Before using a TSE model, we first need to learn its parameters: $\theta^{\text{TSE}} = \{\theta^{\text{Mix}}, \theta^{\text{Clue}}, \theta^{\text{Fusion}}, \theta^{\text{TgtExtractor}}\}$. Most existing studies use fully supervised training, which requires a large amount of training data consisting of the triplets of speech mixture $\mathbf{y}$, target speech signal $\mathbf{x}_s$, and corresponding clue $\mathbf{C}_s$ to learn parameters θ^{TSE}. Since this requires access to a clean target speech signal, such training data are usually simulated by artificially mixing clean speech signals and noise, following the signal model of (1).

Figure 5 illustrates the data generation process using a multispeaker audiovisual speech corpus containing multiple videos for each speaker. First, we generate a mixture by using randomly selected speech signals from the target speaker, the interference speaker, and the background noise. We obtain an audio clue by selecting another speech signal from the target speaker as well as a visual clue from the video signal associated with the target speech.

The training of a neural TSE framework follows the training scheme of NNs with error back propagation. The parameters are estimated by minimizing a training loss function:

$$\theta^{\text{TSE}} = \arg\min_{\theta} \mathcal{L}(\mathbf{x}_s, \hat{\mathbf{x}}_s) \tag{15}$$

where $\mathcal{L}(\cdot)$ is a training loss, which measures how close estimated target speech $\hat{\mathbf{x}}_s = \text{TSE}(\mathbf{y}, \mathbf{C}_s; \theta)$ is to the target source signal $\mathbf{x}_s$. We can use a similar loss as that employed for training noise reduction and BSS systems [14], [39].

Several variants of the losses operating on different domains exist, such as the cross entropy between the oracle and the estimated time-frequency masks and the mean square error loss between the magnitude spectra of the source and the estimated target speech. Recently, a negative signal-to-noise ratio (SNR) measured in the time domain has been widely used [6], [23], [39]:

$$\mathcal{L}^{\text{SNR}}(\mathbf{x}_s, \hat{\mathbf{x}}_s) = -10\log_{10}\left(\frac{\|\mathbf{x}_s\|^2}{\|\mathbf{x}_s - \hat{\mathbf{x}}_s\|^2}\right). \tag{16}$$

The SNR loss is computed directly in the time domain, which forces the TSE system to learn to correctly estimate the magnitude and the phase of the target speech signal. This loss has improved extraction performance [23]. Many works also employ versions of the loss that are invariant to arbitrary scaling, i.e., the scale-invariant SNR (SI-SNR) [39] and linear filtering of the estimated signal, often called the *signal-to-distortion ratio* (*SDR*) [44]. Besides training losses operating on the signal and mask levels, it is also possible to train a TSE system end to end with a loss defined on the output of an ASR system [45]. Such a loss can be particularly effective when targeting ASR applications, as discussed in the "Extension to Other Tasks" section.

The clue encoder can be an NN trained jointly with a speech extraction module [10] and pretrained on a different task, such as speaker identification for audio clue-based TSE [11] and lipreading for visual clue-based TSE [7]. Using a pretrained clue encoder enables the leveraging of large amounts of data to learn robust and highly discriminative embeddings. On the other hand, jointly optimizing the clue encoder allows learning embeddings to be optimized directly for TSE. These two trends can also be combined by fine-tuning the pretrained encoder and using multitask training schemes, which add a loss to the output of the clue embeddings [46].

Considerations when designing a TSE system

We conclude this section with some considerations about the different options for designing a TSE system. In the preceding description, we intentionally ignored the details of the NN architecture used in the speech extraction module, such as the type of layers. Indeed, novel architectures have been and will probably continue to be proposed regularly, leading to gradual performance improvement. For concrete examples, we refer to some public implementations of TSE frameworks presented in the "Resources" section.

Most TSE approaches can borrow a network configuration from architectures proved effective for BSS and noise reduction. One important aspect is that an NN must be able to see enough context in the mixture to identify the target speaker. This has been achieved using such recurrent NN-based architectures as a stack of bidirectional long short-term memory (LSTM) layers [10], convolutional NN (CNN)-based architectures with a stack of convolutional layers that gradually increases the receptive field over the time axis to cover a large context [7], [23], and attention-based architectures [47].

The networks in the mixture encoder and the extraction process generally use a similar architecture. The best performance was reported when using a shallow mixture encoder (typically a single layer/block) and a much deeper extraction network, i.e., where a fusion layer is placed on the lower part of the extraction module. Furthermore, we found in our experiments that the multiplication and FiLM layers usually perform well. However, the impact of the choice of the fusion layer seems rather insignificant.

For the feature extraction, early studies used spectral features computed with the STFT [7], [8], [10]. However, most recent approaches employ a learned feature extraction module, following its success for separation [23], [39]. This approach allows direct optimization of the features for the given task. However, the choice of input features may depend on the acoustic conditions, and some have reported superior performance using the STFT under challenging reverberant conditions [48] and using handcrafted filter banks [49].

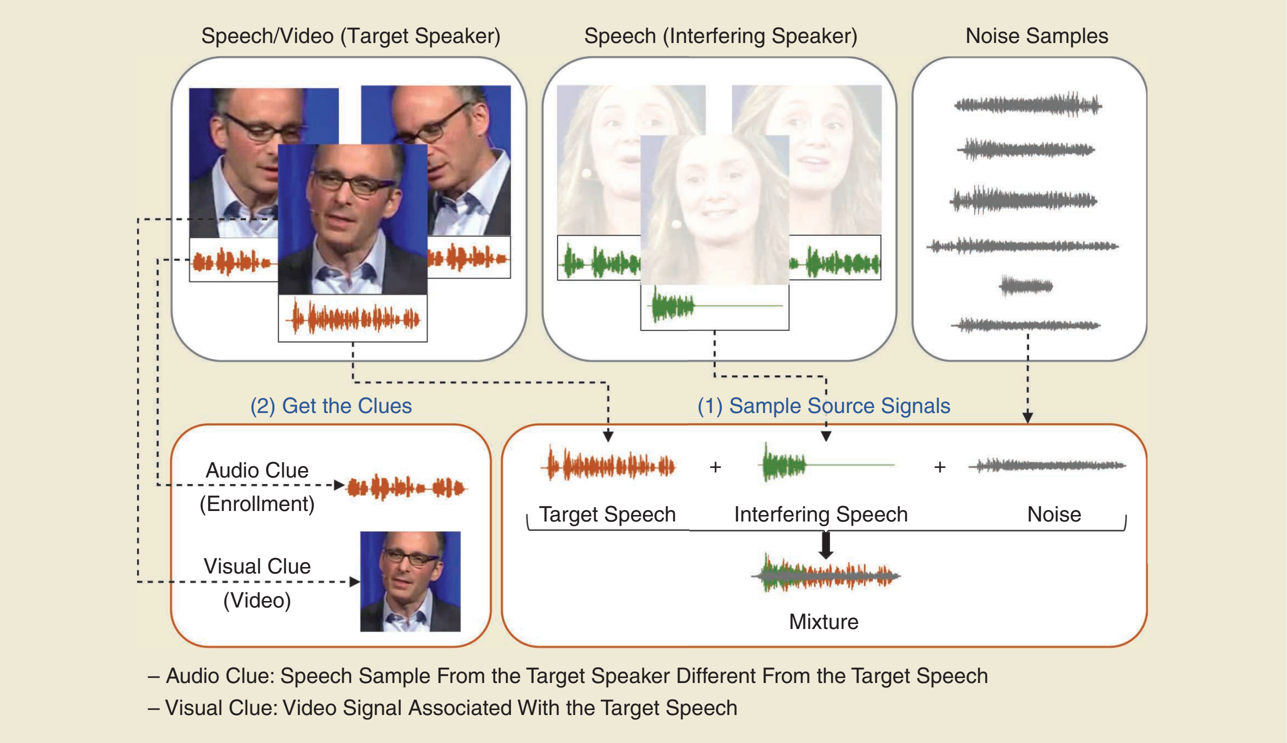

Speech/Video (Target Speaker)
Speech (Interfering Speaker)
Noise Samples
(2) Get the Clues
(1) Sample Source Signals
Audio Clue
(Enrollment)
Visual Clue
(Video)
Target Speech
+
Interfering Speech
+
Noise
Mixture
– Audio Clue: Speech Sample From the Target Speaker Different From the Target Speech
– Visual Clue: Video Signal Associated With the Target Speech

FIGURE 5. An example of generating simulation data for training and testing. This example assumes videos are available so that audio and visual clues can be generated. No video is needed for audio clue-based TSE. For visual clue-based TSE, we do not necessarily need multiple videos from the same speaker.

Except for such general considerations, it is difficult to make solid arguments for a specific network configuration since performance may depend on many factors, such as the task, the type of clue, the training data generation, and the network and training hyperparameters.

Audio-based TSE

In this section, we explain how the general framework introduced in the "General Framework for Neural TSE" section can be applied in the case of audio clues. In particular, we discuss different options to implement the clue encoder, summarize the development of audio-based TSE, and present some representative experimental results.

Audio clue encoder

An audio clue is an utterance spoken by the target speaker from which we derive the characteristics of his or her voice, allowing identification in a mixture. This enrollment utterance can be obtained by prerecording the user of a personal device or with a part of a recording in which a wake-up keyword is uttered. The clue encoder is usually used to extract a single vector that summarizes the entire enrollment utterance.

Since the clue encoder's goal is to extract information that defines the voice characteristics of the target speaker, embeddings from the speaker verification field are often used, such as i-vectors and NN-based embeddings (e.g., d-vectors and x-vectors). Clue encoders trained directly for TSE tasks are also used. Figure 6 describes these three options.

i-Vectors

With their introduction around 2010, i-vectors [50] were the ruling speaker verification paradigm until the rise of NN speaker embeddings. The main idea behind i-vectors is modeling the features of an utterance by using a Gaussian mixture model (GMM), whose means are constrained to a subspace and depend on the speaker and the channel effects. The subspace is defined by the universal background model (UBM), i.e., a GMM trained on a large amount of data from many speakers, and a total variability subspace matrix. The supervector of the means of utterance GMM $\boldsymbol{\mu}$ is decomposed:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{17}$$

where $\mathbf{m}$ is a supervector of the means of the UBM, $\mathbf{T}$ is a low-rank rectangular matrix representing the bases spanning the subspace, and $\mathbf{w}$ is a random variable with standard normal prior distribution. Since an i-vector is the maximum a posteriori estimate of $\mathbf{w}$, it thus consists of values that enable the adaptation of the parameters of the generic UBM speaker model ($\mathbf{m}$) to a specific recording. As a result, it captures the speaker's voice characteristics in the recording.

An important characteristic of i-vectors is that they capture both the speaker and channel variability. This case may be desired in some TSE applications, where we obtain enrollment utterances in identical conditions as the mixed speech. In such a situation, the channel information might also help distinguish the speakers. i-Vectors have also been used in several TSE works [10].

NN-based embeddings

The state-of-the-art speaker verification systems predominantly use NN-based speaker embeddings, which were adopted later for TSE. The common idea is to train an NN for the task of speaker classification. Such an NN contains a "pooling layer" that converts a sequence of features into one vector. The pooling

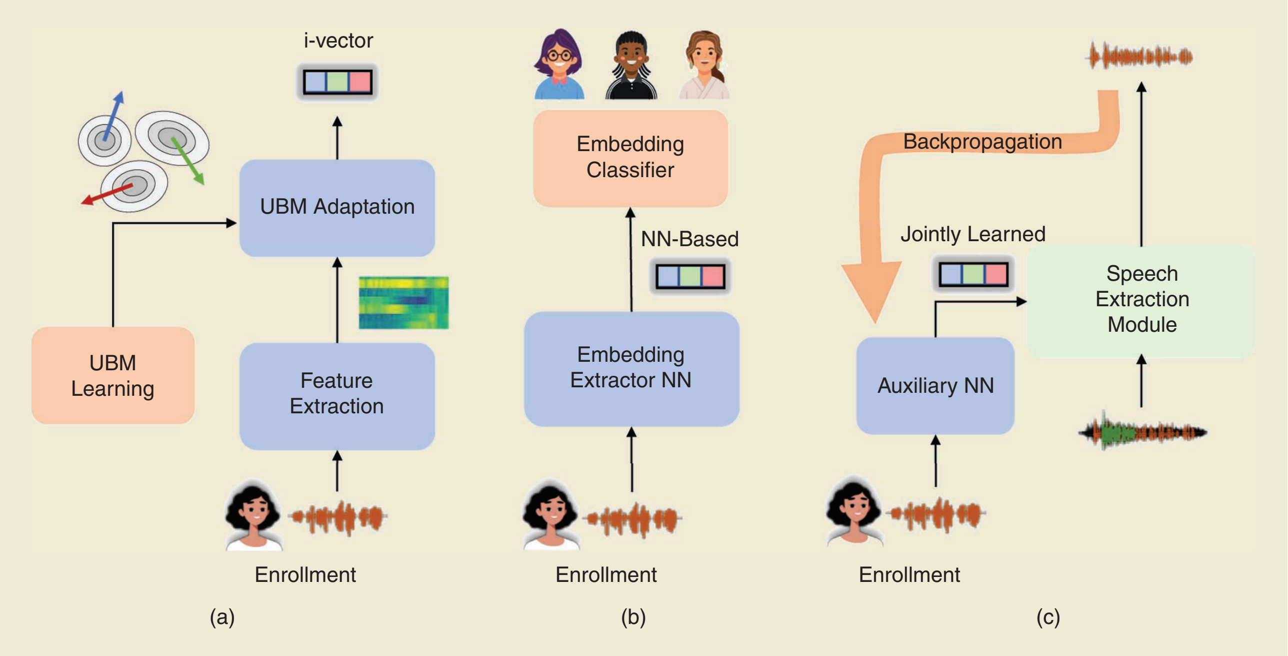


FIGURE 6. Illustration of the different speaker embeddings schemes used for TSE, i.e., (a) i-vector, (b) NN-based embeddings, and (c) jointly learned embeddings. The orange parts are included only in the training stage. UBM: universal background model.

layer computes the mean and, optionally, the standard deviation of the sequence of features over the time dimension. The pooled vector is then classified into speaker classes and used in other loss functions that encourage speaker discrimination. For TSE, the speaker embedding is then the vector of the activation coefficients of one of the last network layers. The most common of such NN-based speaker embeddings are d-vectors and x-vectors [51]. Many TSE works employ d-vectors [11].

Since NNs are trained for speaker classification and related tasks, embeddings are usually highly speaker discriminative. Most other sources of variability are discarded, such as the channel and content. Another advantage of this class of embeddings is that they are usually trained on large corpora with many speakers, noises, and other variations, resulting in very robust embedding extractors. Trained models are often publicly available, and the embeddings can be readily used for TSE tasks.

Jointly learned embeddings

NN-based embeddings, such as x-vectors, are designed and trained for the task of speaker classification. Although this causes them to contain speaker information, it is questionable whether the same representation is optimal for TSE tasks. An alternative is to train the neural embedding extractor jointly with a speech extraction module. The resulting embeddings are thus directly optimized for TSE tasks. This approach has been used for TSE in several works [10], [31].

The NN performing the speaker embedding extraction takes an enrollment utterance $\mathbf{C}_s^{(a)}$ as input and generally contains a pooling layer converting the frame-level features into one vector, similar to the embedding extractors discussed in the preceding. This NN is trained with the main NN, using a common objective function. A second objective function can also be used on the embeddings to improve their speaker discriminability [46].

As mentioned previously, the advantage of such embeddings is that they are trained directly for TSE and thus collect essential information for this task. On the other hand, the pretrained embedding extractors are often trained on larger corpora and may be more robust. A possible middle ground might take a pretrained embedding extractor and fine-tune it jointly with the TSE task. However, this has, to the best of our knowledge, not been done yet.

Existing approaches

The first neural TSE methods were developed around 2017. One of the first published works, SpeakerBeam [10], explored both the single-channel approach, where the target extractor was implemented by time-frequency masking, and the multichannel approach using beamforming. This work also compared different variants of fusion layers and clue encoders. This was followed by such as VoiceFilter [11], which put more emphasis on ASR applications using TSE as a front end and also investigated streaming variants with minimal latency. A slightly modified variant of the task was presented in works on speaker inventory [40], where not one but multiple speakers can be enrolled. Such a setting might be suitable for meeting scenarios. Recently, many works, such as SpEx [31], have started to use time-domain approaches, following their success in BSS [39].

Experiments

An audio clue is a simple way to condition the system for extracting the target speaker. Many works have shown that the speaker information extracted from audio clues is sufficient for satisfactory performance. Demonstrations of many works are available online such as VoiceFilter [11], at https://google.github.io/speaker-id/publications/VoiceFilter/, and SpeakerBeam [10], at https://www.youtube.com/watch?v=7FSHgKip6vI. We present here some results to demonstrate the potential of audio clue-based approaches. The experiments were done with time-domain SpeakerBeam (https://github.com/butspeechfit/speakerbeam), which uses a convolutional architecture, a multiplicative fusion layer, and a jointly learned clue encoder.

The experiments were done on three different datasets (WSJ0-2mix, WHAM!, and WHAMR!) to show the performance in different conditions (clean, noisy, and reverberant, respectively). We describe these datasets in more detail in the "Resources" section. All the experiments were evaluated with the SI-SNR metric and measured the improvements over the SI-SNR of the observed mixture. More details about the experiments can be found in [52].

Figure 7 compares the TSE results with a cascade system, first doing BSS and then independent speaker identification. Speaker identification is done either in an oracle way (selecting the output closest to the reference) or with x-vectors (extracting the x-vectors from all the outputs and the enrollment utterances and selecting the output with the smallest cosine distance as the target). The BSS system uses the same convolutional architecture as TSE, differing only in that it does not have a clue encoder and that the output layer is twice larger, as it outputs two separated speech signals. The direct TSE scheme outperformed the cascade system, especially in more difficult conditions, such as WHAMR!. This difference reflects a couple of causes: 1) the TSE model is directly optimized for the TSE task and does not spend any capacity on extracting other speakers, and 2) the TSE model has additional speaker information.

Figure 8 gives an example of spectrograms obtained using TSE on a recording of two speakers from the WHAMR! database, including noise and reverberation. TSE correctly

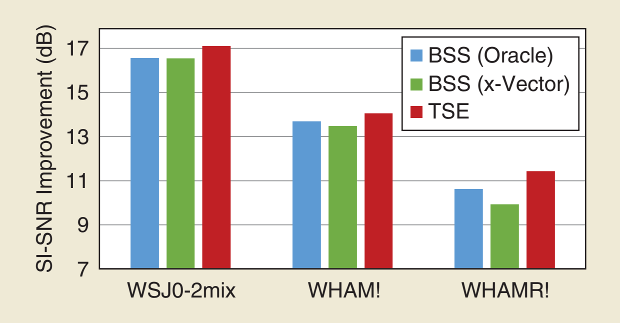


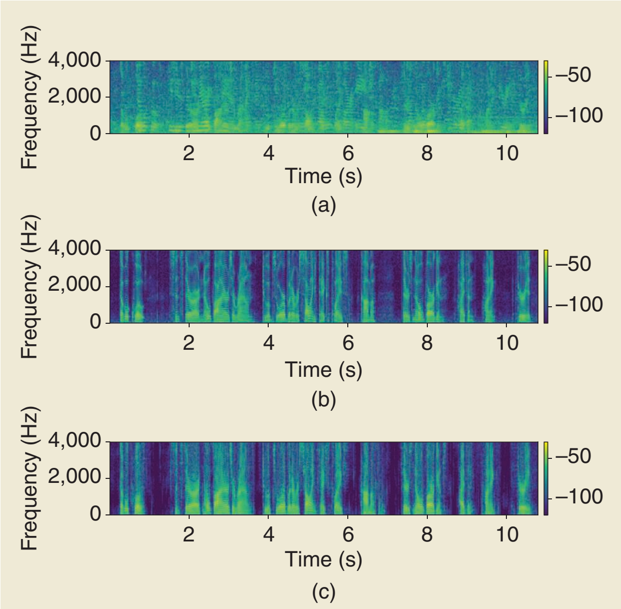
FIGURE 7. A comparison of TSE and cascade BSS systems when using an audio clue in terms of SI-SNR improvement (higher is better) [52].

identifies the target speaker and removes all the interference, including the second speaker, noise, and reverberation.

Limitations and outlook

Using TSE systems conditioned on audio clues is particularly practical due to the simplicity of obtaining the clues; i.e., no additional hardware is needed, such as cameras and multiple microphones. Considering the good performance demonstrated in the literature, these systems are widely applicable. Today, the methods are rapidly evolving and achieving increasingly higher accuracy.

The main challenge in audio clue-based systems is correct identification of the target speaker. The speech signal of the same speaker might have highly different characteristics in different conditions, due to such factors as emotional state, channel effects, and the Lombard effect. TSE systems must be robust enough to such intraspeaker variability. On the other hand, different speakers might have very similar voices, leading to erroneous identification if the TSE system lacks sufficient accuracy.



(a)

(b)

(c)

FIGURE 8. Spectrograms of (a) mixed speech, (b) reference speech, and (c) extracted speech (SI-SNR: 11.56 dB) taken from the WHAMR! database.

Resolving both issues requires precise speaker modeling. In this regard, the TSE methods may draw inspiration from the latest advances in the speaker verification field, including advanced model architectures, realistic datasets with a huge number of speakers for training, and using pretrained features from self-supervised models.

Visual/multimodal clue-based TSE

Visual clue-based TSE assumes that a video camera captures the face of the target speaker who is talking in the mixture [7], [8]. Using visual clues is motivated by psychoacoustic studies (see the references in a previous work [6]) that revealed that humans look at lip movements to understand speech better. Similarly, the visual clues of TSE systems derive hints about the state of the target speech from the lip movements, such as whether the target speaker is speaking or silent as well as more refined information about the phoneme being uttered.

A visual clue, which presents different characteristics than audio clues because it captures information from another modality, is time synchronized with the target speech in the mixture without being corrupted by the interference speakers. Therefore, a visual clue-based TSE can better handle mixtures of speakers with similar voices, such as same-gender mixtures, than audio clue-based systems because the extraction process is not based on the speaker's voice characteristics. Some works can even perform extraction from a mixture of the same speaker's speech [8]. Another potential advantage is that the users may not need to pre-enroll their voice. Video signals are also readily available for many applications, such as video conferencing.

Figure 9 provides a diagram of a visual TSE system that follows the same structure as the general TSE framework introduced in the "General Framework for Neural TSE" section. Only the visual clue encoder part is specific to the task. We describe it in more detail in the following and then introduce a multimodal clue extension. We conclude this section with some experimental results and discussions.

Visual clue encoder

The visual clue encoder computes from the video signal a representation that allows the speech extraction module to identify and extract the target speech in the mixture. This processing involves the steps described in the following:

$$\mathbf{E}_s^{(v)} = \text{Upsample}\left(\text{NN}\left(\text{VFE}\left(\mathbf{C}_s^{(v)}\right), \theta^{\text{v-clue}}\right)\right) \qquad (18)$$

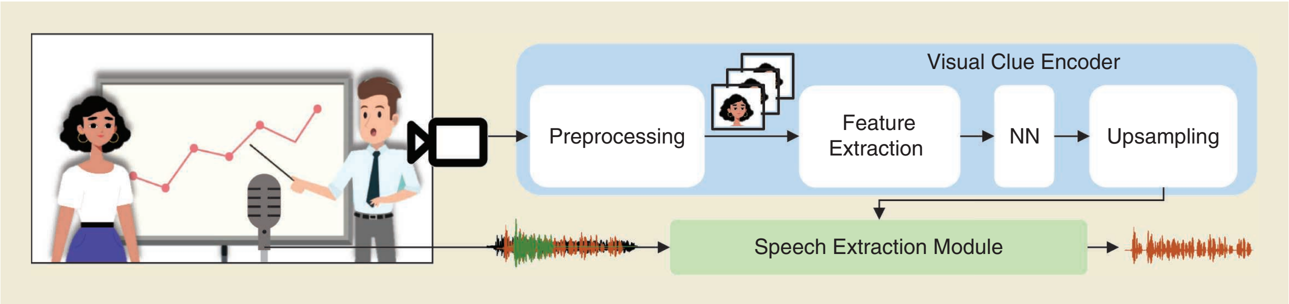


FIGURE 9. The visual clue-based TSE system.

where $\mathbf{E}_s^{(v)} \in \mathbb{R}^{D^{\text{Emb}} \times N}$ represents the sequence of the visual embedding vectors, $\mathbf{C}_s^{(v)}$ is the video signal obtained after preprocessing, VFE(·) is the visual feature extraction module, $\text{NN}(\cdot, \theta^{\text{v-clue}})$ is an NN with parameters $\theta^{\text{v-clue}}$, and Upsample(·) represents the upsampling operation. The latter upsampling step is required because the sampling rates of the audio and video devices are usually different. Upsampling matches the number of frames of the mixture and visual clue encoders.

Preprocessing

First, the video signal captured by the camera requires preprocessing to isolate the face of the target speaker. Depending on the application, this may require detecting and tracking the target speaker's face and cropping the video. These preprocessing steps can be performed using previously well-established video processing algorithms [6].

Visual feature extraction

Similar to an audio clue-based TSE, the visual clue encoder can directly extract embeddings from raw video data and visual features. With the first option, the raw video is processed with a CNN whose parameters are jointly learned with the speech extraction module to enable direct optimization of the features for the extraction task without any loss of information. However, since the video signals are high-dimensional data, achieving joint optimization can be complex. This approach has been used successfully with speaker-close conditions [53]. Extending it to speaker-open conditions might require a considerable amount of data and careful design of the training loss by using, e.g., multitask training to help the visual encoder capture relevant information.

Most visual TSE works use instead a visual feature extractor pretrained on another task to reduce the dimensionality of the data. Such feature extractors can leverage a large amount of image and video data (that do not need to be speech mixtures) to learn representation robust to variations, such as resolution, luminosity, and head orientation. The first option is to use facial landmark points as features. Facial landmarks are the key points on a face that indicate the mouth, eyes, and nose positions and offer a very low-dimension representation of a face, which is interpretable. Moreover, face landmarks can be easily computed with efficient off-the-shelf algorithms [32].

The other option is to use neural embeddings derived from an image/video processing NN trained on a different task, which was proposed in three concurrent works [7], [8], [9]. Ephrat et al. [8] used visual embeddings obtained from an intermediate layer of a face recognition system called FaceNet. This face recognition system is trained so that embeddings derived from photographs of the same person are close and embeddings from different persons are far from one another. It thus requires only a corpus of still images with person identity labels for training the system. However, the embeddings do not capture the lip movement dynamics and are not explicitly related to the acoustic content.

Alternatively, Afouras et al. [7] proposed using embeddings obtained from a network trained to perform lipreading, i.e., where a network is trained to estimate the phoneme or uttered word from the video of the speaker's lips. The resulting embeddings are thus directly related to the acoustic content. However, the training requires video with the associated phoneme and word transcriptions, which are more demanding and costly to obtain.

The third option, introduced by Owens et al. [9], exploits embeddings derived from an NN trained to predict whether the audio and visual tracks of a video are synchronized. This approach enables self-supervised training, where the training data are simply created by randomly shifting the audio track by a few seconds. The embeddings capture information on the association between the lip motions and the timing of the sounds in the audio. All three options [7], [8], [9] can successfully perform a visual TSE.

Transformation and upsampling

Except with joint training approaches, the visual features are (pre)trained on different tasks and thus do not provide a representation optimal for TSE. Besides, since some of the visual features are extracted from the individual frames of a video, the dynamics of lip movements are not captured. Therefore, the visual features are further transformed with an NN, which is jointly trained with the speech extraction module. The NN, which allows learning a representation optimal for TSE, can be implemented with LSTM and convolutional layers across the time dimension to model the time series of the visual features, enabling the lip movement dynamics to be captured. Finally, the visual embeddings are upsampled to match the sampling rate of audio features $\mathbf{Z}_y$.

Audiovisual clue-based TSE

Audio clue- and visual clue-based TSE systems have complementary properties. An audio clue-based TSE is not affected by speaker movements and visual occlusions. In contrast, a visual clue-based TSE is less affected by the voice characteristics of the speakers in the mixture. By combining these approaches, we can build TSE systems that exploit the strengths of both clues for improving the robustness to various conditions [33], [36].

Figure 10 is a diagram of an audiovisual TSE system, which assumes access to the prerecorded enrollment of the target

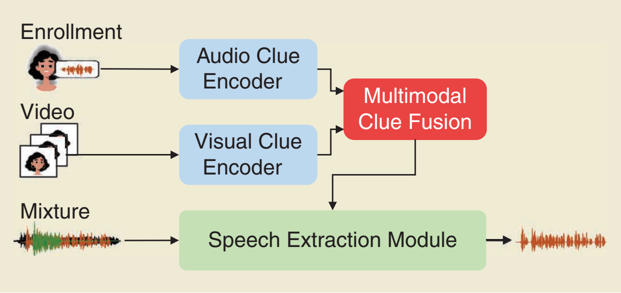


FIGURE 10. The audiovisual clue-based TSE system.

speaker to provide an audio clue and a video camera for a visual clue. The system uses the audio and visual clue encoders described in the “Audio Clue Encoder” and “Visual Clue Encoder” sections and combines these clues into an audiovisual embedding, which is given to the speech extraction module. Audiovisual embeddings can be simply the concatenation [35] or the summation of the audio and visual embeddings, or they can be obtained as a weighted sum [33], [34], where the weights can vary depending on the reliability of each clue. The weighted sum approach can be implemented with an attention layer widely used in machine learning, which enables dynamic weighting of the contribution of each clue.

Experimental results and discussion

Several visual TSE systems have been proposed, which differ mostly by the type of visual features used and the network configuration. These systems have demonstrated astonishing results, which can be attested by the demonstrations available online, e.g., for [9], https://andrewowens.com/multisensory; for [8], https://looking-to-listen.github.io; for [7], https://www.robots.ox.ac.uk/~vgg/demo/theconversation; and for [34], http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html. Here, we briefly describe experiments using the audio, visual, and audiovisual time-domain SpeakerBeam systems [34], which use a similar configuration as the system in the “Audio-based TSE” section. The speech extraction module employs a stack of time-convolutional blocks and a multiplicative fusion layer. The audio clue encoder consists of the jointly learned embeddings described in the “Jointly Learned Embeddings” section. The visual clue encoder uses visual features derived from face recognition, similar to a previous work [8]. The audiovisual system combines the visual and audio clues with an attention layer [34].

The experiments used mixtures of utterances from the LRS3-TED corpus (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html), which consists of single-speaker utterances with associated videos. We analyzed the behavior under various conditions by looking at results from same and different gender mixtures and two examples of clue corruptions (enrollment corrupted with white noise at an SNR of 0 dB and video with a mask on the speaker’s mouth). The details of the experimental setup are available in [34].

Figure 11 compares the extraction performance measured in terms of the SDR improvement for audio, visual, and audiovisual TSE under various mixture and clue conditions. We confirmed that a visual clue-based TSE is less sensitive to the characteristics of the speakers in the mixture since the performance gap between different- and same-gender mixtures is smaller than with an audio clue-based TSE. When using a single clue, performance can be degraded when this clue is corrupted. However, the audiovisual system that exploits both clues can achieve superior extraction performance and is more robust to clue corruption.

Discussions and outlook

Visual clue-based TSE approaches offer an alternative to audio clue-based ones when a camera is available. The idea of using visual clues for TSE is not new [4], [5], although recent neural systems have achieved an impressive level of performance.

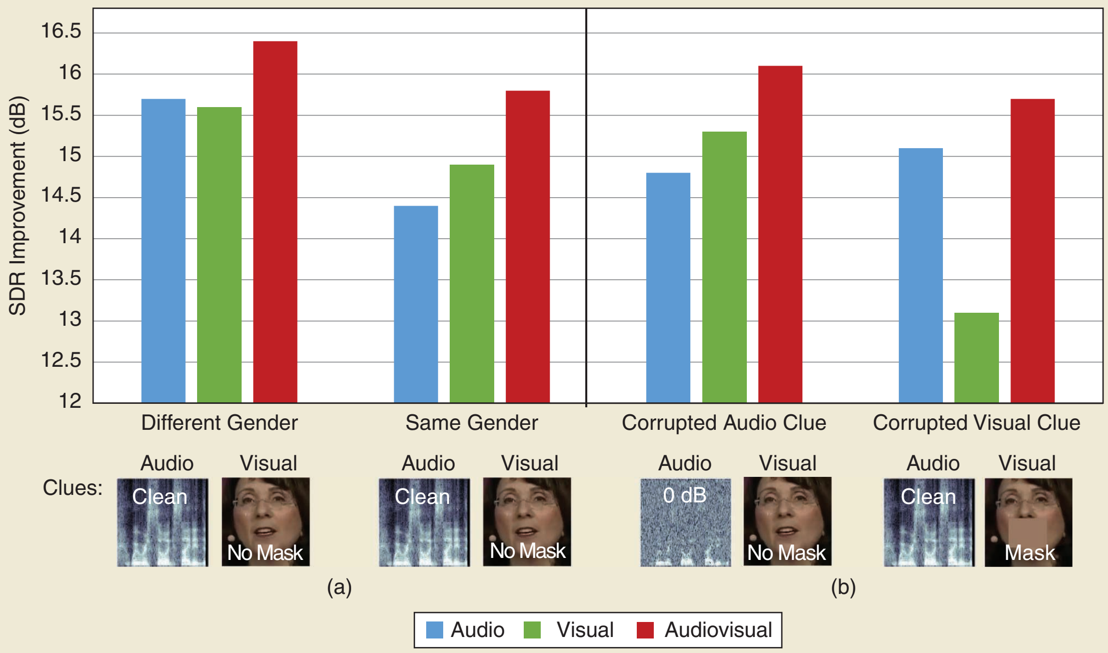


FIGURE 11. The SDR improvement of TSE with audio, visual, and audiovisual clues for (a) mixtures of the same/different gender and (b) corruptions of audio and visual clues. The audio clues were corrupted by adding white noise at an SNR of 0 dB to the enrollment utterance. The video clues were corrupted by masking the mouth region in the video.

This is probably because NNs can effectively model the relationship between the different modalities learned from a large amount of training data.

Issues and research opportunities remain with the current visual clue-based TSE systems. First, most approaches do not consider the speaker tracking problem and assume that the audio and video signals are synchronized. These aspects must be considered when designing and evaluating future TSE systems. Second, video processing involves high computational costs, and more research is needed to develop efficient online systems.

Spatial clue-based TSE

When using a microphone array to record a signal, spatial information can be used to discriminate among sources. In particular, access to multichannel recordings opens the way to extract target speakers based on their location, i.e., using spatial clues (as indicated in Figure 1). This section explains how such spatial clues can be obtained and used in TSE systems. While enhancing speakers from a given direction has a long research history [2], we focus here on neural methods that follow the scope of our overview article.

Note that multichannel signals can also be utilized in the extraction process using beamforming. Such an extraction process can be used in systems with any type of clue, requiring only that the mixed speech be recorded with multiple microphones. This beamforming process was reviewed in the "Integration With Microphone Array Processing" section. In this section, we focus specifically on the processing of spatial clues.

Obtaining spatial clues

In some situations, the target speaker's location is approximately known in advance. For example, for an in-car ASR, the driver's position is limited to a certain region in the car. In other scenarios, we might have access to a multichannel enrollment utterance of the speaker recorded in the same position as the final mixed speech. In such a case, audio source localization methods can be applied. Conventionally, this can be done by methods based on generalized cross correlation and steered-response power, but recently, deep learning methods have also shown success in this task. An alternative is to skip the explicit estimation of the location and directly extract features in which the location is encoded when a multichannel enrollment is available. We detail this approach further in the following section.

Spatial clues can also be obtained from a video by using face detection and tracking systems. A previous work [36] demonstrated this possibility with a 180° wide-angle camera positioned parallel to a linear microphone array. By identifying the target speaker in the video, the azimuth with respect to the microphone array was roughly approximated. Depth cameras can also be used to estimate not only the azimuth but also the elevation and distance of the speaker.

Spatial clue encoder

Figure 12(a) describes the overall structure and the usage of a spatial clue encoder, which usually consists of two parts: the extraction of directional features and an NN postprocessing of them. Two possible forms of spatial clues are dominant in the literature: the angle of the target speaker with respect to the microphone array and a multichannel enrollment utterance recorded in the target location. Both can be encoded into directional features.

When the spatial clue is the DOA, the most commonly used directional features are the angle features, which are computed as the cosine of the difference between the IPD and the target phase difference (TPD):

$$\mathrm{AF}[n,f] = \sum_{m_1,m_2\in\mathcal{M}} \cos\big(\mathrm{TPD}(m_1,m_2,\phi_s,f) - \mathrm{IPD}(m_1,m_2,n,f)\big) \quad (19)$$

$$\mathrm{TPD}(m_1,m_2,\phi_s,f) = \frac{2\pi f F_s}{F}\frac{\cos\phi_s\,\Delta_{m_1,m_2}}{c} \quad (20)$$

$$\mathrm{IPD}(m_1,m_2,n,f) = \angle Y^{m_2}[n,f] - \angle Y^{m_1}[n,f] \quad (21)$$

where $\mathcal{M}$ is a set of pairs of microphones used to compute the feature, F_s is the sampling frequency, ϕ_s is the target direction, c is the sound's velocity, and Δ_{m_1,m_2} is the distance from microphone m_1 to microphone m_2. An example of angle

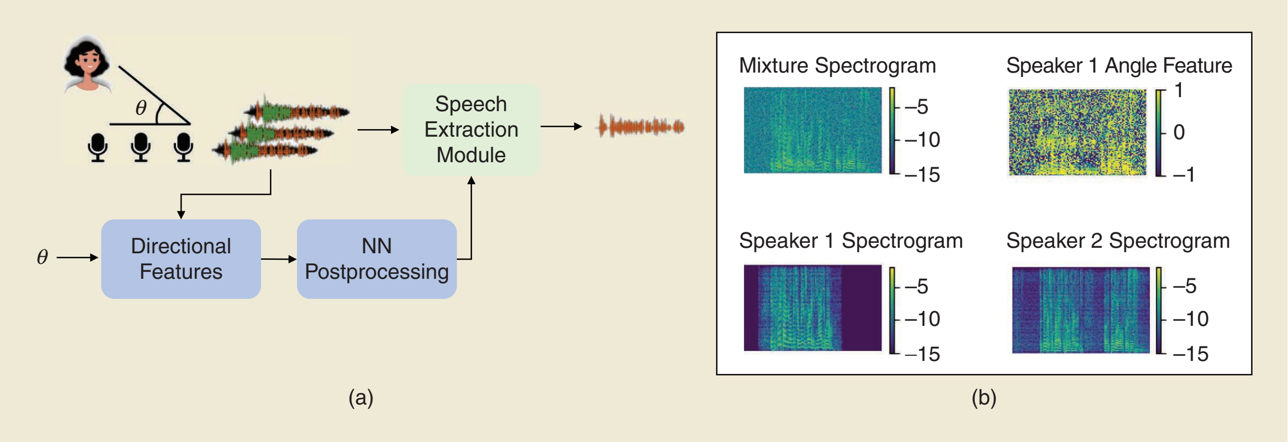


FIGURE 12. The use of (a) a spatial clue encoder and (b) an example of directional features.

features is available in Figure 12(b). For time-frequency bins dominated by the source from direction ϕ_s, the value of the angle feature should be close to one or negative one. Other directional features have been proposed that exploit a grid of fixed beamformers. A directional power ratio measures the ratio between the power of the response of a beamformer steered into the target direction and the power of the beamformer responses steered into all the directions in the grid. In a similar fashion, a directional SNR can also be computed, which compares the response of a beamformer in the target direction with the response of a beamformer in the direction with the strongest interference.

If the spatial clue consists of a multichannel enrollment utterance, the directional feature can be formed as a vector of IPDs computed from the enrollment. Alternatively, the DOA can be estimated from the enrollment, and the spatial features derived from it can be used.

Note that when using a spatial clue to determine the target speaker, the multichannel input of the speech extraction module must also be used. This enables the identification of the speaker coming from the target location in the mixture. Furthermore, a target extractor is often implemented as beamforming, as explained in the "Integration With Microphone Array Processing" section.

Combination with other clues

Although a spatial clue is very informative and generally can lead the TSE system to a correct extraction of the target, it does fail in some instances. Estimation errors of the DOA are harmful to proper extraction. Furthermore, if the spatial separation of the speakers with respect to the microphone array is not significant enough, the spatial clue may not discriminate between them. Combining a spatial clue with audio and visual clues is an option to combat such failure cases.

Experimental results

We next report the results from an experiment with spatial clues [36] that compared the effectiveness of using audio, visual, and spatial clues. The audio clue encoder was trained jointly with the extraction module, and the visual encoder was a pretrained lipreading network. The target speaker's direction was encoded in the angle feature. The spatial and visual embeddings were fused with the extraction network by concatenation and the audio embedding with a factorized layer. The extraction module employed an NN consisting of temporal convolutional layers.

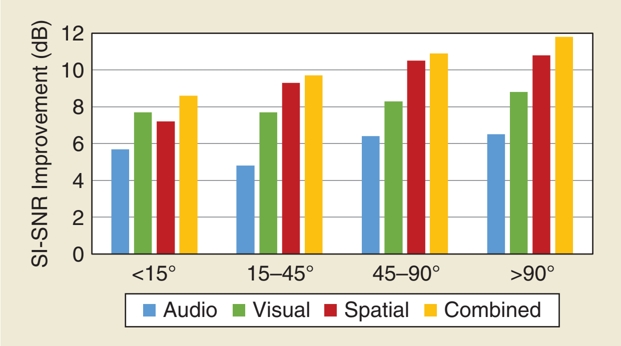


FIGURE 13. The SI-SNR improvement of TSE with audio, visual, and spatial clues in four conditions based on the angle separation among speakers [36].

The experiments were performed on a Mandarin audiovisual dataset containing mixtures of two and three speakers. The results in Figure 13 were divided into several conditions based on the angle separation between the closest speakers. The spatial clue is very effective, although the performance declines when speakers are near one another (<15°). A combination with other modalities outperformed any individual type of clue in all the conditions. Demo samples of [36] can be found online; https://yongxuustc.github.io/grnnbf.

Discussion

Using spatial clues is a powerful way of conditioning a TSE system to extract the target speaker. It relies on the availability of signals from a microphone array and a way to determine the location of the target speaker. Unfortunately, these restrictions limit the applications to some extent. Neural TSE methods with spatial clues follow a long history of research on the topic, such as beamforming techniques, and extend them with nonlinear processing. This approach unifies the methods with those using other clues and allows a straightforward combination of different clues into one system. Such combinations can alleviate the shortcomings of spatial clues, including the failures when the speakers are located in the same direction from the point of view of the microphones.

In most current neural TSE works, the target speaker's location is assumed to be fixed. Although the methods should be easily extended to a dynamic case, investigations of such settings remain relatively rare [24].

Extension to other tasks

The ideas of TSE can be applied to other speech processing tasks, such as ASR and diarization.

TS-ASR

An important application of TSE is TS-ASR, where the goal is to transcribe the target speaker's speech and ignore all the interference speakers. The TSE approaches we described can be naturally used as a front end to an ASR system to achieve TS-ASR. Such a cascade combination allows for a modular system, which offers ease of development and interpretability. However, the TSE system is often optimized with a signal loss, as in (16). Such a TSE system inevitably introduces artifacts caused by the remaining interferences, oversuppression, and other nonlinear processing distortions. These artifacts limit the expected performance improvement from a TSE front end.

One approach to mitigate the effect of such artifacts is to optimize the TSE front end with an ASR criterion [10]. The TSE front end and the ASR back end are NNs and can be interconnected with differentiable operations, such as beamforming and feature extraction. Therefore, a cascade system can be represented with a single computational graph, allowing all parameters to be jointly trained. Such joint training can significantly improve the TS-ASR performance.

Another approach inserts a fusion layer into an ASR system [26], [45] to directly perform clue conditioning. These integrated TS-ASR systems avoid any explicit signal extraction step, a decision that reduces the computational cost, although such systems may be less interpretable than cascade systems.

TS-ASR can use the audio clues provided by prerecorded enrollment utterances [10], [26], [45] and from a keyword (anchor) for a smart device scenario [54], for example. Some works have also exploited visual clues, which can be used for the extraction process and to implement an audiovisual ASR back end, since lipreading also improves ASR performance [55].

TS-VAD and diarization

The problem of speech diarization consists of detecting who spoke when in a multispeaker recording. This technology is essential for achieving, e.g., meeting recognition and analysis systems that can transcribe a discussion among multiple participants. Several works have explored using speaker clues to perform this task [27], [28].

For example, a personalized VAD [27] exploits a speaker embedding vector derived from an enrollment utterance of the target speaker to predict his or her activity, i.e., whether he or she is speaking at a given time. In principle, this can be done with a system such as that presented in the "General Framework for Neural TSE" section, where the output layer performs the binary classification of the speaker activity instead of estimating the target speech signal. Similar systems have also been proposed using visual clues, called *audiovisual VAD* [56]. Predicting the target speaker's activity is arguably a more straightforward task than estimating his or her speech signal. Consequently, personalized VAD can use simpler network architectures, leading to more lightweight processing.

The preceding personalized VAD systems have been extended to simultaneously output the activity of multiple target speakers, which was called TS-VAD [28]. TS-VAD has been used in the systems achieving top performance on evaluation campaigns such as CHiME6 and DIHARD III. (The results of CHiME 6 challenge can be found at: https://chimechallenge.github.io/chime6/results.html, the results of DIHARD III can be found at: https://dihardchallenge.github.io/dihard3/results.)

Remaining issues and outlook

Research toward computational selective hearing has been a long endeavor. Recent developments in TSE have enabled identifying and extracting a target speaker's voice in a mixture by exploiting audio, visual, and spatial clues, which is one step closer to solving the cocktail party problem. Progress in speech processing (speech enhancement and speaker recognition) and image processing (face recognition and lipreading), combined with deep learning technologies to learn models that can effectively condition processing on auxiliary clues, triggered the progress in the TSE field. Some of the works we presented have achieved levels of performance that seemed out of reach just a few years ago and are already being deployed in products. See, for example, the following blog, which details the effort for deploying a visual clue-based TSE system for on-device processing: https://ai.googleblog.com/2020/10/audiovisual-speech-enhancement-in.html. Despite substantial achievements, many opportunities remain for further research, some of which we list in the following.

Deployment of TSE systems

Most of the systems we described operate offline and are computationally expensive. They are also evaluated under controlled (mostly simulated mixture) settings. Deploying such systems introduces engineering and research challenges to reduce computational costs while maintaining high performance under less-controlled recording conditions. We next discuss some of these aspects.

Inactive target speaker

Most TSE systems have been evaluated assuming that the target speaker is actively speaking in the mixture. In practice, we may not know beforehand whether the target speaker will be active. We expect that a TSE system can output no signal when the target speaker is inactive, which may not actually be the case with most current systems that are not explicitly trained to do so. The inactive target speaker problem is specific to TSE. The type of clue used may also greatly impact the difficulty of tackling this problem. For instance, visual VAD [5] might alleviate this issue. However, it is more challenging with audio clues [57], and further research may be required.

Training and evaluation criteria

Most TSE systems are trained and evaluated using such signal-level metrics as the SNR and SDR. Although these metrics are indicative of the extraction performance, their use presents two issues.

First, they may not always be correlated with human perception and intelligibility and with ASR performance. This issue is not specific to TSE; it is common to BSS and noise reduction methods. For ASR, we can train a system end to end, as discussed in the "TS-ASR" section. When targeting applications for human listeners, the problem can be partly addressed using other metrics for training and evaluation that correlate better with human perception, such as short-time objective intelligibility and perceptual evaluation of speech quality [6]. However, controlled listening tests must be conducted to confirm the impact of a TSE on human listeners [6].

Second, unlike BSS and noise reduction, a TSE system needs to identify the target speech, implying other sources of errors. Indeed, failing to identify the target may lead to incorrectly estimating an interference speaker and inaccurately outputting the mixture. Although these errors directly impact the SDR scores, it would be fruitful to agree on the evaluation metrics that separate extraction and identification performance to better reveal the behavior of TSE systems.

Signal-level metrics might not satisfactorily represent the extraction performance for inactive speaker cases. A better understanding of failures might help develop TSE systems that can recognize when they cannot identify the target speech, which is appealing for practical applications. Consequently, developing better training and evaluation criteria is a critical research direction.

Robustness to recording conditions

Training neural TSE systems requires simulated mixtures, as discussed in the "Training a TSE System" section. Applying these systems to real conditions (multispeaker mixtures recorded directly with a microphone) requires that the training data match the application scenario relatively well. For example, the type of noise and reverberation may vary significantly depending on where a system is deployed. This raises questions about the robustness of TSE systems to various recording conditions.

Neural TSE systems trained with a large amount of simulated data have been shown to generalize to real recording conditions [8]. However, exploiting real recordings where no reference target speech signal is available could further improve performance. Real recordings might augment the training data and be used to adapt a TSE system to a new environment. The issue is defining unsupervised training losses correlated with the extraction performance of the target speech without requiring access to the reference target signal.

Another interesting research direction is combining neural TSE systems, which are powerful under matched conditions, with such generative-based approaches as IVE [12], which are adaptive to recording conditions.

Lightweight and low-latency systems

Research on lightweight and low-latency TSE systems is gaining momentum, as the use of teleconferencing systems in noisy environments has risen in response to the COVID-19 pandemic. Other important use cases for TSE are hearing aids and hearables, both of which impose very severe constraints in terms of computation costs and latency. The recent DNS (https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2022/) and Clarity (https://claritychallenge.github.io/clarity_CC_doc/) challenges that target teleconferencing and hearing aid application scenarios include tracks where target speaker clues (enrollment data) can be exploited. This demonstrates the growing interest in practical solutions for TSE.

Since TSE is related to BSS and noise reduction, the development of online and low-latency TSE systems can be inspired from the progress of BSS/noise reduction in that direction. However, TSE must also identify the target speech, which may need specific solutions that exploit the long context of the mixture to reliably and efficiently capture a speaker's identity.

Spatial rendering

For applications of TSE to hearing aids and hearables, sounds must be localized in space after the TSE processing. Therefore, a TSE system must not only extract the target speech but also estimate its direction to allow rendering it so that a listener feels the correct direction of the source.

Self-supervised and cross-modal learning

A TSE system identifies the target speech in a mixture based on the intermediate representation of the mixture and the clue. Naturally, TSE benefits from better intermediate representations. For example, speech models learned with self-supervised learning criteria have gained attention as a way to obtain robust speech representations. They have shown potential for pretraining many speech processing downstream tasks, such as ASR, speaker identification, and BSS. Such self-supervised models could also reveal advantages for TSE since they could improve robustness by allowing efficient pretraining on various acoustic conditions. Moreover, for audio-based TSE, using the same self-supervised pretrained model for the audio clue encoder and the speech extraction module will help to learn the common embedding space between the enrollment and speech signals in the mixture. Similarly, the progress in cross-modal learning, which aims to learn the joint representation of data across modalities, could benefit such multimodal approaches as visual clue-based TSE.

Exploring other clues

We presented three types of clues that have been widely used for TSE. However, other clues can also be considered. For example, recent works have explored other types of spatial clues, such as distance [58]. Moreover, humans do not rely only on physical clues to perform selective hearing. We also use more abstract clues, such as semantic ones. Indeed, we can rapidly focus our attention on a speaker when we hear our name or a topic we are interested in. Reproducing a similar mechanism would require TSE systems that operate with semantic clues, which introduces novel challenges concerning how to represent semantic information and exploit it within a TSE system. Some works have started to explore this direction, such as conditioning on languages [59] and more abstract concepts [60].

Other interesting clues consist of signals that measure a listener's brain activity to guide the extraction process. Indeed, the electroencephalogram (EEG) signal of a listener focusing on a speaker correlates with the envelope of that speaker's speech signal. Ceolini et al. identified the possibility of using EEGs as clues for TSE with a system similar to the one described in the "General Framework for Neural TSE" section [61]. An EEG-guided TSE might open the door for futuristic hearing aids controlled by the user's brain activity, which might automatically emphasize the speaker a user wants to hear. However, research is still needed because developing a system that requires marginal tuning to the listener is especially challenging. Moreover, collecting a large amount of training data is very complicated since it is more difficult to control the quality of such clues. Compared to audio and visual TSE clues, EEG signals are very noisy and affected by changes in the attention of the listener, body movements, and other factors.

Table 3. Some datasets and toolkits.

	Name	Description	Link
Dataset	WSJ0-mix	Mixtures of two or three speakers	https://www.merl.com/demos/deep-clustering
	WHAM!,WHAMR!	Noisy and reverberant versions of WSJ0-mix	https://wham.whisper.ai
	LibriMix	Larger dataset of mixtures of two or three speakers	https://github.com/JorisCos/LibriMix
	LibriCSS	Meeting-like mixtures recorded in a room	https://github.com/chenzhuo1011/libri_css
	MC-WSJ0-mix	Spatialized version of WSJ0-2mix	https://www.merl.com/demos/deep-clustering
	SMS-WSJ	Multichannel corpus based on WSJ	https://github.com/fgnt/sms_wsj
	LRS	Audiovisual corpus from TED and BBC videos	https://www.robots.ox.ac.uk/~vgg/data/lip_reading
	AVSpeech	Very large audiovisual corpus from YouTube videos	https://looking-to-listen.github.io/avspeech
Tools	SpeakerBeam	Time-domain audio-based TSE system	https://github.com/butspeechfit/speakerbeam
	SpEx+	Time-domain audio-based TSE system [31]	https://github.com/xuchenglin28/speaker_extraction_SpEx
	VoiceFilter	Time-domain audio-based TSE system (unofficial) [11]	https://github.com/mindslab-ai/voicefilter
	Multisensory	Visual clue-based TSE [9]	https://github.com/andrewowens/multisensory
	Audiovisual speech enhancement	Face landmark-based visual clue-based TSE [32]	https://github.com/dr-pato/audio_visual_speech_enhancement
	FaceNet	Visual feature extractor used in [8], [33], and [34]	https://github.com/davidsandberg/facenet

Beyond speech

Human selective listening abilities go beyond speech signals. For example, we can focus on listening to the part of an instrument in an orchestra and switch our attention to a siren or a barking dog. In this article, we focused on TSE, but similar extraction problems have also been explored for other audio processing tasks. For example, much research has been performed on extracting the track of an instrument in a piece of music conditioned on, e.g., the type of instrument [62], video of the musician playing [63], and the EEG signal of the listener [64]. These approaches may be important to realize, e.g., audiovisual music analysis [65].

Recently, the problem was extended to the extraction of arbitrary sounds from a mixture [66], [67], e.g., extracting the sound of a siren or a klaxon from a recording of a mixture of street sounds. We can use such systems, as introduced in the "General Framework for Neural TSE" section, to tackle these problems, where the clue can be a class label indicating the type of target sound [66], the enrollment audio of a similar target sound [67], a video of the sound source [9], and a text description of the target sound [68]. Target sound extraction may become an important technology to design, e.g., hearables and hearing aids that could filter out nuisances and emphasize important sounds in our surroundings as well as for audio visual scene analysis [9].

Psychoacoustic studies suggest that humans process speech and music partly by using shared auditory mechanisms and that exposure to music can lead to better discrimination of speech sounds [69]. It would be interesting to explore whether, similar to humans, TSE systems could benefit from exposure to other acoustic signals by training a system to extract target speech, music, and arbitrary sounds.

Resources

We conclude by providing pointers to selected datasets and toolkits available for those motivated to experiment with TSE. TSE works mostly use datasets designed for BSS. These datasets consist generally of artificial mixtures generated from the isolated signals of the individual speakers and background. This allows evaluation of the performance by comparing the estimated signals to the original references. Additionally, TSE methods also require a clue, i.e., an enrollment utterance for the target speaker or video signal. We can obtain enrollment utterances by choosing a random utterance of the target speaker from the same database, provided that the utterance is different from the one in the mixture. For a video clue, it requires using an audiovisual dataset. The top of Table 3 lists some of the most commonly used datasets for audio and visual TSE.

Several implementations of TSE systems are openly available and listed in the lower part of Table 3. Although there are no public implementations for some of the visual TSE systems, they can be reimplemented following the audio TSE toolkits and using openly available visual feature extractors, such as FaceNet, which was used in some previous works [8], [33], [34].

Acknowledgment

This work was partly supported by the Czech Ministry of Education, Youth, and Sports as part of project LTAIN19087 "Multilinguality in Speech Technologies." Computing on the IT4I supercomputer was supported by the Czech Ministry of Education, Youth, and Sports as part of the Large Infrastructures for Research, Experimental Development, and Innovations project "e-Infrastructure CZ–LM2018140." The figures contain elements designed by Pikisuperstar/Freepik.

Authors

Katerina Zmolikova (k.zmolikova@gmail.com) received her Ph.D. degree in information technology from Brno University of Technology in 2022. She is an industrial postdoc at Demant, 2765 Copenhagen, Denmark. She is a recipient of the 2020 Joseph Fourier Award. Her research interests include speech enhancement, speech separation, and robust speech recognition.

Marc Delcroix (marc.delcroix@ieee.org) received his Ph.D. degree from Hokkaido University, Japan. He is a distinguished researcher at NTT Communication Science Laboratories, Kyoto 619-0237, Japan. He is a recipient of the 2006 Student Paper Award from the IEEE Kansai Section, the 2006 Sato Paper Award from the Acoustical Society of Japan, and the 2015 IEEE Automatic Speech Recognition and Understanding Workshop Best Paper Award honorable mention. His research interests include various aspects of speech and audio signal processing, including speech enhancement and robust speech recognition. He is a Senior Member of IEEE.

Tsubasa Ochiai (tsubasa.ochiai.ah@hco.ntt.co.jp) received his Ph.D. degree from Doshisha University. He is a researcher at NTT Communication Science Laboratories, Kyoto 619-0237, Japan. He is a recipient of the 2014 Student Presentation Award from the Acoustical Society of Japan (ASJ), the 2015 Student Paper Award from the IEEE Kansai Section, the 2020 Awaya Prize Young Researcher Award from the ASJ, and the 2021 Itakura Prize Innovative Young Researcher Award from the ASJ. His research interests include speech enhancement, array signal processing, and robust automatic speech recognition. He is a Member of IEEE.

Keisuke Kinoshita (keisuke.kinoshita@ieee.org) received his Ph.D. degree from Sophia University, Tokyo, Japan. He is a research scientist at Google, Tokyo 150-0002, Japan. He is a recipient of the 2006 Institute of Electronics, Information, and Communication Engineers Paper Award; the 2010 Acoustical Society of Japan Outstanding Technical Development Prize; and the 2012 Japan Audio Society Award. His research interests include speech enhancement, speaker diarization, and speech recognition. He is a Senior Member of IEEE.

Jan Černocký (cernocky@but.cz) received his Ph.D. degree in signal processing from Universite Paris XI Orsay and Brno University of Technology (BUT) in 1998. He is a professor in and the head of the Department of Computer Graphics and Multimedia, Faculty of Information Technology (FIT), BUT, Brno 61200, Czech Republic, and he serves as managing director of the BUT Speech@FIT research group. His research interests include artificial intelligence, signal processing, and speech data mining (speech, speaker, and language recognition). He is a Senior Member of IEEE.

Dong Yu (dongyu@ieee.org) received his Ph.D. degree in computer science from the University of Idaho. He is a distinguished scientist and vice general manager at Tencent AI Lab, Seattle, WA 98034, USA. He was a recipient of the IEEE Signal Processing Society Best Paper Award in 2013, 2016, 2020, and 2022 and a recipient of the 2021 North American Chapter of the Association for Computational Linguistics Best Long Paper Award. His research interests include speech recognition and processing and natural language processing. He is a Fellow of IEEE, ACM and ISCA.

References

[1] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Atten. Percept. Psychophys.*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015, doi: 10.3758/s13414-015-0882-9.

[2] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985, doi: 10.1121/1.392786.

[3] R. Gu et al., "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, 2019, pp. 4290–4294, doi: 10.21437/Interspeech.2019-2266.

[4] J. Hershey and M. Casey, "Audio-visual sound separation via Hidden Markov Models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 14, pp. 1173–1180.

[5] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, Apr. 2014, doi: 10.1109/MSP.2013.2296173.

[6] D. Michelsanti et al., "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1368–1396, Mar. 2021, doi: 10.1109/TASLP.2021.3066303.

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248, doi: 10.21437/Interspeech.2018-1400.

[8] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, Aug. 2018, doi: 10.1145/3197517.3201357.

[9] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 631–648.

[10] K. Žmolíková et al., "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, Aug. 2019, doi: 10.1109/JSTSP.2019.2922820.

[11] Q. Wang et al., "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732, doi: 10.21437/Interspeech.2019-1101.

[12] J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, and J. Žd'ánský, "Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 676–680, doi: 10.1109/ICASSP40776.2020.9054693.

[13] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, May 2019, Art. no. e12, doi: 10.1017/ATSIP.2019.5.

[14] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, doi: 10.1109/TASLP.2018.2842159.

[15] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 356–360, doi: 10.1109/ICASSP43922.2022.9746962.

[16] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017, doi: 10.1109/TASLP.2016.2647702.

[17] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, Jan. 2010, doi: 10.1016/j.csl.2008.11.001.

[18] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007, doi: 10.1109/TASL.2006.885253.

[19] A. Ozerov and E. Vincent, "Using the fasst source separation toolbox for noise robust speech recognition," in *Proc. Int. Workshop Mach. Listening Multisource Environ. (CHiME)*, 2011, pp. 1–2.

[20] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.

[21] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 241–245, doi: 10.1109/ICASSP.2017.7952154.

[22] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. IEEE 12th Int. Conf. Signal Process. (ICSP)*, 2014, pp. 473–477, doi: 10.1109/ICOSP.2014.7015050.

[23] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, 2019, pp. 327–334, doi: 10.1109/ASRU46091.2019.9004016.

[24] J. Heitkaemper, T. Fehér, M. Freitag, and R. Haeb-Umbach, "A study on online source extraction in the presence of changing speaker positions," in *Proc. Int. Conf. Statist. Lang. Speech Process.*, Cham, Switzerland: Springer-Verlag, 2019, pp. 198–209, doi: 10.1007/978-3-030-31372-2_17.

[25] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, “Single channel target speaker extraction and recognition with SpeakerBeam,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5554–5558, doi: 10.1109/ICASSP.2018.8462661.

[26] P. Denisov and N. T. Vu, “End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning,” in *Proc. Interspeech*, 2019, pp. 4425–4429, doi: 10.21437/Interspeech.2019-1130.

[27] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, “Personal VAD: Speaker-conditioned voice activity detection,” in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 433–439, doi: 10.21437/Odyssey.2020-62.

[28] I. Medennikov et al., “Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario,” in *Proc. Interspeech*, 2020, pp. 274–278, doi: 10.21437/Interspeech.2020-1602.

[29] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-visual speech enhancement using conditional variational auto-encoders,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1788–1800, Jun. 2020, doi: 10.1109/TASLP.2020.3000593.

[30] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, “Speaker-aware neural network based beamformer for speaker extraction in speech mixtures,” in *Proc. Interspeech*, 2017, pp. 2655–2659, doi: 10.21437/Interspeech.2017-667.

[31] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “SpEx+: A complete time domain speaker extraction network,” in *Proc. Interspeech*, 2020, pp. 1406–1410, doi: 10.21437/Interspeech.2020-1397.

[32] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, “Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 6900–6904, doi: 10.1109/ICASSP.2019.8682061.

[33] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” in *Proc. Interspeech*, 2019, pp. 2718–2722, doi: 10.21437/Interspeech.2019-1513.

[34] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, “Multimodal attention fusion for target speaker extraction,” in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2021, pp. 778–784, doi: 10.1109/SLT48900.2021.9383539.

[35] T. Afouras, J. S. Chung, and A. Zisserman, “My lips are concealed: Audio-visual speech enhancement through obstructions,” in *Proc. Interspeech*, 2019, pp. 4295–4299, doi: 10.21437/Interspeech.2019-3114.

[36] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, “Multi-modal multi-channel target speech separation,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020, doi: 10.1109/JSTSP.2020.2980956.

[37] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “FaceFilter: Audio-visual speech separation using still images,” in *Proc. Interspeech*, 2020, pp. 3481–3485, doi: 10.21437/Interspeech.2020-1065.

[38] M. Maciejewski, G. Sell, Y. Fujita, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, “Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019, pp. 165–169, doi: 10.1109/WASPAA.2019.8937153.

[39] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” in *Proc. IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, Aug. 2019, pp. 1256–1266, doi: 10.1109/TASLP.2019.2915167.

[40] X. Xiao et al., “Single-channel speech extraction using speaker inventory and attention network,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 86–90, doi: 10.1109/ICASSP.2019.8682245.

[41] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, “Learning speaker representation for neural network based multichannel speaker extraction,” in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, 2017, pp. 8–15, doi: 10.1109/ASRU.2017.8268910.

[42] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 196–200, doi: 10.1109/ICASSP.2016.7471664.

[43] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Proc. Interspeech*, 2016, pp. 1981–1985, doi: 10.21437/Interspeech.2016-552.

[44] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Language Process. (2006–2013)*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: 10.1109/TSA.2005.858005.

[45] M. Delcroix et al., “End-to-end speakerbeam for single channel target speech recognition,” in *Proc. Interspeech*, 2019, pp. 451–455, doi: 10.21437/Interspeech.2019-1856.

[46] S. Mun, S. Choe, J. Huh, and J. S. Chung, “The sound of my voice: Speaker representation loss for target voice separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 7289–7293, doi: 10.1109/ICASSP40776.2020.9053521.

[47] X. Li et al., “MIMO self-attentive RNN beamformer for multi-speaker speech separation,” in *Proc. Interspeech*, 2021, pp. 1119–1123, doi: 10.21437/Interspeech.2021-570.

[48] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, “Monaural source separation: From anechoic to reverberant environments,” in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, 2022, pp. 1–5, doi: 10.1109/IWAENC53105.2022.9914794.

[49] D. Ditter and T. Gerkmann, “A multi-phase gammatone filterbank for speech separation via Tasnet,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 36–40, doi: 10.1109/ICASSP40776.2020.9053602.

[50] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Language Process. (2006–2013)*, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[51] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[52] K. Žmolíková, “Neural target speech extraction,” Ph.D. thesis, Brno Univ. Technol., Brno, Czechia, 2022.

[53] A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in *Proc. Interspeech*, 2018, pp. 1170–1174, doi: 10.21437/Interspeech.2018-1955.

[54] B. King et al., “Robust speech recognition via anchor word representations,” in *Proc. Interspeech*, 2017, pp. 2471–2475, doi: 10.21437/Interspeech.2017-1570.

[55] J. Yu et al., “Audio-visual multi-channel integration and recognition of overlapped speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2067–2082, May 2021, doi: 10.1109/TASLP.2021.3078883.

[56] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, “An analysis of visual speech information applied to voice activity detection,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2006, vol. 1, p. I, doi: 10.1109/ICASSP.2006.1660092.

[57] C. Zhang, M. Yu, C. Weng, and D. Yu, “Towards robust speaker verification with target speaker enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 6693–6697, doi: 10.1109/ICASSP39728.2021.9414017.

[58] E. Tzinis, G. Wichern, A. S. Subramanian, P. Smaragdis, and J. Le Roux, “Heterogeneous target speech separation,” in *Proc. Interspeech*, 2022, pp. 1796–1800, doi: 10.21437/Interspeech.2022-10717.

[59] M. Borsdorf, H. Li, and T. Schultz, “Target language extraction at multilingual cocktail parties,” in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, 2021, pp. 717–724, doi: 10.1109/ASRU51503.2021.9688052.

[60] Y. Ohishi et al., “Conceptbeam: Concept driven target speech extraction,” in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 4252–4260, doi: 10.1145/3503161.3548397.

[61] E. Ceolini et al., “Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception,” *NeuroImage*, vol. 223, Dec. 2020, Art. no. 117282, doi: 10.1016/j.neuroimage.2020.117282.

[62] P. Seetharaman, G. Wichern, S. Venkataramani, and J. L. Roux, “Class-conditional embeddings for music source separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 301–305, doi: 10.1109/ICASSP.2019.8683007.

[63] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Sep. 2018, pp. 587–604, doi: 10.1007/978-3-030-01246-5_35.

[64] G. Cantisani, S. Essid, and G. Richard, “Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 36–40, doi: 10.1109/ICASSP39728.2021.9413841.

[65] Z. Duan, S. Essid, C. C. Liem, G. Richard, and G. Sharma, “Audiovisual analysis of music performances: Overview of an emerging field,” *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 63–73, Jan. 2019, doi: 10.1109/MSP.2018.2875511.

[66] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, “Listen to what you want: Neural network-based universal sound selector,” in *Proc. Interspeech*, 2020, pp. 1441–1445, doi: 10.21437/Interspeech.2020-2210.

[67] B. Gfeller, D. Roblek, and M. Tagliasacchi, “One-shot conditional audio filtering of arbitrary sounds,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 501–505, doi: 10.1109/ICASSP39728.2021.9414003.

[68] X. Liu et al., “Separate what you describe: Language-queried audio source separation,” in *Proc. Interspeech*, 2022, pp. 1801–1805, doi: 10.21437/Interspeech.2022-10894.

[69] S. S. Asaridou and J. M. McQueen, “Speech and music shape the listening brain: Evidence for shared domain-general mechanisms,” *Frontiers Psychol.*, vol. 4, Jun. 2013, Art. no. 321, doi: 10.3389/fpsyg.2013.00321.