Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A. Chambers

Audiovisual Speech Source Separation

An overview of key methodologies

The separation of speech signals measured at multiple microphones in noisy and reverberant environments using only the audio modality has limitations because there is generally insufficient information to fully discriminate the different sound sources. Humans mitigate this problem by exploiting the visual modality, which is insensitive to background noise and can provide contextual information about the audio scene. This advantage has inspired the creation of the new field of audiovisual (AV) speech source separation that targets exploiting visual modality alongside the microphone measurements in a machine. Success in this emerging field will expand the application of voice-based machine interfaces, such as Siri, the intelligent personal assistant on the iPhone and iPad, to much more realistic settings and thereby provide more natural human–machine interfaces.

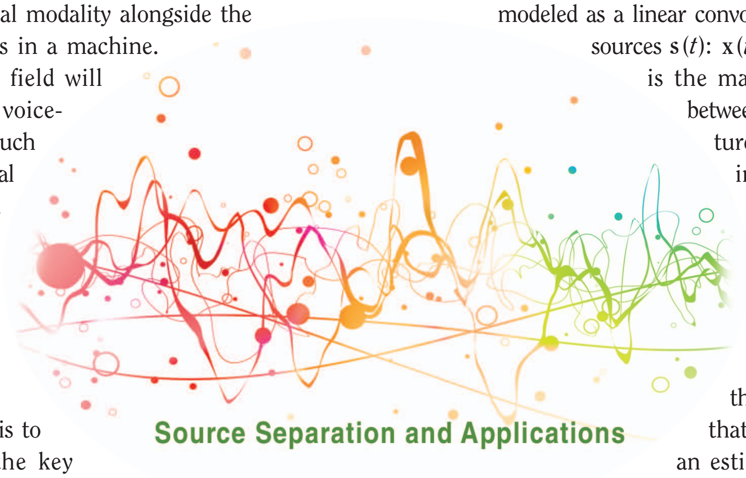


IMAGE LICENSED BY INGRAM PUBLISHING

INTRODUCTION

The purpose of this article is to provide an overview of the key methodologies in AV speech source separation building from early methods that simply use the visual modality to identify speech activity to sophisticated techniques which synthesise a full AV model. New directions in this exciting area of signal processing are also identified.

Separating speech signals that are only observable as mixtures requires techniques such as blind source separation (BSS). This topic has been investigated extensively in the signal processing community during the past two decades and has had impact upon many applications such as speech enhancement and machine audition [1]. A well-known example for demonstrating BSS applications is the so-called cocktail-party problem coined by Cherry [2]. His desire was to build a machine to mimic a human's ability in separating target speech sources from a superposition of multiple sound signals including interfering sounds and background noise, often coupled by sound reflections from room surfaces. This problem is usually addressed within the framework of convolutive BSS taking into account room reverberations in the separation model. In this framework, the vector observations $\mathbf{x}(t)$ are modeled as a linear convolutive mixture of the vector sources $\mathbf{s}(t)$: $\mathbf{x}(t) = H(t) * \mathbf{s}(t)$, where $H(t)$ is the matrix of impulse responses between each source and each mixture, and t is the discrete time index. For simplicity, $H(t)$ is assumed to be square so that the number of microphones and sources is equal, but this is not necessary to achieve separation. The aim is thus to estimate the demixing matrix $W(t)$ so that $\hat{\mathbf{s}}(t) = W(t) * \mathbf{x}(t)$ contains an estimate of each source $s_i(t)$, where subscript i is the index of the source. Alternatively, it is solved in a transform domain by converting the full-band speech mixtures into subband components that are then separated either individually or jointly, leading to a computationally more efficient method, e.g., frequency-domain BSS. In this latter case, assuming static sources, the frequency-domain counterparts of mixing and demixing equations are $\mathbf{x}(m,f) = H(f)\mathbf{s}(m,f)$ and $\hat{\mathbf{s}}(m,f) = W(f)\mathbf{x}(m,f)$, respectively, where $\cdot(m,f)$ is the short-term discrete-time Fourier transform (STFT) of $\cdot(t)$, $\cdot(f)$ is the Fourier transform of $\cdot(t)$, and m and f are the time frame and frequency bin indices, respectively. This, however, introduces the permutation and scaling ambiguity problems due to the potentially inconsistent orders

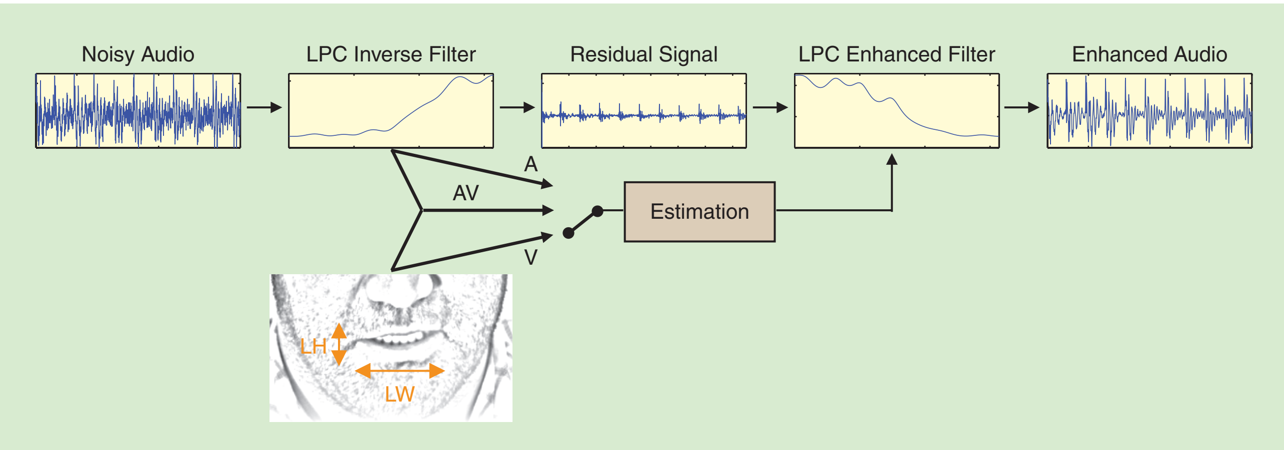
Digital Object Identifier 10.1109/MSP.2013.2296173
Date of publication: 7 April 2014

and scales of the separated source components at the individual frequency bands that are inherent to the instantaneous BSS models: in other words, $W(f) = \Lambda(f)P(f)H(f)^{-1}$, where $\Lambda(f)$ is a diagonal matrix (i.e., modeling the scaling indeterminacy) and $P(f)$ is a permutation matrix (i.e., modeling the permutation ambiguity). Many methods have been developed to mitigate these ambiguities before reconstructing the full-band source signals (more details can be found in [1]). A more recent approach is independent vector analysis (IVA), whereby the permutation problem is mitigated via a coupling of the adaptation across all the frequency bands [3].

These convolutive BSS techniques can be broadly attributed to a category of linear filtering-based methods. Another powerful method for separating convolutive mixtures is based on a form of time-varying filtering using time-frequency (T-F) masking where the aim is to form a probabilistic (soft) or binary (hard) mask $\mathcal{M}(m, f)$ for each source, and then applying the mask to the T-F representation of the mixtures for the extraction of that source: $\hat{s}(m, f) = \mathcal{M}(m, f)x(m, f)$. The mask can be estimated by the evaluation of various cues from the mixtures, such as statistical, spatial, temporal and/or spectral cues, using an expectation maximization (EM) algorithm [4] under a maximum likelihood or a Bayesian framework. The T-F masking techniques can often be applied directly on underdetemined mixtures for the extraction of a larger number of sources than the observed signals. Despite these efforts and the promising progress made in this area, the state-of-the-art algorithms commonly suffer in the following two practical situations: highly reverberant and noisy environments, and when multiple moving sources are present. For example, most existing methods of frequency-domain BSS are practically constrained by the data length limitation, i.e., the number of samples available at each frequency bin is not sufficient for learning algorithms to converge [5], while the various cues, such as the spatial cues that are used to calculate the likelihood of the source being present for the T-F mask estimation, become more ambiguous with the increasing reverberation and background noise. The performance of most existing algorithms degrades substantially in these adverse acoustic environments.

The methods mentioned previously exploit only single modality signals in the audio domain. However, it is now widely accepted that human speech is inherently at least bimodal involving interactions between audio and visual modalities [6]. For example, the uttering activities are often coupled with the visual movements of vocal organs, while reading lip movement can help a human to infer the meaning of a spoken sentence in a noisy environment [7]. The well-known McGurk effect also confirms that visual articulatory information is integrated into the human speech perception process automatically and unconsciously [8]. For example, under certain conditions, a visual /ga/ combined with an auditory /ba/ is often heard as /da/. As also suggested by Cherry [2], fusing the AV information from different sensory measurements would be the best way to address the machine cocktail-party problem. The intrinsic AV coherence has been exploited previously to improve the performance of automatic speech recognition [13] and identification [14]. The term *coherence* is used here to describe the dependency between the audio and visual modalities, to be consistent with the conventional use of the term in previous works in the literature, such as [9]–[12]. As discussed in subsequent sections, the dependency can be modeled as either joint distribution of the AV features or joint AV atoms (i.e., signal components).

In the study in [9], a speech signal corrupted by white noise is enhanced with filters estimated from the video input. The aim is to estimate a time-varying Wiener filter based on a linear regression [linear predictive coding (LPC)] between the audio and visual signals from a regressor trained with a clean database (Figure 1) and therefore is termed an *AV-Wiener filter*. This preliminary study has been shown to be efficient on very simple data (succession of vowels and consonants). For instance, with an input signal-to-noise ratio (SNR) of -18 dB, a simple linear discriminant analysis of the filtered signals leads to a word



[FIG1] The AV estimation of the Wiener filter from [9]. The LPC method is used to model the noisy speech. The audio feature based on the LPC inverse filtered spectrum is fused with the visual features such as the lip width (LW) and height (LH) for enhancing the LPC spectrum of the noisy speech. The enhanced speech signal can therefore be obtained based on this LPC enhanced filter and the residual signal obtained from the inverse filtering of the noisy speech.

classification accuracy (CA) of 40% after the AV enhancement compared to the CA of 10% with the classical audio enhancement while the unfiltered data leads to a CA of 5%. However, due to the complex relationship between audio and video signals, this simple approach is found to have limitations when applied on more complex signals such as natural speech and other noise sources. Nevertheless, this pioneering approach has shown that it can be extremely beneficial to combine video information when dealing with speech enhancement mirroring the advantage gained in automatic speech recognition systems [13].

During the last decade, integrating visual information into an audio-only speech source separation system has been emerging as an exciting new area in signal processing: AV (i.e., multimodal) speech source separation [10]. The activities in this area include robust modeling of AV coherence [12], [15]; fusing of AV coherence with independent component analysis (ICA) or T-F masking [16]; using AV coherence to resolve ambiguities in BSS [15], [17]; employing visual information for the detection of voice activities [18], [19]; exploiting redundancy within the AV data to design efficient speech separation algorithms based on sparse representations [16], [20]; and more recently, AV scene analysis for addressing the challenging problem of speech separation from moving sources [5], [21] or in environments with long reverberation time [22].

A number of different of approaches have clearly been developed to tackle the speech source separation problem using both audio and visual modalities. To present these in a coherent manner, the remainder of this tutorial is organized according to the increasing sophistication in the way in which video is used to help speech source enhancement as summarized in Table 1. The advantages and disadvantages of the methods are also highlighted in the table, and references are added to papers where the full details can be found of experimental studies that present the performance gains achievable by adding video in the processing.

METHODS BASED ON VISUAL VOICE ACTIVITY

A very simple approach to model the link between audio and video signals is to utilize the voice activity of the time domain speech signal. Indeed, there exist pauses during natural speech: for instance, during breathing or before a plosive (such as /p/). Such silences can importantly be partially predicted by the

[TABLE 1] AN OVERVIEW OF AV METHODS FOR SPEECH ENHANCEMENT/SEPARATION. THE METHODS ARE CLASSIFIED ACCORDING TO THE INCREASING SOPHISTICATION IN THE WAY IN WHICH VIDEO IS USED TO HELP SPEECH SOURCE SEPARATION: FROM A COARSE BINARY INDEX (SECTION II) TO FULL JOINT AV MODEL (SECTION IV) INCLUDING VISUAL SCENE ANALYSIS (SECTION III), AND REFERENCES ARE GIVEN WHICH DETAIL COMPARATIVE PERFORMANCE EVALUATION STUDIES.

			METHODS	MAIN ADVANTAGES	MAIN DISADVANTAGES
DETAIL OF AV REPRESENTATION	BINARY	SECTION: "METHODS BASED ON VISUAL VOICE ACTIVITY"	SPECTRAL SUBTRACTION [23], [24]	EFFECTIVE IN NOISE REDUCTION; EASY TO IMPLEMENT	INTRODUCES PROCESSING ARTEFACTS; DIFFICULT TO ESTIMATE THE NOISE POWER FOR NONSTATIONARY SIGNALS
			AV POSTPROCESSING OF AUDIO ICA [17]	LOW COMPUTATIONAL COST; STRENGTH OF ICA FRAMEWORK; CORRECT (ALMOST) ALL PERMUTATIONS	INCREASES DELAY (I.E., LATENCY); POTENTIAL PROCESSING ARTEFACTS
			EXTRACTION BASED ON TEMPORAL VOICE ACTIVITY [18]	LOW COMPUTATIONAL COST; SIMPLE ASSUMPTIONS	LIMITED TO LOW REVERBERATION
	VISUAL SCENE ANALYSIS	SECTION: "VISUAL SCENE ANALYSIS-BASED METHODS"	AV BEAMFORMING/ ICA/IVA [5], [21], [25]–[27]	POTENTIAL FOR SEPARATING MOVING SOURCES; CORRECT THE PERMUTATIONS; IMPROVE CONVERGENCE OF ICA/IVA ALGORITHMS	DEGRADES WITH HIGH REVERBERATIONS
			AV T-F MASKING [22]	EXPLOITS TIME-VARYING PROPERTY OF SOURCES; NOT AFFECTED BY THE PERMUTATION PROBLEM	HIGH COMPUTATIONAL COMPLEXITY; CHALLENGING IN RESOLVING SPECTRAL OVERLAPS
	FULL JOINT AV MODEL	SECTION: "INTRODUCTION"	AV-WIENER FILTER [9]	LOW COMPUTATIONAL COST	LIMITED TO SIMPLE SIGNALS; DIFFICULTY IN LEARNING ACCURATE AV MODEL
		SECTION: "STATISTICAL AV-BASED METHODS"	MAXIMIZATION OF AV LIKELIHOOD [10], [28]	CAN EXTRACT SPEECH SOURCES FROM UNDERDETERMINED MIXTURES	LIMITED TO INSTANTANEOUS MIXTURES; DIFFICULTY IN LEARNING ACCURATE AV MODEL
			AV REGULARIZATION OF ICA [11]	EXPLOITING THE STRENGTH OF THE ICA FRAMEWORK	LIMITED IMPROVEMENT COMPARED TO AUDIO-ONLY ICA IN PARTICULAR FOR CONVOLUTIVE MIXTURES
			AV POSTPROCESSING OF AUDIO ICA [12]	MODERATE COMPUTATIONAL COST	DIFFICULTY OF LEARNING ACCURATE AV MODEL; INCREASES DELAY (I.E., LATENCY)
		SECTION: "SPARSE MODELING"	AVDL + T-F MASKING [16]	CAN CAPTURE THE LOCAL INFORMATION WITHIN THE SIGNALS; NOT AFFECTED BY THE PERMUTATION PROBLEM	HIGH COMPUTATIONAL COMPLEXITY; ONLY BIMODALITY INFORMATIVE PARTS OF THE SIGNALS ARE LEARNED

movements of the lips [19]. Based on this idea, several purely video-based voice activity detectors (V-VADs) have been developed [29]: they are generally based on the velocity of face features, usually motions of the lips. The main advantage of such V-VADs compared to an audio VAD is that they are not corrupted by concurrent audio sources such as environmental noise, or other speakers. It is worth noting that such models do not aim at linking audio and video features, but they try to infer very coarse information on silence [i.e., the probability that speech is present $P(s_i(t) \neq 0 \mid \zeta_i^v(t))$] or not [i.e., $P(s_i(t) = 0 \mid \zeta_i^v(m))$] in the audio modality from the video one, where $\zeta_i^v(t)$ is the visual signal associated with the ith source $s_i(t)$. Examples of speech enhancement methods that exploit a V-VAD are discussed next.

SPECTRAL SUBTRACTION

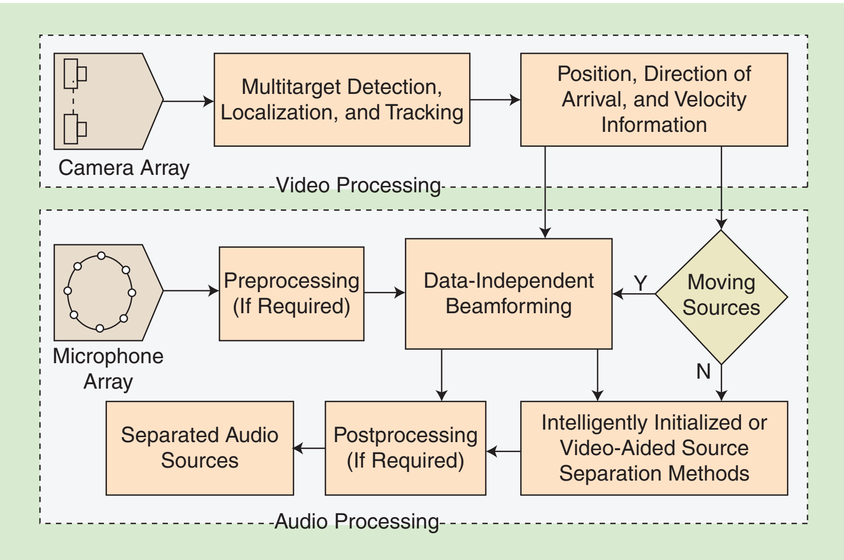
A simple method is to extend classical spectral subtraction by embedding visual features [23], [24]: the spectrum of the enhanced signal is expressed as $|\hat{s}(m,f)|^2 = |x(m,f)|^2 - \alpha |d(m,f)|^2$, where $x(m,f)$ is the STFT of a measured microphone signal $x(t)$, $d(m,f)$ is the estimated interference noise spectrum, and α is a parameter to adjust the subtraction level. The spectrum of the interference noise $d(m,f)$ is estimated from the windows related to the silence of the target source (i.e., the set $\mathcal{T}_i = \{t \mid P(s_i(t) = 0 \mid \zeta_i^v(t))\}$). These windows are efficiently detected by a V-VAD, which is thus not corrupted by the interfering audio noise.

AV POSTPROCESSING OF AUDIO ICA

Another use of such high-level information (speech/nonspeech frames) is to embed it into the efficient ICA framework. In this method, the visual information is used as postprocessing after applying an audio ICA algorithm. Frequency-domain source separation generally suffers from the permutation indeterminacy at each frequency bin: the ICA framework allows the recovery of the sources up to a global permutation (i.e., the order of the estimated sources is arbitrary). As a consequence, to recover the sources, this issue must be solved (i.e., permutation matrices $P(f)$ must be the same for all f). A very intuitive and efficient method is to estimate the permutations in relation to the output power of the sources [17]. Indeed, the V-VAD provides a binary indicator that shows when a specific speaker is silent. Using this information, one can then solve the permutation indeterminacy by simply minimizing the power of the target source during these frames. This method thus exploits the AV dependence, i.e., the joint distribution of AV features, in a very minimal way, but it has been shown to cancel almost all the permutation ambiguities. Compared to purely audio methods, this requires relatively low computational cost to mitigate the permutation ambiguities and allows the extraction of only a specific speech source instead of trying to recover all the sources.

AV EXTRACTION BASED ON TEMPORAL SPEECH ACTIVITY

A more effective use of such high-level information (speech/nonspeech frames) is to directly incorporate it into a separation criterion [18] to extract particular speakers to provide even less computational cost than ICA methods. Indeed, considering a set of time samples $\mathcal{D}$ so that the sources can be split into silent ones $(\forall i \in \mathcal{S}_{\text{silent}}, \forall t \in \mathcal{D}, s_i(t) = 0)$ and active ones $(\forall i \in \mathcal{S}_{\text{active}}, \forall t \in \mathcal{D}, s_i(t) \neq 0)$, purely audio algebraic methods based on generalized eigendecomposition of two covariance matrices can identify 1) the number of silent speakers (i.e., the cardinality of $\mathcal{S}_{\text{silent}}$), and 2) the associated support subspace (i.e., the subspace spanned by $\{s_i\}_{i \in \mathcal{S}_{\text{silent}}}$). In other words, considering any time samples including some not in $\mathcal{D}$ (i.e., the sources in $\mathcal{S}_{\text{silent}}$ can become active), the projection of the audio recordings $\mathbf{x}(t)$ onto the latter identified subspace cancels all sources in $\mathcal{S}_{\text{active}}$ while sources in $\mathcal{S}_{\text{silent}}$ remain unchanged. However, this method can not identify which source is silent and thus cannot be used to extract a specific speaker. To overcome this, a weighted kernel principal component analysis can be used to improve this approach, where the weights are a mixture between the audio probability of silence (given by the eigenvalues) and the video probability of silence provided by the V-VAD for a particular speaker [18]. This simple property provides a very efficient and elegant AV method to extract speech sources.



[FIG2] A block diagram of a visual scene analysis-based method for speech enhancement. Video localization is based on face and head detection. A video tracker is implemented for the tracking of multiple humans and based on the MCMC-PF. The output of the video processing is position, direction of arrival, and/or velocity information. On the basis of the visual scene, the preprocessed audio mixtures are separated either by a data-independent beamformer or intelligently initialized video-aided source separation method. Finally, postprocessing is applied to enhance the separated audio sources.

VISUAL SCENE ANALYSIS-BASED METHODS

In the previous section, AV extraction methods use the visual modality in a very coarse way: simple binary information defining whether or not a specific speaker is silent. In this section, this extra modality is used in a deeper way by visually analyzing the scene for speech enhancement [25], [26]. Such visual scene analysis thereby informs the source separation algorithms of the locations of the speakers, especially when dealing with moving sources, which is a more challenging issue since the mixing filters are now time varying. Thus, the classical ICA framework may be ineffective due to the large number of time samples required to accurately estimate the statistics of the mixtures. These methods are implemented in two stages: mainly video scene analysis (VSA) based on multiple human tracking (MHT) to estimate the position, direction of arrival (DOA), and velocity information of the people in a room or enclosed environment; and audio source separation depending on the scene as illustrated in the schematic diagram in Figure 2.

VIDEO PROCESSING FOR MHT

Video-based face and head detection is applied for multiperson observations from a single image as initialization. Then a Markov chain Monte Carlo-based particle filter (MCMC-PF) is used for MHT in the video. More details of the three important parts of the probabilistic MHT—the state model, the measurement model, and the sampling mechanism—are provided in [5]. Contrary to the V-VAD described in the section "Methods Based on Visual Voice Activity," it is highlighted that the full-frontal close-up views of the faces of the speakers, which are generally not available in a room or an enclosed environment, are not required for these trackers. The above-mentioned MHT methods provide a very good framework for AV scene modeling for source separation: the output of the video-based tracker is the three-dimensional (3-D) position of each speaker p, the elevation (θ_p), and azimuth (β_p) angles of arrival to the center of the microphone array. The direct-path weight vector $\mathbf{d}_p(f, \theta_p, \beta_p)$ can then be computed for frequency bin f and for source of interest $p = 1, \ldots, P$ and the velocity information that can then be used in the AV source separation scheme.

AV SOURCE SEPARATION OF MOVING SOURCES

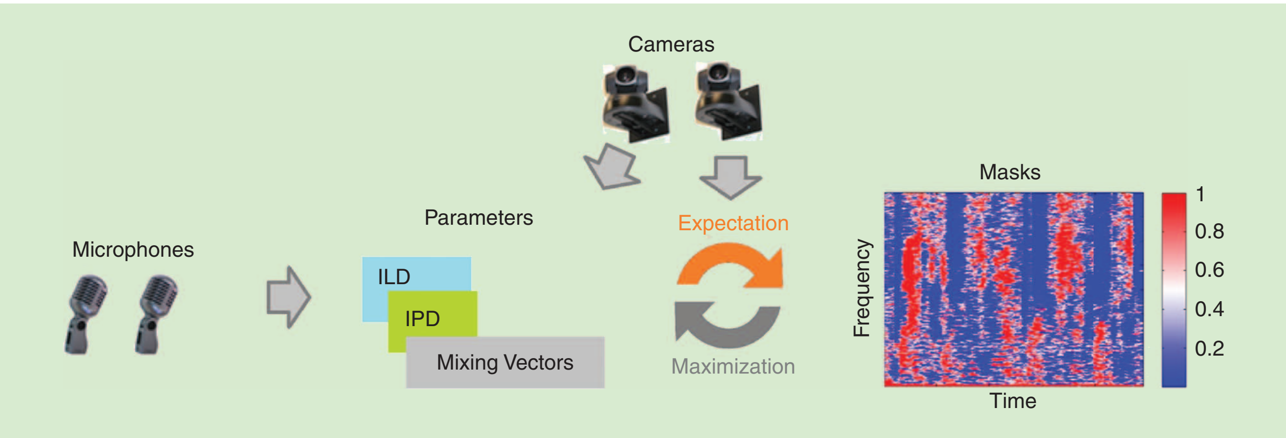
Speech source separation is a challenging issue when dealing with moving sources [26]. The proposed extraction of a particular speaker in an AV context depends on the velocity of this speaker.

PHYSICALLY MOVING SOURCES

After VSA, if the people are moving, then the challenge of separating respective audio sources is that the mixing filters are time varying; as such, the unmixing filters should also be time varying, but these are difficult to determine from only audio measurements. In [21], a multistage method has been developed for speech separation of moving sources based on VSA. This method consists of several stages including the DOA tracking of speech sources based on video processing, separation of sources based on beamforming with the beampatterns generated by the DOAs, and T-F masking as postprocessing. From the video signal, the direct path parameter vector $\mathbf{d}_p$ can be obtained, as discussed above, which is then used for the design of a robust least squares frequency invariant data independent (RLSFIDI) beamformer to separate the audio sources. The T-F masking is used as postprocessing to further improve the separation quality of the beamformer by reducing the interferences to a much lower level. However, such time-varying filtering techniques may introduce musical noise due to the inaccurate estimate of the mask at some T-F points. To overcome this problem, smoothing techniques such as cepstral smoothing may be used as in [21].

PHYSICALLY STATIONARY SOURCES

After video processing, if the speakers are judged to be physically stationary for at least two seconds, then the direct path parameter vector $\mathbf{d}_p$ with the whitening matrix obtained from the audio mixtures is used to intelligently initialize the learning algorithms, such as FastICA/IVA (many learning algorithms are sensitive to initializations) [3], [5], which solves the inherent permutation problem in ICA or block permutation in IVA algorithms and yields improved convergence [27].



[FIG3] Speech recording is obtained from two microphones. The direct path parameter vector is calculated with the help of video cameras. The ILD, the IPD, and the mixing vectors that utilize the direct path parameter vector are used to estimate the model parameters with the EM algorithm. The final probabilistic mask formed from the resulting probabilistic model is used for source separation.

T-F MASKING BASED ON VSA

More recently, a video-aided model-based source separation technique for underdetermined cases when the reverberation time is significant has been proposed [22]. This probabilistic T-F masking approach is motivated by both computational auditory scene analysis (CASA) and BSS, which relies on the assumption of signal sparseness (Figure 3). The interaural level difference (ILD), the interaural phase difference (IPD), and the mixing vectors are modeled as in [30], and the direct path parameter vector $\mathbf{d}_p$ is used as the mean parameter of the mixing vectors that is obtained from video processing. The parameters are updated iteratively with the EM algorithm. Since the EM algorithm is also sensitive to initialization, we initialize the direction vector parameter with the location information of the speakers obtained from video processing.

To form an AV probabilistic T-F mask $\mathcal{M}_i^{av}(m,f)$ for each static source $s_i(t)$, the IPD and ILD models as well as the model for the mixing vectors that utilize the direct-path weight vector obtained with the aid of video are used. It is a hidden maximum-likelihood parameter estimation problem and thus the EM algorithm can provide the solution. Extensive evaluations can be found in [22], which confirm the advantage of exploiting the visual modality to analyze the scene.

FULL JOINT AV MODELING-BASED METHODS

The most sophisticated approach to use the multimodality is then to build a full AV model of speech rather than the binary modeling of V-VAD (see the section "Methods Based on Visual Voice Activity") or the VSA (see the section "Visual Scene Analysis-Based Methods"). Two of these models and their uses for speech extraction are presented in this section: 1) AV statistical models, and 2) AV-dictionary-learning (DL)-based sparse representations models. With the statistical modeling approach, the AV coherence is often established explicitly on a feature space, which provides a holistic representation across all the observation frames of the AV signals [10], [11], [28], [31]. On the other hand, with the sparse representation-based methods, the AV coherence is implicitly modeled through the decomposition of an AV signal as a linear combination of a small number of signal components (i.e., atoms) chosen from a dictionary [16], [32]. The sparse model has shown to be effective in capturing the local information, such as temporal dynamic structures of the AV signals, which otherwise may be lost in the statistical modeling methods, but yet could be crucial for speech perception. Note that we distinguish these two models from the perspectives of modeling and optimization algorithms rather than the property of signals since sparsity can be considered as a statistical property of a signal. The two models could be used together if, e.g., the sparse models are built on a feature space described by some statistical models.

STATISTICAL AV-BASED METHODS

AV MODEL

The coherence between audio and visual modalities can be jointly modeled by, e.g., a Gaussian mixture model (GMM) where the coherence is expressed as a joint AV probability density function (AV-PDF)

$$p_1^{\mathrm{av}}(\zeta^a(m),\zeta^v(m)) = \sum_{k=1}^{K} w_k\, p_G(\zeta^a(m),\zeta^v(m)\,|\,\mu_k^{\mathrm{av}},\Sigma_k^{\mathrm{av}}), \quad (1)$$

where the superscripts a and v refer to the audio and visual modalities, respectively, and $\zeta^a(m)$ and $\zeta^v(m)$ are the audio and visual observation vectors at the mth frame, respectively; μ_k^{av} and Σ_k^{av} are the mean vector and the covariance matrix of the kth Gaussian kernel defined by its probability density function (PDF) $p_G(\cdot\,|\,\mu,\Sigma)$, w_k is the weight of the related kernel, and k is the number of mixture terms. (For simplicity in development, we will use the same notations to denote the AV feature vectors and AV sequence.) Classically, $\zeta^a(m)$ can be chosen as an audio feature vector, such as the modulus of the Fourier transform or the Mel-frequency cepstrum coefficients [33] of a windowed frame signal with frame index m, while $\zeta^v(m)$ is a visual feature vector, containing some shape parameters, e.g., the width and height of the lips or active appearance-based visual features [34]. When dealing with log scale audio parameters in the frequency domain, a more suitable model is the Log–Rayleigh PDF since this PDF explicitly models the nonsymmetric property of the logarithmic scale. The AV-PDF can thus be expressed as [31]

$$p_2^{\mathrm{av}}(\zeta^a(m),\zeta^v(m)) = \sum_{k=1}^{K} w_k\, p_{LR}(\zeta^a(m)\,|\,\Gamma_k^a)\, p_G(\zeta^v(m)\,|\,\mu_k^v,\Sigma_k^v), \quad (2)$$

where $p_{LR}(\zeta^a(m)\,|\,\Gamma)$ is the Log–Rayleigh PDF of localization or power coefficients defined by the diagonal elements of Γ_k^a (see [31] for more details). Such AV-PDFs not only jointly model the two modalities but they can also take into account the ambiguity of speech (i.e., the fact that the same shape of lips can produce several sounds such as /u/ and /y/ in French). The AV-PDF parameters are usually obtained from a clean training AV database using the EM algorithm.

EXTRACTION BY DIRECT AV CRITERIA

One of the first methods for AV source separation [10], [28] was based on the maximization of the AV coherence model described by the joint AV-PDF as in (1)

$$\widehat{\mathbf{b}} = \arg\max_{\mathbf{b}} p_1^{av}(\mathbf{b}^T\mathbf{x}(t),\zeta^v(t)), \quad (3)$$

where $\mathbf{b}$ is the extraction vector for a particular speaker in the instantaneous case, and the superscript $\cdot^T$ denotes the transpose operator. Even though such an approach is shown to be efficient when dealing with the simple succession of vowels and consonants [10], this method suffers from two important drawbacks: 1) a relevant AV probabilistic model is quite difficult to obtain for natural speech and 2) a direct maximization of the AV-PDF becomes rapidly computationally inefficient due to the dimensions of the separation filters when considering reverberant environments.

On the other hand, ICA [1] is an extraordinarily effective framework to separate sources from several mixtures. As a consequence, it is natural to embed AV constraints into a more classical frequency domain ICA criterion $J_{\mathrm{ICA}}(\{B(f)\}_f)$ by defining an AV-penalized ICA criterion [11]: $\{\hat{B}(f)\}_f = \arg\min_{\{B(f)\}_f} J_{\mathrm{ICA}}(\{B(f)\}_f) + P_{\mathrm{AV}}(p_1^{av})$, where the constraint term $P_{\mathrm{AV}}(\cdot)$ is a function of the AV-PDF as in (1). Note that we intentionally keep $J_{\mathrm{ICA}}(\{B(f)\}_f)$ to be general as many frequency-domain ICA criteria defined in the literature, such as in [1], can be used. As one can see, this criterion is a tradeoff between the statistical independence of the estimated sources (first term) based on ICA and the AV coherence of the estimated sources and the video features (second term). This AV constraint only slightly improves the signal-to-interference (SIR) ratio compared to a purely audio criterion [11]: this is mainly due to the difficulty to propose a relevant AV-PDF and appropriate AV constraints.

AV POSTPROCESSING OF AUDIO ICA

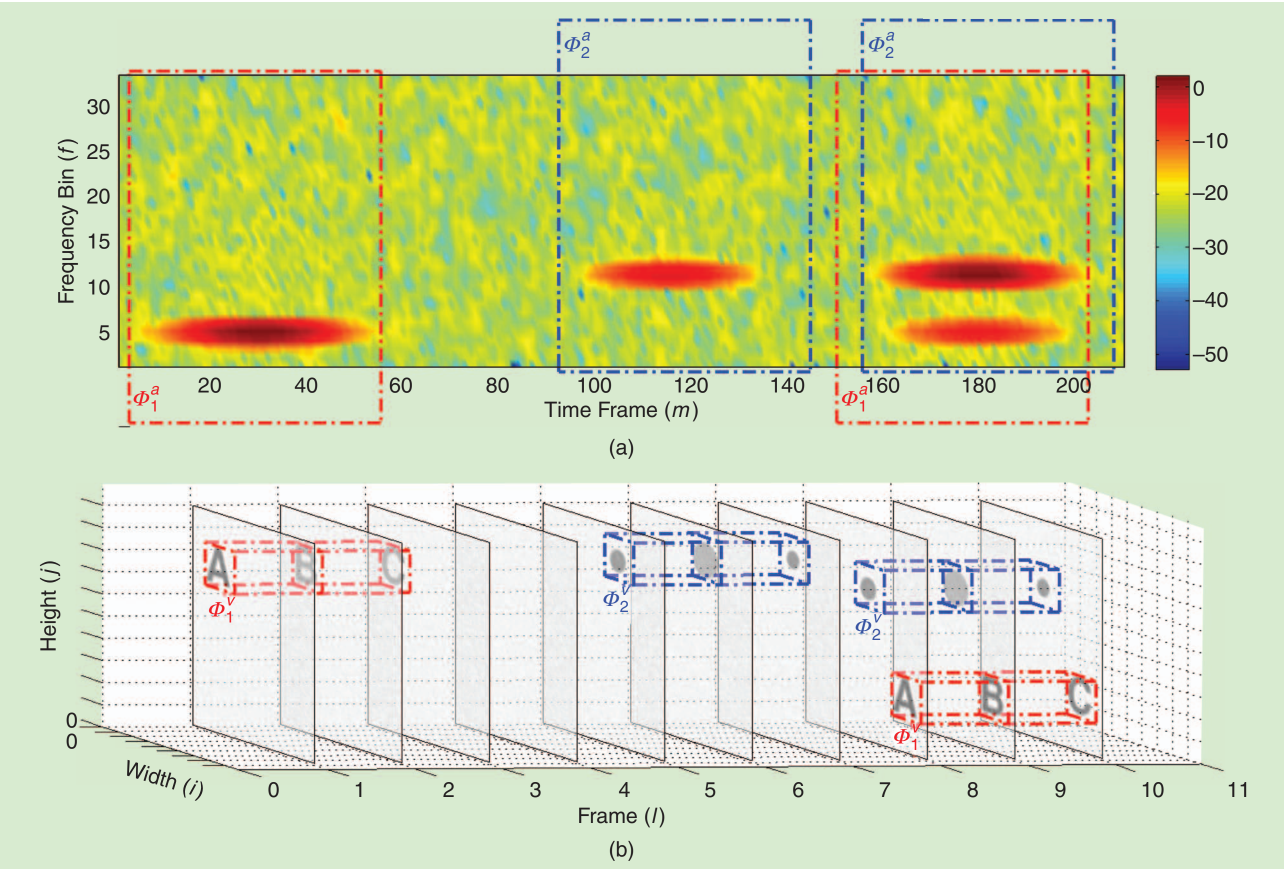
One natural way is to estimate the global permutation by maximizing the AV coherence [12] defined by $p_2^{av}(\cdot,\cdot)$ (2). However, even if these algorithms are shown to be effective to solve the permutation ambiguities, they suffer from their computational costs and from the difficulty to train accurately the statistical parameters that represent all the characteristics of natural speech.

SPARSE MODELING

While the previous methods seem to be quite natural AV extraction methods, the AV coherence used in these methods is often modeled in the feature space from the "global" point of view across all the frames of the AV data. These methods often fail to provide accurate estimation of audio signals due to the difficulty to train a relevant AV statistical model. To address this limitation, an alternative method for capturing the AV coherence has been considered in [16] and [32], using DL-based sparse approximation, which we call *sparse modeling*. As pointed out in [16], this technique can capture the 'local' information, i.e., the interconnection between neighboring samples, which is important for speech perception in a noisy environment.

SPARSE CODING OF AV SIGNALS AND DL OF THE AV ATOMS BASED ON A GENERATIVE AV MODEL

To obtain sparse representation of an AV signal, a generative model [16], [32] can be used, where an AV sequence $\zeta = (\zeta^a; \zeta^v)$



[FIG4] A schematic diagram generated artificially to show the use of a generative model to represent an AV sequence as a linear combination of a small number (two in this case) of atoms. The audio sequence (the spectrogram) is shown in (a) and the video sequence (a series of image frames, depicted as rectangles with solid lines) in (b). The patterns A, B, and C and the dots correspond to the two visual atoms. As highlighted by the rectangles with dot-dashed lines, the AV-coherent part in the sequence is represented by scaling and allocating the atoms at two positions. The audio stream is shown in log scale. The audio atom is a randomly generated spectrogram pattern, rather than a realistic phoneme or word in speech. The figure is modified from [16], where examples of AV sequence and AV atoms from real AV speech data can be found.

is described by a small number of AV atoms $\phi_k = (\phi_k^a; \phi_k^v)$ chosen from an overcomplete dictionary $\mathcal{D} = \{\phi_k\}_{k=1}^K$, where the discrete time index t is omitted here for notational convenience. The audio atoms ϕ_k^a are usually the log-modulus of the STFT of the audio signal component and the video ones ϕ_k^v are the mouth region (i.e. the area in the image frames where the mouth is located) of the video signal. We use a schematic diagram to explain the relationship between the AV sequence and AV atoms as shown in Figure 4, where each audio atom appears in tandem with its corresponding visual atom at a temporal-spatial (TS) position in the related video. In this example, the AV sequence is represented by only two AV atoms with some overlap between the two in a particular TS position.

Given an AV signal and a dictionary $\mathcal{D}$, the coding processing aims to find the sparse coefficients set that leads to a suitable approximation of the original sequence according to a matching criterion. This can be achieved by many algorithms including the greedy algorithms such as the well known matching pursuit (MP) or orthogonal MP algorithms. In [32], the MP algorithm has been extended to an AV-MP version to obtain the coding coefficients, where the matching criterion is defined as the inner product $\langle , \rangle$ between the residue of the AV sequence $(R^n\zeta)$ at the n-th iteration and the translated AV atom ϕ_k:

$$J_1^{av}(R^n\zeta, \phi_k) = |\langle R^n\zeta^a, \mathcal{T}_{\breve{m}}^a \phi_k^a \rangle| + |\langle R^n\zeta^v, \mathcal{T}_{i,j;\breve{l}}^v \phi_k^v \rangle|, \quad (4)$$

where $\mathcal{T}_{\breve{m}}^a$ is the temporal translation operator of the audio atom (i.e., shifting an audio atom by $\breve{m}$ time frames) and $\mathcal{T}_{i,j;\breve{l}}^v$ is the temporal-spatial translation operator of the video atom (i.e., shifting the video atom $\breve{l}$ time frames along the time axis and (i, j) pixels along the horizontal and vertical axes of the image frames). However, as shown in [16], the latter matching criterion may lead to a monomodal criterion due to the imbalance between the two modalities (due to the scale difference). The following criterion is therefore proposed in [16]:

$$J_2^{av}(R^n\zeta, \phi_k) = |\langle R^n\zeta^a, \mathcal{T}_{\breve{m}}^a \phi_k^a \rangle| \times \exp\{\frac{-1}{IJL} \| R^n\zeta^v - \mathcal{T}_{i,j;\breve{l}}^v \phi_k^v \|_1\}, \quad (5)$$

where I and J are the number of width and height pixels of the video atom ϕ_k^v, respectively, and L its time duration; $\| \cdot \|_1$ is the ℓ_1-norm.

The learning process is to adapt the K dictionary atoms $\phi_{k \in \{1,\cdots,K\}}$ to fit the training AV sequence. Several well-known DL algorithms can be used for this purpose, such as singular value decomposition (K-SVD) [35]. In [16], the K-SVD and K-means algorithms are used in each iteration for updating the audio and visual atoms respectively, so as to take into account the different sparsity constraints enforced on these two modalities. The sparse coding and DL stages are often performed in an alternating manner until the predefined criterion such as (5) is optimized.

SPARSE AV-DL-BASED AV SPEECH SEPARATION

From the AV-DL methods, T-F masking-based BSS methods are proposed [16], where the audio T-F mask $\mathcal{M}^a(m,f)$ generated by the purely audio algorithm [36] is fused empirically with a mask $\mathcal{M}^v(m,f)$ defined from the visual modality by the power-law transformation to define an AV T-F mask

$$\mathcal{M}^{av}(m,f) = \mathcal{M}^a(m,f)^{r(\mathcal{M}^v(m,f))}, \quad (6)$$

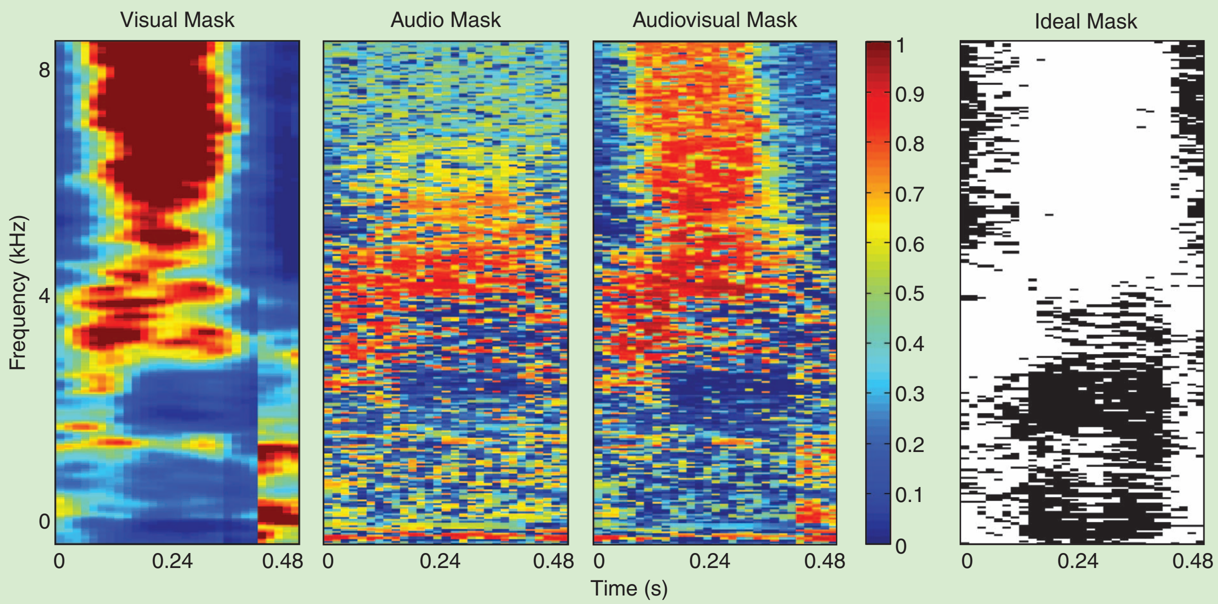
where the power coefficients r are obtained by applying a nonlinear mapping function to $\mathcal{M}^v(m,f)$ based on how confident the visual information is in determining the source occupation likelihood of each T-F point of the mixtures [16]. There are alternative methods for fusing the audio and visual masks, such as a simple linear combination of these two masks. Such a simple scheme is, however, less effective in taking into account the confidence level of the visual information, as compared with the power-law tranformation (more discussions about the motivation of using power-law transformation can be found in [16]). The mask defined from the video can be obtained as

$$\mathcal{M}^v(m,f) = \begin{cases} 1, & \text{if } \hat{\zeta}^a(m,f) > \zeta^a(m,f) \\ \hat{\zeta}^a(m,f)/\zeta^a(m,f), & \text{otherwise.} \end{cases} \quad (7)$$

Here $\hat{\zeta}^a(m,f)$ is the audio signal reconstructed from the speech mixtures by mapping the mixtures (together with the visual sequence) onto the AV dictionary. Note that, even if the latter mask is defined from audio-only sequences $\zeta^a(m,f)$ and $\hat{\zeta}^a(m,f)$, it can be considered as a visually inspired mask since $(\hat{\zeta}^a(m,f)$, is taken from $\left(\hat{\zeta}^a(m,f), \hat{\zeta}^v(y,x,l)\right)$, which represents the AV approximation of the new AV sequence $\zeta = (\zeta^a; \zeta^v)$. In other words, $\hat{\zeta}^a(m,f)$ is the best estimation of the audio signal from the AV sequence ζ obtained from its sparse decomposition on the AV dictionary $\mathcal{D}$. Finally, the noise-robust AV mask $\mathcal{M}^{av}(m,f)$ can be applied to the T-F spectrum of the mixtures for the target speech separation. Figure 5 shows an example of $\mathcal{M}^v(m,f)$, $\mathcal{M}^a(m,f)$, and AV masks $\mathcal{M}^a(m,f)^{r(\mathcal{M}^v(m,f))}$, as compared with the ideal binary mask (IBM). It can be seen that the fused AV masks improve the quality of the audio mask and the resolution of the visual mask. In [16], it is shown that the power-law transform performs better than the average operation, i.e., $(\mathcal{M}^a + \mathcal{M}^v)/2$.

CONCLUSIONS AND FUTURE DIRECTIONS

Over the past decade, AV speech source separation has emerged as a particularly interesting area of research in signal processing. It aims at improving the classical BSS methods for speech extraction by also using information from video and thereby mimicking the multimodal approach of humans. As shown in this article, the bimodality of speech can be used at different levels of sophistication to help audio source separation: from very coarse binary information through to a complete AV model, or from simple joint lip shape parameters to data-dependent acoustic features represented in an AV dictionary. As a result, the methods using the various level of information show different strength and weakness, as highlighted in Table 1. The main advantage of using the extra information from video is to tackle the problems that cannot be easily solved by audio-only algorithms: handling background noise and interference in strongly



[FIG5] A comparison among the visual mask, audio mask, AV mask (power-law), and IBM that shows improved definition in the AV mask. For the IBM, zero is denoted by black, and one by white. Although the visual mask looks smooth, some detailed audio information is missing. By comparing these masks with the IBM, it can be observed that the AV mask provides the best results. The figure is modified from [16], where more quantitative comparisons and analysis can be found.

reverberant environments, together with multiple, potentially moving sources.

There are many directions for further research. The AV coherence based on statistical methods requires high-quality, low-dimensional features for accurate and computationally efficient modeling, therefore emerging methods from manifold or deep learning could be exploited. The current methods in AV DL that attempt to capture the AV informative structure in the bimodal data are computationally expensive due to the intensive numerical operations required in sparse coding algorithms. Low-complexity and robust algorithms are highly desirable and need to be developed. Moreover, to be embedded in everyday devices such as smartphones, real-time approaches must be proposed to overcome the batch nature of many current algorithms. In the longer term, building richer models exploiting psychoacoustic-visual properties on the basis of the fields of brain–science and psychology can potentially further improve the AV speech separation systems, but this presents a particular challenge for future research in this area.

Finally, as speech source separation is clearly profiting from the bimodality of sources, other fields of source separation/extraction should also be explored using multimodal data, for instance, brain imaging, which can record brain activity by electroencephalography, magnetoencephalography, magnetic resonance imaging, and positron emission tomography. The next generation of intelligent multimodal signal processing techniques will combine such information to provide radically improved performance not achievable with methods based on single-modality data.

ACKNOWLEDGMENTS

The authors thank the guest editors and the anonymous reviewers for their helpful comments to drastically improve the quality of this article. The authors thank Dr. Qingju Liu for regenerating and adapting Figures 4 and 5, which originally appeared in [16]. Part of this work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. grant number EP/K014307/1, the U.K. Ministry of Defence University Defence Research Collaboration (MoD UDRC) in Signal Processing, and the European project ERC-2012-AdG-320684-CHESS.

AUTHORS

Bertrand Rivet (berrtand.rivet@gipsa-lab.grenoble-inp.fr) graduated from the Ecole Normale Supérieure de Cachan, France. He received the Agrégation de Physique Appliquée in 2002, the master's degree from the University of Paris-XI, France, in 2003, and the Ph.D. degree from the Grenoble Institute of Technology (Grenoble-INP, GIT), France, in 2006. He is currently an associate professor of signal processing with Physique, Electronique, Matériaux (PHELMA) and a member of GIPSA-lab, Grenoble-INP, France. His research concerns biomedical signal processing, audiovisual speech processing, and BSS.

Wenwu Wang (w.wang@surrey.ac.uk) received his Ph.D. degree from Harbin Engineering University, Harbin, China, in 2002. He is a senior lecturer of signal processing and codirector of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing, at the University of Surrey. He is currently a member of the MoD UDRC in Signal Processing

(2009–present) and the BBC Audio Research Partnership (2011–present). His current research interests include blind signal processing, AV signal processing, sparse representation, and DL. He has over 130 publications in these areas.

Syed Mohsen Naqvi (s.m.r.naqvi@lboro.ac.uk) received his Ph.D. degree in signal processing from Loughborough University, Leicestershire, United Kingdom, in 2009, where he is a lecturer in the School of Electronic, Electrical, and Systems Engineering. He was a postdoctoral research associate on the EPSRC U.K.-funded projects from 2009 to 2012. He contributed over 50 research outputs with main focus on multimodal (AV) speech processing and his research interests include nonlinear filtering, data fusion, and multitarget tracking, all for machine learning. He is a Member of the IEEE and the IEEE Signal Processing Society.

Jonathon A. Chambers (j.a.chambers@lboro.ac.uk) received his Ph.D. degree from the Imperial College, London, United Kingdom, in 1990. He is a professor of communications and signal processing and leads the Advanced Signal Processing Group within the School of Electronic, Electrical, and Systems Engineering at Loughborough University. He is also the associate dean (research) for Loughborough University in London. He is a senior area editor of *IEEE Transactions on Signal Processing* and a member of the IEEE Signal Processing Society Conference Board. He is a Fellow of the IEEE and the Royal Academy of Engineering, United Kingdom.

REFERENCES

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation Independent Component Analysis and Applications*. New York: Academic, 2010.

[2] 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.

[3] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Processing*, accepted for publication.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B.*, vol. 39, no. 1, pp. 1–38, 1977.

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

[6] L. E. Bernstein and C. Benoît, "For speech perception by humans or machines, three senses are better than one," in *Proc. Int. Conf. Spoken Language Processing* (*ICSLP*), Philadelphia, PA, 1996, pp. 1477–1480.

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

[8] A. Liew and S. Wang, Eds., *Visual Speech Recognition: Lip Segmentation and Mapping*. Hershey, PA: IGI Global Press, 2009.

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

[10] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audiovisual speech sources: A new approach exploiting the audiovisual coherence of speech stimuli," *EURASIP J. Appl. Signal Processing,* vol. 2002, no. 11, pp. 1165–1173, 2002.

[11] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP),* Philadelphia, PA, March 2005, pp. v/425–v/428.

[12] 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. Processing*, vol. 15, no. 1, pp. 96–108, 2007.

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[14] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: Evidence for early audiovisual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.

[15] Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve permutation problem in convolutive BSS," *Signal Processing,* vol. 92, no. 8, pp. 1916–1927, 2012.

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

[17] B. Rivet, L. Girin, C. Servière, D.-T. Pham, and C. Jutten, "Using a visual voice activity detector to regularize the permutations in blind source separation of convolutive speech mixtures," in *Proc. Int. Conf. Digital Signal Processing (DSP)*, Cardiff, Wales, U.K., July 2007, pp. 223–226.

[18] 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.

[19] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *J. Acoust. Soc. Amer.*, vol. 125, no. 2, pp. 1184–1196, Feb. 2009.

[20] 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, 2010.

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

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

[23] Q. Liu and W. Wang, "Blind source separation and visual voice activity detection for target speech extraction," in *Proc. Int. Conf. Awareness Science and Technology (iCAST),* 2011, pp. 457–460.

[24] Q. Liu, A. Aubery, and W. Wang, "Interference reduction in reverberant speech separation with visual voice activity detection," submitted for publication.

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

[26] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audiovisual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE J. Select. Topics Signal Processing,* vol. 4, no. 5, pp. 882–894, Oct. 2010.

[27] 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 Processing,* vol. 2012*,* no. 1, 2012.

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

[29] A. J. Aubrey, Y. A. Hicks, and J. A. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing*, vol. 4, no. 6, pp. 463–472, 2010.

[30] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP),* 2011, pp. 209–212.

[31] B. Rivet, L. Girin, and C. Jutten, "Log-Rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 15, no. 3, pp. 796–802, Mar. 2007.

[32] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audiovisual data," *IEEE Trans. Neural Networks*, vol. 20, no. 12, pp. 1898–1910, 2009.

[33] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing—Processing and Perception of Speech and Music*, 2nd ed. Hoboken, NJ: Wiley, 2011.

[34] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 23, no. 6, pp. 681–685, 2001.

[35] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[36] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[SP]