An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation

Daniel Michelsanti, *Member, IEEE*, Zheng-Hua Tan, *Senior Member, IEEE*, Shi-Xiong Zhang, *Member, IEEE*, Yong Xu, *Member, IEEE*, Meng Yu, Dong Yu, *Fellow, IEEE*, and Jesper Jensen, *Member, IEEE*

***Abstract*—*Speech enhancement* and *speech separation* are two related tasks, whose purpose is to extract either one or more target speech signals, respectively, from a mixture of sounds generated by several sources. Traditionally, these tasks have been tackled using signal processing and machine learning techniques applied to the available acoustic signals. Since the visual aspect of speech is essentially unaffected by the acoustic environment, *visual information* from the target speakers, such as lip movements and facial expressions, has also been used for speech enhancement and speech separation systems. In order to efficiently fuse acoustic and visual information, researchers have exploited the flexibility of data-driven approaches, specifically *deep learning*, achieving strong performance. The ceaseless proposal of a large number of techniques to extract features and fuse multimodal information has highlighted the need for an overview that comprehensively describes and discusses audio-visual speech enhancement and separation based on deep learning. In this paper, we provide a systematic survey of this research topic, focusing on the main elements that characterise the systems in the literature: *acoustic features*; *visual features*; *deep learning methods*; *fusion techniques*; *training targets* and *objective functions*. In addition, we review deep-learning-based methods for *speech reconstruction from silent videos* and *audio-visual sound source separation for non-speech signals*, since these methods can be more or less directly applied to audio-visual speech enhancement and separation. Finally, we survey commonly employed *audio-visual speech datasets*, given their central role in the development of data-driven approaches, and *evaluation methods*, because they are generally used to compare different systems and determine their performance.**

***Index Terms*—Audio-visual processing, deep learning, sound source separation, speech enhancement, speech separation, speech synthesis.**

I. Introduction

SPEECH is one of the primary ways in which humans share information. A model that describes *human speech communication* is the so-called *speech chain*, which consists of two stages: *speech production* and *speech perception* [49]. Speech production is the set of voluntary and involuntary actions that allow a person, i.e. a *speaker*, to convert an idea expressed through a linguistic structure into a sound pressure wave. On the other hand, speech perception is the process happening mostly in the auditory system of a *listener*, consisting of interpreting the sound pressure wave coming from the speaker. Some external factors, such as acoustic background noise, can have an impact on the speech chain. Usually, normal-hearing listeners are able to focus on a specific acoustic stimulus, in our case the *target speech* or *speech of interest*, while filtering out other sounds [24], [233]. This well-known phenomenon is called the *cocktail party effect* [33], because it resembles the situation occurring at a cocktail party.

Generally, the presence of high-level acoustic environmental noise or competing speakers poses several challenges to the speech communication effectiveness, especially for hearing-impaired listeners. Similarly, the performance of automatic speech recognition (ASR) systems can be severely impacted by a high level of acoustic noise. Therefore, several signal processing and machine learning techniques to be employed in e.g. hearing aids and ASR front-end units have been developed to perform *speech enhancement* (SE), which is the task of recovering the clean speech of a target speaker immersed in a noisy environment. Especially when the receiver of an enhanced speech signal is a human, SE systems are often designed to improve two perceptual aspects: *speech quality*, concerning how a speech signal sounds, and *speech intelligibility*, concerning the linguistic content of a speech signal. Some applications require the estimation of multiple target signals: this task is known in the literature as *source separation* or *speech separation* (SS), when the signals of interest are all speech signals.

Classical SE and SS approaches (cf. [165], [263] and references therein) make assumptions regarding the statistical characteristics of the signals involved and aim at estimating the underlying target speech signal(s) according to mathematically tractable criteria. More recent methods based on *deep learning* tend to depart from this *knowledge-based* modelling, embracing a *data-driven* paradigm. Most of these approaches treat SE and SS as supervised learning problems[1] [264].

Manuscript received August 24, 2020; revised January 7, 2021 and March 3, 2021; accepted March 9, 2021. Date of publication March 17, 2021; date of current version April 15, 2021. This work was supported by the William Demant Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Timo Gerkmann. *(Corresponding author: Daniel Michelsanti.)*

Daniel Michelsanti and Zheng-Hua Tan are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: danmi@es.aau.dk; zt@es.aau.dk).

Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu are with the Tencent AI Laboratory, Bellevue, WA 98004 USA (e-mail: auszhang@tencent.com; lucayongxu@tencent.com; raymondmyu@tencent.com; dyu@tencent.com).

Jesper Jensen is with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark, and also with Oticon A/S, Smørum 2765, Denmark (e-mail: jje@es.aau.dk).

Digital Object Identifier 10.1109/TASLP.2021.3066303

[1] Sometimes, the approaches used in this context are more properly denoted as self-supervised or unsupervised learning techniques, since they do not use human-annotated datasets to learn representations of the data.

The techniques mentioned above consider only acoustic signals, so we refer to them as audio-only SE (AO-SE) and audio-only SS (AO-SS) systems. However, speech perception is inherently multimodal, in particular audio-visual (AV), because in addition to the acoustic speech signal reaching the ears of the listeners, location and movements of some articulatory organs that contribute to speech production, e.g. tongue, teeth, lips, jaw and facial expressions, may also be visible to the receiver. Studies in neuroscience [78], [206] and speech perception [177], [239] have shown that the visual aspect of speech has a potentially strong impact on the ability of humans to focus their auditory attention on a particular stimulus. Even more importantly for SE and SS, visual information is immune to acoustic noise and competing speakers. This makes vision a reliable cue to exploit in challenging acoustic conditions. These considerations inspired the first audio-visual SE (AV-SE) and audio-visual SS (AV-SS) works [47], [73], which demonstrated the benefit of using features extracted from the video of a speaker. Later, more complex frameworks based on classical statistical approaches have been proposed [2], [14], [138], [158], [162], [174], [190], [191], [216], [217], [236], [237], but they have very recently been outperformed by deep learning methods, such as [7], [10], [12], [55], [66], [77], [85], [99], [123], [129], [167], [168], [181], [199], [227], [247], [277], [278]. In particular, deep learning allowed to overcome the limitations of knowledge-based approaches, making it possible to learn robust representations directly from the data and to jointly process AV signals with more flexibility.

Despite the large amount of recent research and the interest in AV methods, no overview article currently focuses on deep-learning-based AV-SE and AV-SS. The survey article by Wang and Chen [264] is the most extensive overview on deep-learning-based AO-SE and AO-SS for both single-microphone and multi-microphone settings, but it does not cover AV methods. The overview article by Rivet *et al.* [218] surveys AV-SS techniques, but it dates back to 2014, when deep learning was still not adopted for the task. Multimodal methods are also covered by Taha and Hussain [245] in their survey on SE techniques. However, six AV-SE papers are discussed in total, and only one of these is based on deep learning. A limited number of deep learning approaches for AV-SE and AV-SS were described in [215], [293]. In the first case, Rincón-Trujillo and Córdova-Esparza [215] performed an analysis of deep-learning-based SS methods. They considered both AO-SS and AV-SS, with only five AV papers discussed. In the second case, Zhu *et al.* [293] provided a bird's-eye view of several AV tasks, to which deep learning has been applied. Although AV-SE and AV-SS are discussed, the presentation covers only five approaches.

In this paper, we present an extensive survey of recent advances in AV methods for SE and SS, with a specific focus on deep-learning-based techniques. Our goal is to help the reader to navigate through the different approaches in the literature. Given this objective, we try not to recommend one approach over another based on its performance, because a comparison of systems designed for a heterogeneous set of applications might be unfair. Instead, we provide a systematic description of the main ideas and components that characterise deep-learning-based AV-SE and AV-SS systems, hoping to inspire and stimulate new research in the field. This is also the reason why current challenges and possible future directions are presented and discussed throughout the paper. Furthermore, we provide an overview of *speech reconstruction from silent videos* and *audio-visual sound source separation for non-speech signals* because they are strongly related to AV-SE and AV-SS (cf. Fig. 1). Although other tasks may be considered related to AV-SE and AV-SS, their goal is substantially different. For example, AV speech recognition systems have some similarities with AV-SE and AV-SS, but they aim at finding the transcription of a video, not the clean target speech signal(s). We decide not to treat such methods in this overview. Finally, we review AV datasets and evaluation methods, because they are two important elements used to train and assess the performance of the systems, respectively.

A list of resources for datasets, objective measures and several AV approaches can be accessed at the following link: https://github.com/danmic/av-se. There, we provide direct links to available demos and source codes, that would not be possible to include in this paper due to space limitations. Our goal is to allow both beginners and experts in the fields to easily access a collection of relevant resources.

The rest of this paper is organised as follows. Section II presents the basic signal model to provide a formulation of the AV-SE and AV-SS problems. Section III introduces deep-learning-based AV-SE and AV-SS systems as a combination of several elements, described and discussed in the following sections, specifically: acoustic features (in Section IV); visual features (in Section V); deep learning methods (in Section VI); fusion techniques (in Section VII); training targets and objective functions (in Section VIII). Afterwards, Section IX deals with speech reconstruction from silent videos and AV sound source separation for non-speech signals. Section X surveys relevant AV speech datasets that can be used to train deep-learning-based models. Section XI presents a range of methodologies that may be considered for performance assessment. Finally, Section XII provides a conclusion, summarising the principal concepts and the potential future research directions presented throughout the paper.

II. Signal Model and Problem Formulation

Let $h_s[n]$ denote the impulse response from the spatial position of the s-th target source to the microphone, with n indicating a discrete-time index. Furthermore, let $h_s[n] = h_s^e[n] + h_s^l[n]$, where $h_s^e[n]$ is the early part of $h_s[n]$ (containing the direct sound and low-order reflections) and $h_s^l[n]$ is the late part of $h_s[n]$. Assuming a total number of S target speech signals and a number of C additive noise sources, the observed acoustic mixture signal can be modelled as:

$$y[n] = \sum_{s=1}^{S} x_s[n] + d[n] \tag{1}$$

with:

$$x_s[n] = x'_s[n] * h_s^e[n], \tag{2}$$

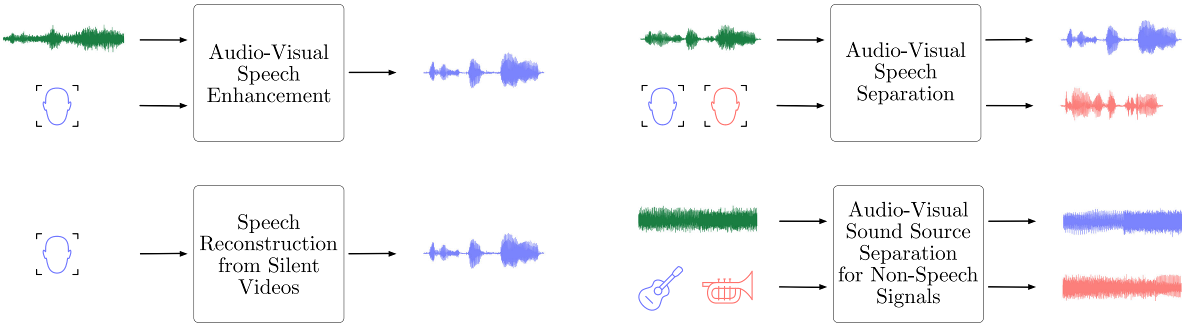

Audio-Visual Speech Enhancement
Audio-Visual Speech Separation
Speech Reconstruction from Silent Videos
Audio-Visual Sound Source Separation for Non-Speech Signals

Fig. 1. Audio-visual sound source separation tasks. In audio-visual speech enhancement, the goal is to extract the target speech signal using a noisy observation of the target speech signal and visual information. Speech reconstruction from silent videos is a special case of audio-visual speech enhancement, where the noisy acoustic input signal is not provided. Audio-visual speech separation aims at extracting multiple target speech signals from a mixture and visual information of the target speakers. When the target sources are not speakers, but, for example, musical instruments, we refer to the task as audio-visual sound source separation for non-speech signals.

$$d[n] = \sum_{s=1}^{S} x'_s[n] * h^l_s[n] + \sum_{c=1}^{C} d_c[n], \qquad (3)$$

where $x'_s[n]$ is the speech signal emitted at the s-th target speaker position, $x_s[n]$ is the clean speech signal from the s-th target speaker at the microphone (including low-order reflections), $d_c[n]$ is the signal from the c-th noise source as observed at the microphone and $d[n]$ indicates the total contribution from noise and late reverberations. Furthermore, let $v[m]$ indicate the observed two-dimensional visual signal, with m denoting a discrete-time index different from n, because the acoustic and the visual signals are usually not sampled with the same sampling rate.

Given $y[n]$ and $v[m]$, the task of AV-SS consists of determining estimates $\hat{x}_s[n]$ of $x_s[n]$,[2] with $s = 1,\ldots,S$. In some setups, additional information is available, for example a speakers' enrolment acoustic signal and a training set collected under time and location different from the recordings of $y[n]$ and $v[m]$.

When $S = 1$, we refer to the task as AV-SE and rewrite Eq. (1) as:

$$y[n] = x[n] + d[n], \qquad (4)$$

with $x[n]$ denoting $x_1[n]$.

Due to the linearity of the short-time Fourier transform (STFT), it is possible to express the acoustic signal model of Eqs. (1) and (4) in the time-frequency (TF) domain as:

$$Y(k,l) = \sum_{s=1}^{S} X_s(k,l) + D(k,l), \qquad (5)$$

for SS, and as:

$$Y(k,l) = X(k,l) + D(k,l), \qquad (6)$$

for SE, where k denotes a frequency bin index, l indicates a time frame index, and $Y(k,l)$, $X_s(k,l)$ and $D(k,l)$ are the short-time Fourier transform (STFT) coefficients of the mixture, the s-th target signal, and the noise, respectively.

The definitions provided above are valid for single-microphone single-camera AV-SE and AV-SS. It is possible to extend all the concepts to the case of multiple acoustic and visual signals. Let F and P be the number of cameras and microphones of a system, respectively. We denote as $v_f[m]$ the observed visual signal with the f-th camera. Assuming S speakers to separate, then the acoustic mixture as received by the p-th microphone can be modelled as:

$$y_p[n] = \sum_{s=1}^{S} x_{ps} + d_p[n]. \qquad (7)$$

with:

$$x_{ps}[n] = x'_s[n] * h^e_{ps}[n] \qquad (8)$$

$$d_p[n] = \sum_{s=1}^{S} x'_s[n] * h^l_{ps}[n] + \sum_{c=1}^{C} d_{pc}[n] \qquad (9)$$

In this case, the SS task consists of determining estimates $\hat{x}_s[n]$ of $x_{p^*s}[n]$ for $s = 1,\ldots,S$, given $v_f[m]$ with $f = 1,\ldots,F$, $y_p[n]$ with $p = 1,\ldots,P$ and any other additional information, assuming that the microphone with index $p = p^*$ is a pre-defined reference microphone.

III. Audio-Visual Speech Enhancement and Separation Systems

The problems of AV-SE and AV-SS have recently been tackled with supervised learning techniques, specifically deep learning methods. Supervised deep-learning-based models can automatically learn how to perform SE or SS after a training procedure, in which pairs of degraded and clean speech signals, together with the video of the speakers, are presented to them. Ideally, deep-learning-based systems should be trained using data that is representative of the settings in which they are deployed. This means that in order to have good performance in a wide variety of settings, very large AV datasets for training and testing need to be collected. In practice, the systems are trained using a large

[2] While preserving early reflections is important in some applications (e.g. hearing aids), in other cases the goal is to determine only estimates of $x'_s[n]$. This observation does not have a big impact on the formulation of the problem, therefore we are not going to make a distinction between the two cases.

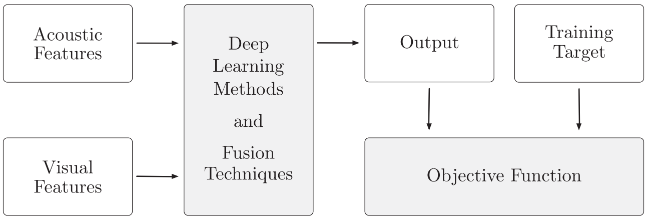


Fig. 2. Interconnections between the main elements of a generic audio-visual speech enhancement/separation system based on deep learning. White boxes represent data, while grey boxes represent processing blocks.

TABLE I
LIST OF ACOUSTIC FEATURES IN AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION PAPERS

Acoustic Features	AV-SE/SS papers
Magnitude spectrogram	[3]–[7], [10], [12], [17], [36], [42], [65], [66], [76], [77] [85], [99], [100], [107], [123], [129], [137], [156], [157] [167], [168], [179], [181], [182], [186], [196], [199] [207], [211], [225]–[227], [247], [266], [278], [283]
Phase[a]	[7], [10], [156]
Complex spectrogram	[55], [107], [109], [172], [242]
Raw waveform	[108], [277]
Speaker embeddings	[10], [85], [172], [196], [211]
IPD | cosIPD | sinIPD	[85], [107], [283] | [107], [247] | [107]
Angle feature	[85], [247], [283]

[a] Only if it is used in processing, not just to reconstruct the signal.

number of complex acoustic scenes that are synthetically generated using a mix-and-separate paradigm [292], where target speech signals are added to signals from sources of interference at several signal to noise ratios (SNRs). This way of generating synthetic training material has empirically shown its effectiveness in both audio-only (AO) and AV settings, since speech signals processed with systems trained in this way improve in terms of both estimated speech quality and intelligibility [7], [55], [144], [286].

In the following sections, we focus on the main elements of deep-learning-based AV-SE and AV-SS systems, i.e.: acoustic features; visual features; deep learning methods; fusion techniques; training targets and objective functions.[3] Fig. 2 provides a conceptual block diagram illustrating the interconnections of these elements.

IV. Acoustic Features

As represented in Fig. 2, acoustic features are one of the main elements of AV-SE and AV-SS systems. In this Section, we report which features are used in the literature, following the list provided in Table I.

A. Single-Microphone Features

AV-SE and AV-SS systems process acoustic information (cf. Fig. 2). As can be seen in Table I, the predominant acoustic input feature is the (potentially transformed) magnitude spectrogram of a single-microphone recording, sometimes in the log mel domain, like in [66]. However, a magnitude spectrogram is generally an incomplete representation of the acoustic signal, because it is computed from STFT coefficients which are complex-valued. Recent works have used as acoustic input to the AV system either the magnitude spectrogram and the respective phase [7], [10], [156], the real and the imaginary parts of the complex spectrogram [55], [107], [109], [172], [242], or directly the raw waveform [108], [277]. Although these approaches allow to incorporate and process the full information of an acoustic signal, research in this area is still active and suggests that there is still room for improvement by exploiting the full information of the noisy speech signal [171], [285].

B. Speaker Embeddings

Since Wang *et al.* [265] showed that an AO system can successfully extract the speech of interest from a mixture signal when conditioned on the *speaker embedding* vector of an enrolment audio signal of the target spreaker, several AV-SE and AV-SS systems have made use of a similar idea. Luo *et al.* [172] showed that i-vectors [48], a low-dimensional representation of a speech signal effective in speaker verification, recognition and diarisation [253], were particularly effective for AV-SS of same gender speakers, obtaining a large improvement over an AV baseline model that did not incorporate speaker embeddings. Afouras *et al.* [10] extracted a compact speaker representation from an enrolment speech signal with the deep-learning-based method in [280] and obtained good performance for mixtures of two and three speakers, especially when face occlusions occurred. In addition, their system could learn the speaker representation on the fly by using the enhanced magnitude spectrogram obtained from a first run of the algorithm without speaker embedding. This essentially bypassed the need for enrolment audio, which is cumbersome or even impossible to collect in certain applications. The approach in [85] also used a pre-trained deep-learning-based model [288] to extract a speaker representation from an additional audio recording. The results indicate that visual information of the speaker's lips is more important than the information contained in the speaker embedding vector, and that their combination led to a general performance improvement. Instead of adopting a pre-trained model, Ochiai *et al.* [196] decided to use a sequence summarising neural network (SSNN) [254], which was jointly trained with the main separation model. Their experiments showed that similar outcomes could be obtained when the enrolment audio and the visual information were used as input in isolation, but better performance was achieved when used at the same time. In general, all these approaches show that speaker embeddings, when extracted from an available additional speech utterance from the target speaker, can be useful, confirming the results obtained in the AO domain [265].

C. Multi-Microphone Features

The spatial information contained in multi-channel acoustic recordings provides an informative cue complementary to spectral information for separating multiple speakers. Specifically, inter-channel phase differences (IPDs) [84], inter-channel time differences (ITDs) [127], inter-channel level differences

[3] Training targets and objective functions are not used during inference.

(ILDs) [127], directional statistics [32] or simply mixture STFT vectors [197] are used in multi-channel deep-learning-based systems to perform SE or SS. Among these features, IPDs are widely applied due to their robustness to reverberation and microphone sensitivities [85]. However, because of the well known issues of spatial aliasing and phase wrapping, IPDs can be the same even for spatially separated sources with different time delays in particular frequencies. This causes fundamental difficulties in separating one source from another. Wang *et al.* [270] proposed to concatenate cosine IPDs (cosIPDs) and sine IPDs (sinIPDs) with log magnitudes as input of their AO system. With this strategy, spectral features can help to resolve the IPDs ambiguity. In addition, the combination of cosIPDs and sinIPDs is preferred over IPDs, because it exhibits a continuous helix structure along frequency due to the Euler formula [269], while IPDs suffer from abrupt discontinuities caused by phase wrapping. In AV-SE and AV-SS, systems used IPDs [85], cosIPDs [107], [247] and sinIPDs [107]. Some AV multi-microphone approaches [85], [247] effectively included also an angle feature [32], which computes the averaged cosine distance between the target speaker steering vector and IPD on all selected microphone pairs.

D. *Shortcomings and Future Research*

As reported above, the vast majority of AV-SE and AV-SS systems use a TF representation of a single-channel acoustic signal as acoustic features. Although a limited number of AV approaches adopt a time-domain signal [108], [277] or multi-microphone cues [85], [107], [247], [283] as acoustic features, there is still room to explore these aspects in future research. In particular, the integration of multi-microphone features with visual information still needs to be investigated further, for example in order to correctly estimate the direction of arrival of the target speech which is hard at low SNRs.

V. Visual Features

Besides acoustic information, AV-SE and AV-SS systems also exploit visual information. In general, the use of vision allows AV systems to obtain a performance improvement over AO systems. A more detailed analysis regarding the actual contribution of vision for AV-SE was conducted in [12]. In particular, visual features were shown to be important to get not only high-level information about speech and silence regions of an utterance, but also fine-grained information about articulation. Although improvements were shown for all *visemes*.[4], sounds that are easier to distinguish visually were the ones that improved the most with an AV-SE system.

The focus of this Section is on visual features, following the list in Table II. Before talking about visual features, we provide information about *face detection* and *tracking*, because a solution of these problems is critical in AV-SE and AV-SS systems.

TABLE II
List of Visual Features in Audio-Visual Speech Enhancement and Separation Papers

Visual Features	AV-SE/SS papers
Raw pixels:	
- Mouth	[12], [66], [76], [77], [85], [99], [123], [129], [167] [168], [179], [181], [182], [227], [247], [266], [278] [225], [226], [283]
- Face	[65]
AAM of mouth region	[137]
2D-DCT of mouth region	[3]–[6]
Optical flow	[17], [65], [157], [167], [168]
Landmark-based features	[100], [157], [186], [207]
Multisensory features	[199]
Face recognition embedding	[55], [109], [172], [196], [242]
VSR embedding	[7], [10], [107]–[109], [156], [227], [277]
Facial appearance embedding	[42], [211]
Compressed mouth frames	[36]
Speaker direction	[85], [247], [283]

A. *Face Detection and Tracking*

Given a video recording, the first step of most AV-SE and AV-SS systems is to determine the number of speakers in it and track their faces across the visual frames. This is usually performed by face detection [141], [163], [257] and tracking [169], [249] algorithms. This approach allows to considerably reduce the dimensionality of the input and, as a consequence, the number of parameters of the SE and the SS models, because only crops of the target faces are considered. In addition, face detection is one way to determine the number of speakers in a scene, an information that can be used by the SS systems that can handle only a fixed number of target speech signals (e.g. [55]), because a priori knowledge of the number of speakers is needed to choose a specific trained multi-speaker model. From these considerations, we can understand the critical importance of face detection and tracking algorithms: if they fail, all the later modules would fail as well. Therefore, robust face tracking, in particular under varying light conditions, occlusions etc. is essential to guarantee high performance in real-world scenarios.

B. *Raw Visual Data*

Once that the video frames of the speaker's face are available, visual features can be used by AV-SE and AV-SS approaches (cf. Table II). Many systems, such as [65] and [66], directly use a crop around the face or the mouth of the target speaker(s) as input, sometimes aligned using an affine transformation [123]. This approach is not always convenient: learning to perform a task from high-dimensional input consisting of raw pixels with a neural network is usually challenging and requires a large amount of data [109], [172]. Hence, several approaches are employed to reduce the input dimensions by extracting different types of features from the raw pixel input, as we report in the following.

C. *Low-Dimensional Visual Features*

Khan *et al.* [137] reduced the dimensionality of the visual information with an active appearance model (AAM) [44], which

[4] A viseme is the basic unit of visual speech and represents what a phoneme is for acoustic speech [176]

is a framework that combines appearance-based and shape-based features through principal component analysis (PCA). Other classical approaches have also been used for visual feature extraction. For example, some works [3]–[6] produced a vector of pixel intensities from the lip region of the speaker with a 2-D discrete cosine transform (DCT). Alternatively, optical flow features were used as an additional input in [65], [157], [167], [168] to explicitly incorporate the motion information in the system.

Research has also been conducted to investigate the use of *facial landmark points*. Hou *et al.* [100] considered a representation of the speaker's mouth consisting of the coordinates of 18 points. Distances for each pair of these points were computed and the 20 elements with the highest variance across an utterance were provided to the SE network. Instead of the distance for each pair of landmark points, Morrone *et al.* [186] obtained a differential motion feature vector by subtracting the face landmark points of a video frame with the points extracted from the previous frame. Motion of landmarks points was also exploited by Li *et al.* [157], who first computed the distance for every symmetric pair of lip landmark points in the vertical and the horizontal directions, and then defined a variation vector of the lip movements consisting of the differences between the distance vectors of two contiguous video frames. This distance-based motion vector was finally combined with aspect ratio features.

A different approach consists of extracting embeddings, i.e. meaningful representations in a typically low dimensional projected space, with a neural network pre-trained on a related task. For example, Owens and Efros [199] proposed to use multisensory features. They designed a deep-learning-based system that could recognise whether the audio and the video streams of a recording were synchronised. The features extracted from such a network provided an AV representation that allows to achieve superior performance compared to an AO-SE approach. Besides multisensory features, embeddings extracted with models trained on face recognition [55] or visual speech recognition (VSR) [7] tasks have been shown to be effective. İnan *et al.* [109] performed a study to evaluate the differences between these two kinds of embeddings. Their results showed that VSR embeddings were able to separate voice activity and silence regions better than face recognition embeddings, which could provide a better distinction between speakers instead. Overall, the performance obtained with VSR embeddings was superior, because they allowed to easier characterise lip movements. Another study [277] further investigated VSR embeddings, showing that the use of features extracted with a model trained for phone-level classification led to better results if compared to the adoption of word-level embeddings.

D. Still Images as Visual Input

Attempts [42], [211] have been made to exploit the information of a still image of the target speaker instead of a video. This approach outperformed a system that used only the audio signals, because there exists a cross-modal relationship between the voice characteristics of a speaker and their facial appearance [140], [198]. This explains why facial features can guide the extraction of the target speech from a mixture. The advantage of using a still image is the reduced complexity of the overall system, although the dynamic information of the video is lost, limiting the system performance considerably.

E. Visual Information in Multi-Microphone Approaches

When the information from *multiple microphones* is available, the location of the target speaker with respect to the microphone array can be used for spatial filtering, i.e. beamforming. In [85], [247], the target direction is estimated with a face detection method. In more complicated scenarios, where people move and turn their heads, face detection might fail over several visual frames. The use of features from the speaker's body might help in building a more robust target source tracker.

F. Shortcomings and Future Research

Current AV-SE and AV-SS approaches only process the visual signal from a single camera. However, previous research on VSR [146], [150] showed that the use of a speaker's profile view can outperform the frontal view. We expect that combining the information from several cameras to capture the different views of a talking face could improve current AV-SE and AV-SS systems. Multi-view input signals were used in approaches for speech reconstruction from silent videos and are reported in Section IX.

Other future challenges include the extraction of features with low complexity algorithms that can be robust to illumination changes, occlusion and pose variations. At the moment, these robustness issues are tackled with a noise-aware training, where the data is artificially modified to include such perturbations [10]. New opportunities to build low-latency systems that are energy-efficient and robust to light changes are given by *event cameras*. In contrast to conventional frame-based cameras, event cameras are asynchronous sensors that output changes in brightness for each pixel only when they occur. They have low latency, high dynamic range and very low power consumption [159]. Arriandiaga *et al.* [17] showed that the SE results obtained with optical flow features, extracted from an event camera, are on par with a frame-based approach. The main limitation of exploiting the full potential of event cameras is that existing image processing algorithms cannot be employed, due to the inherently different nature of the data produced by them. Research in this area is expected to bring novel algorithms and performance improvements.

VI. Deep Learning Methods

As illustrated in Fig. 2, after the feature extraction stage, the actual processing and fusion of acoustic and visual information is performed with a combination of deep neural network models. The main advantage of using these models instead of knowledge-based techniques is the possibility to learn representations of the acoustic and visual modalities at several levels of abstraction and flexibly combine them. Although a detailed exposition of general deep learning architectures and concepts [79] is outside of the scope of this paper, in this Section, we provide a brief

TABLE III
LIST OF DEEP LEARNING METHODS IN AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION PAPERS

Deep Learning Methods	AV-SE/SS papers
FFNN	[3]–[6], [10], [12], [36], [42], [55], [65], [66], [76], [77] [99], [100], [107], [109], [137], [156], [157], [167], [168] [172], [179], [181], [182], [186], [196], [211], [225]–[227] [242], [247], [266], [278]
CNN	[3], [5], [7], [10], [12], [36], [42], [55], [66], [76], [77] [85], [99], [107]–[109], [123], [156], [157], [167], [168] [172], [179], [181], [182], [196], [199], [211], [242], [247] [266], [277], [278], [283]
AE	[36], [66], [107], [123], [129], [179], [181], [182], [199] [225]–[227]
LSTM	[3]–[6], [12], [36], [76], [77], [109], [129], [266]
BiLSTM	[10], [17], [55], [107], [123], [157], [167], [168], [172] [186], [196], [207], [211], [242], [247], [278]
Skip connections	[107], [123], [129], [179], [181], [182], [199]
Residual connections	[7], [10], [42], [65], [85], [107], [108], [123], [129], [156] [247], [277], [283]

presentation of the deep neural network models used in AV-SE and AV-SS systems, as listed in Table III.

A. Feedforward Neural Networks

One of the most used architectures is the feedforward fully-connected neural network (FFNN), also known as multilayer perceptron (MLP). A FFNN consists of several artificial neurons, or *nodes*, organised into a number of *layers*. The network is fully-connected because each node shares a connection with every node belonging to the previous layer. In addition, it is feedforward since the information flows only in one direction from the input layer to the output layer, through the intermediate layers, called *hidden layers*. In order to act as a universal approximator [45], [97], [98], i.e. being able to approximate arbitrarily well any function which maps intervals of real numbers to some real interval, a FFNN needs also to include activation functions, like sigmoid or ReLU, which allow to model potential non-linearities of the function to approximate.

Another kind of feedforward network is the convolutional neural network (CNN) [154]. While in FFNNs each node is connected with all the nodes of the previous layer, CNNs are based on the *convolution operation*, which leverages *sparse connectivity*, *parameter sharing* and *equivariance to translation* [79]. Sometimes, a convolutional layer is followed by a *pooling operation*, which performs a downsampling, for example by local maximisation, to reduce the amount of parameters and obtain *invariance to local transformations*. In AV-SE and AV-SS systems, CNNs are generally used to process the visual frames and automatically extract visual features [278]. They are also adopted for the acoustic signals, to process either the spectrogram [66] or the raw waveform [277]. Since in SE and SS the acoustic input and the output shares a similar structure, some approaches, such as [123], [179], [199], adopted a convolutional autoencoder (AE) architecture, sometimes including skip-connections like in U-Net [221] to allow the information to flow despite the bottleneck.

The training of feedforward neural networks, i.e. the update of the network parameters, is performed e.g. using stochastic gradient descent (SGD) [139], [220] to minimise an objective function (see Section VIII for further details) using the backpropagation algorithm [224] for gradient computation. Variations of SGD are also adopted, in particular RmsProp [248] and Adam [142]. Although increasing the number of hidden layers, i.e. the network *depth*, usually leads to a performance increase [234], two issues often arise: *vanishing/exploding gradient* [22], [74] and *degradation problem* [91]. These issues are generally addressed with *batch normalisation* [111] and *residual connections* [91], respectively, both extensively adopted in AV-SE and AV-SS systems.

B. Recurrent Neural Networks

When dealing with speech signals, a different family of neural networks is also used: recurrent neural networks (RNNs) [224]. The reason is that RNNs were designed to process sequential data. Therefore, they are particularly suitable for speech signals, in which the temporal dimension is important. The training of RNNs is performed with backpropagation through time [273] and, similarly to feedforward neural networks, vanishing/exploding gradient issues are common. The most effective solution to the problem is to introduce paths in which the gradient could flow through time and regulate the propagation of information with *gates*. This class of networks are called gated RNNs, and among them the most adopted are long short-term memory (LSTM) [72], [96] and gated recurrent unit (GRU) [34]. Although these models have a causal structure, architectures in which the output at a given time step depends on the whole sequence, including past and future observations, are also common, and they are known as bidirectional RNNs (BiRNNs) [230], bidirectional LSTMs (BiLSTMs) and bidirectional GRUs (BiGRUs).

C. Shortcomings and Future Research

Compared to knowledge-based approaches, deep learning methods have some disadvantages that we expect to be addressed in future works. First of all, neural network architectures need to be trained with a large amount of data to generalise well to a wide variety of speakers, languages, noise types, SNRs, illumination conditions and face poses. A big step in the evolution of AV-SE and AV-SS systems occurred when researchers started to train the models with large-scale AV datasets [7], [55], [199]. An interesting research direction would be to study the possibility of training deep-learning-based systems with a smaller amount of data without degrading the performance in unknown scenarios [76], [77]. In this context, it would be relevant to explore unsupervised learning techniques, such as the one proposed by Sadeghi et al. [225]–[227], who extended a previous work on AO-SE [155] and adopted variational auto-encoders (VAEs) for AV-SE. In their approach, there is no need of mixing many different noise types with the speech of interest at several SNRs, because the system models directly the clean speech. Despite this attempt, a supervised learning approach that learns a mapping from noisy to clean speech or from a mixture to separated speech signals is still the preferred way to tackle AV-SE and AV-SS, because it allows to reach state-of-the-art performance.

Furthermore, typical paradigms employed for training AV-SE and AV-SS systems assume that the sound sources of a scene are independent from each other. This assumption is adopted for convenience, because collecting actual speech in noise data is costly. However, it is often wrong, since speakers tend to change the way they speak, when they are immersed in a noisy environment, in order to make their speech more intelligible. This phenomenon is known in the literature as *Lombard effect* [25], [166]. Recent work [181], [182] investigated the impact of this effect on data-driven AV-SE models, showing that training a system with Lombard speech is beneficial especially at low SNRs. Therefore, the performance of most deep-learning-based AV-SE systems is affected by the fact that data used for training does not match real conditions.

Another issue especially for low-resource devices is that deep learning models are usually computationally expensive, because data needs to be processed with an algorithm consisting of millions of parameters in order to achieve satisfactory performance. It is important to explore novel ways to reduce the model complexity without reducing the speech quality and intelligibility of the processed signals.

VII. Fusion Techniques

As previously mentioned, AV-SE and AV-SS systems typically consist of a combination of the neural network architectures presented above, which allows to fuse the acoustic and visual information in several ways. In this Section, we present several fusion strategies used in the literature. However, we first introduce the problem of *AV synchronisation*, which is relevant when acoustic and visual data need to be integrated.

A. Audio-Visual Synchronisation

When acoustic and visual signals are recorded with different equipment, an AV synchronisation problem might occur. In other words, audio and video might not be temporally aligned. Humans can detect a lack of synchronisation when the audio leads the video by more than 45 ms or when the video is advanced with respect to the audio by more than 125 ms [114]. AV synchronisation has an impact also on AV-SE and AV-SS performance [7]. Since, in most existing works, the datasets used to train and evaluate AV-SE and AV-SS approaches are properly synchronised, the problem of temporal alignment for AV signals is usually not addressed. In fact, when the audio leads the video or vice versa, it is possible to pre-process the data using the approach proposed in [40]. However, this method might fail at low SNRs [7].

Even when AV signals are temporally aligned, there might still be a need to synchronise acoustic and visual features because the two signals are sampled at different rates. As reported in Section IV, most AV-SE and AV-SS systems use a TF representation of the acoustic signals. In this case, the audio frame rate is determined by the window size and the hop length chosen for the STFT and usually differs from the video frame rate. A common way to solve this problem is to upsample the video frames to match the temporal dimension of the acoustic features [10]. In this respect, the use of time-domain acoustic signals, as done in recent end-to-end deep-learning-based systems, might be beneficial, since it poses fewer constraints than the STFT.

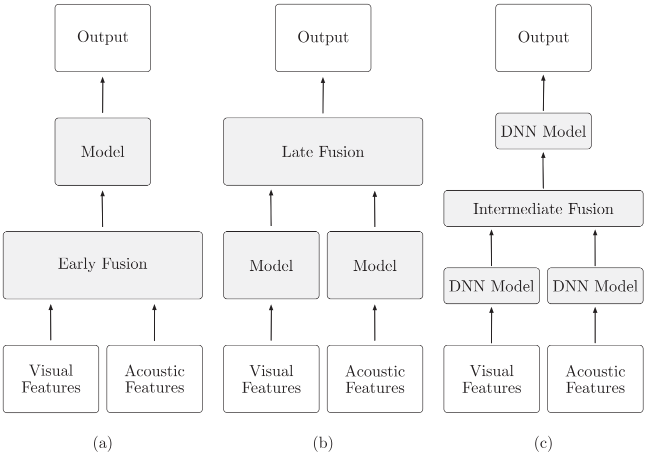


Fig. 3. AV fusion paradigms. (a) Early fusion. (b) Late fusion. (c) Intermediate fusion. DNN model indicates a generic deep neural network model.

B. Traditional Fusion Paradigms

The traditional multimodal fusion approaches are generally grouped into two classes, based on the processing level at which the fusion occurs [161], [212]: *early fusion* and *late fusion*.

As shown in Fig. 3, early fusion consists of combining the information of the different modalities into a joint representation at the feature level. The main advantage is that the correlation between audio and video can be exploited with a single model at a very early stage, making the system more robust if compared to another one that processes the two modalities separately and combines them only at a later stage. Evidence in speech perception suggests that also in humans the AV integration occurs at a very early stage [231]. The disadvantage of early fusion is that usually the features of the two modalities are inherently different. Therefore, appropriate techniques for feature normalisation, transformation and synchronisation need to be developed.

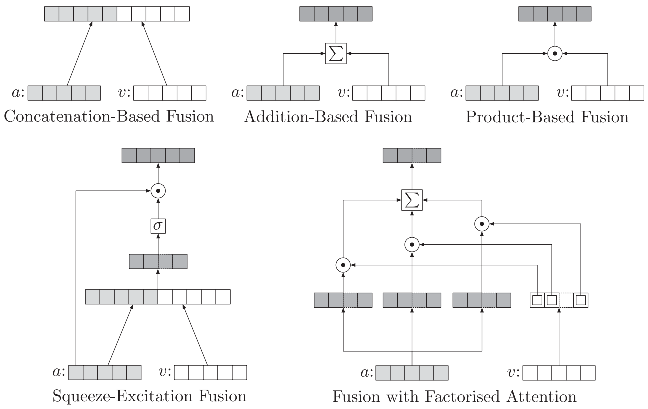
Late fusion, on the other hand, consists of combining the modalities only at the decision level, after that the acoustic and visual information is processed separately with two different models (cf. Fig. 3). Although, from a theoretical perspective, early fusion would be preferable for the reasons mentioned above, late fusion is often used in practice for two reasons: it is possible to use unimodal models designed and validated over the years to achieve the best performance for each modality [129]; it is easy to perform late fusion, because the data processed from the two modalities belongs to the same domain, being different estimates of the same quantity.

C. Fusion Paradigms With Deep Learning

Although some AV-SE and AV-SS works showed that deep learning offers the possibility to perform both early [186] and late [65], [137] fusion, the majority of existing systems (e.g. [7], [55], [66], [129]) exploited the flexibility of deep learning

TABLE IV
LIST OF FUSION TECHNIQUES IN AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION PAPERS

Fusion Techniques	AV-SE/SS papers
Concatenation-based	[3], [5], [7], [10], [12], [17], [36], [42], [55], [76] [77], [85], [99], [100], [107]–[109], [156], [157] [167], [168], [172], [179], [181], [182], [186] [199], [207], [211], [226], [227], [242], [247] [277], [278], [283]
Addition-based	[10], [137]
Product-based	[167], [196], [266]
Squeeze-excitation fusion	[123], [129]
Attention-based	[42], [85], [156], [196], [242]
Integration within a Wiener filtering framework	[3]–[6]



Concatenation-Based Fusion

Addition-Based Fusion

Product-Based Fusion

Squeeze-Excitation Fusion

Fusion with Factorised Attention

Fig. 4. Simplified graphical representation of the main fusion techniques used in audio-visual speech enhancement and separation systems. More details regarding the specific operations in squeeze-excitation fusion and fusion with factorised attention can be found in [123] and in [85], respectively. The symbols a and v indicate acoustic and visual feature vectors, respectively.

techniques and fused the different unimodal representations into a single hidden layer. This fusion strategy is known as *intermediate fusion* [212] (cf. Fig. 3).

Besides the level at which the AV integration occurs, it is important to consider the way in which this integration is performed. Table IV reports a list of the fusion techniques used in the literature, and the most important ones are represented in Fig. 4. The preferred way to fuse the information in AV-SE and AV-SS systems is through *concatenation*. Although this approach is easy to implement, it comes with some potential problems. When two modalities are concatenated, the system uses them simultaneously and treats them in the same way. This means that although, in principle, a deep-learning-based system trained with a very large amount of data should be able to distinguish the cases in which the two modalities are complementary or in conflict [161], in practice we often experience that one modality (not necessarily the most reliable in a given scenario) tends to dominate over the other [62], [66], causing a performance degradation. In AV-SE and AV-SS the acoustic modality is the one that dominates [66], [99]. This is something that might happen also for the approaches that employ an *addition-based fusion*, in which the representations of the multimodal signals are added, with or without weights, not dealing explicitly with the aforementioned issues. Research has been conducted to investigate several possible methods to avoid that one modality dominates over the other. We provide some examples in the following.

Hou *et al.* [99] adopted two strategies. First, they forced the system under development to use both modalities by learning the target speech and the video frames of the speaker mouth at the same time. However, this approach alone does not guarantee that the network discovers AV correlations: it might happen that the network automatically learns to use some hidden nodes to process only the audio modality, and other nodes to process only the video modality. To avoid this selective behaviour, the second strategy adopted in [99] was a multi-style training approach [41], [192], in which one of the input modalities could be randomly zeroed out. Gabbay *et al.* [66] introduced a new training procedure, which consisted of including training samples, in which the noise signal added to the target speech was, in fact, another utterance from the target speaker. Since it is hard to separate overlapping sentences from the same speaker using only the acoustic modality, the network learned to exploit the visual features better. Morrone *et al.* [186] proposed a two-stage training procedure: first, a network was forced to use visual information because it was trained to learn a mapping between the visual features and a target mask to be applied to the noisy spectrogram; then, a new network used the acoustic features together with the visually-enhanced spectrogram obtained from the previous stage to further enhance the speech signal. Wang *et al.* [266] trained two networks separately for each modality to learn target masks and used a gating network to perform a *product-based* fusion, keeping the system performance lower-bounded by the results of the AO network. This approach guaranteed good performance also at high SNRs, where many AV systems fail because acoustic information, which is very strong, and visual information, which is rather weak, is strongly coupled with early or intermediate fusion [266]. Joze *et al.* [129] and Iuzzolino and Koishida [123] proposed the use of squeeze-excitation blocks which generalised the work in [102] for multimodal applications. In particular, each block consisted of two units [129]: a squeeze unit that provided a joint representation of the features from each modality; an excitation unit which emphasised or suppressed the multimodal features from the joint representation based on their importance.

In order to softly select the more informative modality for AV-SE and AV-SS, *attention-based fusion* mechanisms have also been investigated in several works [42], [85], [156], [196], [242]. The attention mechanism [18] was introduced in the field of natural language processing to improve sequence-to-sequence models [34], [243] for neural machine translation. A sequence-to-sequence architecture consists of RNNs organised in an encoder, which reads an input sequence and compresses it into a context vector of a fixed length, and a decoder, which produces an output (i.e. the translated input sequence) considering the context vector generated by the encoder. Such a model fails when the input sequence is long, because the fixed-length context vector acts as a bottleneck. Therefore, Bahdanau *et al.* [18] proposed to use a context vector that preserved the information of all the encoder hidden cells and allowed to align source and target sequences. In this case, the model could attend to salient parts of the input. Besides neural machine translation [18], [173],

[252], attention was later successfully applied to various tasks, like image captioning [256], [281], speech recognition [35] and speaker verification [289]. In the context of AV-SE and AV-SS, two representative works are [42] and [85]. In [42], temporal attention [160] was used, motivated by the fact that different acoustic frames need different degrees of separation. For example, the frames where only the target speech is present should be treated differently from the frames containing overlapped speech or only the interfering speech. In [85], a rule-based attention mechanism [83] was employed to take into account the fact that the significance of each information cue depended on the specific situation that the system needed to analyse. For example, when the speakers were close to each other, spatial and directional features did not provide high discriminability. Therefore, when the angle difference between the speakers was small, the attention weights allowed the model to selectively attend to the more salient cues, i.e. the spectral content of the audio and the lip movements. In addition, a factorised attention was adopted to fuse spatial information, speaker characteristics and lip information at embedding level. The model first factorised the acoustic embeddings into a set of subspaces (e.g., phone and speaker subspaces) and then used information from other cues to fuse them with selective attention.

For completeness, it is relevant to mention approaches that tried to leverage both deep-learning-based and knowledge-based models. For example, Adeel *et al.* [6] used a deep-learning-based model to learn a mapping between the video frames of the target speaker and the filterbank audio features of the clean speech. The estimated speech features were subsequently used in a Wiener filtering framework to get enhanced short-time magnitude spectra of the speech of interest. This approach was extended in [5], where both acoustic and visual modalities were used to estimate the filterbank audio features of the clean speech to be employed by the Wiener filter. The combination of deep-learning-based and knowledge-based approaches was leveraged not only in a single-microphone setup, but also for multi-microphone AV-SS. In [283], a jointly trained combination of a deep learning model and a beamforming module was used. Specifically, a multi-tap minimum variance distortionless response (MVDR) was proposed with the goal of reducing the nonlinear speech distortions that are avoided with a MVDR beamformer [27], but inevitable for pure neural-network-based methods. With the jointly trained multi-tap MVDR, significant improvements of ASR accuracy could be achieved compared to the pure neural-network-based methods for the AV-SS task.

D. Shortcomings and Future Research

The fusion strategies and the design of neural network architectures experimented by researchers still require a lot of expertise. This means that, despite the number of works on AV-SE and AV-SS, researchers might not have explored the best architectures for data fusion. A way to deal with this issue is to investigate the possibility for a more general learning paradigm that focuses not only on determining the parameters of a model, but also on automatically exploring the space of the possible fusion architectures [212].

TABLE V
List of Training Targets and Objective Functions in Audio-Visual Speech Enhancement and Separation Papers

Training Targets	AV-SE/SS papers		
Magnitude spectrogram (DM)	[3]–[6], [36], [66], [99], [100], [179], [199], [207], [247] [278]		
Phase	[7], [10], [156], [199]		
Mask:	MA:	IM:	Other:
- IBM	[76], [77]	–	[65], [137] [167], [168]
- TBM	[186]	–	[65]
- PBM	[266]	–	–
- IRM	[12], [266]	–	[65], [137]
- IAM	[179], [182] [181]	[7], [10], [17], [42] [55], [123], [129] [156], [179], [186] [196], [211]	–
- Ratio mask	–	–	[85], [247] [283]
- PSM	[179]	[107], [157], [179]	–
- CRM	[172]	[55], [107], [109] [242]	[283]
Waveform	[108], [277]		
Mouth frames	[99]		
Compressed mouth frames	[36]		
Objective Functions	**AV-SE/SS papers**		
MSE	[3]–[6], [12], [17], [36], [42], [66], [99], [100], [107], [109] [137], [157], [172], [179], [181], [182], [186], [196], [207] [242], [247], [266], [278]		
MAE	[7], [10], [12], [123], [129], [199], [202]		
Cosine distance/similarity	[7], [10], [12], [156]		
Cross entropy	[76], [77], [107], [186], [266]		
SI-SDR[a]	[85], [108], [247], [277], [283]		
Multitask learning	[42], [99], [196], [207], [266]		
CTC loss	[207]		
Speaker representation loss	[42]		
PIT	[107], [199]		
Deep clustering	[167], [168]		
Triplet loss	[167]		

[a] Applied to the time-domain signal.

In addition, future work should focus on techniques that take into account possible temporal misalignments of AV signals, which might make multimodality fusion critical.

VIII. Training Targets and Objective Functions

As shown in Fig. 2, two other important elements of AV-SE and AV-SS systems are training targets, i.e. the desired outputs of deep-learning-based models, and objective functions, which provide a measure of the distance between the training targets and the actual outputs of the systems. Here, we discuss the adoption of the various training targets and objective functions for AV-SE and AV-SS comprehensively listed in Table V, using the taxonomy proposed in [179].

A. Direct Mapping

Following the terminology of Eq. (6) introduced in Section II (the extension to SS is straightforward), let $A_{k,l} = |X(k,l)|$, $V_{k,l} = |D(k,l)|$ and $R_{k,l} = |Y(k,l)|$ indicate the magnitude of the STFT coefficients for the clean speech, the noise and the noisy speech signals, respectively. A common way to perform the enhancement is by *direct mapping* (DM) [241] (cf. Fig. 5): a system is trained to minimise an objective function reflecting the difference between the output, $\widehat{A}_{k,l}$, and the ground truth, $A_{k,l}$. The most frequently used objective function is the *mean squared error* (MSE), whose minimisation is equivalent

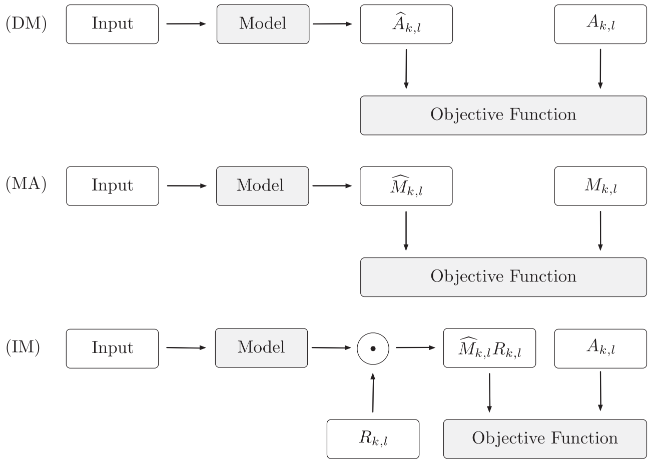


Fig. 5. Illustration of direct mapping (DM), mask approximation (MA) and indirect mapping (IM) approaches. In the specific case of IM, the figure shows the estimation of the ideal amplitude mask. Similar illustrations can be made for different masks.

to maximising the likelihood of the data under the assumption of normal distribution of the errors. Alternatively, some AV models, such as [199], have been trained with the *mean absolute error* (MAE), experimentally proved to increase the spectral detail of the estimates and obtain higher performance if compared to MSE [178], [202].

In order to reconstruct the time-domain signal, an estimate of the target short-time phase is also needed. The noisy phase is usually combined with $\widehat{A}_{k,l}$, since it is the optimal estimator of the target short-time phase [54], under the assumption of Gaussian distribution of speech and noise. However, choosing the noisy phase for speech reconstruction poses limitations to the achievable performance of a system. Iuzzolino and Koishida [123] reported a significant improvement in terms of PESQ and STOI when their system used the target phase instead of the noisy phase to reconstruct the signal. This suggests that modelling the phase could be important in AV applications and some research [7], [10], [156], [199] has moved towards this direction. Specifically, Owens and Efros [199] predicted both the target magnitude log spectrogram and the target phase with their model. Afouras *et al.* [7] designed a sub-network to specifically predict a residual which, when added to the noisy phase, allowed to estimate the target phase. In this case, the phase sub-network was trained to maximise the *cosine similarity* between the prediction and the target phase, in order to take into account the angle between the two. The experiments showed that using the phase estimate was better than using the phase of the input mixture, although there was still room for improvements to match the performance obtained with the ground truth phase.

B. Mask Approximation

An alternative approach to DM consists of using a deep-learning-based model to get an estimate $\widehat{M}_{k,l}$ of a *mask*, $M_{k,l}$. To reconstruct the clean speech signal during inference, $\widehat{M}_{k,l}$ needs to be element-wise multiplied with a TF representation of the noisy signal [179], [267]. This approach is known as *mask approximation* (MA), and an illustration of it is shown in Fig. 5.

In the literature, several masks have been defined in the context of AO-SE [264], [267] and then adopted for AV-SE and AV-SS. One way to build a TF mask is by setting its TF units to binary values according to some criterion. An example is the *ideal binary mask* (IBM) [267], defined as:

$$M_{k,l}^{IBM} = \begin{cases} 1 & \frac{A_{k,l}}{V_{k,l}} > \Gamma(k) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\Gamma(k)$ indicates a predefined threshold. Later, other binary masks have been defined, such as the *target binary mask* (TBM) [143], [267] and the *power binary mask* (PBM) [266]. They have all been adopted as training targets in AV approaches [76], [77], [186], [266] using the *cross entropy loss* as objective function.

Besides binary masks, which are based on the principle of classifying each TF unit of a spectrogram as speech or noise dominated, continuous masks have been introduced for soft decisions. An example is the *ideal ratio mask* (IRM) [267]:

$$M_{k,l}^{IRM} = \left(\frac{A_{k,l}^2}{A_{k,l}^2 + V_{k,l}^2} \right)^{\beta}, \tag{11}$$

where β is a scaling parameter. It is worth mentioning that this mask is heuristically motivated, although its form for $\beta = 1$ has some resemblance with the Wiener filter [165], [264]. IRM has been adopted as training target for a few AV models [12], [266], using either MSE or MAE as objective function. Aldeneh *et al.* [12] proposed the use of a hybrid loss which combined MAE and cosine distance to overcome the limitations of MSE, getting sharp results and bypassing the assumption of statistical independence of the IRM components that the use of MSE or MAE alone would imply.

The IRM does not allow to perfectly recover the magnitude spectrogram of the target speech signal when multiplied with the noisy spectrogram. Hence, the *ideal amplitude mask* (IAM) [267] was introduced:

$$M_{k,l}^{IAM} = \frac{A_{k,l}}{R_{k,l}}. \tag{12}$$

As we discussed previously, the noisy phase is often used to reconstruct the time-domain speech signal. All the masks that we mentioned above do not take the phase mismatch between noisy and target signals into account. Therefore, the *phase sensitive mask* (PSM) [58], [264] and the *complex ratio mask* (CRM) [264], [275] have been proposed. PSM is defined as:

$$M_{k,l}^{\text{PSM}} = \frac{A_{k,l}}{R_{k,l}} \cos(\theta_{k,l}), \tag{13}$$

and tries to compensate for the phase mismatch by introducing a factor, $\cos(\theta_{k,l})$, which is the cosine of the phase difference between the noisy and the clean signals. CRM is the only mask that allows to perfectly reconstruct the complex spectrogram of the clean speech when applied to the complex noisy spectrogram, i.e.:

$$X(k,l) = M_{k,l}^{\text{CRM}} * Y(k,l), \tag{14}$$

where $*$ denotes the complex multiplication, and $M_{k,l}^{\text{CRM}}$ indicates the CRM. IAM, PSM and CRM can be found in several AV systems [172], [179], [181], [182], adopting MSE as objective function.

MA is usually preferred to DM. The reason is that a mask is easier to estimate with a neural network [59], [179], [267]. An exception is AV speech dereverberation (SD), which is addressed only by Tan *et al.* [247], although reverberations have an impact on the signal at the receiver end (cf. the signal model presented in Section II). In this case, the use of a mask, specifically a ratio mask, is discouraged for the two reasons reported in [247]. First, as seen in Eq. (3), reverberation is a convolutive distortion and as such it does not justify the use of ratio masking, which assumes that target speech and interference are uncorrelated [264]. In addition, if a system consists of a cascade of SS and SD modules, such as [247], a ratio mask applied in the SD stage would not be able to easily reduce the artefacts often introduced by SS, because they are correlated with the target speech signal [247].

C. Indirect Mapping

An attempt to exploit the advantages of DM and MA at the same time is done by *indirect mapping* (IM) [241], [272] (cf. Fig. 5). In IM, the model outputs a mask, as in MA, because it is easier to estimate than a spectrogram as mentioned above, but the objective function is defined in the signal domain, as in DM. A comparison between DM, MA and IM for AV-SE was conducted in [179]. In contrast to what one might expect, the results showed that IM did not obtain the best performance among the three paradigms, as observed also in [241], [272] for AO systems. Weninger *et al.* [272] experimentally showed for AO-SS that IM alone performed worse than MA, but it was beneficial when used to fine-tune a system previously trained with the MA objective. Despite these results, AV-SE and AV-SS systems were often trained from scratch with the IM paradigm (cf. Table V) obtaining good results. The reason is probably the use of large-scale datasets, which allowed an optimal convergence of the models.

D. Other Paradigms for Training Targets Estimation

Researchers experimented also with other ways than DM, MA and IM to estimate training targets with a neural network model. For example, Gabbay *et al.* [65] and Khan *et al.* [137] used an estimate of the clean magnitude spectrogram, obtained from visual features with a deep-learning-based model, to build a binary mask that could be applied to the noisy spectrogram. The approaches in [85], [247] can be considered an extension of IM for a time-domain objective. Specifically, a system was trained to output a TF ratio mask using SI-SDR as objective function applied to the reconstructed time-domain signals. The ratio mask obtained with this approach was different from the IAM, because it was not necessarily the one that allowed a perfect reconstruction of the clean magnitude spectrogram. In [247], the system was also trained with an objective that combined MSE on the magnitude spectrograms and SI-SDR on the waveform signals. An objective function in the time domain was also used in [108], [277]. In these cases, a system, inspired by Conv-TasNet [171], was used to directly estimate the waveform of the target speech signal with the SI-SDR training objective.

E. Multitask Learning

Other AV systems [42], [99], [196], [207] tried to improve SE and SS performance with *multitask learning* (MTL) [28], which consists of training a learning model to perform multiple related tasks. Pasa *et al.* [207] investigated MTL using a joint system for AV-SE and ASR. They tried to either jointly minimise a SE objective, MSE, and an ASR objective, *connectionist temporal classification* (CTC) *loss* [80], or alternate the training between an AV-SE stage and an ASR stage. The alternated training was reported to be the most effective. Chung *et al.* [42] used two objective functions to train their system: one was the MSE on magnitude spectrograms and the other was the *speaker representation loss* [188] on embeddings from a network that extracted the speaker identity. Ochiai *et al.* [196] used a combination of losses that allowed their system to work even when either acoustic or visual cues of the speaker were not available. Finally, Hou *et al.* [99] trained their system to both perform AV-SE and reconstruct visual features. As we explained before in Section VII, this approach forces the system to use visual information from the input.

F. Source Permutation

A typical issue for SS is the so-called *source permutation* [92], [286]. This problem occurs in speaker-independent SS systems and it is characterised by an inconsistent assignment over time of the separated speech signals to the sources. Two solutions have been proposed in AO settings: *permutation invariant training* (PIT) [145], [286] and *deep clustering* (DC) [31], [92], [112], [170]. The idea behind PIT is to calculate the objective function for all the possible permutations of the sources and use the permutation associated with the lowest error to update the model parameters. In DC, an embedding vector is learned for each TF unit of the mixture spectrogram and is used to perform clustering to learn an IBM for SS. An extension of DC is the *deep attractor network* [31], which creates attractor points in the embedding space learned from the TF representation of the signal and estimates a soft mask from the similarity between the attractor points and the TF embeddings. Although some AV-SS systems used PIT or DC (cf. Table V), source permutation is less of a problem in deep-learning-based AV-SS, assuming that the target speakers are visible while they talk: visual information is a strong guidance for the systems and allows to automatically assign the separated speech signals to the correct sources.

G. Shortcomings and Future Research

Although many training targets and objective functions have already been investigated for AV-SE and AV-SS, we expect further improvements following several research directions, such as: the use of perceptually motivated objective functions; the estimation of binaural cues to preserve the spatial dimension also at the receiver end; a greater effort for designing and estimating time-domain training targets to perform end-to-end training.

IX. Related Research

In this section, we consider two problems, *speech reconstruction from silent videos* and *audio-visual sound source separation for non-speech signals*, because the first is a special case of AV-SE in which the acoustic input is missing, while the second is the complementary task of AV-SS within the AV sound source separation problem space, considering that the target signals are not speech signals, but, for example, sounds from musical instruments.[5]

Models used for speech reconstruction from silent videos can easily be adopted to estimate a mask for AV-SE and AV-SS. An example is the one presented in [65], where TF masks, obtained by thresholding the reconstructed speech spectrograms, are used to filter the noisy spectrogram. On the other hand, sound source separation for non-speech signals techniques can be adopted also for speech signals by re-training the deep-learning-based models on an AV speech dataset. In some cases, these techniques are domain-specific, such as [67], making the adoption to the speech domain hard. Nevertheless, the ways in which multimodal data is processed and fused can be of inspiration also for AV-SE and AV-SS.

A. Speech Reconstruction From Silent Videos

In some circumstances, the only reliable and accessible modality to understand the speech of interest is the visual one. Real-world scenarios of this kind include, for example: conversations in acoustically demanding situations like the ones occurring during a concert, where the sound from the loudspeakers tends to dominate over the target speech; teleconferences, in which sound segments are missing, e.g. due to audio packet loss [187]; surveillance videos, generally recorded in a situation where the target speaker is acoustically shielded (e.g. with a window) from camera(s) and microphone(s). All these scenarios might be considered as an extreme case of AV-SE where the goal is to estimate the speech of interest from the silent video of a talking face.

In the literature, the problem of estimating speech from visual information is known as speech reconstruction from silent videos. This task is hard because the information that can be captured by a frontal or a side camera is incomplete and cannot include: the excitation signal, i.e. the airflow signal immediately after the vocal chords; most of the the tongue movements. In particular, not having access to tongue movements is critical for speech synthesis, because they are very important for the generation of several speech sounds. For this reason, attempts were made to exploit silent articulations not visible with regular cameras using other sensors[6] [50], [105], [128], [136]. In this subsection we will not consider such silent speech interfaces because they greatly differ from the main topic of this overview. Instead, we focus on deep-learning-based approaches, as listed in Table VI, that directly perform a mapping from videos captured with regular cameras to speech signals.

TABLE VI
Chronological List Of Deep-Learning-Based Approaches for Speech Reconstruction From Silent Videos. MV: Multi-View. SI: speaker-Independent. VSR: Visual Speech Recognition

Paper	Year	Input	Output	Model Info	MV	SI	VSR
[151]	2015	2-D DCT / AAM mouth	LPC or mel-filterbank amplitudes	GMM / FFNN	✗	✗	✗
[152]	2017	AAM mouth	Codebook entries (mel-filterbank amplitudes)	FFNN / RNN	✗	✗	✗
[57]	2017	Raw pixels face	LSP of LPC	CNN, FFNN	✗	✗	✗
[56]	2017	Raw pixels, optical flow face	Mel-scale and linear-scale spectrograms	CNN, FFNN, BiGRU	✗	✗	✗
[11]	2018	Raw pixels face	AE features, spectrogram	CNN, LSTM, FFNN, AE	✗	✗	✗
[147]	2018	Raw pixels mouth	LSP of LPC	CNN, LSTM, FFNN	✓	✗	✗
[149]	2018	Raw pixels mouth	LSP of LPC	CNN, BiGRU, FFNN	✓	✗	✗
[148]	2019	Raw pixels mouth	LSP of LPC	CNN, BiGRU, FFNN	✓	✓	✓
[246]	2019	Raw pixels mouth	WORLD spectrum	CNN, FFNN	✗	✗	✗
[259]	2019	Raw pixels mouth	Raw waveform	GAN, CNN, GRU	✗	✓	✗
[250]	2019	Raw pixels mouth	AE features, spectrogram	CNN, LSTM FFNN, AE	✓	✓	✗
[180]	2020	Raw pixels mouth / face	WORLD features	CNN, GRU, FFNN	✗	✓	✓
[209]	2020	Raw pixels face	mel-scale spectrogram	CNN, LSTM	✓	✗	✗

Le Cornu and Milner [151] were the first to employ a neural network to estimate a speech signal using only the silent video of a speaker's frontal face. They decided to base their system on STRAIGHT [135], a vocoder which allows to perform speech synthesis from a set of time-varying parameters describing fundamental aspects of a given speech signal: fundamental frequency (F0), aperiodicity (AP) and spectral envelope (SP). Supported by the results of some previous works [15], [19], [284], they assumed that only SP could be inferred from visual features. Therefore, AP and F0 were not estimated from the silent video, but artificially produced without taking the visual information into account, while SP was estimated with a Gaussian mixture model (GMM) and FFNN within a regression-based framework. As input to the models, two different visual features were considered, 2-D DCT and AAM, while the explored SP representations were linear predictive coding (LPC) coefficients and mel-filterbank amplitudes. While the choice of visual features did not have a big impact on the results, the use of mel-filterbank amplitudes allowed to outperform the systems based on LPC coefficients.

This work was extended in [152], where two improvements were proposed. First, instead of adopting a regression framework, visual features were used to predict a class label, which in turn was used to estimate audio features from a codebook. Secondly, the influence of temporal information was explored from a feature-level point of view, by grouping multiple frames, and from a model-level point of view, by using RNNs. The obtained improvement in terms of intelligibility was substantial, but the speech quality was still low, mainly because the excitation parameters, i.e. F0 and AP, were produced without exploiting visual cues.

Ephrat and Peleg [57] moved away from a classification-based method as the one presented in [152] and went back

[5] Sometimes, the target signals include singing voices, which typically have different characteristics from speech.

[6] The data used in these works include (but are not limited to) recordings obtained with electromagnetic articulography, electropalatography and laryngography sensors.

to a regression-based framework. Their approach consisted of predicting a line spectrum pairs (LSP) representation of LPC coefficients directly from raw visual data with a CNN, followed by two fully connected layers. Their findings demonstrated that: no hand-crafted visual features were needed to reconstruct the speaker's voice; using the whole face instead of the mouth area as input improved the performance of the system; a regression-based method was effective in reconstructing out-of-vocabulary words. Although the results were promising in terms of intelligibility, the signals sounded unnatural because Gaussian white noise was used as excitation to reconstruct the waveform from LPC features. Therefore, a subsequent study [56] focused on speech quality improvements. In particular, the proposed system was designed to get a linear-scale spectrogram from a learned mel-scale one with the post-processing network in [268]. The time-domain signal was then reconstructed combining an example-based technique similar to [200] with the Griffin-Lim algorithm [81]. Furthermore, a marginal performance improvement was obtained by providing not only raw video frames as input, but also optical flow fields computed from the visual feed.

Another system was developed by Akbari *et al.* [11], who tried to reconstruct natural sounding speech by learning a mapping between the speaker's face and speech-related features extracted by a pre-trained deep AE. The approach was effective and outperformed the method in [57] in terms of speech quality and intelligibility.

The main limitation of these techniques was that they were employed to reconstruct speech of talkers observed by the model at training time. The first step towards a system that could generate speech from various speakers was taken by Takashima *et al.* [246]. They proposed an exemplar-based approach, where a CNN was trained to learn a high-level acoustic representation from visual frames. This representation was used to estimate the target spectrogram with the help of an audio dictionary. The approach could generate a different voice without re-training the neural network model, but by simply changing the dictionary with that of another speaker.

Prajwal *et al.* [209] developed a sequence-to-sequence system adapted from Tacotron 2 [232]. Although their goal was to learn speech patterns of a specific speaker from videos recorded on unconstrained settings, obtaining state-of-the-art performance, they also proposed a multi-speaker approach. In particular, they conditioned their system on speaker embeddings extracted from a reference speech signal as in [126]. Although they could synthesise speech of different speakers, prior information was needed to get speaker embeddings. Therefore, this method cannot be considered a speaker-independent approach, but a speaker-adaptive one.

The challenge of building a speaker-independent system was addressed by Vougioukas *et al.* [259], who developed a generative adversarial network (GAN) that could directly estimate time-domain speech signals from the video frames of the talker's mouth region. Although this approach was capable of reconstructing intelligible speech also in a speaker independent scenario, the speech quality estimated with PESQ was lower than that in [11]. The generated speech signals were characterised by a low-power hum, presumably because the model output was a raw waveform, for which suitable loss functions are hard to find [57].

The method proposed in [180] intended to still be able to reconstruct speech in a speaker independent scenario, but also to avoid artefacts similar to the ones introduced by the model in [259]. Therefore, vocoder features were used as training target instead of raw waveforms. Differently from [151], [152], the system adopted WORLD vocoder [185], which was proved to achieve better performance than STRAIGHT [184], and was trained to predict all the vocoder parameters, instead of SP only. In addition, it also provided a VSR module, useful for all those applications requiring captions. The results showed that a MTL approach, where VSR and speech reconstruction were combined, was beneficial for both the estimated quality and the estimated intelligibility of the generated speech signal.

Most of the systems described above assumed that the speaker constantly faced the camera. This is reasonable in some applications, e.g. teleconferences. Other situations may require a robustness to multiple views and face poses. Kumar *et al.* [147] were the first to make experiments in this direction. Their model was designed to take as input multiple views of the talker's mouth and to estimate a LSP representation of LPC coefficients for the audio feed. The best results in terms of estimated speech quality were obtained when two different views were used as input. The work was extended in [149], where results from extensive experiments with a model adopting several view combinations were reported. The best performance was achieved with the combination of three angles of view (0°, 45° and 60°).

The systems in [147] and in [149] were personalised, meaning that they were trained and deployed for a particular speaker. Multi-view speaker-independent approaches were proposed in [148] and [250]. In both cases, a classifier took as input the multi-view videos of a talker and determined the angles of view from a discrete set of lip poses. Then, a decision network chose the best view combination and the reconstruction model to generate the speech signal. The main difference between the two systems was the audio representation used. While Uttam *et al.* [250] decided to work with features extracted by a pre-trained deep AE, similarly to [11], the approach in [148] estimated a LSP representation of LPC coefficients. In addition, Kumar *et al.* [148] provided a VSR module, as in [180]. However, this module was trained separately from the main system and was designed to provide only one among ten possible sentence transcriptions, making it database-dependent and not feasible for real-time applications.

Despite the research done in this area, several critical points need to be addressed before speech reconstruction from silent videos reaches the maturity required for a commercial deployment. All the approaches in the literature except [209] presented experiments conducted in controlled environments. Real-world situations pose many challenges that need to be taken into account, e.g. the variety of lighting conditions and occlusions. Furthermore, before a practical system can be employed for unseen speakers, performance needs to improve considerably. At the moment, the results for the speaker-independent case are unsatisfactory, probably due to the limited number of speakers used in the training phase.

TABLE VII
CHRONOLOGICAL LIST OF DEEP-LEARNING-BASED APPROACHES FOR AUDIO-VISUAL SOURCE SEPARATION FOR NON-SPEECH SIGNALS. L: LOCALISATION

Paper	Year	Key Idea	L
[68]	2018	Guide source separation with audio frequency bases learned with a framework that maps to visual objects.	✗
[292]	2018	Separate audio sources into components that can be localised in the video frames.	✓
[223]	2019	Perform independent image co-segmentation and sound source separation for not synchronised data.	✓
[69]	2019	Use predicted binaural audio to aid sound source separation.	✓
[205]	2019	Use of a multiple instance learning paradigm for separation and localisation of weakly-labeled data.	✓
[291]	2019	Incorporate temporal motion information and employ a curriculum learning scheme for training.	✓
[282]	2019	Do not separate the sounds independently to avoid that acoustic components from the original mixture get lost.	✓
[70]	2019	Devise a new paradigm to use videos with multiple (correlated) sounds during training.	✗
[235]	2020	Explore conditioning techniques with video stream and weak labels.	✗
[67]	2020	Use keypoint-based structured visual representations to model human-object interactions.	✗
[295]	2020	Refine the separated sounds with cascaded opponent filtering.	✓
[294]	2020	Use an appearance attention module for separation.	✓

B. Audio-Visual Sound Source Separation for Non-Speech Signals

Sound source separation might involve signals different from speech. Imagine, for example, the task of extracting the individual sounds coming from different music instruments playing together. Although the signal of interest is not speech in this case, the approaches developed in this area can provide useful insights also for AV-SE and AV-SS.

Several works addressed AV source separation for non-speech signals. Similarly to other fields, classical methods [20], [29], [203], [204], [210] were recently replaced by deep-learning-based approaches, that we listed in Table VII. The first two works that concurrently proposed deep processing stages for the task under analysis were [68] and [292].

In [68], a novel neural network for multi-instance multi-label learning (MIML) was used to learn a mapping between audio frequency bases and visual object categories. Disentangled audio bases were used to guide a non-negative matrix factorisation (NMF) framework for source separation. The method was successfully employed for in-the-wild videos containing a broad set of object sounds, such as musical instruments, animals and vehicles. NMF was also adopted in a later work by Parekh *et al.* [205], where both audio frequency bases and their activations were used, leveraging temporal information. In contrast to [68], the system could also perform visual *localisation*, which is the task of detecting the sound sources in the visual input.

In [292], audio and video information were jointly used by a deep system called PixelPlayer to simultaneously localise the sound sources in the visual frames and acoustically separate them. The results of this technique sparked a particular interest in the research community, causing the development of several methods aiming at improving it further.

First of all, Rouditchenko *et al.* [223] extended the work in [292] for unsynchronised audio and video data. Their approach consisted of a network that learned disentangled acoustic and visual representations to independently perform visual object co-segmentation and sound source separation.

Then, PixelPlayer only considered semantic features extracted from the video frames. Appearance information is important as highlighted in [294], where the separation was guided with a single image, but higher performance is expected to be achieved when also motion information is exploited. Zhao *et al.* [291] proposed to combine trajectory and semantic features to condition a source separation network. The system was trained with a curriculum learning scheme, consisting of three consecutive stages characterised by increasing levels of difficulty. This approach showed its effectiveness even for separating sounds of the same kind of musical instruments, an achievement not possible in [292]. However, the trajectory motion cues are not able to accurately model the interactions between a human and an object, e.g. a musical instrument. For this reason, Gan *et al.* [67] proposed to use keypoint-based structured visual representations together with the visual semantic context. In this way, they were able to achieve state-of-the-art performance. Motion information was also used in the form of optical flow and dynamic image [23] by Zhu and Rahtu [295]. Their approach refined the separated sounds in multiple stages within a framework called cascaded opponent filter (COF). In addition, they could achieve accurate sound source localisation with a sound source location masking (SSLM) network, following the idea in [103].

Especially when dealing with musical instruments, having a priori knowledge of the presence or absence of a particular instrument in a recording, i.e. weak labels, might be advantageous. Slizovskaia *et al.* [235] studied the problem of source separation conditioned with additional information, which included not only visual cues but also weak labels. Their investigation covered, among other aspects, neural network architectures (either U-Net [221] or multi-head U-Net (MHU-Net) [51]), conditioning strategies (either feature-wise linear modulation (FiLM) [52] or multiplicative conditioning), places of conditioning (at the bottleneck, at all the encoder layers or at the final decoder layer), context vectors (static visual context vector, visual-motion context vector and binary indicator vector encoding the instruments in the mixture) and training targets (binary mask or ratio mask).

The audio signals used in these systems are generally monaural. Inspired by the fact that humans benefit from binaural cues [90], Gao and Grauman [69] proposed a method to exploit visual information with the aim of converting monaural audio into binaural audio. This conversion allowed to expose acoustic spatial cues that turned out to be helpful for sound source separation.

Zhao *et al.* [292] separated the sound sources in the observed mixture assuming that they were independent. This assumption can generate two main issues. The first is that the sum of the separated sounds might be different from the actual mixture, i.e. some acoustic components of the actual mixture might not be found in any outputs of the separation system. Therefore, Xu *et al.* [282] proposed a novel method called MinusPlus network. The idea was to have a two-stage system in which: a minus stage recursively identified the sound with the highest energy and removed it from the mixture; a plus stage refined the removed sounds. The recursive procedure based on sound energy allowed to automatically handle a variable number of sound sources and made the sounds with less energy emerge. The second issue is related to the fact that training is usually performed following a paradigm in which distinct AV clips are randomly mixed. However, sounds that appear in the same scene are usually correlated, e.g. two musical instruments playing the same song. The use of training materials consisting of independent videos might hinder a deep network from capturing such correlations. Hence, Gao and Grauman [70] introduced a new training paradigm, called co-separation, in which an association between consistent sounds and visual objects across pairs of training videos was learned. Exploring this aspect further and possibly overcoming the supervision paradigm used in most of the works in the literature by using real-world recordings and not only synthetic mixtures for training is an interesting future research direction that can easily be adopted also for AV-SE and AV-SS.

X. Audio-Visual Speech Corpora

One of the key aspects that allowed the recent progress and adoption of deep learning techniques for a range of different tasks is the availability of large-scale datasets. Therefore, the choice of a database is critical and it is determined by the specific purpose of the research that needs to be conducted. With this Section, our goal is to provide a non-exhaustive overview of existing resources which will hopefully help the reader to choose AV datasets that suit their purpose. In Table VIII, we provide information regarding AV speech datasets, such as: year of publication, number of speakers, linguistic content, video resolution and frame rate, audio sample frequency, additional characteristics (e.g. recording settings) and the AV-SE and AV-SS papers in which the datasets are used for the experiments.

Most of the datasets provide clean signals of several speakers. Researchers use these signals to create synthetic scenes with overlapping speakers and/or acoustic background noise from the AV speech databases or other sources. We expect that effort will be put in providing AV datasets containing speaker mixtures combined with real noise signals to have a benchmark for AV-SS in noise, like in AO-SS [274].

A. Data in Controlled Environment

From Table VIII, we notice that the two most commonly used databases in the area of deep-learning-based AV-SE and AV-SS are GRID [43] and TCD-TIMIT [89]. GRID consists of audio and video recordings where 34 speakers (18 males and 16 females) pronounce 1000 sentences each. The data was collected in a *controlled environment:* the speakers were placed in front of a plain blue wall inside an acoustically isolated booth and their face was uniformly illuminated. A GRID sentence has the following structure: <command(4)> <color(4)> <preposition(4)> <letter(25)> <digit(10)> <adverb(4)>, where the number of choices for each word is indicated in parentheses. Although the number of possible command combinations using such a sentence structure is high, the vocabulary is small, with only 51 words. This may pose limitations to the generalisation performance of a deep learning model trained with this database. Similar to GRID, TCD-TIMIT consists of recordings in a controlled environment, where the speaker is in front of a plain green wall and their face is evenly illuminated. Compared to GRID, TCD-TIMIT has more speakers, 62 in total (32 males and 30 females, three of which are lipspeakers.[7]), and they pronounce a phonetically balanced group of sentences from the TIMIT corpus.

Other databases have characteristics similar to GRID and TCD-TIMIT (e.g. [1], [13], [16], [208], [228]), but these two are still the most adopted ones, probably for two reasons: the amount of data in them is suitable to train reasonably large deep-learning-based models; their adoption in early AV-SE and AV-SS techniques have made them benchmark datasets for these tasks.

Datasets collected in controlled environments are a good choice for training a prototype designed for a specific purpose or for studying a particular problem. Examples of databases useful in this sense are: TCD-TIMIT [89] and OuluVS2 [16], to study the influence of several angles of view; MODALITY [46] and OuluVS2 [16], to determine the effect of different video frame rates; Lombard GRID [13], to understand the impact of the Lombard effect, also from several angles of view; RAVDESS [164], to perform a study of emotions in the context of SE and SS; KinectDigits [229] and MODALITY [46], to determine the importance that supplementary information from the depth modality might have; ASPIRE [77], to evaluate the systems in real noisy environments.

B. Data in the Wild

More recently, an effort of the research community has been put to gather *data in the wild*, in other words recordings from different sources without the careful planning and setting of the controlled environment used in conventional datasets, like the already mentioned GRID and TCD-TIMIT. The goal of collecting such large-scale datasets, characterised by a vast variety of speakers, sentences, languages and visual/auditory environments, not necessarily in a controlled lab setup, is to have data that resemble real-world recordings. One of the first AV speech in-the-wild databases is LRW [39]. LRW was specifically collected for VSR and consists of around 170 hours of AV material from British television programs. The utterances in the dataset are spoken by hundreds of speakers and are divided into 500 classes. Each sentence of a class contains a non-isolated keyword between 5 and 10 characters. The trend of collecting larger datasets has continued in subsequent collections which consist of materials from British television programs [8], [38], [41], generic YouTube videos [37], [189], TED talks [9], [55],

[7] Lipspeakers are professionals trained to make their mouth movements more distinctive. They silently repeat a spoken talk, making lipreading easier for hearing impaired listeners [89]

TABLE VIII
CHRONOLOGICAL LIST OF THE MAIN AUDIO-VISUAL SPEECH DATASETS. THE LAST COLUMN INDICATES THE AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION ARTICLES WHERE THE DATABASE HAS BEEN USED

Dataset	Year	# Speakers	Linguistic Content	Video	Audio	Additional Characteristics	AV-SE/SS Papers
CUAVE [208]	2002	36 (19 males) and 20 pairs	Connected and isolated digits (7,000 utterances)	720×480 29.97 FPS	Stereo 44 kHz Mono 16 kHz	Controlled environment Speaker movements Simultaneous speakers	[55]
GRID [43]	2006	34 (18 males)	Command sentences (1,000 of 3 seconds per speaker)	720×576 25 FPS	50 kHz	Controlled environment Frontal face	[3], [5], [6], [17], [65], [66], [76], [77] [108], [137], [157], [167], [168], [266] [179], [186], [199], [207], [242]
OuluVS [290]	2009	20 (17 males)	10 everyday greetings (817 sequences)	720×576 25 FPS	48 kHz	Controlled environment Rotating head movements	–
LDC2009V01 [214]	2009	14 (4 males)	Single words and full sentences (7 hours)	720×480 29.97 FPS	48 kHz	Controlled environment Frontal face	[278]
TCD-TIMIT [89]	2015	62 (32 males) 3 lipspeakers	Phonetically rich sentences (13,826 clips)	1920×1080 30 FPS	48 kHz	Controlled environment Straight and 30° camera	[55], [65], [66], [77], [168], [186] [207]
OuluVS2 [16]	2015	53 (40 males)	Continuous digits and sentences	1920×1080 30 FPS 640×480 (front) 100 FPS (front)	48 kHz	Controlled environment Five views	–
KinectDigits [229]	2016	30 (15 males)	English digits 0 – 9	104×80 30 FPS	Four-channel 16 kHz	Controlled environment RGB and depth frames of the mouth region Mic. array recordings	–
LRW [39]	2016	Hundreds	Utterances of 500 different words (173 hours)	256×256 25 FPS	16 kHz	Videos in the wild Recordings from BBC Mostly frontal faces	[107], [108]
Small Mandarin Sentences Corpus [100]	2016	1 male	40 utterances (3-4 seconds each)	320×240	48 kHz	Controlled environment Frontal face Mandarin	[100]
MODALITY [46]	2017	35 (26 males)	Separated commands Continuous sentences (31 hours)	1920×1080 100 FPS 320×240 (ToF) 60 FPS (ToF)	Eight-channel + phone mic. 44.1 kHz	Controlled environment RGB and depth frames Clean and noisy conditions Mic. array recordings	–
NTCD-TIMIT [1]	2017	Extension of TCD-TIMIT obtained by adding six noise types to the corpus: white, babble, car, living room, street and cafe					[157], [225]–[227]
LRS [38]	2017	Several (Not specified[a])	Continuous sentences (75.5 hours)	Not specified[a]	Not specified[a]	Videos in the wild Recordings from BBC Mostly frontal faces	–
MV-LRS [41]	2017	Several (Not specified[a])	Continuous sentences (777.2 hours)	Not specified[a]	Not specified[a]	Videos in the wild Recordings from BBC Multiview	[10]
VoxCeleb [189]	2017	1,251 (690 males)	Continuous sentences (153,516 utterances, 352 hours)	Not specified[b]	Not specified[b]	Videos in the wild from Youtube Challenging multi-speaker acoustic environments	[199]
Mandarin Sentences Corpus [99]	2018	1 male	320 utterances (3-4 seconds each)	1920×1080 30 FPS	48 kHz	Controlled environment Frontal face Mandarin	[55], [66], [77], [99]
Obama Weekly Addresses [66]	2018	1 male	Continuous sentences (300 videos of 2-3 minutes long)	Not specified[c]	Not specified[c]	Wide variety of lighting, face pose, background, scaling and audio recording conditions	[66]
Lombard GRID [13]	2018	54 (24 males)	Command sentences (50 Lombard and 50 plain per speaker)	720×480 (front) 24 FPS (front) 864×480 (side) 30 FPS (side)	48 kHz	Controlled environment Frontal face Lombard effect recordings Straight and side camera	[181], [182]
RAVDESS [164]	2018	24 actors (12 males)	Continuous sentences and songs (2452 audio-visual clips)	1920×1080 30 FPS	48 kHz	Controlled environment Emotional speech and singing at two levels of intensity	–
VoxCeleb2 [37]	2018	6,112 (3,761 males)	Continuous sentences (1,128,246 utterances, 2,442 hours)	Not specified[b]	Not specified[b]	Videos in the wild from Youtube Challenging visual and auditory environments	[7], [42], [123], [129], [156], [172]
LRS2 [8]	2018	Hundreds	Continuous sentences (up to 100 characters each - 224.5 hours)	Not specified[b]	Not specified[b]	Videos in the wild Recordings from BBC	[7], [10], [107], [156], [277]
LRS3 [9]	2018	Around 5,000	Continuous sentences (438 hours)	224×224 25 FPS	16 kHz	Videos from TED and TEDx YouTube channels	[10], [196], [211]
AVSpeech [55]	2018	150,000	Continuous sentences (4,700 hours)	Not specified[b]	Not specified[b]	Videos in the wild Wide variety of people, languages and face poses	[55], [109]
AV Chinese Mandarin [247]	2019	Several (Not specified)	Continuous sentences (155 hours)	Not specified[b]	Not specified[b]	Mandarin lectures from YouTube Grayscale frames of lips	[85], [247], [283]
AVA-ActiveSpeaker [82], [222]	2019	Several (Not specified)	Continuous sentences (38.5 hours)	Not specified[b]	Not specified[b]	Movie and TV videos from YouTube Human-labelled frames	–
ASPIRE [77]	2019	3 (1 male)	Command sentences (6,000 utterances)	1920×1080 30 FPS	44.1 kHz Binaural	Recordings in real noisy places and isolated booth	[75], [77]

[a] We could not get this information because the database is not available to the public due to license restrictions. [b] Since the material is from YouTube, we can expect variable video resolution and audio sample rate. [c] Weekly addresses from The Obama White House YouTube channel. Original video resolution of 1920×1080 at 30 FPS and audio sample rate of 44.1 kHz.

movies [222] or lectures [55], [247]. Among them, AVSpeech is the largest dataset used for AV-SE and AV-SS, with its 4700 hours of AV material. It consists of a wide range of speakers (150 000 in total), languages (mostly English, but also Portuguese, Russian, Spanish, German and others) and head poses (with different pan and tilt angles). Each video clip contains only one talking person and does not have acoustic background interferences.

The large-scale in-the-wild databases, as opposed to the ones containing recordings in controlled environments, are particularly suitable for training deep models that must perform robustly in real-world situations. However, although some in-the-wild datasets are more used than others (as can be seen in the last column of Table VIII), there is not a standard benchmark dataset used to perform the experiments in unconstrained conditions.

XI. Performance Assessment

The main aspects generally of interest for SE and SS are *quality* and *intelligibility*. Speech quality is largely subjective [49], [165] and can be defined as the result of the judgement based on the characteristics that allow to perceive speech according to the expectations of a listener [124]. Given the high number of dimensions that the quality attribute possesses and the different subjective concept of what is high and low quality for every person, a large variability is usually observed in the results of speech quality assessments [165]. On the other hand, intelligibility can be considered a more objective attribute, because it refers to the speech content [165]. Still, a variability in the results of intelligibility assessments can be observed due to the individual speech perception, which has an impact on the ability of recognising words and/or phonemes in different situations.

In the rest of this Section, we review how AV-SE and AV-SS systems are evaluated, with a particular focus on speech quality and speech intelligibility. A summary of the different methods and measures used in the literature are shown in Table X.

A. Listening Tests

A proper assessment of SE and SS systems should be conducted on the actual receiver of the signals. In many scenarios (e.g. hearing assistive devices, teleconferences etc.), the receiver is a human user and *listening tests*, performed with a population of the expected end users, are the most reliable way for the evaluation.

The tests that are currently employed for the assessment of AV-SE and AV-SS systems are typically adopted from the AO domain, i.e. they follow procedures validated for AO-SE and AO-SS techniques. Although different kinds of listening tests exist, some general recommendations include:

- Several subjects are required to be part of the assessment. The number depends on the task, the listeners' experience and the magnitude of the performance differences (e.g. between a system under development and its predecessor) that one wishes to detect. Generally, fewer subjects are required, if they are expert listeners.
- Before the actual test, a training phase allows the subjects to familiarise themselves with the material and the task.
- The speech signals are presented to the listeners in a random order.
- To reduce the impact of listening fatigue, long test sessions are avoided.

TABLE IX
Signal (SIG), Background (BAK) and Overall (OVRL) Quality Rating Scales According to [119]. The Overall Quality Scale is the Same as the Mean Opinion Score Scale

Rating	SIG	BAK	OVRL
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

The most common method used in SE [165] to assess speech quality is the *mean opinion score* (MOS) test [113], [116], [117]. This test is characterised by a five-point rating scale (cf. the 'OVRL' column of Table IX) and was adopted in three AV works [3], [5], [6]. However, the MOS scale was originally designed for speech coders, which introduce different distortions than the ones found in SE [165]. Therefore, an extended standard [119] was proposed and five-point discrete scales were used to rate not only the overall (OVRL) quality (like in the MOS test), but also the signal (SIG) distortion and the background (BAK) noise intrusiveness (cf. Table IX). This kind of assessment was adopted to evaluate the AV system in [99].

A distinct quality assessment procedure, the *multi stimulus test with hidden reference and anchor* (MUSHRA) [115], was used in [75], [77], [181]. In this case, the listeners are presented with speech signals to be rated using a continuous scale from 0 to 100, consisting of 5 equal intervals labelled as 'bad,' 'poor,' 'fair,' 'good,' and 'excellent'. The test is divided into several sessions. In each session, the subjects are asked to rate a fixed number of signals under test (processed and/or noisy speech signals), one hidden reference (the underlying clean speech signal) and at least one hidden anchor (a low-quality version of the reference). In addition, the clean speech signal (i.e. the unhidden reference) is provided. The hidden reference allows to understand whether the subject is able to detect the artefacts of the processed signals, while the hidden anchor provides a lowest-quality fixed-point in the MUSHRA scale, determining the dynamic range of the test. Having the possibility to switch among the signals at will, the listeners can make comparisons with a high degree of resolution.

Together with speech quality, also intelligibility should be assessed with appropriate listening tests. It is possible to group the intelligibility tests into three classes, based on the speech material adopted [165]:

- *Nonsense syllable tests*- Listeners need to recognise nonsense syllables drawn from a list [64], [183]. Usually, it is hard to build such a list of syllables where each item is equally difficult to be identified by the subjects, hence these tests are not very common.
- *Word tests*- Listeners are asked to identify words drawn from a phonetically balanced list [53] or rhyming words [60], [101], [258]. Among these tests, the *diagnostic*

rhyme test (DRT) [258] is extensively adopted to evaluate speech coders [165]. The main criticism about word tests is that they may be unable to predict the intelligibility in real-world scenarios, where a listener is usually exposed to sentences, not single words.

- *Sentence tests*- Listeners are presented with sentences and are asked to identify keywords or recognise the whole utterances. It is possible to distinguish these tests between the ones that use everyday sentences [130], [195] and the ones that use sentences with a fixed syntactical structure, known as *matrix tests* [86], [95], [201], [260]–[262]. One of the most commonly used sentence tests is the *hearing in noise test* (HINT) [195], also adapted for different languages, including Canadian-French [251], Cantonese [276], Danish [193], [194] and Swedish [87].

A simple way to quantify the intelligibility for the previously mentioned tests is by calculating the so-called *percentage intelligibility* [165]. This measure indicates the percentage of correctly identified syllables, words or sentences at a fixed SNR. The main drawback is that it might be hard to find the SNR at which the test can be optimally performed, because floor or ceiling effects might occur if the listeners' task is too hard or too easy. This issue can be mitigated by testing the system at several SNR within a pre-determined range, at the expense of the time needed to conduct the listening experiments. As an alternative, speech intelligibility can be measured in terms of the so-called *speech reception threshold* (SRT), which is the SNR at which listeners correctly identify the material they are exposed to with a 50% accuracy[8] [165]. The SRT is determined with an adaptive procedure, where the SNR of the presented stimuli increases or decreases by a fixed amount at every trial based on the subject's previous response. In this case, the main drawback is that the test is not informative for SNRs that substantially differ from the determined SRT.

Speech intelligibility tests are yet to be adopted by the AV-SE and AV-SS community. In fact, an intelligibility evaluation involving human subjects for AV-SE can only be found in [181]. There, listeners were exposed to speech signals from the Lombard GRID corpus [13] processed with several systems and were asked to determine three keywords in each sentence. The results were reported in terms of percentage intelligibility for four different SNRs distributed in uniform steps between -20 dB and -5 dB.

An important element to consider is the modality in which the stimuli are presented. The listening tests conducted in AV-SE works, like [5], [6], [75], [77], [99], generally used AO signals. Although simpler to conduct, this kind of setting has the disadvantage of completely ignoring the visual information, which has an impact on speech perception [177], [239]. Moreover, it is important to perform tests under the same conditions in which the systems are used in practice. In the situation of human-receiver devices, the user is often exposed to both auditory as well as visual stimuli. Consequently, such systems ought to be tested in natural conditions with human subjects receiving both auditory and corresponding visual stimuli. This is the reason why the tests in [181] were performed with AV signals. However, an AV setup entails a challenging interpretation of the results due to several factors, as highlighted in [106], [181]:

- There is a big difference among individuals in lip-reading abilities. This difference is not reflected in the variation in auditory perception skills [240].
- The per-subject fusion response to discrepancies between the auditory and the visual syllables is large and unpredictable [175].
- The availability of visual information makes ceiling effects more probable to occur.

These considerations suggest a strong need for exploration and development of ecologically valid paradigms for AV listening tests [106], which should reduce the variability of the results and provide a robust and reliable estimation of the performance in real-world scenarios. A first step towards achieving this goal is to perform tests in which the subjects are carefully selected within a homogeneous group and exposed to AV speech signals that resemble actual conversational settings from a visual and an acoustic perspective.

B. Objective Measures

Listening tests are ideal in the assessment performance of SE and SS systems. However, conducting such tests can be time consuming and costly [165], in addition to requiring access to a representative group of end users. Therefore, researches developed algorithmic methods for repeatable and fast evaluation, able to estimate the results of listening tests without listening fatigue effects. Such methods are often called *objective measures* and most of them exploit the knowledge from low-level (e.g. psychoacoustics) and high-level (e.g. linguistics) human processing of speech [165] (cf. Table X).

The most widely used objective measure to assess speech quality for AV-SE and AV-SS is the *perceptual evaluation of speech quality* (PESQ) measure [118], [120], [121], [219]. PESQ was originally designed for telephone networks and codecs. It is a fairly complex algorithm consisting of several components, including level equalisation, pre-processing filtering, time alignment, perceptual filtering, disturbance processing and time averaging. All these steps are used to take into account relevant psychoacoustic principles:

- The frequency resolution of the human auditory system is not uniform, showing a higher discrimination for low frequencies [238].
- Human loudness perception is not linear, meaning that the ability to perceive changes in sound level varies with frequency [296].
- Masking effects might hinder the perception of weak sounds [71].

The output of PESQ is supposed to approximate the MOS score and it is a value generally ranging between 1 and 4.5, although a lower score can be observed for extremely distorted speech signals. Rix *et al.* [219] reported a high correlation with listening tests in several conditions, i.e. mobile, fixed, voice over IP (VoIP) and multiple type networks. A later study [104]

[8] Variants exist where a different percentage is used.

TABLE X

LIST OF THE MAIN PERFORMANCE ASSESSMENT METHODS FOR AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION. THE LAST COLUMN INDICATES THE AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION ARTICLES WHERE THE EVALUATION METHOD HAS BEEN USED

Type	Evaluation Method	Year	Main Characteristics	AV-SE/SS papers
Listening tests for speech quality assessment	MOS [113], [116], [117]	–	Audio-only listening test with 5-point rating scale	[3], [5], [6]
	SIG / BAK / OVRL [119]	2003	Extension of MOS considering signal distorsion and noise intrusiveness	[99]
	MUSHRA [115]	2003	Audio-only listening test with continuous rating scale	[75], [77]
	MUSHRA-like audio-visual test [115], [181]	2019	MUSHRA test using audio-visual stimuli	[181]
Listening tests for speech intelligibility assessment	DRT [258]	1983	Audio-only listening test using rhyming words	–
	HINT [195]	1994	Audio-only listening test using everyday sentences	–
	Matrix-like audio-visual test [181]	2019	Matrix test using audio-visual stimuli [13]	[181]
Estimators of speech quality based on perceptual models	PESQ [118], [120], [121], [219]	2001	Designed to assess quality across a wide range of codecs and network conditions mostly for telephony	[3], [5]–[7], [12], [17], [36], [55], [65] [66], [76], [77], [85], [99], [107], [108] [109], [123], [129], [137], [156], [157] [179], [181], [182], [186], [225]–[227] [242], [247], [266], [278], [283]
	CSIG / CBAK / COVRL [104]	2007	Composite measures which combine basic objective measures	[108]
	HASQI [132], [134]	2010	Specifically designed for hearing-impaired listeners	[99], [100]
	POLQA [122]	2011	PESQ successor	–
	ViSQOL [93], [94]	2012	Specifically designed for voice over IP transmission	[55], [186]
Estimators of speech quality based on energy ratios	SNR (Signal-to-Noise Ratio)	–	It does not provide a proper estimation of speech distortion	[12], [65], [66], [109]
	SSNR / SSNRI (Segmental SNR) (SSNR Improvement)	–	Assessment of short-time behaviour	[100], [108], [242]
	SDI [30]	2006	It provides a rough distortion measure	[99], [100]
	SDR [255]	2006	Specifically designed for blind audio source separation	[7], [10], [17], [42], [55], [65], [85] [107]–[109], [137], [156], [157], [172] [167], [168], [186], [196], [199], [207] [211], [225]–[227]
	SIR [255]	2006	Specifically designed for blind audio source separation	[7], [65], [107], [137], [167], [168] [199]
	SAR [255]	2006	Specifically designed for blind audio source separation	[65], [107], [137], [167], [168], [199]
	SI-SDR [153]	2019	Extension of SDR to make it scale-invariant	[77], [85], [108], [247], [277]
Estimators of speech intelligibility	SII [110]	1997	Used for additive stationary noise or bandwidth reduction	[108]
	CSII [131]	2004	Extension of SII for broadband peak-clipping and center-clipping distortion	[108]
	ESII [213]	2005	Extension of SII for fluctuating noise	[108]
	STOI [244]	2011	Able to predict quite accurately speech intelligibility in several situations	[7], [36], [55], [77], [85], [108], [109] [99], [123], [129], [137]
	HASPI [133]	2014	Specifically designed for hearing-impaired listeners	[99], [100]
	ESTOI [125]	2016	Extension of STOI for highly modulated noise sources	[107], [108], [179], [181], [182], [247]
Automatic speech recognition performance	WER (Word Error Rate)	–	Word-level comparison	[7], [10], [85], [157], [211], [283]
	PER (Phone Error Rate)	–	Phone-level comparison	[207]
Computational efficiency	RTF (Real-Time Factor)	–	Ratio between GPU processing time and audio time.	[85]

showed that PESQ correlates well also with the overall quality of signals processed with common SE algorithms.

As new network and headset technologies were introduced, PESQ was not able to accurately predict speech quality. Therefore, a new measure, the perceptual objective listening quality assessment (POLQA) [122], was introduced. POLQA is considered the successor of PESQ and it is particularly recommended in scenarios where its predecessor performs poorly or cannot be used, e.g. for high background noise, super-wideband speech, variable delay and time scaling. Although POLQA correlates well with listening test results, outperforming PESQ [21], [122], it has not been used to evaluate AV-SE and AV-SS systems yet.

For SS techniques, assessing the overall quality of the processed signals might not be sufficient, because it is desirable to have measures that characterise different speech quality degradation factors. For this reason, the majority of AV-SS systems are evaluated using a set of measures contained in the *blind source separation* (BSS) *Eval* toolkit [255]. The computation of these measures consists of two steps. First, each of the processed signals is decomposed into four terms, representing the components perceived as coming from: the desired speaker, other target speakers (generating cross-talk artefacts), noise sources and other causes (e.g. processing artefacts). The second step provides performance criteria from the computation of energy ratios related to the previous four terms: *source to distortion ratio* (SDR), *source to interferences ratio* (SIR), *sources to noise ratio* and *sources to artefacts ratio* (SAR). Although a reasonable correlation was found between SIR and human ratings of interference [271], other experiments [26], [271] showed that energy-based measures are not ideal for determining perceptual sound quality for SS algorithms.

Besides speech quality estimators, objective intelligibility measures have also been developed. Among them, the *short-time objective intelligibility* (STOI) measure [244] is the most commonly used for AV-SE and AV-SS. STOI is based on the computation of a correlation coefficient between the short-time overlapping temporal envelope segments of the clean and the degraded/processed speech signals. It has been shown that STOI correlates well with the results of intelligibility listening experiments [61], [244], [279]. An extension of STOI, ESTOI, was later proposed [125] to provide a more accurate prediction of speech intelligibility in presence of highly modulated noise sources.

Table X indicates also other measures that we have not presented above, because they are less adopted in AV-SE and AV-SS works. However, it is worth mentioning some of them, since they can be used by researchers to evaluate the systems for specific purposes. For example, the *hearing-aid speech quality index* (HASQI) [132], [134] and the *hearing-aid speech perception index* (HASPI) [133] are two measures that have been specifically designed to evaluate speech quality and and intelligibility as perceived by hearing-impaired listeners. Sometimes, the evaluation of a system is expressed in terms of word error rate (WER) as measured by an ASR system (cf. Table X). This measure assumes that the receiver of the signals is a machine, not a human, and it provides additional performance information for specific applications, e.g. video captioning for teleconferences or augmented reality.

Most of the objective measures used to evaluate AV-SE and AV-SS systems have two main limitations to be desirably addressed in future works. First, they require the target speech signal(s) in order to produce a quality or an intelligibility estimate of the degraded/processed signal(s). These measures are known as *intrusive estimators*. For algorithm development, where clean speech reference is readily available, this assumption is reasonable. However, for in-the-wild tests, it is not possible to collect reference signals and intrusive estimators cannot be adopted.

The other limitation is the use of AO signals in all the objective measures. As already pointed out for listening tests, ignoring the visual component of speech may cause an erroneous estimation of the system performance in many real-world situations, where the listener is able to look at the speaker. In order to develop new predictors of quality and intelligibility in an AV context, a substantial amount of data from AV listening tests is required. When such AV data is available, it would be possible to understand the factors influencing human AV perception of processed speech and properly design and validate new objective measures.

C. Beyond Speech Quality and Intelligibility

When considering SE and SS systems, aspects other than speech quality and intelligibility might be of interest to assess. Some systems, like hearing assistive devices and teleconference systems, have a low-latency requirement, because they need to deliver processed signals to allow real-time conversations. In this case, it might be relevant to report a measure of the *computational efficiency* of the approach under analysis. An example is the so-called real-time factor (RTF), used in [85] and defined as the ratio between the processing time and the duration of the signal.

Sometimes, a given processed speech signal could be fully intelligible, but the effort that the listener must put into the listening task could be substantial in order to be able to understand the speech content. Therefore, it might be important to measure the energy that a subject needs to invest in a listening task, i.e. the *listening effort*. As for speech quality and intelligibility, the listening effort may be measured with listening tests [63], [287].

Moreover, speech carries a lot of additional information, e.g. about the speaker, including *gender*, *age*, *emotional state*, *mood*, their *location* in the scene, etc. These aspects might be important and SE or SS systems should ideally preserve them even after the processing of a heavily corrupted speech signal (cf. [88], in which the proposed system is specifically tested for its ability to preserve spatial cues). Standardised methods for the assessment of these aspects of AV speech are currently lacking, but they would be important to develop in order to guarantee high performance to the end users.

XII. Conclusion

In this paper, we presented an overview of deep-learning-based approaches for audio-visual speech enhancement (AV-SE) and audio-visual speech separation (AV-SS). As expected, visual information provides a benefit for both speech enhancement and separation. In particular, AV-SE systems either outperform their audio-only counterpart for very low signal-to-noise ratios (SNRs) or show similar performance at high SNRs. Performance improvements can be seen across all visemes, with better results for sounds easier to be distinguished visually [12]. Regarding speech separation, audio-visual systems not only outperform their audio-only counterpart, but, since vision is a strong guidance, they are also unaffected by the source permutation problem, occurring when the separated speech signals are assigned inconsistently to the sources.

Throughout the paper, we surveyed a large number of approaches, deliberately avoiding to advocate a method over another based on their performance. This choice was motivated by the fact that a fair comparison of AV-SE and AV-SS approaches is hard, given the wide range of possible applications, each of them having different requirements (e.g. regarding latency, computational complexity and real-time processing). In addition, the lack of standardised audio-visual evaluation procedures, either in the form of listening tests or objective measures, makes the results obtained from such a comparison hardly interpretable and not representative of actual real-world conditions. The design of an audio-visual evaluation framework (cf. the proposal in [106]) would be extremely valuable for the community, but it is clearly outside of the scope of this overview. Instead, we leave the reader to decide which methods to use and further investigate, based on the provided description and discussion of the main elements characterising state-of-the-art systems, namely: acoustic features; visual features; deep learning methods; fusion techniques; training targets and objective functions.

We saw that AV-SE and AV-SS systems generally use the short-time magnitude spectrogram as acoustic input. Since the short-time magnitude spectrogram is not a complete representation of the acoustic signal, some methods exploit the phase information, the complex spectrogram or directly the time-domain signal. Regarding the visual input, although raw data is often used, low-dimensional features are preferred in several works. This choice serves two purposes: first, it allows to reduce the complexity of AV-SE and AV-SS algorithms, since the dimensionality of the data to process is lower; secondly, it makes it possible to train deep learning models for AV-SE and AV-SS with less data, because the low-dimensional features usually adopted are already somewhat robust to several factors, such as illumination conditions, face poses, etc. Future systems, where development data and computational resources may be abundant, might aim for end-to-end training using directly raw visual and acoustic signals as input.

In state-of-the-art AV-SE and AV-SS systems, the actual data processing is obtained with deep-learning-based techniques. Generally, acoustic and visual features are processed separately using two neural network models. Then, the output vectors of these models are fused, often by concatenation, and, afterwards, used as input to another deep learning model. This strategy is convenient, because it is very easy to implement. However, it comes with a major drawback: a simple concatenation does not allow to control how the information from the acoustic and the visual modalities is treated. As a consequence, one of the two modalities may dominate over the other, determining a decrease in the system's performance. Among the strategies adopted to tackle this problem, attention-based mechanisms, which allow the systems to attend to relevant parts of the input, mitigate the potential unbalance caused by concatenation-based fusion.

The last two elements of AV-SE and AV-SS systems are training targets, i.e. the desired output of a deep learning model, and objective functions, i.e. functions that measure the distance between the desired output of a model and its actual output. Although a few approaches tried to directly approximate the target speech signal(s) in time domain, more often a time-frequency (TF) representation of the signals is used. In particular, the deep-learning-based systems are generally trained to minimise the mean squared error (MSE) between the network output and the (potentially transformed) TF coefficients of the training target, which can be either the clean magnitude spectrogram or a mask that is applied to the noisy spectrogram to obtain an enhanced speech signal. Among the two training targets, the latter is usually preferred, because a mask has been empirically found to be easier to estimate with deep learning if compared to the clean magnitude spectrogram.

We also presented three other aspects related to AV-SE and AV-SS, since they can provide additional insights. First, we described deep-learning-based methods used to solve two related tasks: speech reconstruction from silent videos and audio-visual sound source separation for non-speech signals. In particular, we reported a chronological evolution of these fields, because they influenced the first AV-SE and AV-SS approaches and they may still provide a source of inspiration for AV-SE and AV-SS research and vice versa.

Second, we surveyed audio-visual speech datasets, since data-driven methods, like the ones based on deep learning, heavily rely on them. We saw that AV-SE and AV-SS research can still benefit from data collected in a controlled environment to study specific phenomena, like Lombard effect. The general tendency, however, is to use large-scale in-the-wild datasets to make the deep-learning-based systems robust to the variety of conditions that may be present in real-world applications.

Third, we reviewed the principal methodologies used to assess the performance of AV-SE and AV-SS. Specifically, we considered listening tests and objective measures. The former represent the ideal way to assess processed speech signals and must be employed eventually for a realistic system evaluation. However, they are generally time-consuming and costly to conduct. The latter allow to estimate some speech aspects, like quality and intelligibility, in a quick and easily repeatable way, which is highly desirable in the development phase of audio-visual systems. Although many objective measures exist, it might be reasonable to choose the ones that are widely adopted, to make comparisons with previous approaches, and that are reported to correlate well with listening test results. Examples include PESQ, STOI (or its extended version, ESTOI), and SDR (or its scale-invariant definition, SI-SDR). Currently, such objective measures are audio-only. This is in contrast to human communication, which is generally audio-visual.

Finally, we identified several future research directions. Some of them address aspects, such as robustness to a variety of acoustic and visual conditions, to be applied e.g. in teleconferences, and reduction of the computational complexity of deep learning algorithms, especially relevant for low-resource devices like hearing aids. Others, like the investigation of new paradigms for audio-visual fusion, are more focused on a better exploitation of properties and constraints in multimodal systems, and they could, for example, further affect audio-visual speech recognition, audio-visual emotion recognition and audio-visual temporal synchronisation.

ACKNOWLEDGMENT

The authors would like to thank Laurent Girin and the anonymous reviewers for their comments, which helped in improving the overall quality of the article.

REFERENCES

[1] A. H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," in *Proc. INTERSPEECH*, 2017, pp. 3752–3756.

[2] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cogn. Comput.*, vol. 6, no. 2, pp. 200–217, 2014.

[3] A. Adeel, J. Ahmad, H. Larijani, and A. Hussain, “A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids,” *Cogn. Comput*, vol. 12, no. 3, pp. 589–601, 2019.

[4] A. Adeel, M. Gogate, and A. Hussain, “Towards next-generation lip-reading driven hearing-aids: A preliminary prototype demo,” in *Proc. Int. Workshop Challenges Hearing Assistive Technol.*, 2017, pp. 61–64.

[5] A. Adeel, M. Gogate, and A. Hussain, “Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments,” *Inf. Fusion*, vol. 59, pp. 163–170, 2020.

[6] A. Adeel, M. Gogate, A. Hussain, and W. M. Whitmer, “Lip-reading driven deep learning approach for speech enhancement,” *IEEE Trans. Emerg. Topics Comput. Intell.*, to be published, doi: 10.1109/TETCI.2019.2917039.

[7] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” *Proc. INTERSPEECH*, 2018, pp. 3244–3248.

[8] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, [Online]. Available: https://ieeexplore.ieee.org/document/8585066

[9] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: A large-scale dataset for visual speech recognition,” 2018, *arXiv:1809.00496*.

[10] 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.

[11] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “Lip2Audspec: Speech reconstruction from silent lip movements video,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2516–2520.

[12] Z. Aldeneh *et al.*, “Self-supervised learning of visual speech features with audiovisual speech enhancement,” 2020, *arXiv:2004.12031*.

[13] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, “A corpus of audio-visual lombard speech with frontal and profile views,” *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[14] I. Almajai and B. Milner, “Visually derived Wiener filters for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[15] I. Almajai, B. Milner, and J. Darch, “Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise,” in *Proc. INTERSPEECH*, 2006, pp. 2470–2473.

[16] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, “OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis,” in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–5.

[17] A. Arriandiaga, G. Morrone, L. Pasa, L. Badino, and C. Bartolozzi, “Audio-visual target speaker extraction on multi-talker environment using event-driven cameras,” in *Proc. IEEE Int. Symp. Circuits Syst.*, to be published.

[18] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[19] J. P. Barker and F. Berthommier, “Evidence of correlation between acoustic and visual features of speech,” in *Proc. Int. Congr. Phonetic Sci.*, 1999, pp. 199–202.

[20] Z. Barzelay and Y. Y. Schechner, “Harmony in motion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[21] J. G. Beerends *et al.*, “Perceptual objective listening quality assessment (POLQA), the 3rd generation ITU-T standard for end-to-end speech quality measurement part II - Perceptual model,” *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 385–402, 2013.

[22] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[23] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3034–3042.

[24] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acta Acustica United Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[25] H. Brumm and S. A. Zollinger, “The evolution of the Lombard effect: 100 years of psychoacoustic research,” *Behaviour*, vol. 148, no. 11–13, pp. 1173–1198, 2011.

[26] E. Cano, D. FitzGerald, and K. Brandenburg, “Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics,” in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 1758–1762.

[27] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[28] R. Caruana, “Multitask learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[29] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, “Blind audiovisual source separation based on sparse redundant representations,” *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, Aug. 2010.

[30] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[31] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.

[32] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 558–565.

[33] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[34] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[35] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[36] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, “Lite audio-visual speech enhancement,” in *Proc. INTERSPEECH*, 2020, pp. 1131–1135.

[37] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” *Proc. INTERSPEECH*, 2018, pp. 1086–1090.

[38] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3444–3453.

[39] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 87–103.

[40] J. S. Chung and A. Zisserman “Out of time: Automated lip sync in the wild,” in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 251–263.

[41] J. S. Chung and A. Zisserman “Lip reading in profile,” in *Proc. British Mach. Vis. Conf.*, 2017, pp. 1–11.

[42] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “FaceFilter: Audio-visual speech separation using still images,” *Proc. INTERSPEECH*, 2020, pp. 3481-3485.

[43] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[44] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[45] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[46] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, “An audio-visual corpus for multimodal automatic speech recognition,” *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 167–192, 2017.

[47] T. Darrell, J. W. Fisher, and P. Viola, “Audio-visual segmentation and “the cocktail party effect”,” in *Proc. Int. Conf. Multimodal Interfaces*, 2000, pp. 32–40.

[48] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[49] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[50] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[51] C. S. Doire and O. Okubadejo, “Interleaved multitask learning for audio source separation with independent databases,” 2019, *arXiv:1908.05182*.

[52] V. Dumoulin *et al.*, “Feature-wise transformations,*” Distill*, vol. 3, no. 7, p. e 11, 2018.

[53] J. P. Egan, “Articulation testing methods,” *Laryngoscope*, vol. 58, no. 9, pp. 955–991, 1948.

[54] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[55] 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. 112:1–112:11, 2018.

[56] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 455–462.

[57] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 5095–5099.

[58] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[59] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux "Deep recurrent networks for separation and recognition of single-channel speech in non-stationary background audio," in *New Era for Robust Speech Recognition*, Springer, 2017, pp. 165–186.

[60] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *J. Acoust. Soc. Amer.*, vol. 30, no. 7, pp. 596–600, 1958.

[61] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[62] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[63] J. F. Feuerstein, "Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort," *Ear Hearing*, vol. 13, no. 2, pp. 80–86, 1992.

[64] H. Fletcher and J. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, no. 4, pp. 806–854, 1929.

[65] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 3051–3055.

[66] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 1170–1174.

[67] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10475–10484.

[68] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–53.

[69] R. Gao and K. Grauman, "2.5D visual sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 324–333.

[70] R. Gao and K. Grauman "Co-separating sounds of visual objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3878–3887.

[71] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. Boca Raton, FL, USA: CRC Press, 2016.

[72] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.

[73] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Amer.*, vol. 109, no. 6, pp. 3007–3020, 2001.

[74] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[75] M. Gogate, A. Adeel, K. Dashtipour, P. Derleth, and A. Hussain, "AV speech enhancement challenge using a real noisy corpus," 2019, *arXiv:1910.00424*.

[76] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "DNN driven speaker independent audio-visual mask estimation for speech separation," in *Proc. INTERSPEECH*, 2018, pp. 2723–2727.

[77] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "Cochleanet: A robust language-independent audio-visual model for speech enhancement," *Inf. Fusion*, vol. 63, pp. 273–285, 2020.

[78] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party'," *J. Neurosci.*, vol. 33, no. 4, pp. 1417–1426, 2013.

[79] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016.

[80] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[81] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[82] C. Gu *et al.*, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6047–6056.

[83] R. Gu *et al.*, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. INTERSPEECH*, 2019, pp. 4290–4294.

[84] R. Gu *et al.*, "End-to-end multi-channel speech separation," 2019, *arXiv:1905.06286*.

[85] 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.

[86] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scand. Audiol.*, vol. 11, no. 2, pp. 79–87, 1982.

[87] M. Hällgren, B. Larsby, and S. Arlinger, "A Swedish version of the hearing in noise test (HINT) for measurement of speech recognition," *Int. J. Audiol.*, vol. 45, no. 4, pp. 227–237, 2006.

[88] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6404–6408.

[89] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[90] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Amer.*, vol. 115, no. 2, pp. 833–843, 2004.

[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[92] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[93] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The virtual speech quality objective listener," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2012, pp. 1–4.

[94] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, pp. 1–18, 2015.

[95] S. Hochmuth, T. Brand, M. A. Zokoll, F. Z. Castro, N. Wardenga, and B. Kollmeier, "A spanish matrix sentence test for assessing speech reception thresholds in noise," *Int. J. Audiol.*, vol. 51, no. 7, pp. 536–544, 2012.

[96] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[97] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.

[98] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[99] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[100] J.-C. Hou *et al.*, "Audio-visual speech enhancement using deep neural networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[101] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, no. 1, pp. 158–166, 1965.

[102] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[103] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3868–3877.

[104] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[105] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Comput. Speech Lang.*, vol. 36, pp. 274–293, 2016.

[106] A. Hussain *et al.*, "Towards multi-modal hearing aid design and evaluation in realistic audio-visual settings: Challenges and opportunities," in *Proc. 1st Int. Conf. Challenges Hearing Assistive Technol.*, 2017, pp. 29–34.

[107] E. Ideli, "Audio-visual speech processing using deep learning techniques." M.S. thesis, Dept. Appl. Sci.: Sch. of Eng. Sci., 2019.

[108] E. Ideli, B. Sharpe, I. V. Bajić, and R. G. Vaughan, "Visually assisted time-domain speech enhancement," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2019, pp. 1–5.

[109] B. İnan, M. Cernak, H. Grabner, H. P. Tukuljac, R. C. Pena, and B. Ricaud, "Evaluating audiovisual source separation in the context of video conferencing," in *Proc. INTERSPEECH*, 2019, pp. 4579–4583.

[110] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[111] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[112] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *Proc. INTERSPEECH*, 2016, pp. 545–549.

[113] *Subjective Assessment of Sound Quality*, Rec. ITU-R BS.562, International Telecommunications Union, Geneva, Switzerland, 1990.

[114] *Relative Timing of Sound and Vision for Broadcasting*, Rec. ITU-R BT.1359-1, International Telecommunications Union, Geneva, Switzerland, 1998.

[115] *Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*, Rec. ITU-R BS.1534-1, International Telecommunications Union, Geneva, Switzerland, 2003.

[116] *General Methods for the Subjective Assessment of Sound Quality*, Rec. ITU-R BS.1284-2, International Telecommunications Union, Geneva, Switzerland, 2019.

[117] *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*, Rec. ITU-T P.830, International Telecommunications Union, Geneva, Switzerland, 1996.

[118] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Rec. ITU-T P.862, International Telecommunications Union, Geneva, Switzerland, 2001.

[119] *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm*, Rec. ITU-T P.835, International Telecommunications Union, Geneva, Switzerland, 2003.

[120] *Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, Rec. ITU-T P.862.1, International Telecommunications Union, Geneva, Switzerland, 2003.

[121] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, Rec. ITU-T P.862.2, International Telecommunications Union, Geneva, Switzerland, 2005.

[122] *Perceptual Objective Listening Quality Assessment*, Rec. ITU-T P.863, International Telecommunications Union, Geneva, Switzerland, 2011.

[123] M. L. Iuzzolino and K. Koishida, "AV(SE)^2: Audio-visual squeeze-excite speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7539–7543.

[124] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Berlin/Heidelberg: Springer, 2006.

[125] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[126] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 4485–4495.

[127] Y. Jiang and R. Liu, "Binaural deep neural network for robust speech enhancement," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput.*, 2014, pp. 692–695.

[128] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Commun.*, vol. 52, no. 4, pp. 354–366, 2010.

[129] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13286–13296.

[130] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[131] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2604–2604, 2004.

[132] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, 2010.

[133] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, 2014.

[134] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, 2014.

[135] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[136] C. T. Kello and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2354–2364, 2004.

[137] F. U. Khan, B. P. Milner, and T. Le Cornu, "Using visual speech information in masking methods for audio speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1742–1754, Oct. 2018.

[138] M. S. Khan, S. M. Naqvi, W. Wang, A. ur-Rehman, and J. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1900–1912, Sep. 2013.

[139] J. Kiefer, J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466, 1952.

[140] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 276–292.

[141] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[142] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[143] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[144] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[145] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[146] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. IV-429–IV-432.

[147] Y. Kumar, M. Aggarwal, P. Nawal, S. Satoh, R. R. Shah, and R. Zimmermann, "Harnessing AI for speech reconstruction using multi-view silent video feed," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1976–1983.

[148] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 2588–2595.

[149] Y. Kumar, R. Jain, M. Salik, R. R. Shah, R. Zimmermann, and Y. Yin, "MyLipper: A personalized system for speech reconstruction using multi-view visual feeds," in *Proc. IEEE Int. Symp. Multimedia*, 2018, pp. 159–166.

[150] Y. Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2012, pp. 432–437.

[151] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. INTERSPEECH*, 2015, pp. 3355–3359.

[152] T. Le Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 9, pp. 1751–1761, Sep. 2017.

[153] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR-half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[154] Y. LeCun, "Generalization and network design strategies," *Connectionism Perspective*, vol. 19, pp. 143–155, 1989.

[155] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process.*, 2018, pp. 1–6.

[156] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7314–7318.

[157] Y. Li, Z. Liu, Y. Na, Z. Wang, B. Tian, and Q. Fu, "A visual-pilot deep fusion for target speech separation in multitalker noisy environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4442–4446.

[158] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, p. 183, 2012.

[159] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[160] Z. Lin *et al.*, "A structured self-attentive sentence embedding," *Proc. Int. Conf. Learn. Representations*, 2017.

[161] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *Proc. KDD BigMine*, 2018.

[162] Q. Liu, W. Wang, P. J. Jackson, M. Barnard, J. Kittler, and J. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5520–5535, Nov. 2013.

[163] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[164] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, 2018, Art. no. e0196391.

[165] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2013.

[166] E. Lombard, "Le signe de l'elevation de la voix," *Annales des Mal. de L'Oreille et du Larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[167] R. Lu, Z. Duan, and C. Zhang, "Listen and look: Audio-visual matching assisted speech source separation," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1315–1319, Sep. 2018.

[168] R. Lu, Z. Duan, and C. Zhang, "Audio-visual deep clustering for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1697–1712, Nov. 2019.

[169] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[170] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.

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

[172] Y. Luo, J. Wang, X. Wang, L. Wen, and L. Wang, "Audio-visual speech separation using i-Vectors," in *Proc. IEEE 2nd Int. Conf. Inf. Commun. Signal Process.*, 2019, pp. 276–280.

[173] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[174] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2257–2269, Nov. 2007.

[175] D. B. Mallick, J. F. Magnotti, and M. S. Beauchamp, "Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type," *Psychon. Bull. Rev.*, vol. 22, no. 5, pp. 1299–1307, 2015.

[176] D. W. Massaro and J. A. Simpson, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. New York/London: Psychology Press, 2014.

[177] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[178] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. INTERSPEECH*, 2017, pp. 2008–2012.

[179] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "On training targets and objective functions for deep-learning-based audio-visual speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8077–8081.

[180] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-based speech synthesis from silent videos," in *Proc. INTERSPEECH*, 2020, pp. 3530–3534.

[181] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect," *Speech Commun.*, vol. 115, pp. 38–50, 2019.

[182] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Effects of Lombard reflex on the performance of deep-learning-based audio-visual speech enhancement systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6615–6619.

[183] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.

[184] M. Morise and Y. Watanabe, "Sound quality comparison among high-quality vocoders by using re-synthesized speech," *Acoust. Sci. Technol.*, vol. 39, no. 3, pp. 263–265, 2018.

[185] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[186] 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.*, 2019, pp. 6900–6904.

[187] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, "Audio-visual speech inpainting with deep learning," *Proc. Int. Conf. Acoust., Speech Signal Process.*, to be published.

[188] 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.*, 2020, pp. 7289–7293.

[189] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017, pp. 2616–2620.

[190] S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. A. Chambers, "Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking," *IET Signal Process.*, vol. 6, no. 5, pp. 466–477, 2012.

[191] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 895–910, Oct. 2010.

[192] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28rd Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[193] J. B. Nielsen and T. Dau, "Development of a Danish speech intelligibility test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, 2009.

[194] J. B. Nielsen and T. Dau, "The Danish hearing in noise test," *Int. J. Audiol.*, vol. 50, no. 3, pp. 202–208, 2011.

[195] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[196] 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.

[197] T. Ochiai, S. Watanabe, and S. Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[198] T.-H. Oh *et al.*, "Speech2Face: Learning the face behind a voice," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7531–7540.

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

[200] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2405–2413.

[201] E. Ozimek, A. Warzybok, and D. Kutzner, "Polish sentence matrix test for speech intelligibility measurement in noise," *Int. J. Audiol.*, vol. 49, no. 6, pp. 444–454, 2010.

[202] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5414–5418.

[203] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Guiding audio source separation by video object information," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 61–65.

[204] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6–10.

[205] S. Parekh, A. Ozerov, S. Essid, N. Q. Duong, P. Pérez, and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 268–272.

[206] S. Partan and P. Marler, "Communication goes multimodal," *Science*, vol. 283, no. 5406, pp. 1272–1273, 1999.

[207] L. Pasa, G. Morrone, and L. Badino, "An analysis of speech enhancement and recognition losses in limited resources multi-talker single channel audio-visual ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7309–7313.

[208] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2002, pp. II-2017–II-2020.

[209] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13793–13802.

[210] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2901–2905.

[211] L. Qu, C. Weber, and S. Wermter, "Multimodal target speech separation with voice and face references," *Proc. INTERSPEECH*, 2020, pp. 1416–1420.

[212] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[213] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[214] C. Richie, S. Warburton, and M. Carter, *Audiovisual Database of Spoken American English.* Linguistic Data Consortium, 2009.

[215] J. Rincón-Trujillo and D. M. Córdova-Esparza, "Analysis of speech separation methods based on deep learning," *Int. J. Comput. Appl.*, vol. 148, no. 9, pp. 21–29, 2019.

[216] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[217] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Commun.*, vol. 49, no. 7–8, pp. 667–677, 2007.

[218] 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, May 2014.

[219] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[220] H. Robbins and S. Monro, "A stochastic approximation method," Ann. Math. Statist., vol. 22, no. 3, pp. 400–407, 1951.

[221] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Assist. Interv.*, 2015, pp. 234–241.

[222] J. Roth *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4492–4496.

[223] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2357–2361.

[224] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[225] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for VAE-based audio-visual speech enhancement," 2019, *arXiv:1912.10647*.

[226] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7534–7538.

[227] 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, Lang. Process.*, vol. 28, pp. 1788–1800, 2020.

[228] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognit.*, vol. 36, no. 2, pp. 293–302, 2003.

[229] L. Schönherr, D. Orth, M. Heckmann, and D. Kolossa, "Environmentally robust audio-visual speaker identification," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 312–318.

[230] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[231] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Audio-visual scene analysis: Evidence for a "very-early" integration process in audio-visual speech perception," in *Proc. 7th Int. Conf. Spoken Lang. Process. - INTERSPEECH*, 2002, pp. 1937–1940.

[232] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[233] B. G. Shinn-Cunningham and V. Best, "Selective attention in normal and impaired hearing," *Trends Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[234] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[235] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for music instrument performances," 2020, *arXiv:2004.03873*.

[236] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Commun.*, vol. 44, no. 1–4, pp. 113–125, 2004.

[237] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Adv. Signal Process.*, no. 1, pp. 1165–1173, 2002.

[238] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[239] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, 1954.

[240] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. Roy. Soc. London. Ser. B: Biol. Sci.*, vol. 335, no. 1273, pp. 71–78, 1992.

[241] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.

[242] Z. Sun, Y. Wang, and L. Cao, "An attention based speaker-independent audio-visual deep learning model for speech enhancement," in *Proc. of Int. Conf. Multimedia Model.*, 2020, pp. 722–728.

[243] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[244] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[245] T. M. F. Taha and A. Hussain, "A survey on techniques for enhancing speech," *Int. J. Comput. Appl.*, vol. 179, no. 17, pp. 1–14, 2018.

[246] Y. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based lip-to-speech synthesis using convolutional neural networks," in *Proc. IW-FCV*, 2019. [Online]. Available: https://mr.hanyang.ac.kr/IW-FCV2019/

[247] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 542–553, Mar. 2020.

[248] T. Tieleman and G. Hinton, "Lecture 6.5 - RmsProp: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[249] C. Tomasi and T. Kanade, "Detection and tracking of point features," Tech. Rep. CMU-CS-91-132, 1991.

[250] S. Uttam *et al.*, "Hush-hush speak: Speech reconstruction using silent videos," in *Proc. INTERSPEECH*, 2019, pp. 136–140.

[251] V. Vaillancourt *et al.*, "Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations," *Int. J. Audiol.*, vol. 44, no. 6, pp. 358–361, 2005.

[252] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[253] P. Verma and P. K. Das, "i-Vectors in speech processing applications: A survey," *Int. J. Speech Technol.*, vol. 18, no. 4, pp. 529–546, 2015.

[254] K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5315–5319.

[255] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[256] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A. neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[257] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[258] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," Speech Technol., vol. 1, pp. 30–39, 1983.

[259] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," in *Proc. INTERSPEECH*, 2019, pp. 4125–4129.

[260] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und evaluation eines satztests in deutscher sprache - Teil II: Optimierung des Oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, no. 38, pp. 44–56, 1999.

[261] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und evaluation eines satztests in deutscher sprache - Teil III: Evaluierung des Oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, no. 38, pp. 86–95, 1999.

[262] K. Wagener, V. Kühnel, and B. Kollmeier, "Entwicklung und evaluation eines satztests in deutscher sprache - Teil I: Design des Oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, no. 38, pp. 4–15, 1999.

[263] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Piscataway: Wiley-IEEE Press, 2006.

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

[265] Q. Wang *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*, 2019, pp. 2728–2732.

[266] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7529–7533.

[267] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[268] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[269] Z.-Q. Wang, "Deep learning based array processing for speech separation, localization, and recognition," Ph.D. dissertation, The Ohio State Univ., 2020.

[270] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1–5.

[271] D. Ward, H. Wierstorf, R. D. Mason, E. M. Grais, and M. D. Plumbley, "BSS Eval or PEASS? Predicting the perception of singing-voice separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 596–600.

[272] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.

[273] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[274] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. INTERSPEECH*, 2019, pp. 1368–1372.

[275] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[276] L. L. Wong and S. D. Soli, "Development of the Cantonese hearing in noise test (CHINT)," *Ear Hearing*, vol. 26, no. 3, pp. 276–289, 2005.

[277] J. Wu *et al.*, "Time domain audio visual speech separation," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 667–673.

[278] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-modal hybrid deep neural network for speech enhancement," 2016, *arXiv:1606.04750*.

[279] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in Mandarin," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4465–4468.

[280] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5791–5795.

[281] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[282] X. Xu, B. Dai, and D. Lin, "Recursive visual sound separation using minus-plus net," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 882–891.

[283] Y. Xu *et al.*, "Neural spatio-temporal beamformer for target speech separation," in *Proc. INTERSPEECH*, 2020, pp. 56–60.

[284] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1, pp. 23–43, 1998.

[285] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9458–9465.

[286] D. Yu, M. Kolbæk, 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.*, 2017, pp. 241–245.

[287] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," *Ear Hearing*, vol. 31, no. 4, pp. 480–490, 2010.

[288] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.

[289] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 171–178.

[290] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.

[291] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1735–1744.

[292] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 570–586.

[293] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," 2020, *arXiv:2001.04758*.

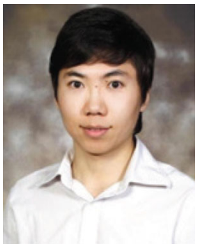
[294] L. Zhu and E. Rahtu, "Separating sounds from a single image," 2020, *arXiv:2007.07984*.

[295] L. Zhu and E. Rahtu, "Visually guided sound source separation using cascaded opponent filter network," *Proc. Asian Conf. Comput. Vis.*, 2020.

[296] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, vol. 22. Springer Science & Business Media, 2013.

Daniel Michelsanti (Member, IEEE) received the B.Sc. degree in computer science and electronic engineering (cum laude) from the University of Perugia, Perugia, Italy, in 2014, and the M.Sc. degree in vision, graphics and interactive systems and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2017 and 2021, respectively. He is currently a Research Assistant with the section for Artificial Intelligence and Sound, Department of Electronic Systems, Aalborg University. His research interests include multimodal speech enhancement and machine learning, specifically deep learning.

Zheng-Hua Tan (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 1999. He is currently a Professor with the Department of Electronic Systems and the Co-Head of the Centre for Acoustic Signal Processing Research with Aalborg University, Aalborg, Denmark. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA, an Associate Professor with SJTU, Shanghai, China, and a Postdoctoral Fellow with KAIST, Daejeon, Korea. He has authored or coauthored more than 200 refereed publications. His research interests include machine learning, deep learning, pattern recognition, speech and speaker recognition, noise-robust speech processing, multimodal signal processing, and social robotics. He is the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He was an Editorial Board Member for Computer Speech and Language and was the Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and *Neurocomputing*. He was the General Chair of IEEE MLSP 2018 and the TPC Co-Chair for IEEE SLT 2016.



Shi-Xiong Zhang (Member, IEEE) received the M.Phil. degree in electronic and information engineering from The Hong Kong Polytechnic University, Hong Kong, in 2008 and the Ph.D. degree from Machine Intelligence Laboratory, Engineering Department, Cambridge University, Cambridge, U.K., in 2014. From 2014 to 2018, he was a Senior Speech Scientist with Speech Group, Microsoft. He is currently a Principal Researcher with Tencent America. His research interests include speech recognition, speaker verification, speech separation, multimodal learning, and machine learning particularly structured prediction, graphical models, kernel methods, and Bayesian nonparametric methods. He was the recipient of the IC Greatness Award in Microsoft in 2015 and the Best Paper Award in 2008 IEEE Signal Processing Postgraduate Forum for the paper "Articulatory-Feature based Sequence Kernel For High-Level Speaker Verification." He was nominated a 2011 Interspeech Best Student Paper Award for the paper "Structured Support Vector Machines for Noise Robust Continuous Speech Recognition."

Yong Xu (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2015. He is currently a Senior Researcher with Tencent AI lab, Bellevue, WA, USA. From April 2016 to May 2018, he was a Research Fellow with the University of Surrey, Guildford, U.K. From September 2014 to May 2015, he was with the Georgia Institute of Technology, Atlanta, GA, USA. From 2015 to 2016, he was with iFLYTEK as a Researcher. His current research interests include deep learning based speech enhancement, speech separation, noise robust speech recognition, and sound event detection. He was the recipient of the IEEE Signal Processing Society 2018 Best Paper Award for the work on deep neural networks based speech enhancement.

Meng Yu received the B.S. degree in computational mathematics from Peking University, Beijing, China, in 2007 and the Ph.D. degree in applied mathematics from the University of California Irvine, Irvine, CA, USA, in 2012. He is currently a Principal Research Scientist with Tencent AI Lab, Bellevue, WA, USA, working on far field frontend speech processing, deep learning based speech enhancement and separation, and their joint optimization with keyword spotting, speaker verification, and acoustic model of speech recognition.

Dong Yu (Fellow, IEEE) is currently a Distinguished Scientist and the Vice General Manager with Tencent AI Lab, Bellevue, WA, USA. Prior to joining Tencent in 2017, he was a Principal Researcher with Microsoft Research (Redmond), Microsoft, where he joined in 1998. He has authored or coauthored two monographs and more than 250 papers. His current research focuses on speech recognition and processing. His works have been cited for more than 30 000 times per Google Scholar and have been recognized by the prestigious IEEE Signal Processing Society 2013 and 2016 Best Paper Award. He is currently the Vice Chair of the IEEE Speech and Language Processing Technical Committee. He was a Member of the IEEE SLPTC from 2013 to 2018, a Distinguished Lecturer of APSIPA from 2017 to 2018, an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2011 to 2015, an Associate Editor for the *IEEE Signal Processing Magazine* from 2008 to 2011, the Lead Guest Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING - Special Issue on Deep Learning for Speech and Language Processing from 2010 to 2011, the Guest Editor of the IEEE/CAA JOURNAL OF AUTOMATICA SINICA - Special Issue on Deep Learning in Audio, Image, and Text Processing from 2015 to 2016, and Member of organization and technical committees of many conferences and workshops.

Jesper Jensen (Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation, Aalborg University, as a Ph.D. Student and an Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and an Assistant Professor with the Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. He is currently a Senior Principal Scientist with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is a Professor with the Section for Artificial Intelligence and Sound, Department of Electronic Systems, Aalborg University. He is also the Co-Founder of the Centre for Acoustic Signal Processing Research, Aalborg University. His main research interests include acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.