Source Separation of Convolutive and Noisy Mixtures Using Audio-Visual Dictionary Learning and Probabilistic Time-Frequency Masking

Qingju Liu, Wenwu Wang, *Senior Member, IEEE*, Philip J. B. Jackson, Mark Barnard, Josef Kittler, and Jonathon Chambers, *Fellow, IEEE*

***Abstract*—In existing audio-visual blind source separation (AV-BSS) algorithms, the AV coherence is usually established through statistical modelling, using e.g., Gaussian mixture models (GMMs). These methods often operate in a low-dimensional feature space, rendering an effective global representation of the data. The local information, which is important in capturing the temporal structure of the data, however, has not been explicitly exploited. In this paper, we propose a new method for capturing such local information, based on audio-visual dictionary learning (AVDL). We address several challenges associated with AVDL, including cross-modality differences in size, dimension and sampling rate, as well as the issues of scalability and computational complexity. Following a commonly employed bootstrap coding-learning process, we have developed a new AVDL algorithm which features, a bimodality balanced and scalable matching criterion, a size and dimension adaptive dictionary, a fast search index for efficient coding, and cross-modality diverse sparsity. We also show how the proposed AVDL can be incorporated into a BSS algorithm. As an example, we consider binaural mixtures, mimicking aspects of human binaural hearing, and derive a new noise-robust AV-BSS algorithm by combining the proposed AVDL algorithm with Mandel's BSS method, which is a state-of-the-art audio-domain method using time-frequency masking. We have systematically evaluated the proposed AVDL and AV-BSS algorithms, and show their advantages over the corresponding baseline methods, using both synthetic data and visual speech data from the multimodal LILiR Twotalk corpus.**

***Index Terms*—Audio-visual coherence, blind source separation, convolutive mixtures, dictionary learning, noisy mixtures.**

I. Introduction

A. BSS and AV-BSS

In complex auditory scenes, humans with normal hearing ability are generally capable of listening selectively to a particular sound source from a mixture of sounds including competing sound and background noise. This is known as the cocktail party problem [1]. To replicate such capabilities with machines is however extremely challenging, and a popular method for addressing this challenge is to model the problem under the framework of blind source separation (BSS), where the mixtures are usually assumed to be convolutive (or reverberant), in order to model the surface reflections of sounds in an enclosed room environment [2]. Many algorithms have been developed to recover the original unknown source signals from such mixtures [3]–[12] Independent component analysis (ICA)[3]–[6] exploits the statistical independence between source signals to address the BSS problem. Beamforming techniques separate sources by exploiting the geometric information of source positions, assuming a far- and free-field propagation model [7], [8] for the acoustic environment. With the sparsity assumption in an auxiliary transform domain, masking techniques [9]–[12] can be applied to separate sources by evaluating various cues such as inter-aural phase difference (IPD) and inter-aural level difference (ILD) [11]–[13].

Even though the above methods achieve good performance in near ideal conditions, i.e., noise free or low noise level with relatively low level reverberation, they degrade steadily in adverse conditions, e.g., in the presence of a high level of noise and reverberation and interfering sounds. To improve the intelligibility of noise corrupted speech, it is beneficial to introduce additional information that is robust to acoustic noise. One such type of information comes from visual cues (visual movement or lip-reading) associated with the concurrent sound source production and its perception. Studies show that human brains interconnect auditory with visual cues instead of dealing with the sound in isolation [14]–[18], which considerably improves speech intelligibility in a noisy environment. Such a relationship, known as audio-visual (AV) coherence, has been exploited to improve the performance of automatic speech recognition [19], identification [18] and source separation [20]–[28] in a noisy environment. Using the AV coherence to enhance BSS is known as AV-BSS [20]–[28].

One key challenge however is to model reliably the AV coherence, which, in AV-BSS, is established by two levels of fusion approaches:

- **Decision level fusion**. The source signals are first localised using visual tracking, then the characteristics of the separation filters are analysed based on either beamforming or convolutive BSS [24]–[26].
- **Feature level fusion** AV sequences are mapped into the feature space, and a statistical model is applied for AV feature fusion. Coherence maximization techniques are used

Manuscript received January 27, 2013; revised June 06, 2013 and July 29, 2013; accepted July 31, 2013. Date of publication August 08, 2013; date of current version October 10, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Antonio Napolitano. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK (Grant number EP/H012842/1) and the MOD University Defence Research Centre on Signal Processing (UDRC).

Q. Liu, W. Wang, P. J. B. Jackson, M. Barnard, and J. Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, U.K. (e-mail: q.liu@surrey.ac.uk; w.wang@surrey.ac.uk; p.jackson@surrey.ac.uk; mark.barnard@surrey.ac.uk; j.kittler@surrey.ac.uk).

J. Chambers is with School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, U.K. (e-mail: j.a.chambers@lboro.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2277834

to assist BSS, via directly estimating the separation matrix [20], [21], or overcoming the limitations of traditional BSS such as permutation ambiguities [22], [27], or providing audio activity information [29].

In the above methods, the AV coherence is modelled from the 'global' point of view across all observation frames, assuming the sampling distribution exhaustively represents all the AV information. However, the 'local' representation that exploits the temporal structures, i.e., the interconnection between neighbouring samples, is not considered. As a consequence of ignoring local features, essential information that locally describes a signal, is lost. Yet, this information plays an important role in speech perception. For instance, several consecutive visual frames focusing on a region with the mouth opening widely and sustaining for 200 ms (i.e., 5 frames) may indicate a long open vowel utterance, such as /a/; the presence of only one wide open frame of the above visual snippet more likely indicates a short vowel or a transition. To address this limitation, a better representation method for capturing the AV coherence should be considered, using e.g., dictionary learning (DL), as discussed next.

B. Dictionary Learning (DL) and Audio-Visual DL

Inspired by the AV fusion framework in [23], [30], we use an AV dictionary to model the AV coherence, where each atom in the dictionary has a bimodality-informative temporal-spatial (TS) inseparable [31] structure, which contains coherent events in both modalities, e.g., a visual lip movement lasting for several frames and a concurrent audio utterance, and the term 'spatial' denotes a position in the visual data. The new framework focuses on the structure of local features that might occur at any TS location of an AV sequence, rather than the sample characteristics of the whole AV sequence. Therefore, our AV fusion framework, to some extent, resembles locality-constrained linear coding (LLC) [32].

As opposed to using some pre-defined dictionary such as the Haar wavelets and the Fourier basis, our dictionary is adapted for a specific scenario, e.g, a person speaking, which needs to be learned via dictionary learning. DL is closely linked to sparse representation, whose aim is to describe a signal with a small number of atoms chosen from a redundant dictionary, where the number of atoms in the dictionary is greater than the dimension of the input signal. In other words, DL aims to find the optimal dictionary that best fits the input data [30], [31], [33]–[37], with a high level of sparsity and a low level of error. DL methods often employ a bootstrap process iterating between two stages: sparse coding and dictionary updating [31], [33], [34]. Firstly, the coding coefficients are obtained given the data and the dictionary via e.g., greedy techniques such as matching pursuit (MP) [38], orthogonal matching pursuit (OMP) [39], and convex relaxation methods such as basis pursuit [40]. Secondly, the dictionary atoms are updated to fit the input data via e.g., the least squares solutions to optimal directions [41], iterative gradient descent [42], singular value decomposition (K-SVD) [35], and more recently simultaneous codeword optimisation (SimCO) [43].

The above methods have been successfully applied to monomodal data such as images. Yet, little work has been undertaken for multimodal data, e.g., AV data as considered here. Tropp [36] proposed a simultaneous orthogonal matching pursuit method for multi-sensor data of the same type, which however is unsuitable for audio and visual streams that have different dimensions and temporal resolutions. Monaci *et al.* [37] proposed an iterative bootstrap coding and learning process between audio and visual streams with de-correlation constraints. This algorithm is fast and flexible. However, it may result in *spurious* AV atoms, i.e., physically-meaningless atoms. They proposed an improved audio-visual matching pursuit algorithm [30], with a more consistent joint coding process, and successfully applied it to speaker localisation in the presence of acoustic interference and visual distractors. This method is nevertheless constrained by the following limitations. *Firstly*, the weights of the two modalities used in the objective function may be unbalanced. *Secondly*, due to the high-dimensional data used for learning AV atoms, it is prone to errors induced by outliers including convolutive audio noise, and may result in over-fitting. *Thirdly*, the computational complexity is very high. To overcome these limitations, we propose a more robust, efficient and size-adaptive audio-visual dictionary learning (AVDL) method, which we then use to derive a new AV-BSS algorithm based on probabilistic time-frequency masking [11].

C. Our Contributions

1) AVDL: We build upon the DL approaches in [30], [35]–[37], [41], [42] and our method is also a bootstrap coding-learning process. With a similar AV structure as in [30], each AV atom in our dictionary contains an audio atom and a coherent visual atom spanning the same temporal length. The audio atom is the magnitude spectrum of a snippet of an audio signal, which is robust to convolutive noise (as observed in our experiments in Section V). The visual atom is composed of several consecutive frames of image patches, focusing on the movement of the whole mouth region, rather than highlighting the activity of a small part of the lips as used in [30]. The preliminary version of our work has been presented in [28], and we have the following main contributions:

- A new generative model and a new objective function are proposed to balance the audio and visual modalities, and to accommodate the different size, dimension, sampling rate, and the degrees of sparsity for the two modalities.
- The AV training sequence is mapped into a low-dimensional space to avoid the over-fitting problem and to improve the robustness of the AVDL algorithm to convolutive audio noise.
- A fast scanning and thresholding scheme is proposed for the coding stage to reduce the computational complexity of the dictionary learning algorithm.

2) AVDL-Incorporated BSS: We also demonstrate how the proposed AVDL algorithm can be used to improve the performance of BSS algorithms for audio mixtures corrupted by noise. To this end, we consider binaural auditory mixtures, mimicking aspects of human binaural hearing. Mandel's state-of-the-art method [11] is used for this purpose, where the spatial cues of IPD and ILD are exploited to generate an audio mask for source separation in the time-frequency (TF) domain. To improve the confidence of the audio mask in adverse conditions, we incorporate the AVDL into Mandel's method by re-weighting each TF point of the mask, to derive a noise-robust AV-BSS algorithm, whose framework is introduced in the next section.

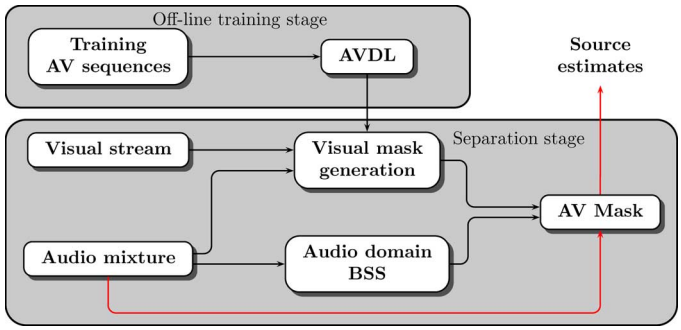


Fig. 1. The flow of our proposed AV-BSS system: AVDL-incorporated BSS. It contains two parallel mask generation processes: an audio mask generation process via a conventional audio-domain BSS, and a visual mask generation process via matching the corrupted AV sequence to the AV dictionary, which is learned in the off-line training stage via AVDL, thereby aiming to model the AV coherence. The integration of the above two masks is used to separate the target speech from the mixtures.

The remainder of the paper is organised as follows. Section II introduces the overall framework of our proposed AV-BSS. Section III describes in detail our proposed AVDL method for modelling the AV coherence, which is used in the off-line training stage of our AV-BSS method. Section IV explains in detail the separation stage of AV-BSS. Experimental results are shown in Section V, followed by the conclusions.

II. The Overall System

We present the block diagram of our proposed AV-BSS in Fig. 1, which contains the off-line training stage (upper shaded box) and the separation stage (lower shaded box).

In the training stage, we apply the proposed AVDL to the training AV sequence associated with a target speaker, and the learned dictionary is then used to model the AV coherence. The challenges in this stage include dealing with the cross-modality differences in size, dimension and sampling rate.

In the separation stage, two parallel TF mask generation processes are combined to derive an AV mask for separating the target source from the mixtures. One of the mask generation processes operates in the audio domain, exploiting binaural cues to cluster each TF point of the audio spectrum coming from different sources statistically, i.e., Mandel's method [11] as mentioned above; the other process exploits the AV coherence modelled in the off-line training stage, by mapping the corrupted AV sequence to the learned AV dictionary using the MP algorithm, to approximate a visually-constrained audio estimate for generating the visual mask (a mask for audio separation based on video). Finally, the audio mask and the visual mask are integrated to obtain a noise-robust AV mask, which is applied to the TF representation of the audio mixtures for the target source extraction. These two main stages will be discussed in detail in the next two sections.

III. Audio-Visual Dictionary Learning (AVDL)

AVDL is used in the off-line training stage of AV-BSS, which aims to learn the bimodality-coherent parts from AV sequences, resembling the *joint* receptive field of human vision and hearing [14], [16]. For presentation convenience, we divide this section into four parts.

- The generative model of an AV sequence with which AVDL is derived.
- The coding stage of AVDL, which, given a dictionary, decomposes an AV sequence using the generative model.
- The learning stage of AVDL, which updates the dictionary to better fit the data.
- The computational complexity of AVDL.

A. Generative Model

We denote an AV sequence as $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$ where the superscripts a and v denote audio and visual modalities respectively. It contains an audio signal $\boldsymbol{\psi}^a$ and a visual signal $\boldsymbol{\psi}^v$. Since using the audio magnitude spectrum tends to be more robust to noise and convolutive filters as compared to the time-domain audio signal (observed in Section V), hereafter, we transform the time-domain audio signal to the TF domain (magnitude spectrum) via the short-time Fourier transform (STFT), using the same notation $\boldsymbol{\psi}^a$:

$$\boldsymbol{\psi}^a = (\psi^a(m,\omega)) \in \mathbb{R}^{\tilde{M}\times\tilde{W}},$$
$$\boldsymbol{\psi}^v = (\psi^v(y,x,l)) \in \mathbb{R}^{\tilde{Y}\times\tilde{X}\times\tilde{L}},$$

where m and l are the discrete audio time (block) and visual time (frame) indices at different sampling rates f_s^a and f_s^v, y and x denote the pixel coordinates, ω is the frequency index of the audio spectrum. The upper-case letters with tilde define the size of the AV sequence[1]. In our proposed AVDL algorithm, the same operations will be applied to all the frequency bins of the audio modality, therefore we will intentionally drop the frequency index ω in this AVDL section for notational simplicity: $\boldsymbol{\psi}^a = (\psi^a(m))$.

In the same way, we define an AV dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^K$, with each atom denoted as $\boldsymbol{\phi}_k = (\boldsymbol{\phi}_k^a; \boldsymbol{\phi}_k^v)$ where $\boldsymbol{\phi}_k^a$ has a unit Euclidean norm. Each atom has a similar structure to an AV sequence:

$$\boldsymbol{\phi}_k^a = (\phi_k^a(m)) \in \mathbb{R}^{M\times W},$$
$$\boldsymbol{\phi}_k^v = (\phi_k^v(y,x,l)) \in \mathbb{R}^{Y\times X\times L},$$

where the upper-case letters define the atom size, which is much smaller compared to the AV sequence size ($\tilde{Y} \geq Y, \tilde{X} \geq X, \tilde{L} \gg L, \tilde{M} \gg M$). Note that, $\tilde{W} = W$ in our implementation.

The bimodality-coherent parts of the AV sequence $\boldsymbol{\psi}$ can be described as a linear superposition of atoms in the dictionary $\mathcal{D}$, each of which is convolved with a TS-varying signal, as demonstrated in Fig. 2. Assuming $f_s^a > f_s^v$, there are $Y_s X_s L_s$ possible TS positions indexed by $(\breve{y}, \breve{x}, \breve{l})$ for each visual atom where $Y_s = \tilde{Y} - Y + 1, X_s = \tilde{X} - X + 1, L_s = \tilde{L} - L + 1$, and M_s possible fine time positions indexed by $\breve{m}$ for each audio atom where $M_s = \tilde{M} - M + 1$. The generative model is given as

$$[\psi^a(m), \psi^v(y,x,l)]^T \approx [\hat{\psi}^a(m), \hat{\psi}^v(y,x,l)]^T = \sum_{k=1}^{K} \begin{pmatrix} \sum_{\breve{m}=1}^{M_s} c_{k\breve{m}} \phi_k^a(m-\breve{m}) \\ \sum_{\breve{y}=1,\breve{x}=1,\breve{l}=1}^{Y_s,X_s,L_s} b_{k\breve{y}\breve{x}\breve{l}} \phi_k^v(y-\breve{y}, x-\breve{x}, l-\breve{l}) \end{pmatrix} \quad (1)$$

where the superscript T denotes the transpose operator; $c_{k\breve{m}}$ is the audio coefficient and $b_{k\breve{y}\breve{x}\breve{l}}$ the visual coefficient, which together comprise the TS-varying coefficients being convolved

[1]We use the tilde to distinguish the sequence size from the size of the dictionary atom defined in the next paragraph.

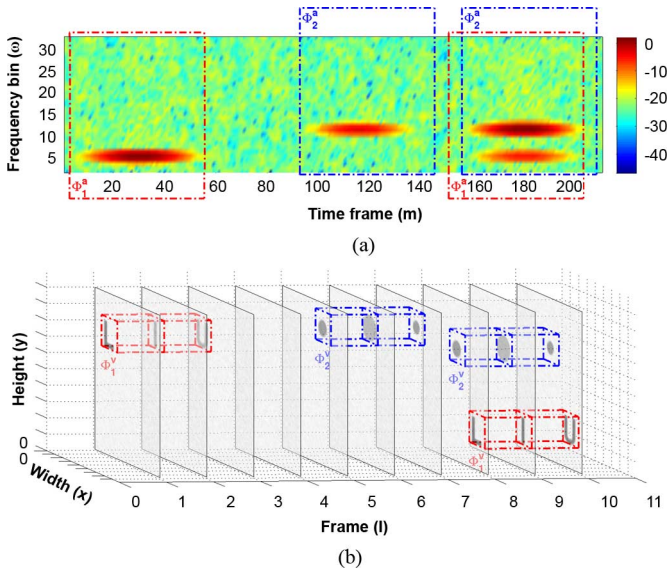


Fig. 2. Demonstration of the generative model in (1). The bimodality-coherent part of an AV sequence $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$, is represented as a linear superposition of atoms in the dictionary $\mathcal{D}$. Each dictionary atom $\boldsymbol{\phi}_k$ contains an audio atom $\boldsymbol{\phi}_k^a$ and a visual atom $\boldsymbol{\phi}_k^v$. We show two atoms in this example. Each atom is scaled and allocated at two places to represent the AV-coherent part in the sequence, as highlighted in the rectangles. For demonstration purposes, the audio stream is shown on the decibel scale, and similarly the following audio magnitude spectrograms. (a) Audio stream $\boldsymbol{\psi}^a$; (b) Visual stream $\boldsymbol{\psi}^v$.

with $\boldsymbol{\phi}_k$ to represent the AV signal. The AV atoms are inseparable, i.e., each audio atom and its associated visual atom always appear in pairs at a TS position of the AV sequence. As a result, if a visual atom $\boldsymbol{\phi}_k^v$ appears at the TS position $(\breve{y}, \breve{x}, \breve{l})$, i.e., $b_{k\breve{y}\breve{x}\breve{l}} \neq 0$, there exists a corresponding none-zero coefficient in the set $\{c_{k\breve{m}}\}$, subject to

$$\breve{m} \in \left\{ \left\lceil (f_s^a/f_s^v)(\breve{l}-1) \right\rceil + 1, \ldots, \left\lceil (f_s^a/f_s^v)\breve{l} \right\rceil \right\}. \quad (2)$$

In the above set, $\lceil \cdot \rceil$ rounds a number to its nearest integer; $\breve{m}$ denotes the same temporal position as $\breve{l}$ with a finer resolution. The coarse TS position $(\breve{y}, \breve{x}, \breve{l})$ and $\breve{m}$ comprise a fine TS position $(\breve{y}, \breve{x}, \breve{l}, \breve{m})$. We denote the approximation error (i.e., the residual) as $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}^a; \boldsymbol{\upsilon}^v) = \boldsymbol{\psi} - \hat{\boldsymbol{\psi}}$.

In (1), each AV atom $\boldsymbol{\phi}_k$ can be considered as an event that may sparsely occur (activate) at the TS position $(\breve{y}, \breve{x}, \breve{l})$ of $\boldsymbol{\psi}$. For the visual atom, we have the sparsity constraint that the visual activity (visual coding coefficient $b_{k\breve{y}\breve{x}\breve{l}}$) is binary with value either 1 or 0, depending on whether or not $\boldsymbol{\phi}_k^v$ occurs at $(\breve{y}, \breve{x}, \breve{l})$. For the audio atom, a more explicit sparsity constraint is enforced. We need to evaluate whether or not the event $\boldsymbol{\phi}_k^a$ occurs at a TS position, as well as how active (loud) it is. Therefore, the audio activity (audio coding coefficient $c_{k\breve{m}}$) is non-negative, and its value reflects the energy contribution of $\boldsymbol{\phi}_k^a$ at the temporal position $\breve{m}$.

We denote $\Omega = \{\mathbf{B}, \mathbf{C}\}$ as the coding parameter set, where $\mathbf{B} = (b_{k\breve{y}\breve{x}\breve{l}}) \in \mathbb{R}^{K \times Y_s \times X_s \times L_s}$, $\mathbf{C} = (c_{k\breve{m}}) \in \mathbb{R}^{K \times M_s}$. We enforce the sparseness constraint that there are only N non-zero elements in Ω with $N \ll KY_sX_sL_s$ or KM_s:

$$N = \sum_{k=1}^{K} N_k, \text{with} N_k = \|\mathbf{B}(k,:,:,:)\|_p = \|\mathbf{C}(k,:)\|_p, \quad (3)$$

where $\|\cdot\|_p$ is the ℓ_p norm, with $p = 0$ in this specific situation, and N_k gives the number of non-zero elements for $\boldsymbol{\phi}_k$.

To learn a dictionary that best fits the generative model in (1), a novel AVDL algorithm is developed, which, similar to [30], [35]–[37], [41], [42], contains a bootstrap coding-learning process, as shown in Algorithm 1. The learning process is stopped when the maximum iteration is achieved, or a robust dictionary is obtained, i.e., for two successive iterations, the coding parameters Ω stay the same or highly similar to each other. The coding and learning stages are detailed in the following two subsections respectively.

Algorithm 1: Framework of the Proposed AVDL

Input: A training AV sequence $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$, an initial $\mathcal{D}$ with K atoms, and the number of non-zero coefficients N

Output: An AV dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^K$

Initialization: $iter = 1$, $MaxIter$
while $iter \leq MaxIter$ **do**
%**Coding stage**
Given $\mathcal{D}$, decompose $\boldsymbol{\psi}$ using **(1)** to obtain Ω.
%**Learning stage**
Given Ω and the residual $\boldsymbol{\upsilon}$, update $\mathcal{D} = \{\phi_k\}$ for $k = 1, 2, \ldots, K$ to fit model **(1)**.
$iter = iter + 1$

B. Coding Stage

We use matching pursuit (MP) [38] in our sparse coding stage, which is a greedy method that iteratively chooses the optimal atom from the dictionary to approximate the signal. This facilitates the numerical comparisons with the baseline method [30], whose coding stage also adopts the MP algorithm. The MP is performed as follows: in the n-th iteration, the atom that has the highest value of the matching criterion with the training signal, is chosen to approximate the signal, whose contribution is then removed from the residual (i.e., approximation error). In the $(n+1)$-th iteration, we continue to find the next optimal atom and remove its contribution from the residual. In total, N iterations are applied.

1) New Matching Criterion: The 'matching criterion' measures how well an atom fits the signal in the MP algorithm, which is composed of the audio matching criterion and the visual matching criterion for AV signals. It is calculated between an AV atom and an AV segment extracted from the sequence (i.e., the updated AV residual $\boldsymbol{\upsilon}$). For simplicity, we define a segment extracted from $\boldsymbol{\upsilon}$ at the TS position $(\breve{y}, \breve{x}, \breve{l}, \breve{m})$ as $\bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}\breve{m}} = (\bar{\boldsymbol{\upsilon}}_{\breve{m}}^a; \bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}}^v)$, of which $\bar{\boldsymbol{\upsilon}}_{\breve{m}}^a \in \mathbb{R}^{M \times W}$ and $\bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}}^v \in \mathbb{R}^{Y \times X \times L}$. The short bar on the top distinguishes the residual segment from the complete residual sequence. In Monaci *et al.* [30], the following matching criterion J_{Mon}^{av} has been used:

$$J_{\text{Mon}}^{av}(\bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k) = J_{\text{Mon}}^{a}(\bar{\boldsymbol{\upsilon}}_{\breve{m}}^a, \boldsymbol{\phi}_k^a) + J_{\text{Mon}}^{v}(\bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}}^v, \boldsymbol{\phi}_k^v), \quad (4)$$

where $J_{\text{Mon}}^a = |\langle \bar{\boldsymbol{\upsilon}}_{\breve{m}}^a, \boldsymbol{\phi}_k^a \rangle|$, $J_{\text{Mon}}^v = |\langle \bar{\boldsymbol{\upsilon}}_{\breve{y}\breve{x}\breve{l}}^v, \boldsymbol{\phi}_k^v \rangle|$, with $|\cdot|$ being a modulus operator and $\langle \cdot, \cdot \rangle$ the inner product.

The above criterion, however, is not balanced between the two modalities. For example, if we scale the audio signal by a factor γ, the matching criterion between $\boldsymbol{\upsilon}^a$ and translated $\boldsymbol{\phi}_k^a$ will be proportionally scaled by γ, while the visual matching

criterion remains the same. As a result, the overall audio-visual criterion does not proportionally change, leading possibly to a monomodality criterion.

Another limitation lies in the visual matching criterion. In [30], the video sequence is pre-whitened to highlight moving object edges, resembling the motion-selective receptive field of the human vision system. J^v_{Mon} applied to the whitened video is not adaptive to the differences in shape and intensity of the visual objects that might be matched to a visual atom.

To address the above limitations, we propose a new overall matching criterion together with a new visual matching criterion:

$$J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k) = J^a(\bar{\boldsymbol{v}}^a_{\breve{m}}, \boldsymbol{\phi}^a_k) J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k), \tag{5}$$

where $J^a = J^a_{\text{Mon}}$. In the new matching criterion, any change of a monomodality criterion will proportionally scale the overall criterion. The visual criterion is defined as:

$$J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k) = \exp\left\{ \frac{-1}{YXL} \left\| \bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}} - \boldsymbol{\phi}^v_k \right\|_1 \right\}. \tag{6}$$

In the above visual criterion, we consider the absolute difference between a visual atom and a visual segment. The segment value does not directly affect the visual matching criterion. The exponential operation enlarges the variance of the visual criterion, which prevents the overall criterion from being dominated by the audio modality. Another advantage with (6) is that the low-dimensional visual feature extracted at the TS position $(\breve{y}, \breve{x}, \breve{l})$ can be used to replace $\bar{v}^v_{\breve{y}\breve{x}\breve{l}}$, e.g., the lip contour used in our experiments, which greatly reduces the computational complexity.

2) Optimisation Method: In the MP method, we use the matching criterion maximization as our objective function. To optimise it, first we need to calculate the overall matching criterion $J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k)$, $\forall(k, \breve{y}, \breve{x}, \breve{l}, \breve{m})$ using (5), with $\breve{m}$ being tied with $\breve{l}$ via set (2). In the n-th iteration of the coding stage, the optimal atom index k_n and the associated translation can therefore be found by maximizing the following objective function:

$$[k_n, y_n, x_n, l_n, m_n] = \underset{[k,\breve{y},\breve{x},\breve{l},\breve{m}]}{\arg\max}\ J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k), \tag{7}$$

where $\breve{m}$ is associated with $\breve{l}$ as defined in set (2). Then we can set values in the parameter set Ω:

$$\begin{aligned} B(k_n, y_n, x_n, l_n) &= 1 \\ C(k_n, m_n) &= J^a(\bar{\boldsymbol{v}}^a_{m_n}, \boldsymbol{\phi}^a_{k_n}). \end{aligned} \tag{8}$$

Finally, the residual [2] will be updated via:

$$\bar{\boldsymbol{v}}^a_{l_n} \leftarrow \bar{\boldsymbol{v}}^a_{l_n} - C(k_n, l_n)\boldsymbol{\phi}^a_{k_n}. \tag{9}$$

There are N iterations in total. We summarise the coding stage in Algorithm 2, where the scanning index $\mathcal{S}^{av}$, described in Section III-B.3, is used to improve its computational efficiency.

[2]To accommodate the visual sparsity constraint, the K-means technique is used to learn the visual atom, and therefore we do not need to calculate the visual residual as for the visual modality.

Algorithm 2: The Coding State of the Proposed AVDL

Input: An AV sequence $\boldsymbol{\psi}$, the dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^K$, the threshold δ, the number of non-zero coefficients N
Output: The coding parameter set $\Omega = \{\mathbf{B}, \mathbf{C}\}$ and residual $\boldsymbol{v}$
Initialization: Set Ω with zero tensors,
$\boldsymbol{v} = \boldsymbol{\psi}, n = 1, J_{opt} = J_{\max} = 0$
Calculate $\mathcal{S}^{av}$ using **(10)** to **(13)**.
while $n \le N$ *and* $J_{opt} \ge \delta J_{\max}$ **do**
  % Projection
  $\mathcal{L} = \begin{cases} \{1 : L_s\}, & \text{n=1} \\ l_{n-1} + \{1 - L : L - 1\}, & \text{otherwise} \end{cases}$
  for $k \leftarrow 1$ **to** K do
    foreach $\breve{l} \in \mathcal{L}$ **do**
      Calculate $J^a(\bar{\boldsymbol{v}}^a_{\breve{m}}, \boldsymbol{\phi}^a_k)$, where $\breve{m}$ is tied with $\breve{l}$ via set **(2)**.
      foreach $(\breve{y}, \breve{x}), \breve{y} \in \{1 : Y_s\}, \breve{x} \in \{1 : X_s\}$ **do**
        if $\mathcal{S}^{av}(\breve{y}, \breve{x}, \breve{l}) = 1$ **then**
          Obtain $J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k)$ via **(6)**
          and $J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k)$ via **(5)**.
  % Selection
  Obtain $[y_n, x_n, l_n, k_n, m_n]$ via **(7)**.
  Update Ω via **(8)**.
  Residual calculation via **(9)**.
  $J_{opt} = J^{av}(\bar{\boldsymbol{v}}_{y_n x_n l_n m_n}, \boldsymbol{\phi}_{k_n})$
  if $n = 1$ **then**
  $J_{\max} = J^{av}(\bar{\boldsymbol{v}}_{y_1 x_1 l_1 m_1}, \boldsymbol{\phi}_{k_1})$
  $n = n + 1$

For the convergence of our AVDL, we consider the coding process to be complete when either of the following two conditions is satisfied: 1) when the iteration number reaches the predefined number N, 2) when the maximum matching criterion $J^{av}(\bar{\boldsymbol{v}}_{y_n x_n l_n m_n}, \boldsymbol{\phi}_{k_n})$ in the current iteration is smaller than $\delta J_{\max}$, where $J_{\max}$ is the maximum matching criterion in the first iteration and δ is a selected threshold.

Note we do not use the residual energy $\|\boldsymbol{v}\|_2$ as the stopping condition, since in our coding stage, we aim to approximate the AV-coherent parts, whose energy is not proportional to the AV coherence. This residual may contain some AV-irrelevant parts with high energy, which are not approximated.

3) Fast Searching Factor for Better Convergence: A limitation in both the proposed algorithm and the baseline algorithm [30] is the computational complexity of the coding stage. In this section, we describe a novel method for improving its computational efficiency using a logical scanning index $\mathcal{S}(\breve{y}, \breve{x}, \breve{l})$, computed as:

$$\mathcal{S}(\breve{y}, \breve{x}, \breve{l}) = \mathcal{S}^v(\breve{y}, \breve{x}, \breve{l}) \cdot \mathcal{S}^a(\breve{l}), \tag{10}$$

where $\cdot$ denotes logical conjunction (AND or product), and $\mathcal{S}^a$ and $\mathcal{S}^v$ are the audio and visual scanning indices respectively.

The audio scanning index $\mathcal{S}^a$ is obtained by thresholding the short-term energy (of an audio segment having the same length as the audio atom), i.e., $E^a(\breve{l})$), as follows,

$$\mathcal{S}^a(\breve{l}) = \begin{cases} 1 & \text{if } E^a(\breve{l}) > \delta^a \overline{\boldsymbol{E}^a} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where δ^a is the audio threshold, $\boldsymbol{E}^a = (E^a(\breve{l})) \in \mathbb{R}^{L_s}, \breve{l} = 1, 2, \ldots L_s$, and $\overline{\boldsymbol{E}^a}$ denotes the mean value of $\boldsymbol{E}^a$.

The visual scanning index $\mathcal{S}^v$ is obtained similarly to the audio index, i.e., by thresholding the energy of the video image block after whitening. The whitening process is to highlight the dynamic lip region and to remove the static background in the images. First, after applying the Fourier transform over the third dimension (i.e., along l) of $\psi^v(y, x, l)$, we equalise the spectrum (i.e., whitening) with a high-pass filter to highlight *the dynamic visual parts* of the lip region. Then, we transform it back to the time domain, $\tilde{\psi}^v(y, x, l)$, which is then smoothed with three simple moving average filters f_y, f_x and f_l that contain Y, X and L elements respectively:

$$E^v(y, x, l) = f_l\left(f_x\left(f_y(\tilde{\psi}^v(y, x, l))\right)\right).$$

We then crop a block of video from $\boldsymbol{E}^v = (E^v(y, x, l)) \in \mathbb{R}^{\tilde{Y}\times\tilde{X}\times\tilde{L}}$, starting from (Y, X, L), denoted as $\breve{\boldsymbol{E}}^v = (E^v(\breve{y}, \breve{x}, \breve{l})) \in \mathbb{R}^{Y_s\times X_s\times L_s}$. We then focus on the peaky dynamic (i.e., mouth) region in each frame by removing most of the irrelevant positions:

$$E^v(\breve{y}, \breve{x}, \breve{l}) \leftarrow \begin{cases} 0 & \text{if } E^v(\breve{y}, \breve{x}, \breve{l}) < 0.8 \max\left(\breve{\boldsymbol{E}}^v(:, :, \breve{l})\right) \\ E^v(\breve{y}, \breve{x}, \breve{l}) & \text{otherwise.} \end{cases} \tag{12}$$

We obtain $\mathcal{S}^v$ by using a visual threshold δ^v:

$$\mathcal{S}^v(\breve{y}, \breve{x}, \breve{l}) = \begin{cases} 1 & \text{if } E^v(\breve{y}, \breve{x}, \breve{l}) > \delta^v \overline{\breve{\boldsymbol{E}}^v_{\neq 0}} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $\breve{\boldsymbol{E}}^v_{\neq 0}$ takes the non-zero elements in $\breve{\boldsymbol{E}}^v$.

We skip the matching criterion calculation at the TS position $(\breve{y}, \breve{x}, \breve{l})$ for $k = 1, 2, \ldots K$ if $\mathcal{S}(\breve{y}, \breve{x}, \breve{l}) = 0$, thereby reducing the computational cost of the coding stage significantly.

Using the scanning index, we have assumed implicitly that the physically meaningful AV information lies in the active parts of the AV sequence. This is particularly true in real-world audio-visual data in which the visual activities are often accompanied by the audio activities, and vice versa. As such, using the scanning index can significantly improve the computational efficiency of the coding process, without losing important information or compromising the performance. According to the evaluations on a set of 50 synthetic signals as used in Section V-A.1, we found that the proposed scanning index reduces the number of valid TS positions to 6.3%, while retaining 90.7% of the AV information.

C. Learning Stage

In this stage, we adapt the dictionary atom $\boldsymbol{\phi}_k, k = 1, 2, \ldots K$ to fit the AV sequence. Due to the different sparsity constraints of the audio and visual modalities, the K-SVD [35] and K-means algorithms are used for learning the audio and visual atoms respectively.

To demonstrate the K-SVD for the audio modality, we first introduce the notation $\mathbf{vec}(\cdot)$ that represents the vectorisation of a tensor and $\mathbf{ivec}(\cdot|\boldsymbol{\phi}^a_k)$ that reshapes a vector to the same size as the tensor $\boldsymbol{\phi}^a_k$. To apply K-SVD, the contribution of $\boldsymbol{\phi}_k$ should be added back to the residual,

$$\bar{\boldsymbol{v}}^a_{\breve{m}} \leftarrow \bar{\boldsymbol{v}}^a_{\breve{m}} + c_{k\breve{m}}\boldsymbol{\phi}^a_k, \ \forall \breve{m}. \tag{14}$$

We then build a matrix $\Upsilon_k \in \mathbb{R}^{MW\times N_k}$ whose columns are $\mathbf{vec}\left(\bar{\boldsymbol{v}}^a_{\breve{m}}\right), \text{ subject to } c_{k\breve{m}} \neq 0, \ \forall \breve{m}$. After that, we apply the SVD to Υ_k to obtain the first principal component:

$$\Upsilon_k \approx \lambda_k \mathbf{u}_k \mathbf{v}_k^T, \tag{15}$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are the two singular vectors associated with the largest singular value λ_k. Then $\boldsymbol{\phi}^a_k$ can be updated via

$$\boldsymbol{\phi}^a_k \leftarrow \mathbf{ivec}(\mathbf{u}_k|\boldsymbol{\phi}^a_k). \tag{16}$$

The non-zero elements in $\mathbf{C}$ associated with the k-th atom will be updated as the elements in the row vector $\lambda_k \mathbf{v}_k^T$. The residual at the associated positions will be updated[3] as:

$$\bar{\boldsymbol{v}}^a_{\breve{m}} \leftarrow \bar{\boldsymbol{v}}^a_{\breve{m}} - c_{k\breve{m}}\boldsymbol{\phi}^a_k, \ \forall \breve{m}. \tag{17}$$

Each visual coefficient in $\mathbf{B}$ is either 1 or 0, which simply includes or excludes one visual segment to the class $\boldsymbol{\phi}^v_k$. Therefore, we use K-means to update the visual atom $\boldsymbol{\phi}^v_k$.

The detailed learning stage is depicted in Algorithm 3.

Algorithm 3: The Learning Stage of the Proposed AVDL.

Input: The parameter set $\Omega = \{\mathbf{B}, \mathbf{C}\}$, the residual $\boldsymbol{v}$, the old dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^K$
Output: A new dictionary $\mathcal{D}$
1 **Initialization:** $k = 1$
2 **while** $k \leq K$ **do**
3 Update $\boldsymbol{\phi}^a_k$, $\mathbf{C}$ and $\boldsymbol{v}$ via K-SVD using **(14)** to **(17)**.
4 Update $\boldsymbol{\phi}^v_k$ via the K-means algorithm
5 $\boldsymbol{\phi}^v_k = \text{Mean}\left(b_{k\breve{y}\breve{x}\breve{l}}\bar{\boldsymbol{v}}^v_{k\breve{y}\breve{x}\breve{l}}\right), \text{ subject to } b_{k\breve{y}\breve{x}\breve{l}} \neq 0, \ \forall(\breve{y}, \breve{x}, \breve{l})$
6 $k = k + 1$

D. Complexity

The computational complexity of our proposed AVDL is dominated by the coding stage, as for the baseline method by Monaci *et al.* [30]. They are compared in Table I, where the complex operations include divisions, multiplications and logarithmic operations, while simple operations include summations and subtractions. We can observe that, for the audio modality, the proposed AVDL is faster than the baseline method by a factor of $8/N_{fft}$, assuming a 0.75-overlap STFT is imposed when applying the AVDL. For the visual modality, due to the proposed new matching function, the calculation load is greatly reduced in terms of complex operations, at the expense of importing additional simple operations. Note that, this comparison does not include the computational savings introduced by the proposed scanning index.

IV. AV-BSS

In this section, we describe in detail the three blocks in the separation stage of our proposed AV-BSS system in Fig. 1:

[3]This step is necessary in case two allocated atoms overlap with each other.

TABLE I
COMPUTATIONAL COMPLEXITY QUANTIZATION FOR THE PROPOSED AVDL AND MONACI'S METHOD

Monaci	Complex operations	Simple operations
Audio	$KL(f_s^a/f_s^v)^2(\tilde{L}+2NL)$	$KL(f_s^a/f_s^v)^2(\tilde{L}+2NL)$
Visual	$KYXL(Y_sX_sL_s+2NYXL)$	$KYXL(Y_sX_sL_s+2NYXL)$
AVDL	Complex operations	Simple operations
Audio	$8KL(f_s^a/f_s^v)^2(\tilde{L}+2NL)/N_{fft}$	$8KL(f_s^a/f_s^v)^2(\tilde{L}+2NL)/N_{fft}$
Visual	$K(Y_sX_sL_s+2NYXL)$	$2KYXL(Y_sX_sL_s+2NYXL)$

- The audio-domain BSS, i.e., to generate an audio-domain TF mask for source separation.
- The parallel noise-robust visual mask generation process, using the AV coherence modelled by AVDL.
- The integration of these two masks for the target speech separation.

A. Audio Mask Generation Using Binaural Cues

The proposed AVDL method can be flexibly combined with many existing BSS methods in the literature, and can in principle be applied to any number of mixtures. As an application example, however, we consider binaural mixtures, which mimic human binaural perception. Mandel's method, which exploits the spatial cues of IPD and ILD, is shown to produce the state-of-the-art results[11]–[13], and is therefore chosen as the audio-domain BSS to be combined with AVDL. The principles of Mandel's method are as follows.

A source signal arrives at two ears with different time delays and attenuations, exhibiting:

$$L(m,\omega)/R(m,\omega) = 10^{\frac{\alpha(m,\omega)}{20}} e^{j\beta(m,\omega)}, \tag{18}$$

where $L(m,\omega)$ and $R(m,\omega)$ are the STFTs of the left and right ear signals respectively, at the TF point (m,ω). The ILD is $\alpha(m,\omega)$ and the IPD is $\beta(m,\omega)$, which can be statistically modelled with mixtures of Gaussian distributions for different sources and time delays. The model parameters can be estimated iteratively via the expectation maximization (EM) algorithm based on a maximum likelihood framework. Based on these models, each TF point can be associated probabilistically to the source signals, i.e., to generate audio-domain separation masks. We denote the TF mask that contributes to the reconstruction of the target speech as $\mathcal{M}^a(m,\omega)$, which can be applied to either of the binaural signals for target source estimation, or to both of the binaural signals to obtain their average result as the source estimation, as done in our experiments.

B. Visual Mask Generation Using AVDL

In this section, we generate a visual mask from the noise-corrupted audio signal (i.e., speech mixtures possibly with additional noise) and the associated clean video signal, given a dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}$ that has been trained on the target speaker. Previously, in the dictionary learning section, we intentionally dropped the frequency index ω for the audio modality since there is no difference in the operations between different frequency channels. In this section, we aim to obtain a frequency-dependent separation mask for separating the target speech, so hereafter we denote the elements in the audio modality with both temporal m and frequency ω indices. For example, we denote $\psi^a(m,\omega) = (|L(m,\omega)| + |R(m,\omega)|)/2$ as the average magnitude spectrum from the noise-corrupted mixtures. Suppose $\boldsymbol{\psi}^v$ is the clean visual stream related to the target source signal, we can first approximate the new AV sequence $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$ using (1), via the same MP method as used in the coding stage of AVDL, and obtain the AV approximation denoted as $\left(\hat{\psi}^a(m,\omega), \hat{\psi}^v(y,x,l)\right)$.

In the coding stage, the audio matching criterion is affected by interference and noise. The target speech information may be corrupted or masked by the interference information, which often occurs at a TF position when the distortion energy is higher than the target speech energy. Yet, the audio matching criterion can approximate the contribution of the target speech in the matched frames. For the visual modality, the visual matching criterion is not affected by acoustic noise, and this avoids 'fake' matches caused by audio outliers. Here, we consider interference and background noise as generators of audio outliers with respect to the expected audio from the target. Therefore, the audio approximation $\hat{\psi}^a(m,\omega)$ gives an estimate of the contribution of the clean target speech in the matched TS positions, which is robust to acoustic noise. Comparing the reconstructed audio sequence with the corrupted audio sequence, we can obtain a visual mask in the TF domain:

$$\mathcal{M}^v(m,\omega) = \begin{cases} 1, & \text{if } \hat{\psi}^a(m,\omega) > \psi^a(m,\omega) \\ \hat{\psi}^a(m,\omega)/\psi^a(m,\omega), & \\ & \text{otherwise.} \end{cases} \tag{19}$$

We set 1 as the upper-bound since we aim to recover the information embedded in the mixture that comes directly from the target speaker. Hence, the reconstructed source magnitude should not be greater than that of the mixture. For those temporal positions m where no AV atom matches the AV sequence, the visual mask $\mathcal{M}^v(m,:)$ is set as 0.5. Since the reconstructed audio stream $\hat{\psi}^a(m,\omega)$ is obtained by mapping the corrupted AV sequence to the AV dictionary, which encodes the 'clean' AV coherence information associated with the target speaker, the 'fake' matches can be effectively suppressed.

C. Audio-Visual Mask Fusion for BSS

The probabilistic audio mask obtained by using the interaural spatial cues works well when the noise level in the mixtures is relatively low. However, with the increase of the noise level in the mixtures, the quality of the probabilistic mask starts to deteriorate, mainly because the confidence of assigning the TF point of mixtures to a particular source is reduced due to the noise corruption which essentially makes the binaural cues estimated from the mixtures in the audio domain increasingly ambiguous.

To increase the confidence of the TF assignment when generating the TF mask for source separation, we propose an empirical method for audio-visual fusion based on the power-law transformation as follows,

$$\mathcal{M}^{av}(m,\omega) = \mathcal{M}^a(m,\omega)^{r(\mathcal{M}^v(m,\omega))}, \tag{20}$$

where the power coefficients are obtained by applying a nonlinear mapping to $\mathcal{M}^v(m,\omega)$ shown in Fig. 3 We fix several of the power coefficients, and the other values of $r(\mathcal{M}^v(m,\omega))$ are obtained via curve fitting techniques, e.g., the spline interpolation used in our method.

In particular, the visual information is likely to increase the confidence of TF assignment in the situation where the audio mask has a low confidence, i.e., the source occupation likelihood determined via IPD and ILD is in the range around 0.5 (for two-source scenario), since in this case, the algorithm is not certain which source the TF point of the mixture belongs to. The power-law transformation, however, increases the discrimination confidence by either increasing or alternatively decreasing the occupation likelihood based on the information from the visual mask, so as to assign the TF point to the target or the interfering source. The principles for adjusting the occupation likelihood using the visual mask are as follows. The higher the confidence that the visual mask has, the more likely the occupation likelihood will be adjusted towards the value 1 or 0. In addition, when the visual mask has a very low confidence, i.e., 0.5, we retain the audio mask without being modified by the visual mask. This is also the situation when the mismatches happen in the AV sparse coding, which means none of the learned dictionary atoms occurs in this frame. A mismatch does not mean that the target speaker is silent in this period. Thus, we set the visual mask with value 0.5 rather than 0 for the mismatched frames.

Fig. 10(b) illustrates the process as the visual mask adjusts the noise-corrupted audio mask towards the ground-truth ideal mask using our proposed AV fusion method. The power-law transformation, in terms of our observation and evaluations, works well for incorporating the visual information, discussions and illustration of an alternative method of using the simple linear combination can also be found in Section V-B.4.

Finally, the noise-robust AV mask is used for the target source separation on both $L(m,\omega)$ and $R(m,\omega)$ to obtain their average result. The proposed AV-BSS is summarised in Algorithm 4.

Algorithm 4: Summary of the proposed AV-BSS.

Input: The AV dictionary $\mathcal{D}$, the binaural mixtures $L(m,\omega), R(m,\omega)$, the video ψ^v
Output: The target source estimate
1 % **Audio mask generation**
2 Obtain the audio mask $\mathcal{M}^a(m,\omega)$ with Mandel's method.
3 % **Visual mask generation**
4 Reconstruct $\left(\hat{\psi}^a(m,\omega), \hat{\psi}^v(y,x,l)\right)$ via MP using $\mathcal{D}$.
5 $\mathcal{M}^v(m,\omega)$ calculation via equation **(19)**.
6 % **Audio-visual mask generation**
7 $\mathcal{M}^{av}(m,\omega)$ calculation via equation **(20)**.
8 Apply $\mathcal{M}^{av}(m,\omega)$ to the binaural signal for source separation

V. Experimental Evaluations

This section contains two parts: evaluations of the proposed AVDL and evaluations of the proposed AV-BSS method. In the AVDL evaluation part, we used both synthetic AV data and short speech signals. For comparison purposes, we also implemented Monaci's method [30], in which we used the 'K-SVD C_1' type, i.e., ℓ_1 norm in the objective function, and K-SVD in the learning process to update dictionary atoms. We have quantified the performance of the AVDL in terms of approximation error rates for both audio and visual modalities. Examples of the learned dictionary atoms for synthetic and real speech data are analysed to demonstrate our proposed AVDL method.

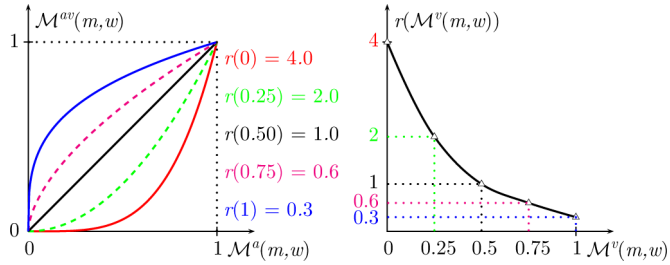


Fig. 3. Combine $\mathcal{M}^a(m,\omega)$ and $\mathcal{M}^v(m,\omega)$ to obtain $\mathcal{M}^{av}(m,\omega)$. The power coefficients are determined by a non-linear interpolation with pre-defined values. Considering the extreme situation where the audio mask values for both the target signal and the interference are 0.5, these pre-defined values are chosen to minimise the potential distortion due to processing artefacts. If only the target speaker is silent (ideally, $\mathcal{M}^v(m,\omega) = 0$), the value 4 is chosen to attenuate the target mask within 10 percent of the overall mask ($0.5^4 < 0.1$). If only the target speaker is active (ideally, $\mathcal{M}^v(m,\omega) = 1$), the value 0.2 is chosen so that the target mask spans 90 percent of the overall mask ($0.5^{0.2} \approx 0.9$). We slightly decrease the visual influence by replacing 0.2 with 0.3, considering that the hard upper bound threshold in equation ((19)) introduces some artificial distortion. When $\mathcal{M}^v(m,\omega) = 0.5$, the value 1 is chosen so that the visual mask does not alter the audio mask. When $\mathcal{M}^v(m,\omega)$ is 0.75 (.resp 0.25), the value is set to 0.6 (.resp 2), so that the change from the audio mask value 0.5 to the AV mask value, is half of that when $\mathcal{M}^v(m,\omega)$ is 0 (.resp 1).

In the AV-BSS evaluation part, we have compared our proposed method, denoted as AVDL-BSS, with four competing methods, of which two BSS methods are in the audio-visual domain, one in the audio domain and another in the visual domain. We evaluate the separation performance with the overall perceptual score using the PEASS tool-kit[44], which is specially designed for perceptual objective quality assessment of audio source separation.

A. AVDL Evaluations

In this subsection, we test our proposed AVDL algorithm, for both synthetic data and speech signals. For demonstration purposes, we use short AV sequences. To obtain a computationally feasible algorithm, we also apply our proposed scanning index to Monaci's baseline method.

1) AVDL for Synthetic Data:

Data, Parameter Setup and Performance Metrics: Similar to [30], we also generate a synthetic AV sequence, which lasts 40 s, with $f_s^a = 16$ kHz, $f_s^v = 30$ Hz. The video size is $\tilde{Y} \times \tilde{X} \times \tilde{L} = 40 \times 40 \times 1200$, while the audio length is 640000 samples (4 s). The synthetic data is generated by scaling and allocating five AV generative atoms at $32 \times 5 = 160$ randomly-chosen TS positions. Each generative atom contains a moving object on a white background as the visual atom and a snippet of audio vowels as the audio atom, including /a/, /i/, /o/, /u/. Of the five generative atoms, three atoms contain both audio and visual information, one atom is audio-only and one is visual-only, as shown in the upper row of Fig. 4. Part of the synthetic AV sequence lasting one second is shown in the lower row of Fig. 4. When we generate the synthetic data, the TS positions of the chosen atoms are randomly placed, and two allocated atoms are allowed to overlap with each other. To simulate the noise in a real AV sequence caused by background noise and image aliasing, and to test the robustness of our proposed AVDL to noise, 10 dB signal-to-noise ratio (SNR) audio noise and 20 dB peak signal-to-noise ratio (PSNR) visual noise are added, both in the form of Gaussian white noise.

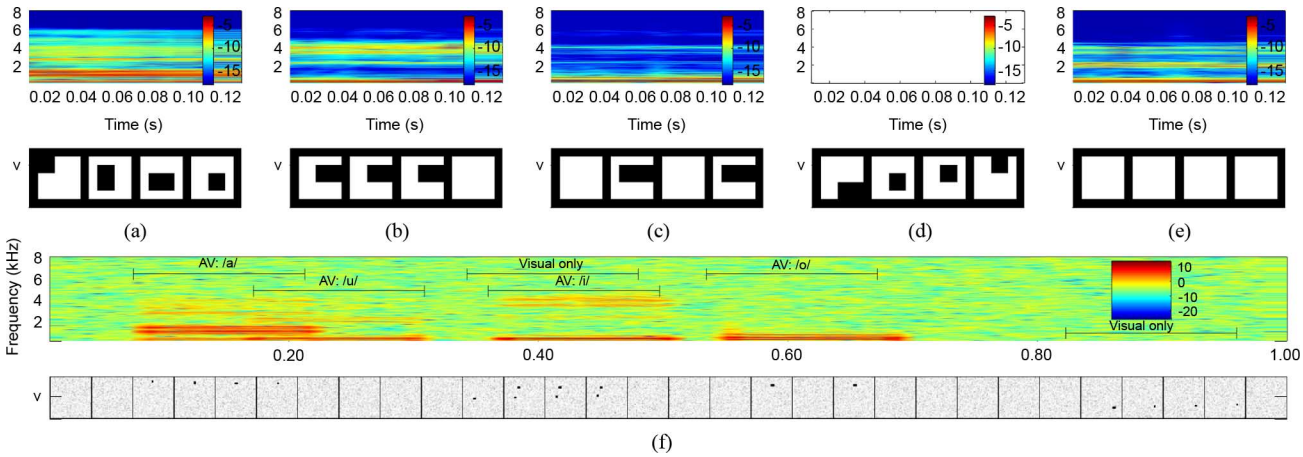


Fig. 4. The generative atoms and the synthetic data generated via the model (1). The white pixels have the minimal value 0 and the black pixels have the maximal value 1. Some atoms may have similar partial feature, e.g., AV atom: /i/ and audio only atom: /u/ have very similar audio structures. The audio sequence is normalized, e.g., the maximal magnitude is 1. (a) AV: /a/; (b) AV: /i/; (c) AV: /o/; (d) Visual only; (e) Audio only: /u/; (f) The generated AV synthetic sequence (only one second data is shown).

To implement the AVDL method, the STFT is first applied to the audio stream to obtain the audio spectrum $\boldsymbol{\psi}^a$, with a Hamming window size $N_{fft} = 512$ (32 ms) equal to the FFT size and a hop-size of 128 (8 ms), leading to a 75% overlap between the neighbouring windows. To synchronize with the video stream, spectrum $\boldsymbol{\psi}^a(m,\omega)$ was repeat-padded at the beginning and the end, and the audio stream is hence downsampled to $f_s^a = 125$ Hz. We set the dictionary size $K = 5$, the visual atom size $Y \times X \times L = 5 \times 5 \times 6$. Therefore, we have the audio atom size $W \times M$, where $W = N_{fft}/2 + 1 = 257$ and $M = \lceil (f_s^a/f_s^v)L \rceil = 30$. N in the sparse generative model was set to 100. To calculate the scanning index, we set $\delta^a = 0.05$ and $\delta^v = 0.1$. For the convergence of AVDL, we set the coding threshold $\delta = 0.01$ (.resp $0.05, 0.1$) in the second (.resp fifth, tenth) bootstrap iteration, and the maximal iteration number $MaxIter = 200$. In addition, we set the specific 'evolutionary' TS constraint as follows. In the first coding-learning iteration ($iter = 1$), two visual atoms are not allowed to have any overlap (i.e., by setting $\mathcal{S}(y_n + [1-Y : Y-1], x_n + [1-X : X-1], l_n + [1-L : L-1]) = 0$ after finding the n-th optimal atom in the coding stage). From the fifth iteration, two visual atoms may have at most half overlap, and from the tenth iteration when the dictionary atoms already tend to converge, two atoms are allowed to have full overlap (i.e., keep $\mathcal{S}$ unchanged).

To evaluate the performance of the two dictionary learning methods, we use the approximation error as the quantitative metric. We first generate five different training sequences as above, to train five different pairs of AV dictionaries via AVDL and Monaci's method. For each dictionary, 10 testing AV sequences with each lasting 40 s are generated, and the learned dictionary is used to approximate these testing sequences, with the approximated sequence denoted as $\hat{\psi} = [\hat{\psi}^a(m), \hat{\psi}^v(y,x,l)]^T$. Comparing $\hat{\psi}$ with the ground-truth signal ψ_g, which contains only the AV-coherent parts contributed by the AV atoms (the first three generative atoms):

$$\psi_g = [\psi_g^a(m), \psi_g^v(y,x,l)]^T \sum_k^3 \begin{pmatrix} \sum_{\breve{m}}^{M_s} c_{k\breve{m}} \phi_k^a(m-\breve{m}) \\ \sum_{\breve{y},\breve{x},\breve{l}}^{Y_s,X_s,L_s} b_{k\breve{y}\breve{x}\breve{l}} \phi_k^v(y-\breve{y}, x-\breve{x}, l-\breve{l}) \end{pmatrix}$$

we can obtain the audio approximation error $\mathcal{E}^a$ and the visual approximation error $\mathcal{E}^v$ separately:

$$\mathcal{E}^a = \frac{\sum_m |\psi_g^a(m) - \hat{\psi}^a(m)|^2}{\sum_m |\psi_g^a(m)|^2},$$
$$\mathcal{E}^v = \frac{\sum_{y,x,l} |\psi_g^v(y,x,l) - \hat{\psi}^v(y,x,l)|^2}{\sum_{y,x,l} |\psi_g^v(y,x,l)|^2}.$$

Results Comparison and Analysis: After 10 iterations, both algorithms successfully converge to three AV atoms, while ignoring the audio-only and the visual-only atoms. The upper row in Fig. 5 shows the AV atoms obtained via AVDL, while the bottom row shows the AV atoms obtained via Monaci's method.

However, if we amplify the audio sequence, or re-sample the audio sequence with a new temporal resolution or re-sample the video with a new spatial resolution, Monaci's algorithm may fail to converge to the correct AV atoms, due to its sensitivity to the size change of the AV atom. For instance, Monaci's method converges to four AV atoms with the four vowels as audio atoms if we increase the audio amplitude by a factor of 10, while the visual atoms are blurred by noise. In other words, Monaci's method becomes an audio-only dictionary learning method in this specific situation. However, our method still converges to the AV atoms accurately, since changes of the criterion in one modality proportionally change the overall AV matching criterion.

Moreover, our method is robust to convolutive noise encountered in a real acoustic environment where sounds reaching the sensors are filtered by room impulse responses. Hence we ran another independent test. A training sequence was generated with the same parameter setup, except that it was convolved with a time-varying FIR filter with 100 taps (~ 6 ms). i.e., at each TS position where a generative AV atom is allocated, a 100-tap filter whose coefficients are randomly chosen is generated and convolved with the allocated atom. Both dictionary learning algorithms are applied to this training sequence corrupted by time-varying convolutive filters. After the convergence of both algorithms, we notice that our AVDL still successfully learns the three AV atoms. However, of the four atoms

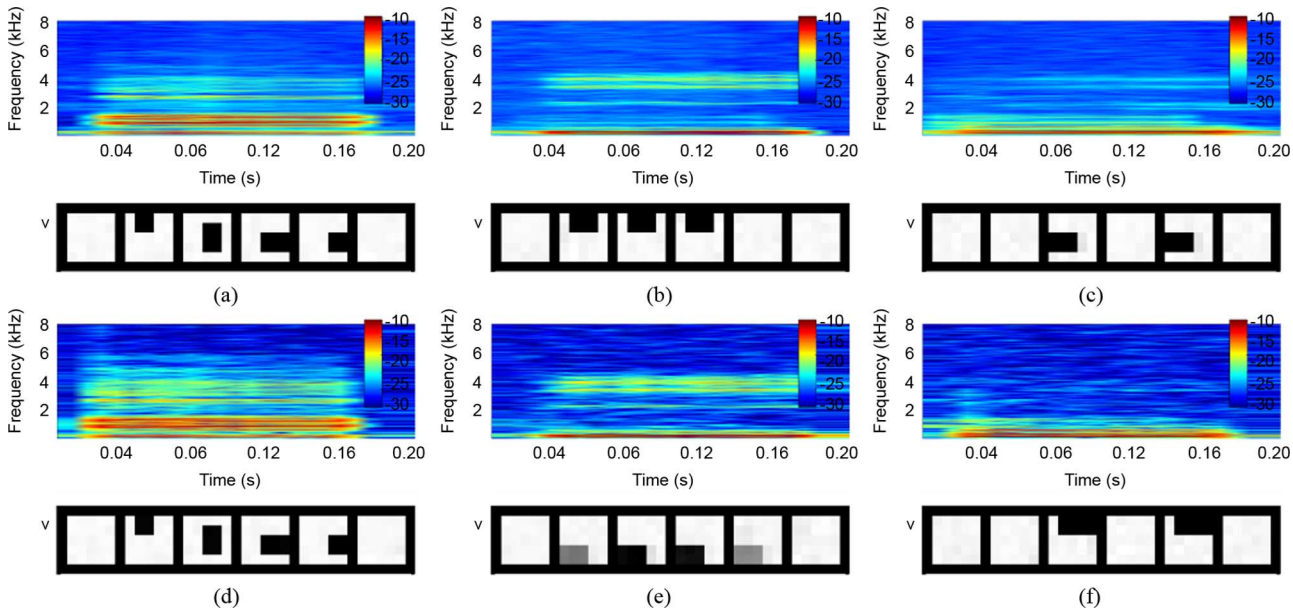


Fig. 5. The converged AV atoms using our proposed AVDL (the upper row) and the competing method (the bottom row). Both algorithms successfully converge to the right AV atoms in a few iterations. We have converted the audio atoms via Monaci's algorithm into the TF spectrum for ease of comparison. (a) AVDL: /a/; (b) AVDL: /i/; (c) AVDL: /o/; (d) Monaci: /a/; (e) Monaci: /i/; (f) Monaci: /o/.

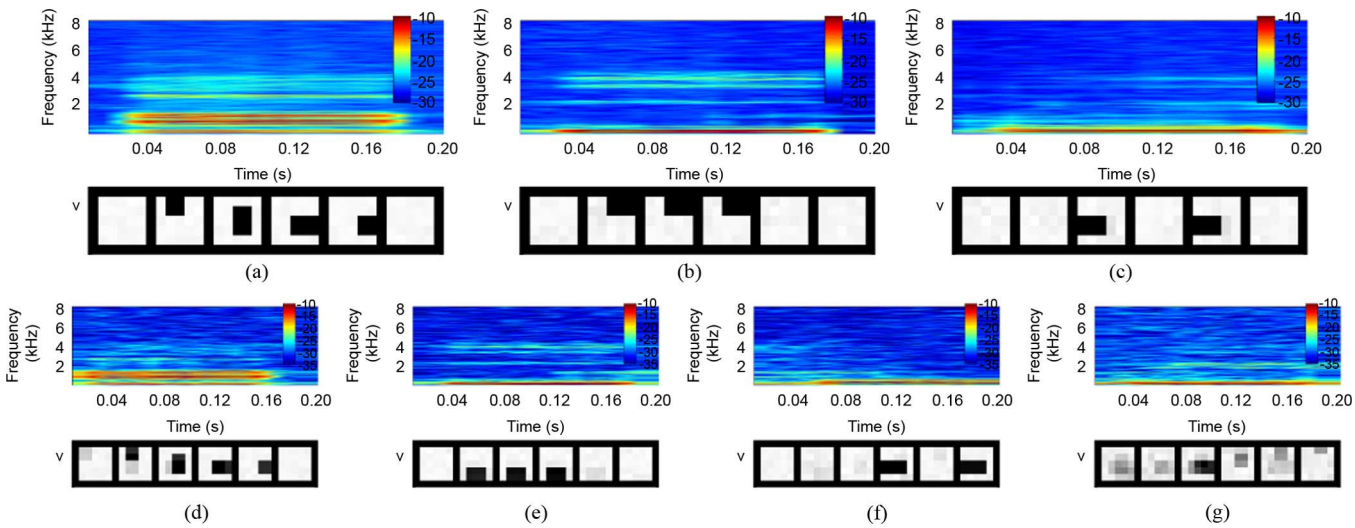


Fig. 6. The converged AV atoms using our proposed AVDL and Monaci's method when there is extra convolutive noise applied to the audio sequence. Our method successfully learns the three AV atoms, while the baseline method learns two accurate AV atoms and two spurious atoms (the last two). For the first spurious atom, the visual atom is from the visual-only outliers, and the audio atom is the combination of vowel /u/ and /o/. The second one contains the visual atom for the third generative AV atom and a distorted audio snippet. (a) AVDL1; (b) AVDL2; (c) AVDL3; (d) Monaci1; (e) Monaci2; (f) Monaci3; (g) Monaci4.

converged via the baseline method, two are spurious AV atoms shown in the last two atoms in Fig. 6.

We then quantitatively evaluate the performance of our proposed AVDL via the objective metrics of $\mathcal{E}^a$ and $\mathcal{E}^v$. Fig. 7 demonstrates the performance comparison, which shows that the proposed AVDL outperforms the baseline approach, giving an average of 33% improvement for the audio modality, from a set of 50 independent tests, together with a 26% improvement for the visual modality.

2) AVDL for Short Speech Data: To demonstrate our proposed method on real speech signals, we applied our AVDL and Monaci's baseline method on the multimodal LILiR Twotalk dataset [45], which was recorded with each subject uttering continuous speech. Sequences were obtained from 6 recordings, with each lasting from 210 s to 240 s, sampled at 16 kHz. For demonstration purposes, only a one-minute AV sequence data was used.

Data and Parameter Setup: For the visual stream, we used the lip contour coordinates to represent the video stream instead of the raw video for computational complexity reduction. A 38-point lip contour extraction algorithm [46] was applied for both inner and outer lips. Then we normalized the lip region to make the outer lip contour have a unit size, at the sampling rate of $f_s^v = 25$ Hz. After that, a new visual stream $\boldsymbol{\psi}^v = (\psi^v(y,x,l)) \in \mathbb{R}^{\tilde{Y}\times\tilde{X}\times\tilde{L}} = \mathbb{R}^{38\times2\times1500}$ was obtained. For the audio stream, we still used the same STFT parameters and synchronisation with $N_{fft} = 512$ and a hop-size of 128, and the audio sampling rate again became $f_s^a = 125$ Hz. We set

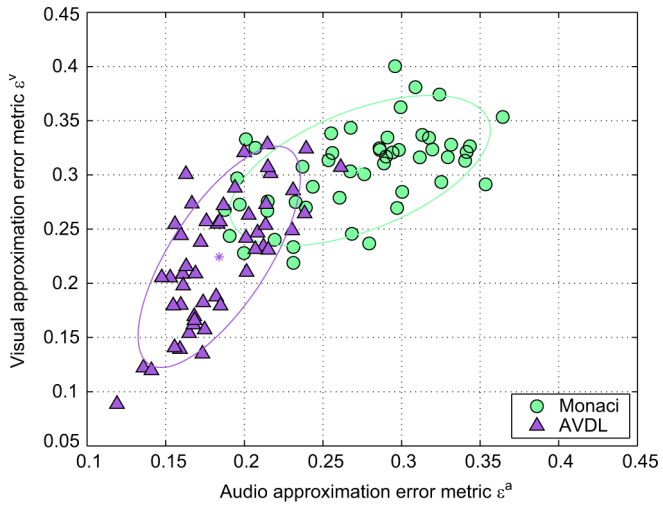


Fig. 7. The approximation error metrics comparison of AVDL and Monaci's method over 50 independent tests.

the dictionary size $K = 20$, the visual atom size $Y \times X \times L = 38 \times 2 \times 10$. Therefore, we had the audio atom size $W \times M$, where $W = N_{fft}/2 + 1 = 257$ and $M = \lceil (f_s^a / f_s^v) L \rceil = 50$. $N = 150$ was set for the sparsity. The other parameters were the same as those used in the previous subsection for synthetic data.

We applied Monaci's method for comparison. For the visual stream, we first manually cropped a rough lip region from the original video as the visual sequence with the size of $\tilde{Y} \times \tilde{X} \times \tilde{L} = 80 \times 120 \times 1500$. We then 'pre-whitened' the cropped video data to highlight the moving object edges, i.e., pixels in the neighbourhood of lip contours, and the 3D whitening technique [16] was applied. For the audio stream, a normalized audio sequence with a unit maximum magnitude and a sampling rate of $f_s^a = 16$ kHz was used. We set the visual atom size $Y \times X \times L = 64 \times 96 \times 10$. Therefore, the audio atom size was $1 \times M$, where $M = \lceil (f_s^a / f_s^v) L \rceil = 6400$. To balance the audio and visual modalities, we amplified the audio stream with a factor[4] of 10000. The other parameters were set the same as for AVDL.

Results Comparison and Analysis: Both dictionary learning methods converged after 15 iterations. Our proposed AVDL produced 13 AV atoms, while the baseline method produced only 7 AV atoms. We show one converged AV atom for each algorithm in Fig. 8, which correspond to the same generative AV atom.

From Fig. 8, we notice that the visual atom obtained via AVDL has a distinct outline, while the one obtained via Monaci's method is blurred, which means the learned visual atom via the baseline method tends to be distorted by other visual atoms or outliers. For the audio atom, some useful signal parts are truncated with Monaci's method, compared to that learned via AVDL. The cause of this problem is that some audio segments with high magnitude (and energy) are matched with the AV atom although their audio structures are not very similar, i.e., an outlier might be incorrectly matched

[4]This factor is not adaptive to size changes, and it was empirically chosen in case the baseline method is reduced to audio-only or visual-only. This factor was only effective with the predefined parameter setup. Any parameter change, e.g., the visual atom size set to $32 \times 32 \times 10$ resulted in the failure of the baseline method in our tests.

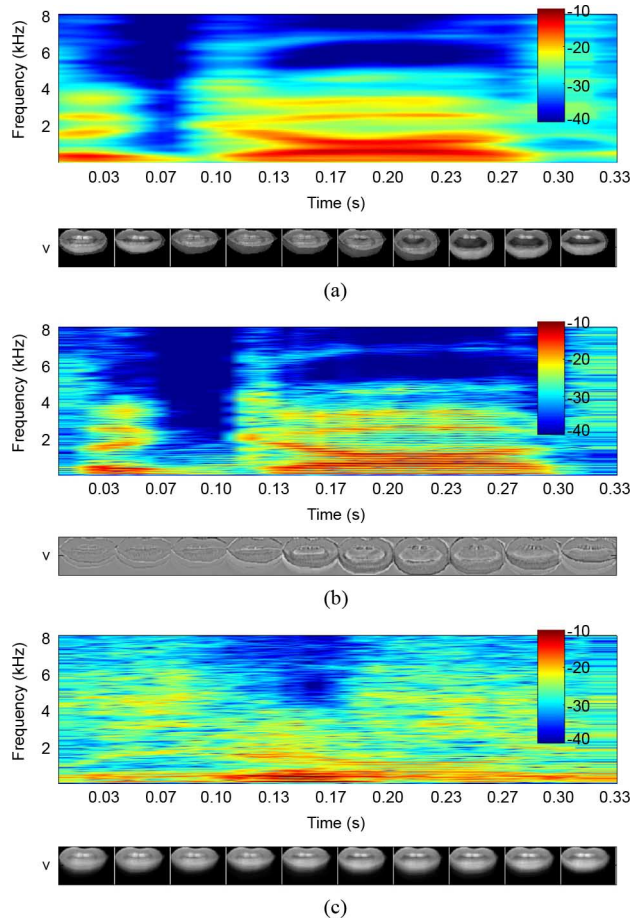


Fig. 8. The converged AV atoms after applying our proposed AVDL algorithm (a) and the competing method (b) to AV data from LILiR Twotalk database. The visual atom in AVDL contains only the normalized coordinates of the lip region. For ease of visual comparison, we reconstruct the lip intensity images, by mapping the TS positions of a learned visual atom in the lip contour data to the original video, and calculating the mean of the projected regions. To compare our proposed algorithm with the competing method on a fair basis, we also implemented Monaci's algorithm using the extracted lip features as used in our algorithm, rather than the 3D-filtered visual raw data as stated in their original paper. The competing algorithm converged to atoms with ambiguous visual contours and spurious audio spectrum, as demonstrated in the bottom row (c), which shows that Monaci's method is limited in this circumstance. The reason is that the inner product used to calculate the visual matching criterion in (4) can not distinguish different lip contour features very well. For example, a relatively high-valued inner product can still be calculated between two lip contours associated with different utterances, which may result in a high AV-valued matching criterion. To avoid this situation, Monaci's algorithm is applied to the pre-processed raw data for the following experiments, as introduced previously. (a) AVDL; (b) Monaci; (c) Monaci-Lip Feature.

with the AV atom. The updated AV atom, i.e., the first principal component of all the matched audio segments associated with the AV atom, is affected by the outlier, and therefore suffers from the information loss and distortion. It is worth noting that, due to the lack of ground truth AV atoms for real speech data, quantitative evaluations of the quality of the learned dictionary by the proposed method in comparison with Monaci's method become difficult. This offers an interesting point for future investigation.

B. AV-BSS Evaluations

1) Off-Line Training: The LILiR Twotalk corpus is also used here for training the AV dictionary, to improve the performance

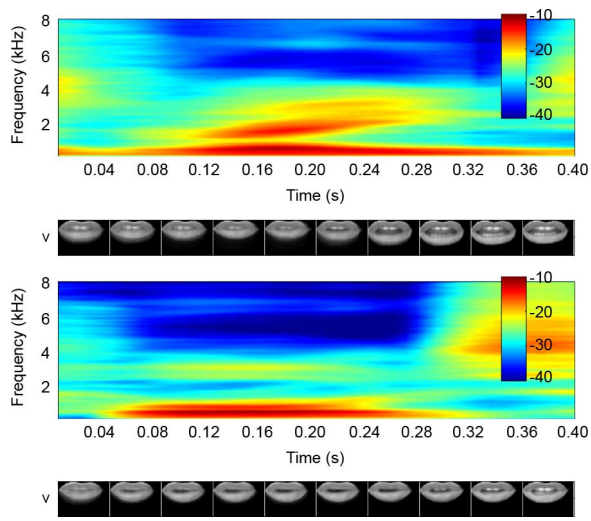


Fig. 9. Converged dictionary atoms using our proposed AVDL. Each learned AV atom contains an audio atom and an associated visual atom. The audio atom is the magnitude spectrum, while the visual atom contains the 38 lip contour coordinates in 10 consecutive frames. In the above figure, we reconstruct the lip regions via the lip contour coordinates to demonstrate the learned visual atoms. The first AV atom represents the utterance 'marine' /mari:n/ while the second one denotes the utterance 'port' /por:t/.

of the BSS. The first four of six sequences lasting 23278 frames in total (approximately 931 s) were concatenated for training. We enlarged the dictionary size to $K = 100$ to represent the AV sequence more explicitly. $N = 2328$ was set for the sparsity. For the other parameters, we used the same setup as in the previous Section V-A.2.

After the dictionary learning, $K = 80$ meaningful dictionary atoms were learned, of which two are shown in Fig. 9.

The converged atoms can represent about 7380 frames of the training sequence in the coding process, excluding about 6400 silence frames[5]. Therefore, there were still 9500 frames that were not properly represented by the learned dictionary. The reason behind this is that human speech produces a very complex signal, and a limited number of atoms (80 in our experiment) can hardly represent all the utterance variations. Also, some utterances appeared only once or too few times, therefore AVDL may consider them as outliers. Furthermore, we need to stress that we aim to learn and reconstruct the most bimodality-informative structures of the AV sequence, rather than fully reconstructing it. Monaci's method is also applied to train another dictionary, which is used in AV-BSS for comparison.

2) Data, Parameter Setup for Separation: We used the other two video sequences associated with the target speaker for testing. The interfering speech came from another speaker. We considered real-environment auditory mixtures, assuming a time-invariant mixing process. The binaural room impulse responses (BRIRs) measured by Hummersone [47] were used, which were recorded using a dummy head in four reverberant rooms in the University of Surrey, indexed by A, B, C and D. The average reverberation times are $[320 470 680 890]$ ms respectively. To simulate the room mixtures, we set the target speaker to an azimuth of zero degrees, i.e., in front of the dummy head, and we changed the azimuth of the interfering speaker on the right hand side, varying from 15° to 90° with an increment of 15°. In each of the six interference angles, 15 pairs of source signals were randomly chosen from the two testing sequences associated with the target speaker, as well as sequences associated with the interfering speaker, each lasting 10 s. The source signals were passed through the BRIRs to generate the mixtures. To test the robustness of our algorithm, Gaussian white noise was added to the mixtures at different SNRs.

We compared our proposed AV-BSS method, i.e., 'AVDL-BSS', with four competing algorithms that we implemented. The first one is Mandel's state-of-the-art audio-domain method, as introduced in Section IV-A, which we denote as 'Mandel'. We then incorporated the learned dictionary via Monaci's method using the proposed separation method, denoted as 'Monaci-BSS'. Since the audio atoms learned via Monaci's method are in the time domain, while the separation is mainly in the TF domain, we first transformed these atoms into the TF domain. We also compared the results of another AV-BSS method that we proposed previously based on ICA [27], where the AV coherence is modelled by Gaussian mixture models and coherence maximization is used to solve the permutation problem caused by ICA techniques, which we denote as 'AV-LIU'. However, we neglected the feature selection process, since the visual sampling rate for the dataset was 25 Hz, which was relatively low compared to the data used in [27]. To demonstrate how each part of our proposed AV-BSS works, we added an intermediate experiment, where the visual masks generated by the AV sparse coding, as introduced in Section IV-B, are applied directly to the binaural mixtures for source separation. This is denoted as 'Visual-only'.

3) Demonstration of TF Mask Fusion in AV-BSS: To demonstrate the AV fusion process, where the visual mask constrains the audio mask to produce a noise-robust AV mask, we compare the audio mask, the visual mask, and the audio-visual mask with an ideal binary mask (IBM) [48]. Supposing the source signals are known in advance, the IBM [48] can be calculated and used as a bench-mark for speech separation performance evaluation. For demonstration purposes, masks spanning a block of 30 time-frames are shown in Fig. 10(b). We also plotted the log-spectra of the two original source signals, and the binaural mixture signals in the associated time frames.

From Fig. 10(b), we notice that the IBM gives an accurate description of the target speech (source 1) at each TF point. The boundary region distinguishes the target signal from the interfering signal in perfect detail. The audio mask presents a relatively accurate approximation of the target signal. However its accuracy is affected by the competing signal, which is particularly evident in those TF points having a very low confidence with values around 0.5. The visual mask gives a rough approximation of the target signal, which however suffers greatly from information loss, especially the detailed information. The AV mask is generated by adjusting the audio mask with the visual mask, which keeps the detailed information of the audio mask, and enhances the TF points with low audio confidence towards the IBM based on the visual mask.

[5]After the coding process, the AV atoms from the dictionary are sparsely located at 738 TS positions, where each AV atom spans 10 frames. Therefore, 7380 frames are approximated ignoring the potential overlap between two allocated AV atoms. The approximate number of silence frames, i.e., 6400, is obtained in terms of $\mathcal{S}^{\alpha}(\check{l}) = 0$, as defined in (11).

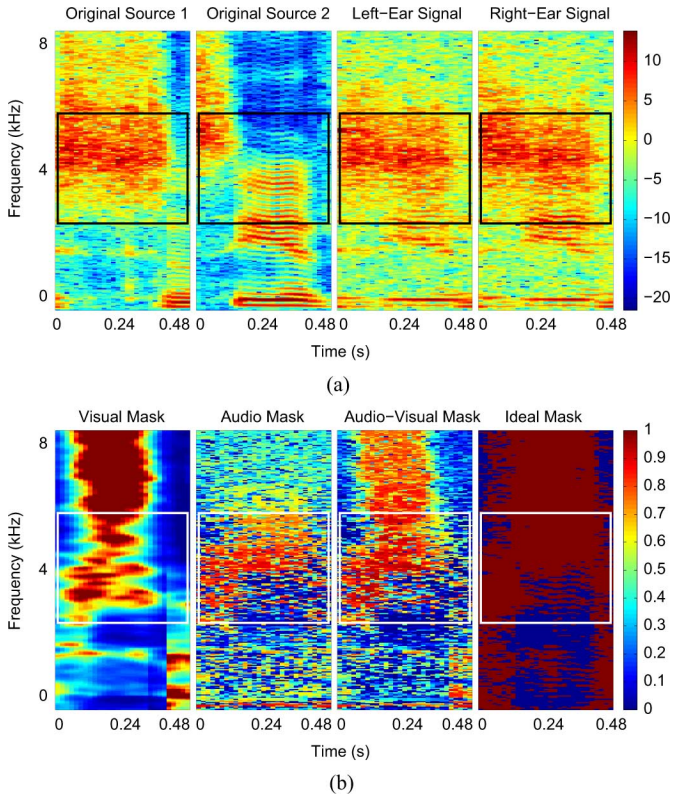


Fig. 10. (a) Spectrograms of source and mixture signals. (b) Comparison of the audio mask, the visual mask, the AV mask with the ground-truth IBM. Our proposed algorithm amplifies the audio mask when the confidence of the visual-only mask is high, and attenuates the audio mask when its confidence is low. The power law regularisation pushes the AV mask towards the IBM. The highlighted rectangles are further analysed in Fig. 12. (a) Sources and mixtures; (b) Masks.

4) Experimental Results of AV-BSS: To evaluate the performance of the BSS methods[6], we used a new perceptual objective quality assessment of audio source separation using PEASS [44] tool-kit. In PEASS [44], the overall-perceptual-score (OPS) has a high coherence with subjective perceptual evaluation, which we denote as OPS-PEASS, and it was used as the perceptual evaluation metric.

From Fig. 11, we found that our method suffered from an average of 3 points loss compared to Mandel's method in the noise-free condition. We believe the reason lies in the imperfect match between atoms from the learned dictionary with the testing sequence. Since one learned atom resembles the common characteristics of one AV event that occurs at different TS positions, which is not identical with any new occurrence of the same AV event, some artificial distortion is incurred. In an ideal noise-free environment, the audio-domain method already successfully generates an accurate audio mask in the TF domain, and further processing using the visual mask may introduce artificial distortions that degrade the accuracy to some extent. However in adverse conditions, our method shows some advantages over Mandel's method with an average of 2 point improvement. Even though the improvement was modest, it demonstrates that our learned dictionary inherited the underlying audio-visual coherence, and each converged AV atom could be used for separation (and potentially other

[6]We also performed the evaluations based on the signal to distortion ratio (SDR). The results are omitted due to space limitation.

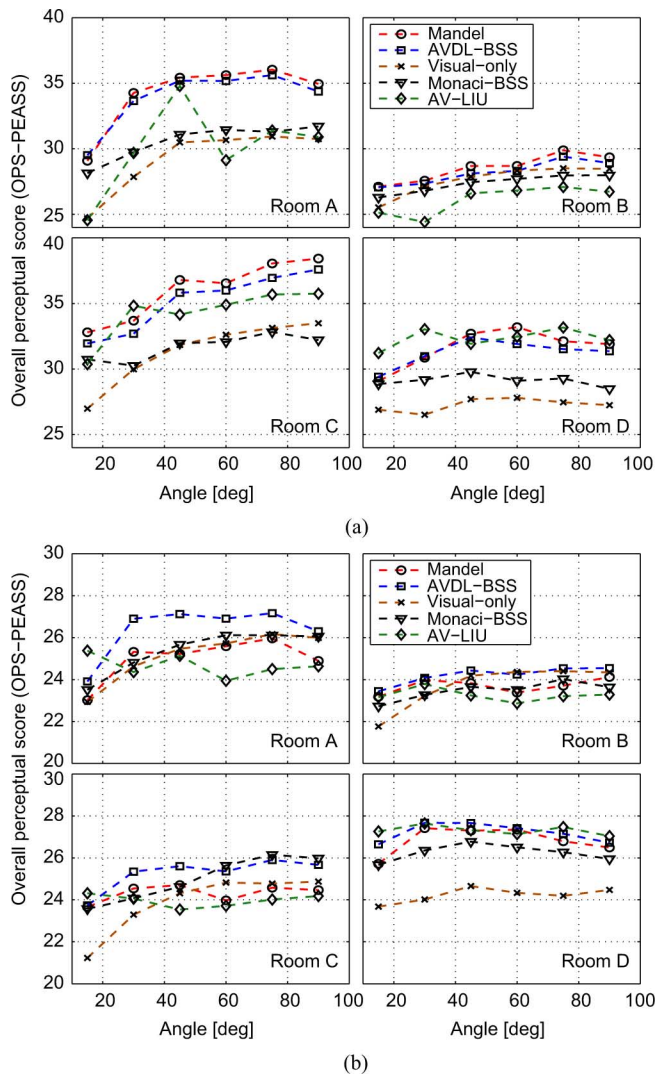


Fig. 11. OPS-PEASS evaluations without noise (upper) and with 10 dB Gaussian noise (bottom). (a) OPS-PEASS evaluations without noise; (b) OPS-PEASS evaluations with 10 dB Gaussian noise.

applications in the field of AV signal processing such as localisation, verification and recognition). Using the AV dictionary learnt via Monaci's method, Monaci-BSS cannot compete with our proposed algorithm. This is because Monaci's dictionary learning method introduces more distortion, since the same AV mask fusion process was applied in both Monaci-BSS and AVDL-BSS which is also consistent with the results presented in Section V-A.1.

Using only AV sparse coding, i.e., Visual-only, however, can not achieve satisfactory results for source separation tasks. From Fig. 11, Visual-only shows the worst results, for both noise-free and 10 dB noise-corrupted situations. This is because the visual mask affects only the matched frames. In mismatched frames, the visual mask cannot determine the audio information.

We also note when the two sources are very near to each other (input angle is small), all the methods fail to produce satisfactory results, since the mixing filters exhibit strong singularity, and hence give similar directions of arrival (DOA) as well as near-zero IPDs and ILDs. This is because in the AV mask fusion process, the visual mask is used to modify the audio mask, rather

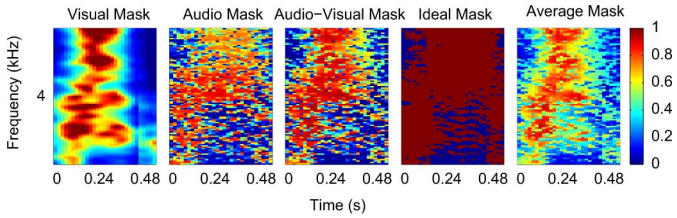


Fig. 12. The highlighted areas of the masks shown in Fig. 10(b) are compared with the same area extracted from the average AV mask, obtained by a symmetrical linear audio and visual mask combination $(\mathcal{M}^a + \mathcal{M}^v)/2$. The average AV mask is degraded due to the low confidence of the visual mask, whose accuracy is lower than the audio mask. The structure of the average mask is further pushed away from that of the IBM, compared to the audio mask. However, using our proposed AV mask fusion method with the power law transformation, the generated AV mask resembles the IBM better than the average AV mask.

than a symmetrical combination where the visual mask can take over when the audio mask fails. The symmetrical combination is not used since the visual mask fails to outperform the audio mask in our experiments, and a visual mask of low confidence is likely to degrade the overall AV mask confidence, if they are fused using a linear superposition. We illustrated a linear fusion in Fig. 12 where the overall AV mask (on the rightmost) is the average of the audio and visual masks.

Interestingly, the AV-LIU method shows the highest OPS score for the most reverberant room D. There are three possible reasons behind this observation. First, the sigmoid function that is used in PEASS to non-linearly evaluate the target distortion, the interference distortion and the artefact distortion, gives less priority to the artefact distortion mainly caused by background noise. Second, the long reverberation blurs the TF spectrum, which exhibits consistency in the ICA-based separation process and therefore suppresses the interference distortion. Third, the separation is not dependent solely on the reverberation time, it is affected by direct-to-reverberant ratios (DRRs) as well, where DRRs for the four different reverberant rooms are $[6.09\ 5.31\ 8.82\ 6.12]$ dB respectively. The complex relationship of the ICA performance with RT60 and DRR needs further study.

VI. Conclusions

We have developed an audio-visual dictionary learning (AVDL) algorithm that can capture the most AV-coherent structures of an AV sequence. The dictionary learned via AVDL implicitly inherits the AV coherence robust to acoustic noise, and therefore can be used to improve the performance of traditional audio domain BSS methods in noisy environments. In our proposed AV-BSS system, a visual mask is generated by matching the corrupted AV sequence to the learned AV dictionary. Considering the binaural room mixtures, an audio mask is generated in parallel using the spatial cues of IPDs and ILDs. Integrating the above two masks, a visually constrained noise-robust mask is generated for separating the target speech signal. We have tested our proposed AVDL on both synthetic data and the LILiR Twotalk corpus, and numerical results show the advantages of our method, with a greatly reduced computational load and a smaller approximation error rate, compared to another baseline audio-visual dictionary learning method. We have also tested our proposed AV-BSS method using a dictionary learned from the LILiR Twotalk corpus, which shows a performance improvement in noisy reverberant room environments in term of overall-perceptual-score (OPS) using the PEASS tool-kit.

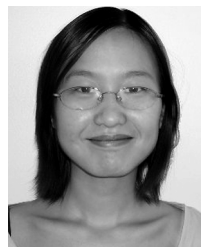
Acknowledgment

The authors thank the Associate Editor and the anonymous reviewers for their contributions to improving the quality of the paper.

References

[1] 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.
[2] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Commun.*, vol. 53, no. 4, pp. 524–539, 2011.
[3] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.
[4] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
[5] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc. F Radar Signal Process.*, vol. 140, no. 6, pp. 362–370, 1993.
[6] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
[7] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.
[8] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1135–1146, 2003.
[9] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, vol. 5, pp. 2985–2988.
[10] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.
[11] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.
[12] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," *Proc. ICASSP*, pp. 209–212, 2011.
[13] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 5, pp. 1503–1512, May 2012.
[14] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[15] M. Sams, R. Aulanko, M. Hämäläinen, R. Hari, O. V. Lounasmaa, S.-T. Lu, and J. Simola, "Seeing speech: visual information from lip movements modifies activity in the human auditory cortex," *Neurosci. Lett.*, vol. 127, no. 1, pp. 141–145, 1991.
[16] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Curr. Opin. Neurobiol.*, vol. 11, no. 4, pp. 505–509, 2001.
[17] D. A. Bulkin and J. M. Groh, "Seeing sounds: visual and auditory interactions in the brain," *IEEE Trans. Neural Netw*, vol. 16, no. 4, pp. 415–419, Apr. 2006.
[18] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.
[19] 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, Sep. 2003.
[20] 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. Appl. Signal Process*, vol. 2002, no. 11, pp. 1165–1173, 2002.
[21] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. Chambers, "Video assisted speech source separation," *Proc. ICASSP*, pp. 425–428, 2005.

[22] 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, 2007.
[23] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimed.*, vol. 12, no. 5, pp. 358–371, 2010.
[24] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks, "A geometrically constrained multimodal approach for convolutive blind source separation," *Proc. ICASSP*, vol. 3, pp. III969–III972, 2007.
[25] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach for blind source separation of moving sources," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 5, pp. 895–910, May 2010.
[26] 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 Process.*, vol. 6, no. 5, pp. 466–477, 2012.
[27] Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve the permutation problem in convolutive BSS," *Signal Process.*, vol. 92, no. 8, pp. 1916–1927, 2012.
[28] Q. Liu, W. Wang, P. Jackson, and M. Barnard, "Reverberant speech separation based on audio-visual dictionary learning and binaural cues," in *Proc. IEEE Statist. Signal Process. Workshop (SSP)*, 2012, pp. 664–667.
[29] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," *Proc. Sensor Signal Process. Defence*, 2011.
[30] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio-visual data," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, Dec. 2009.
[31] B. A. Olshausen, "Sparse coding of time-varying natural images," *Proc. ICA*, pp. 603–608, 2000.
[32] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Comput. Soc. Conf. on Comput. Vision Pattern Recogn.*, 2010.
[33] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.
[34] M. S. Lewicki, T. J. Sejnowski, and H. Hughes, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 1998.
[35] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, 2006.
[36] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.
[37] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval, "Learning multi-modal dictionaries," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2272–2283, 2007.
[38] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.
[39] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Conf. Rec. 27th Asilomar Conf. on Signals, Syst. Comput.*, 1993, pp. 40–44.
[40] D. L. Donoho, S. S. Chen, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
[41] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," *Proc. ICASSP*, vol. 5, pp. 2443–2446, 1999.
[42] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, pp. 349–396, 2003.
[43] W. Dai, T. Xu, and W. Wang, "Dictionary learning and update based on simultaneous codeword optimization (SimCO)," *Proc. ICASSP*, pp. 2037–2040, 2012.
[44] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.
[45] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Cultural factors in the regression of non-verbal communication perception," in *Proc. IEEE Int. Conf. Comput. Vision Workshops (ICCV)*, 2011, pp. 1242–1249.
[46] E.-J. Ong and R. Bowden, "Robust lip-tracking using rigid flocks of selected linear predictors," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recogn.*, 2008.
[47] C. Hummersone, "A Psychoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments," Ph.D., Univ. Surrey, Surrey, 2011.
[48] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York: Springer, 2005, ch. 12, pp. 181–197.



Qingju Liu received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China, in 2008. After that she took one year study in communication and information systems in the Graduate School of Information Science and Engineering in Shandong University.

Since 2009, she has been pursuing the Ph.D. degree under the supervision of Dr. Wenwu Wang within the Machine Audition Group at the Centre for Vision, Speech and Signal Processing in University of Surrey, Guildford, U.K. Her current research interests include audiovisual signal processing and machine learning.

Wenwu Wang (M'02–SM'11) received the B.Sc. degree in automatic control in 1997, the M.E. degree in control science and control engineering in 2000, and the Ph.D. degree in navigation guidance and control in 2002, all from Harbin Engineering University, Harbin, China.

He then joined King's College, London, U.K., in May 2002, as a Postdoctoral Research Associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group Ltd. (now Antix Labs Ltd.), Reading, U.K., as a DSP engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an R&D engineer, working on 3D spatial audio for mobile devices. Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Senior Lecturer, and a Co-Director of the Machine Audition Lab. He is also a member of the MoD UDRC in Signal Processing and the BBC Audio Research Partnership. During spring 2008, he has been a visiting scholar at the Perception and Neurodynamics Lab and the Center for Cognitive Science, The Ohio State University. His current research interests include blind signal processing, sparse signal processing, audiovisual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He has (co-)authored over 130 publications in these areas, including a book entitled *Machine Audition: Principles, Algorithms and Systems* (IGI Global).

Dr. Wang has been a regular reviewer for many IEEE journals including IEEE TRANSACTIONS ON SIGNAL PROCESSING. He was a Local Arrangement Co-Chair of the 2013 IEEE International Workshop on Machine Learning for Signal Processing, an Area Chair of the 2012 European Signal Processing Conference (EUSIPCO), a Track Chair and Publicity Co-Chair of 2009 IEEE Statistical Signal Processing Workshop, Program Co-Chair of the 2009 IEEE Global Congress on Intelligent Systems. He has been a Session Chair for numerous conferences including 2012 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) and the EUSIPCO 2012. He was a Tutorial Co-Speaker of "Dictionary Learning for Sparse Representations: Algorithms and Applications" on ICASSP 2013.

Philip J. B. Jackson received the M.A. degree in engineering from Cambridge University, U.K. and the Ph.D. degree in electronic engineering from the University of Southampton, U.K.

He joined the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K., in November 2002, as an academic with expertise in speech and audio processing. Together with Dr. W. Wang in CVSSP, he leads the Machine Audition Group of a dozen research fellows and students. His research in audio and speech processing has contributed to projects (e.g., BALTHASAR, DANSA, Dynamic Faces, QESTRAL, UDRC

and POSZ) in active noise control for aircraft, acoustics of speech production, source separation for automatic speech recognition (ASR), use of articulatory models for ASR, audiovisual processing for speech enhancement and visual speech synthesis, and spatial aspects of subjective sound quality evaluation. He has more than 100 publications in high-quality academic journals, conference proceedings and books.

Dr. Jackson is a reviewer for journals and conferences such as the *Journal of the Acoustical Society of America*, IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, *InterSpeech*, and *ICASSP*, and is an Associate Editor for *Computer Speech and Language* (Elsevier).

Mark Barnard received the Ph.D. degree (Docteur es Science) from Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland in November 2005.

While completing his Ph.D. degree, he was with The IDIAP Research Institute as a research assistant. His thesis was entitled "Multimedia Event Modeling and Recognition." In 2006, he joined the Machine Vision Group at the University of Oulu, Finland, where he completed three years as a Postdoctoral Researcher. He is currently a Research Fellow at the Centre for Vision, Speech and Signal processing at the University of Surrey, U.K. His current research interests include, audio-visual tracking, dictionary based image representation, and audio head pose estimation. He has published numerous papers in international journals and conferences.

Dr. Barnard has also served as a reviewer for many top-level journals and conferences.

Josef Kittler is a Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. He conducts research with a focus on biometrics, video and image database retrieval, automatic inspection, medical data analysis, and cognitive vision. He published a textbook, *Pattern Recognition: A Statistical Approach* (Englewood Cliffs, NJ: Prentice-Hall), and more than 170 journal papers.

Dr. Kittler serves as Series Editor for Springer-Verlag *Lecture Notes in Computer Science*.

Jonathon Chambers (S'83–M'90–SM'98–F'11) received the Ph.D. degree in signal processing from the Imperial College of Science, Technology and Medicine, Imperial College London, London, U.K., in 1990.

From 1991 to 1994, he was a Research Scientist with Schlumberger Cambridge Research Center, Cambridge, U.K. In 1994, he returned to Imperial College London, as a Lecturer in signal processing and was promoted as a Reader (Associate Professor) in 1998. From 2001 to 2004, he was the Director of the Centre for Digital Signal Processing and a Professor of signal processing with the Division of Engineering, King's College London. From 2004 to 2007, he was a Cardiff Professorial Research Fellow with the School of Engineering, Cardiff University, Wales, U.K. In 2007, he joined the Department of Electronic and Electrical Engineering, Loughborough University (LU), Loughborough, U.K., where he currently heads the Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering and serves as the Associate Dean (Research) LU in London. He is a coauthor of the books *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability* (Wiley, 2001) and *EEG Signal Processing* (Wiley, 2007). He has advised more than 50 researchers through to Ph.D. graduation and published more than 400 conference and journal articles, many of which are in IEEE journals. His research interests include adaptive and blind signal processing and their applications.

Dr. Chambers is a Fellow of the Royal Academy of Engineering, U.K., and the Institution of Electrical Engineers. He was the Technical Program Chair of the 15th International Conference on Digital Signal Processing (2007) and the 2009 IEEE Workshop on Statistical Signal Processing, both held in Cardiff, U.K., and a Technical Program Cochair for the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (2011), Prague, Czech Republic. He is the recipient of the first QinetiQ Visiting Fellowship in 2007 "for his outstanding contributions to adaptive signal processing and his contributions to QinetiQ" as a result of his successful industrial collaboration with the international defense systems company QinetiQ. He has served on the IEEE Signal Processing Theory and Methods Technical Committee for six years and the IEEE Signal Processing Society Awards Board and the European Signal Processing Society Best Paper Awards Selection Panel. He has also served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING for three terms over the periods 1997–1999, 2004–2007, and 2011– (and is currently a Senior Area Editor).