Blind Audio–Visual Localization and Separation via Low-Rank and Sparsity

Jie Pu, *Student Member, IEEE*, Yannis Panagakis, Stavros Petridis, *Member, IEEE*, Jie Shen, and Maja Pantic, *Fellow, IEEE*

***Abstract*—The ability to localize visual objects that are associated with an audio source and at the same time to separate the audio signal is a cornerstone in audio–visual signal-processing applications. However, available methods mainly focus on localizing only the visual objects, without audio separation abilities. Besides that, these methods often rely on either laborious preprocessing steps to segment video frames into semantic regions, or additional supervisions to guide their localization. In this paper, we aim to address the problem of visual source localization and audio separation in an unsupervised manner and avoid all preprocessing or post-processing steps. To this end, we devise a novel structured matrix decomposition method that decomposes the data matrix of each modality as a superposition of three terms: 1) a low-rank matrix capturing the background information; 2) a sparse matrix capturing the correlated components among the two modalities and, hence, uncovering the sound source in visual modality and the associated sound in audio modality; and 3) a third sparse matrix accounting for uncorrelated components, such as distracting objects in visual modality and irrelevant sound in audio modality. The generality of the proposed method is demonstrated by applying it onto three applications, namely: 1) visual localization of a sound source; 2) visually assisted audio separation; and 3) active speaker detection. Experimental results indicate the effectiveness of the proposed method on these application domains.**

***Index Terms*—Audio separation, audio–visual localization, low-rank, multimodal analysis, sparsity.**

I. Introduction

CROSS-MODAL analysis has recently received increasing attention from the signal-processing and computer vision communities, enabling the development of a wide range of applications, such as automatic speech recognition [1]; multimodal speaker diarization [23], [24]; audio–visual scene analysis [20]; and audio–visual object tracking [25], [26], to name but a few. In these tasks, the fusion of audio and visual modalities is crucial, providing information which is inaccessible when audio and visual data are analyzed independently.

In this paper, we investigate the problem of blind audio–visual localization and separation in real-world settings. The goal is to localize visual objects (by means of detecting pixels) associated with an audio signal and simultaneously separate the audio signal from irrelevant audio components and noise. More specifically, we focus on audio–visual data (i.e., videos) captured in real-world conditions (with distracting motions and noise, e.g., pedestrians in the street), using only one camera and one microphone (i.e., without microphone arrays and multiple cameras to provide additional spatial information). Clearly, this task is intrinsically difficult, which justifies the fact that the vast majority of the related methods mainly focus on either sound-producing object localization (without audio separation) [3], [5]–[8], [21] or audio separation (without object localization) [10], [32]–[35]. Only two methods, namely, [2] and [4], have been proposed for joint audio–visual localization and separation.

Distinct from the previous methods, we propose a novel method for unsupervised audio–visual source localization and separation, using a robust matrix decomposition [9], [16], [31], [36]. Our decomposition is based on two intrinsic properties of audio–visual data, namely, the low-rank of the background visual/audio information and the sparsity of the foreground components. That is, we assume that the background of the video lies in a low-dimensional subspace, while the moving foreground objects are considered as sparse within the image sequence. Similarly, a time-frequency distribution (e.g., spectrogram) of the audio signal is assumed to be a superposition of low-rank and sparse parts, corresponding to the spectrogram of the background and the foreground audio. Such assumptions are common in low-rank and sparse models and are indeed reasonable for most realistic videos, with successful applications in robust face recognition [41], image alignment [31], background subtraction [42], turbulence detection [43], and monaural audio separation [10].

Consequently, here we seek to express visual (i.e., simple pixel intensities) and audio (i.e., the magnitude of the spectrogram) modalities as a sum of three terms: 1) a low-rank matrix capturing the uncorrelated background components

Manuscript received April 13, 2018; revised August 31, 2018 and November 10, 2018; accepted November 12, 2018. Date of publication December 13, 2018; date of current version April 15, 2020. This work was supported by the European Community Horizon 2020 through SEWA under Grant 645094 and through DE-ENIGMA under Grant 688835. The work of Y. Panagakis was supported by EPSRC Project (FACER2VM) under Grant EP/N007743/1. This paper was recommended by Associate Editor S. Cruces. *(Corresponding author: Jie Pu.)*

J. Pu is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: jie.pu15@imperial.ac.uk).

Y. Panagakis is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., also with the Department of Computer Science, Middlesex University London, London NW4 4BT, U.K., and also with Samsung AI Centre, Cambridge CB1 2JB, U.K.

S. Petridis, J. Shen, and M. Pantic are with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with Samsung AI Centre, Cambridge CB1 2JB, U.K.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

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/TCYB.2018.2883607

(background image and background sound); 2) a sparse matrix accounting for the correlated foreground component (the sound source in visual modality or the associated sound in audio modality); and 3) a sparse matrix modeling the uncorrelated foreground component (distracting moving objects or other sound). An overview of the proposed method is depicted in Fig. 1. The proposed method is coined as *coupled low-rank and sparse* (CLS) matrix decomposition.

This paper makes three main contributions summarized as follows.

1) A novel method for audio–visual localization and separation is developed by employing simple representations of audio and video modalities, that is, the spectrogram and pixel intensities, respectively.
2) The heart of the proposed method is a novel coupled structured matrix factorization and its algorithmic framework that facilitates finding correlated components across modalities, while at the same time accounts for background noise and irrelevant motions/sounds.
3) A new manually annotated dataset, that is, the *sound of pixels* (SOPs) dataset, is introduced for the task under study. The SOP dataset consists of 20 challenging videos (5465 frames in total) captured in the wild. This is the first publicly available[1] dataset for the task with annotations, which enable quantitative evaluations.

To thoroughly evaluate the performance of the method, a series of experiments have been conducted in three different applications: 1) visual localization of sound source; 2) visually assisted audio separation; and 3) active speaker detection. The experimental results indicate the effectiveness of the proposed approach.

The remainder of this paper is organized as follows. Section II reviews the related work. In Section III, the proposed CLS method is detailed. The SOP dataset is introduced in Section IV along with the experimental results. Section V concludes this paper.

II. Related Work

In this section, an overview of methods designed for simultaneous audio–visual localization and separation as well as the simpler tasks of audio–visual localization and audio separation, is provided.

Audio–Visual Localization and Separation Methods: As already mentioned, there are only two methods proposed for simultaneous audio–visual localization and separation [2], [4]. Both methods extract visual features from regions of images, and employ several correlation measures. Concretely, Barzelay and Schechner [4] represented the visual and audio signals as onsets and used the coincidence between audio and visual onsets to find the correlated audio–visual components. In [2], audio and video modalities are first decomposed into basic structures using redundant representations, and then their synchrony is quantified and used for audio–visual association. However, the localization results of these methods are not precise, while quantitative evaluation is absent. Besides that, they are designed to work with recordings captured in controlled settings and thus cannot deal with noise information in challenging real-world settings (e.g., distracting movements in visual modality and environmental noise in audio modality). Our preliminary work [36] also attempts to solve the problem of simultaneous audio–visual localization and separation, but it decomposes each modality into only two terms (either low-rank background or sparse correlated component) and does not consider any uncorrelated information. It can be viewed as a special case (simplified version) of the proposed method here and is not able to handle noise information in real-world conditions.

Audio–Visual Localization Methods: Several multimodal signal processing methods focusing only on the localization of the sound-producing visual objects (without audio separation) have been proposed [3], [5]–[8], [21]. These methods can be categorized in two categories based on the type of visual features employed: 1) pixel-level localization [3], [5], [6], [21], which directly use pixel values as visual features and 2) object-level localization [7], [8], which extract visual features from regions of images. In pixel-level approaches, Kidron *et al.* [3] used sparse canonical correlation analysis (SCCA) to find the correlation between the audio and video modalities and output correlated pixels as the result of localization. Casanovas and Vandergheynst [5] used nonlinear diffusion to capture the pixels whose motion is most consistent with changes of audio energy. A major drawback of pixel-level approaches is that the localization result is just sporadic pixels and carries little high-level semantic meaning. In object-level approaches, Izadinia *et al.* [7] extracted visual features (velocity and acceleration) for each visual object and then used CCA to identify the audio-associated visual object. Li *et al.* [8] proposed a region tracking algorithm to extract visual features, where it consists of two segmentation processes and two clustering processes. After the extraction of visual features, a nonlinear correlation measure (namely, the Winner-take-all hash) is implemented to search the most correlated visual object. The major drawback of object-level approaches is that they used hand-crafted visual features which require a laborious extraction process and it is sensitive to the change of parameters during the extraction.

Audio Separation Methods: Existing studies for audio separation fall into two categories according to the number of microphones they used: 1) multichannel audio separation with multiple microphones [27]–[29], [38]–[40] and 2) monaural audio separation with one microphone [10], [32]–[35]. In the multichannel audio separation, these microphones provide additional information for separation and hence produce better results. It is clear that more information makes the separation task easier, and then integrating visual information into this audio separation task becomes a promising idea. So circular microphone arrays and camera arrays are used to collect as much information as possible. On the other hand, studies on monaural audio separation deal with a more difficult problem, with audio information from only one microphone. These studies can be either supervised or unsupervised. The supervised systems tend to use deep neural

[1] https://ibug.doc.ic.ac.uk/resources/SOP/

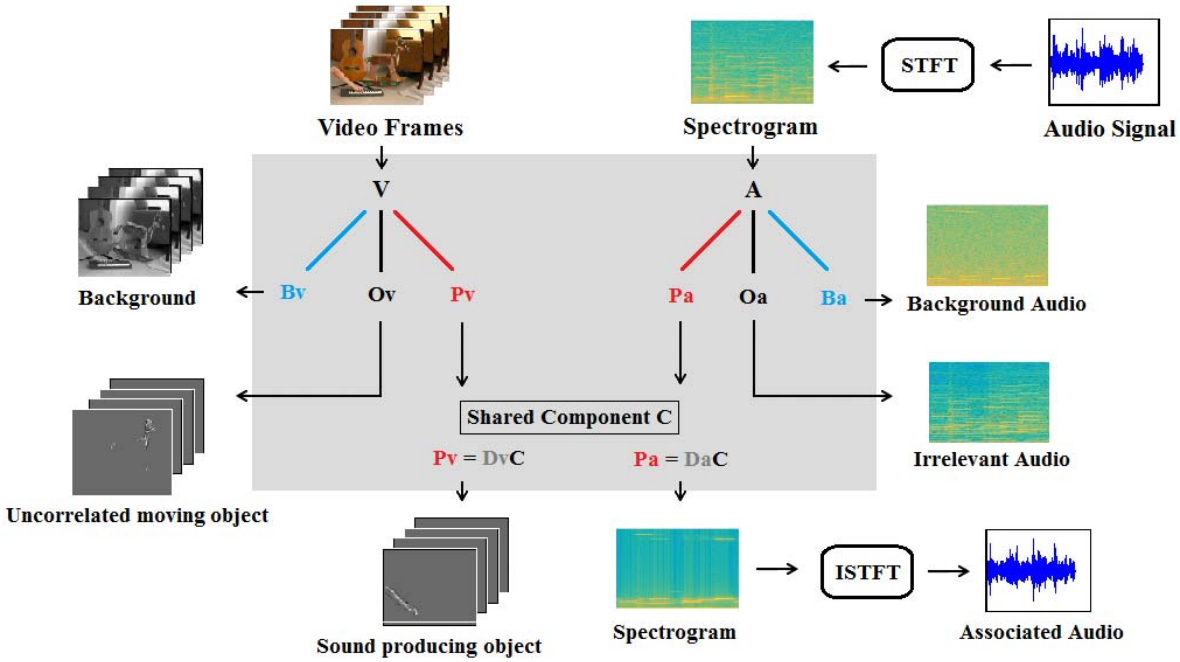


Fig. 1. Overview of the proposed audio–visual localization and separation method, in which the "STFT" represents the short-term Fourier transform and "ISTFT" is the inverse STFT. Matrix $\mathbf{V}$ contains the vectorized video frames in its columns while $\mathbf{A}$ represents the magnitude of the spectrogram, obtained by applying the STFT to the audio signal. The proposed method decomposes $\mathbf{V}$ and $\mathbf{A}$ as superposition of low-rank and sparse parts, where the low-rank matrix $\mathbf{B}_v$ captures the background uncorrelated visual information in video, the low-rank matrix $\mathbf{B}_a$ captures the background audio distraction, while the sparse matrices $\mathbf{P}_v$ and $\mathbf{P}_a$ capture the correlation among visual and acoustic modalities, revealing the location of pixels associated with the sound-producing visual object as well as its associated spectrogram. Two sparse matrices $\mathbf{O}_v$ and $\mathbf{O}_a$ represent the uncorrelated moving objects and irrelevant audio signals in visual and acoustic modalities, respectively.

networks [32], [33] or non-negative matrix factorization with pretrained dictionaries [34], [35], while unsupervised monaural audio separation methods often use low-rank and sparse models [10].

In the surge of deep learning era, there are methods [44]–[46] attempting to learn a common embedding between visual and audio modalities using deep neural networks. The common embedding learning is a more general task than sound-source localization and separation. The main difference between these methods and ours is that they require a huge amount of data for training (self-supervised), while the proposed method does not rely on any training data and is conducted in a completely unsupervised manner.

Distinct from the existing methodologies mentioned above, the proposed method in Section III can simultaneously localize the sound-producing object and separate its produced audio signal in real-life scenarios. Unlike [2] and [4] that are limited to controlled settings, our method works well in the presence of distracting motions, environmental noise, and sometimes other sounds. Unlike [7] and [8] that use hand-crafted visual features which require a laborious extraction process and are sensitive to the change of parameters during extraction, we use simple pixel values for visual representation and a spectrogram for audio representation, which leads to stable results. Unlike [44]–[46] which rely on a huge amount of data for training, our method is unsupervised and does not require any training data.

III. Proposed Methodology

In this section, the proposed CLS method along its solver is developed. To begin with, consider $\mathbf{V} \in \mathbb{R}^{I_1 \times T}$ and $\mathbf{A} \in \mathbb{R}^{I_2 \times T}$ representing the visual and audio modalities, respectively, where T is the number of frames in the video. Each column of $\mathbf{V}$ contains the vectorized image pixels at one frame, so I_1 is the number of pixels. As for the audio, we first transform the signal into a spectrogram using the parameterized short-term Fourier transform (STFT), where the parameters of the STFT are chosen in order to make the dimensionality of the spectrogram matrix having the same number as the video frames. Then we get the matrix $\mathbf{A}$ by keeping only the magnitude of spectrogram. So each column of $\mathbf{A}$ stands for the magnitude of spectrogram corresponding to one video frame, and I_2 is the dimensionality of the magnitude vector. It is worth mentioning that the visual frame and audio frame should be synchronized beforehand.

Let us now define the notions used. Throughout this paper, matrices (vectors) are denoted by uppercase (lowercase) boldface letters, e.g., $\mathbf{X}$ and ($\mathbf{x}$). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. $\mathbf{0}$ is the zero matrix. The ith column of matrix $\mathbf{X}$ is denoted as $\mathbf{x}_i$, and the entry of $\mathbf{X}$ at position (i, j) is denoted by x_{ij}. For the set of real numbers, the symbol $\mathbb{R}$ is used. Regarding matrix norms, $\|\mathbf{X}\|_*$ denotes the nuclear norm and it is defined as the sum of its singular values; the matrix ℓ_1-norm is denoted by $\|\mathbf{X}\|_1 \doteq \sum_i \sum_j |x_{ij}|$, where $|\cdot|$ represents the absolute value operator. $\|\mathbf{X}\|_F \doteq \sqrt{\sum_i \sum_j x_{ij}^2} =$

$\sqrt{\text{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, where tr(·) denotes the trace of matrices.

A. Model Formulation

In order to localize the visual object that produces sound and separates its associated audio signal, we seek to decompose each matrix into three terms

$$\begin{aligned}\mathbf{V} &= \mathbf{B}_v + \mathbf{P}_v + \mathbf{O}_v \\ \mathbf{A} &= \mathbf{B}_a + \mathbf{P}_a + \mathbf{O}_a\end{aligned} \tag{1}$$

where $\mathbf{B}_v \in \mathbb{R}^{I_1 \times T}$ and $\mathbf{B}_a \in \mathbb{R}^{I_2 \times T}$ are the low-rank components, capturing the information about background images and background sounds, respectively. $\mathbf{P}_v \in \mathbb{R}^{I_1 \times T}$ and $\mathbf{P}_a \in \mathbb{R}^{I_2 \times T}$ are the sparse matrices accounting for the foreground sound source in images and the correlated part of sounds, respectively. $\mathbf{O}_v \in \mathbb{R}^{I_1 \times T}$ and $\mathbf{O}_a \in \mathbb{R}^{I_2 \times T}$ are the sparse matrices accounting for uncorrelated moving objects and irrelevant sounds.

To ensure that $\mathbf{P}_v$ and $\mathbf{P}_a$ are maximally correlated, they are further decomposed as follows:

$$\begin{aligned}\mathbf{P}_v &= \mathbf{D}_v \cdot \mathbf{C} \\ \mathbf{P}_a &= \mathbf{D}_a \cdot \mathbf{C}\end{aligned} \tag{2}$$

where matrices $\mathbf{D}_v \in \mathbb{R}^{I_1 \times K}$, $\mathbf{D}_a \in \mathbb{R}^{I_2 \times K}$, and $\mathbf{C} \in \mathbb{R}^{K \times T}$. The matrix $\mathbf{C}$ represents a common (shared) low-dimensional subspace between the matrices $\mathbf{P}_v$ and $\mathbf{P}_a$, while the matrices $\mathbf{D}_v$ and $\mathbf{D}_a$ map the low-dimensional subspace to the high-dimensional pixel domain or spectrogram domain. In analogy to the multimodal dictionary learning [11], this underlying subspace, represented by $\mathbf{C}$, can be seen as a new feature representation, and its shared property enforces a latent relationship between $\mathbf{P}_v$ and $\mathbf{P}_a$. The latent relationship herein says that both $\mathbf{P}_v$ and $\mathbf{P}_a$ can be represented in a common feature space, where the correlation between audio and visual modalities is maximized. The parameter K (the number of rows in matrix $\mathbf{C}$) denotes the number of correlated components kept in the correlation between the visual and audio modalities.

In order to perfectly disentangle the sparse correlated components $\mathbf{P}_v$, $\mathbf{P}_a$ and the sparse uncorrelated components $\mathbf{O}_v$, $\mathbf{O}_a$, we establish a mutual orthogonality between $\mathbf{C}$ and $\mathbf{O}_v$, $\mathbf{O}_a$, respectively. That is to say, both $\mathbf{O}_v$ and $\mathbf{O}_a$ should be orthogonal to the shared low-dimensional subspace $\mathbf{C}$. The mutual orthogonality enforces $\mathbf{O}_v$, $\mathbf{O}_a$ to capture the individual and unique information in either visual or audio modality, which is intrinsically different from the correlated information in $\mathbf{P}_v$, $\mathbf{P}_a$. Then the orthogonal constraints are formulated as

$$\begin{aligned}\mathbf{C} \cdot \mathbf{O}_v{}^T &= \mathbf{0} \\ \mathbf{C} \cdot \mathbf{O}_a{}^T &= \mathbf{0}.\end{aligned} \tag{3}$$

A natural estimator accounting for the low rank of the $\mathbf{B}_v$, $\mathbf{B}_a$ components is to minimize their ranks. For the sparsity of the $\mathbf{P}_v$, $\mathbf{P}_a$, $\mathbf{O}_v$, $\mathbf{O}_a$, it is straightforward to minimize the number of nonzero entries, which can be measured by the ℓ_0-norm [12]. Nevertheless, due to the discrete nature of the rank and the ℓ_0-norm, their minimizations are NP-hard [13], [14] and thus intractable. Then, we use the nuclear norm $\|.\|_*$ and the ℓ_1-norm to serve as convex surrogation of the rank and ℓ_0-norm, respectively. Therefore, the objective function of our novel algorithm can be defined as follows:

$$\begin{aligned}\mathcal{F}(\mathcal{V}) = \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1\|\mathbf{P}_v\|_1 + \lambda_2\|\mathbf{P}_a\|_1 \\ + \lambda_3\|\mathbf{O}_v\|_1 + \lambda_4\|\mathbf{O}_a\|_1\end{aligned} \tag{4}$$

where the unknown matrices are collected in the set $\mathcal{V} \doteq \{\mathbf{B}_v, \mathbf{B}_a, \mathbf{P}_v, \mathbf{P}_a, \mathbf{O}_v, \mathbf{O}_a\}$, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the positive parameters to balance the significance of minimizing the sparsity compared to the minimization of ranks. In other words, for a larger value of λ, the optimal solution tends to obtain a sparser matrix and a less low-rank (higher rank) matrix, while a smaller λ works the other way around.

The slowness principle in slow feature analysis [47] enables us to obtain a smoother solution for the shared coding matrix $\mathbf{C}$. Since objects in the world are persistent and their appearance change with time in a continuous fashion, the shared coding matrix $\mathbf{C}$ that represents the correlation between two modalities should also change in a continuous way. Moreover, the videos that we are working on are short (average at 10 s), and then their accumulated temporal change of $\mathbf{C}$ will be small. Therefore, we can confine the temporal change of the matrix $\mathbf{C}$ and enforce its temporal closeness. To achieve this, a temporal Laplacian regularization function $\mathcal{G}(\mathbf{C})$ is adopted from [15] and defined as follows:

$$\mathcal{G}(\mathbf{C}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|c_i - c_j\|_2^2 = \text{tr}(\mathbf{C} \cdot \mathbf{L} \cdot \mathbf{C}^T)$$

where c_i is the ith column in the shared coding matrix $\mathbf{C}$, and $\mathbf{L}$ is a temporal Laplacian matrix. The intuition behind the temporal Laplacian function is that the temporal closeness of $\mathbf{C}$ is measured by the difference between its columns, where each column represents the correlated subspace in one frame. In detail, we define $\mathbf{L} = \hat{\mathbf{D}} - \mathbf{W}$, $\hat{\mathbf{D}}_{ii} = \sum_{j=1}^{n} w_{ij}$, and $\mathbf{W}$ is the weight matrix that captures the sequential relationships in data. Let s denote the number of sequential neighbors, and the element in $\mathbf{W}$ is calculated as

$$w_{ij} = \begin{cases} 1, & \text{if } |i-j| \le \frac{s}{2} \\ 0, & \text{if } |i-j| > \frac{s}{2}. \end{cases}$$

With the temporal regularization function $\mathcal{G}(\mathbf{C})$, the objective function of our algorithm in (4) becomes

$$\begin{aligned}\mathcal{F}(\mathcal{V}) = \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1\|\mathbf{P}_v\|_1 + \lambda_2\|\mathbf{P}_a\|_1 \\ + \lambda_3\|\mathbf{O}_v\|_1 + \lambda_4\|\mathbf{O}_a\|_1 + \lambda_5\text{tr}(\mathbf{C} \cdot \mathbf{L} \cdot \mathbf{C}^T)\end{aligned} \tag{5}$$

where the λ_5 is a positive parameter to indicate the significance of the temporal regularization function $\mathcal{G}(\mathbf{C})$. Therefore, we formalize the complete constrained optimization problem as follows:

$$\begin{aligned}&\underset{\mathcal{V}}{\text{minimize}}\ \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1\|\mathbf{P}_v\|_1 + \lambda_2\|\mathbf{P}_a\|_1 \\ &\qquad + \lambda_3\|\mathbf{O}_v\|_1 + \lambda_4\|\mathbf{O}_a\|_1 + \lambda_5\text{tr}(\mathbf{C} \cdot \mathbf{L} \cdot \mathbf{C}^T) \\ &\text{s.t. } \mathbf{V} = \mathbf{B}_v + \mathbf{D}_v\mathbf{C} + \mathbf{O}_v, \quad \mathbf{A} = \mathbf{B}_a + \mathbf{D}_a\mathbf{C} + \mathbf{O}_a \\ &\qquad \mathbf{P}_v = \mathbf{D}_v \cdot \mathbf{C}, \qquad \mathbf{P}_a = \mathbf{D}_a \cdot \mathbf{C} \\ &\qquad \mathbf{C} \cdot \mathbf{O}_v{}^T = \mathbf{0}, \qquad \mathbf{C} \cdot \mathbf{O}_a{}^T = \mathbf{0}\end{aligned} \tag{6}$$

where the set of unknown primary variables in $\mathcal{V}' \doteq \{\mathbf{B}_v, \mathbf{B}_a, \mathbf{P}_v, \mathbf{P}_a, \mathbf{D}_v, \mathbf{D}_a, \mathbf{O}_v, \mathbf{O}_a, \mathbf{C}\}$, all of λ is the positive regularization parameter to balance the significance of minimization, and the $\mathbf{L}$ is a predefined Laplacian matrix used to confine the temporal change of $\mathbf{C}$.

B. Optimization Algorithm

To solve the optimization problem (6), we developed an algorithm based on the alternating direction method of multiplier (ADMM) [30]. In particular, the inexact augmented Lagrange multipliers (ALMs) method is employed [48]. The method is a simple but powerful variant of ADMM that is suitable for large-scale optimization problems. It breaks a large global optimization problem into smaller pieces by iteratively solving for one variable with others fixed.

To this end, by incorporating all the constraints in (6), the augmented Lagrangian function is as follows:

$$\begin{aligned}\mathcal{L}(\mathcal{V}', \mathcal{M}) = &\|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1\|\mathbf{P}_v\|_1 + \lambda_2\|\mathbf{P}_a\|_1 \\ &+ \lambda_3\|\mathbf{O}_v\|_1 + \lambda_4\|\mathbf{O}_a\|_1 + \lambda_5 \mathrm{tr}\big(\mathbf{C}\cdot\mathbf{L}\cdot\mathbf{C}^T\big) \\ &+ \langle \mathbf{Y}, \mathbf{V} - \mathbf{B}_v - \mathbf{D}_v\mathbf{C} - \mathbf{O}_v\rangle \\ &+ \frac{\mu}{2}\|\mathbf{V} - \mathbf{B}_v - \mathbf{D}_v\mathbf{C} - \mathbf{O}_v\|_F^2 \\ &+ \langle \mathbf{Z}, \mathbf{A} - \mathbf{B}_a - \mathbf{D}_a\mathbf{C} - \mathbf{O}_a\rangle \\ &+ \frac{\mu}{2}\|\mathbf{A} - \mathbf{B}_a - \mathbf{D}_a\mathbf{C} - \mathbf{O}_a\|_F^2 \\ &+ \langle \mathbf{G}, \mathbf{D}_v\cdot\mathbf{C} - \mathbf{P}_v\rangle + \frac{\mu}{2}\|\mathbf{D}_v\cdot\mathbf{C} - \mathbf{P}_v\|_F^2 \\ &+ \langle \mathbf{F}, \mathbf{D}_a\cdot\mathbf{C} - \mathbf{P}_a\rangle + \frac{\mu}{2}\|\mathbf{D}_a\cdot\mathbf{C} - \mathbf{P}_a\|_F^2 \\ &+ \frac{\mu}{2}\Big(\|\mathbf{C}\cdot\mathbf{O}_v{}^T\|_F^2 + \|\mathbf{C}\cdot\mathbf{O}_a{}^T\|_F^2\Big) \end{aligned} \tag{7}$$

where $\mathcal{M} \doteq \{\mathbf{Y}, \mathbf{Z}, \mathbf{G}, \mathbf{F}\}$ gathers the Lagrange multipliers associated with the first four constraints in (6). $\mu > 0$ is a positive penalty parameter.

The inexact ALM method minimizes the $\mathcal{L}(\mathcal{V}', \mathcal{M})$ with respect to each variable in an alternating fashion, and then the Lagrange multipliers get updated at each iteration. The complete procedure is summarized in Algorithm 1. A detailed derivation of the algorithms is provided in Appendix A in the supplementary material.

Let us define the operators used in Algorithm 1. The shrinkage operator $\mathcal{S}_\tau(x)$ [9] is defined as $\mathcal{S}_\tau(x) = \mathrm{sgn}(x)\max(|x| - \tau, 0)$ and extend to every element in matrices. The singular value thresholding operator $\mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^*$ [17] and $X = U\Sigma V^*$ is any singular value decomposition. For the Sylvester equation of $\mathbf{C}$

$$\begin{aligned}\mathbf{M} &\coloneqq 2\mu\big(\mathbf{D}_v^T\mathbf{D}_v + \mathbf{D}_a^T\mathbf{D}_a\big) \\ \mathbf{N} &\coloneqq 2\lambda_5\mathbf{L} + \mu\big(\mathbf{O}_v^T\mathbf{O}_v + \mathbf{O}_a^T\mathbf{O}_a\big) \\ \mathbf{K} &\coloneqq -\mu\mathbf{D}_v^T\Big(\mathbf{V} - \mathbf{B}_v - \mathbf{O}_v + \frac{1}{\mu}\mathbf{Y} + \mathbf{P}_v - \frac{1}{\mu}\mathbf{G}\Big) \\ &\quad - \mu\mathbf{D}_a^T\Big(\mathbf{A} - \mathbf{B}_a - \mathbf{O}_a + \frac{1}{\mu}\mathbf{Z} + \mathbf{P}_a - \frac{1}{\mu}\mathbf{F}\Big).\end{aligned}$$

Algorithm 1 Inexact ALM Solver for (6)

Input: Data: the visual matrix $\mathbf{V}$ and the audio matrix $\mathbf{A}$. Parameters: $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 > 0$. K: the number of rows in matrix $\mathbf{C}$. The Laplacian matrix $\mathbf{L}$.

Initialize: Set the $\{\mathbf{B}_v[0], \mathbf{B}_a[0], \mathbf{P}_v[0], \mathbf{P}_a[0], \mathbf{D}_v[0], \mathbf{D}_a[0], \mathbf{O}_v[0], \mathbf{O}_a[0], \mathbf{C}[0], \mathbf{Y}[0], \mathbf{Z}[0], \mathbf{G}[0], \mathbf{F}[0]\}$ all to zero matrices, and $\mu > 0, \rho > 0, \theta > 0, t = 0$

while not converged **do**

$\mathbf{B}_v[t+1] \leftarrow \mathcal{D}_{\frac{1}{\mu}}[\mathbf{V} - \mathbf{D}_v[t]\mathbf{C}[t] - \mathbf{O}_v[t] + \frac{1}{\mu}\mathbf{Y}[t]]$

$\mathbf{B}_a[t+1] \leftarrow \mathcal{D}_{\frac{1}{\mu}}[\mathbf{A} - \mathbf{D}_a[t]\mathbf{C}[t] - \mathbf{O}_a[t] + \frac{1}{\mu}\mathbf{Z}[t]]$

$\mathbf{P}_v[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_1}{\mu}}(\mathbf{D}_v[t]\cdot\mathbf{C}[t] + \frac{1}{\mu}\mathbf{G}[t])$

$\mathbf{P}_a[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_2}{\mu}}(\mathbf{D}_a[t]\cdot\mathbf{C}[t] + \frac{1}{\mu}\mathbf{F}[t])$

$\mathbf{D}_v[t+1] \leftarrow \frac{1}{2}(\mathbf{V} - \mathbf{B}_v[t+1] - \mathbf{O}_v[t] + \frac{1}{\mu}\mathbf{Y}[t] + \mathbf{P}_v[t+1] - \frac{1}{\mu}\mathbf{G}[t])\cdot\mathbf{C}[t]^T\cdot\big(\mathbf{C}[t]\cdot\mathbf{C}[t]^T\big)^{-1}$

$\mathbf{D}_a[t+1] \leftarrow \frac{1}{2}(\mathbf{A} - \mathbf{B}_a[t+1] - \mathbf{O}_a[t] + \frac{1}{\mu}\mathbf{Z}[t] + \mathbf{P}_a[t+1] - \frac{1}{\mu}\mathbf{F}[t])\cdot\mathbf{C}[t]^T\cdot\big(\mathbf{C}[t]\cdot\mathbf{C}[t]^T\big)^{-1}$

$\mathbf{O}_v[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_3}{\mu\theta}}\{\mathbf{O}_v[t] + \frac{1}{\theta}[\mathbf{V} - \mathbf{B}_v[t+1] - \mathbf{D}_v[t+1]\mathbf{C}[t] + \frac{1}{\mu}\mathbf{Y}[t] - \mathbf{O}_v[t](\mathbf{I} + \mathbf{C}[t]^T\mathbf{C}[t])]\}$

$\mathbf{O}_a[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_4}{\mu\theta}}\{\mathbf{O}_a[t] + \frac{1}{\theta}[\mathbf{A} - \mathbf{B}_a[t+1] - \mathbf{D}_a[t+1]\mathbf{C}[t] + \frac{1}{\mu}\mathbf{Z}[t] - \mathbf{O}_a[t](\mathbf{I} + \mathbf{C}[t]^T\mathbf{C}[t])]\}$

$\mathbf{C}[t+1] \leftarrow$ solve the Sylvester equation: $\mathbf{M}\mathbf{C}[t+1] + \mathbf{C}[t+1]\mathbf{N} + \mathbf{K} = \mathbf{0}$

Update Lagrange multipliers

Update μ by $\mu \leftarrow \min\big(\rho\cdot\mu, 10^{18}\big)$

$t \leftarrow t + 1$

end while

Output: Background low-rank components $\{\mathbf{B}_v, \mathbf{B}_a\}$, sparse correlated components $\{\mathbf{P}_v, \mathbf{P}_a\}$ and sparse uncorrelated components $\{\mathbf{O}_v, \mathbf{O}_a\}$

C. Computational Complexity and Convergence

The dominant cost of each iteration in Algorithm 1 is the computation of the thresholding operator when updating $\mathbf{B}_v$ and $\mathbf{B}_a$. Thus, the complexity of each iteration is $\mathcal{O}(\max(I_1^2\cdot T, I_2^2\cdot T))$. Regarding the convergence of Algorithm 1, there is no established convergence proof of the inexact ALM to local minima when employed to solve nonconvex problems [30], [31]. A systematic convergence proof goes beyond the scope of this paper, yet for the proof of the weak convergence of Algorithm 1, one can follow the approach in [12]. In practice, the experiments in Section IV indicate that the proposed algorithm has stable convergence.

In detail, the convergence criterion is based on satisfying the constraints of the optimization problem (6). In each iteration, we monitor the errors of constraints and decide whether to terminate the algorithm by comparing with a

TABLE I
MAIN SPECIFICATIONS AND CONTENTS OF TEST VIDEO SEQUENCES

Index	Resolution H*W	Frame rate fps	No. of frames	Video content	Distracting objects	Audio content	Noise
V1	360*480	25.00	278	Playing harp	-	Harp sound	-
V2	360*450	25.00	278	Playing piano in the crowd	Pedestrians	Piano sound	Environment noise
V3	720*1280	25.00	363	Playing guitar in the street	Pedestrians	Guitar sound	Environment noise
V4	720*1280	29.97	301	Playing xylophone	-	Xylophone sound	-
V5	720*1280	25.00	310	Playing xylophone	-	Xylophone sound	-
V6	720*1280	25.00	338	Playing guitar	-	Guitar sound	-
V7	240*320	25.00	129	Playing violin	-	Violin sound	-
V8	384*480	24.87	101	Playing keyboards	Wooden horse	Keyboards sound	-
V9	720*1280	25.00	308	Playing keyboards	-	Keyboards sound	-
V10	720*1280	29.97	300	Interviewing two soldiers	Silent soldier	One soldier's speech	-
V11	720*1280	23.98	240	Speaking	-	Speaking	Environment noise
V12	288*512	29.97	300	Speaking	-	Speaking	Environment noise
V13	360*640	29.97	301	Interviewing one rescuer	Pedestrian	Speech (interviewee)	Speech (interviewer)
V14	720*1280	29.97	333	Playing guitar in the street	Pedestrian	Guitar sound	Environment noise
V15	720*1280	25.00	263	Playing guitar	-	Guitar sound	-
V16	360*640	23.97	156	Playing guitar	-	Guitar sound	-
V17	720*1280	25.00	255	Playing piano in the street	Cars	Piano sound	Environment noise
V18	360*480	25.00	323	Playing violin	-	Violin sound	-
V19	720*1280	25.00	250	Playing cello in the street	Pedestrians	Cello sound	Environment noise
V20	720*1280	25.00	338	Playing cello	-	Cello sound	-

predefined threshold δ. Since the orthogonal constraints have been included during the inexact ALM process, we only consider the first four constraints in (6). Then their corresponding error terms (E_1, E_2, E_3, and E_4) at the end of iteration t are

$$E_1[t] = \mathbf{V} - (\mathbf{B}_v[t] + \mathbf{D}_v[t]\mathbf{C}[t] + \mathbf{O}_v[t])$$
$$E_2[t] = \mathbf{A} - (\mathbf{B}_a[t] + \mathbf{D}_a[t]\mathbf{C}[t] + \mathbf{O}_a[t])$$
$$E_3[t] = \mathbf{P}_v[t] - \mathbf{D}_v[t] \cdot \mathbf{C}[t]$$
$$E_4[t] = \mathbf{P}_a[t] - \mathbf{D}_a[t] \cdot \mathbf{C}[t].$$

The error terms will be measured in the $\|\cdot\|_F$ norm, and the algorithm terminates only if all of them are no bigger than the threshold δ

$$\max(\|E_1[t]\|_F, \|E_2[t]\|_F, \|E_3[t]\|_F, \|E_4[t]\|_F) \leq \delta.$$

IV. EXPERIMENTAL EVALUATION

This section provides a thorough experimental evaluation of the proposed CLS method in practical applications. Some of the experimental demonstrations can be found on our website.[2] Three sets of experiments are conducted which are summarized as follows.

1) *Visual Localization of Sound Source:* The performance of the proposed method is first assessed in the task of sound source localization, where test videos come from real-life scenarios. Most videos contain more than one moving objects, but only one of them (the sound source) produces sound. The audio modality contains the audio signal associated with the moving object, environmental noise, and sometimes other sounds. The proposed method produces two sparse matrices $\mathbf{P}_v$ and $\mathbf{P}_a$, where $\mathbf{P}_v$ indicates the location of the moving sound source and $\mathbf{P}_a$ represents the associated audio signal. For the experimental evaluation, we can only evaluate the localization result for $\mathbf{P}_v$ although the audio separation process for $\mathbf{P}_a$ is conducted at the same time with localization. The reason is that there is no way to obtain the ground truth for the associated audio signal as the real-world audio is mixed naturally. The localization results are compared with those obtained by two state-of-the-art methods: one [8] for the object-level approach and the other [21] for the pixel-level approach.
2) *Visually Assisted Audio Separation:* To evaluate the capability of the proposed method in audio separation, we synthesize noisy audio signals by corrupting the original audio with either white noise or background music and then keep the original audio as the ground truth for audio separation. In this case, the synthetic noisy audio is used for the audio modality and the visual modality remains the same as the first set of experiments (visual localization). After obtaining the sparse matrix $\mathbf{P}_a$, we can now evaluate the audio separation process. Specifically, the result of audio separation is compared with the state-of-the-art unsupervised monaural audio separation method [10].
3) *Active Speaker Detection:* To demonstrate the generality of the proposed method, we exploit the result of visual localization and use it for the task of speaker detection. By calculating the energy distribution of localization results, the active speaker is identified as the one with the highest energy of pixels. The comparison is made against the widely used open-source toolbox LIUM [22], which has demonstrated the state-of-the-art performance for broadcast news diarization.

[2]https://sites.google.com/view/blindaudiovisual

A. Dataset

The new dataset, referred to as SOP, consists of 20 audio–visual recordings of sound sources, such as talking faces or music instruments. It contains 5465 frames in total, and every frame is manually annotated. Most audio–visual recordings are videos from Youtube except for video V8, which is from [3]. Besides, video V7 was used in [7] and [8], and V16 was used in [8]. Details of the video sequences are listed in Table I.

The SOP dataset is the first dataset for the task with annotations, which enable quantitative evaluations. There are bigger datasets [44], [46] used by deep neural networks, but they do not contain related annotations. For the studies [8] and [21] that contain quantitative evaluations, their methods are evaluated on two [21] and six [8] videos, respectively, which justifies the importance of introducing the SOP dataset.

The videos in the dataset have an average duration of 10 s, and they are all recorded by one camera and one microphone. The audio signals are sampled at 16 kHz for V7, V8, and V16 and 44.1 kHz for the rest. The video frames contain the sound-making object (sound source) and distracting objects (e.g., pedestrian on the street), while the audio signals consist of the sound produced by the sound source (human speech or instrumental music), environmental noise, and sometimes other sounds. The ground truth of visual localization has been annotated by annotators. The key principle is to annotate the significant movement that triggers the sound.

In the first set of experiments (visual localization), the proposed method is evaluated on all 20 videos in the dataset and gets compared with state-of-the-art methods. In the second set of experiments (audio separation), we use 12 videos (V1, V4–V10, V15, V16, V18, V20) out of the 20, where their audio signals are relatively clean and can be kept as the ground truth for audio separation. In the third set of experiments (active speaker detection), we created ten additional videos (V21–V30) from the SEWA database.[3] The original videos in SEWA contain two persons engaged in a dialogue. The dialogue is to discuss one advertisement they watched together. We merge two synchronous frames of the two speakers during a dialogue into one frame and use the proposed method to find "Who is speaking and When" in the dialogue. The annotation is provided by the SEWA database, where annotators watch through the videos and record the speaking time of each person. The ten test videos are sampled at 50 frames/s and have 8000 frames in length. They are videos corresponding to real-life scenarios where the dialogues contain frequently changed turns of speaking, unexpected silence, and simultaneous speech.

B. Quantitative Evaluation Criteria

1) Visual Localization Evaluation: Different evaluation criteria may favor different localization approaches. The object-level approach tends to produce one complete region with a well-defined boundary, while the pixel-level approach often produces isolated pixels with concentrated energy. The F_1-measure in [7] favors the object-level approaches since it measures the overlapping between the ground truth and the detected region. The L_c criterion defined in [3] is more suitable for pixel-level approaches since it provides the evaluation from an energy perspective. To evaluate the visual localization results more objectively, we combine the evaluation framework in [3] and [7] and report both the F_1 and L_c criteria.

First, we manually segmented the video images into the two regions: one region R_t (ground truth) that contains the sound source and the remaining region R_u that is uncorrelated to the audio signal. Then the precision and recall metrics are computed as follows [7]:

$$\text{Precision} = \frac{(R_t \cap R_d)}{R_d}; \quad \text{Recall} = \frac{(R_t \cap R_d)}{R_t}$$

where intersection $\cap$ stands for overlapping, and R_d is the region detected by the proposed method. The size of the detected region R_d is controlled by a predefined threshold, since only the pixels with bigger values than the predefined threshold are considered as "detected." We set the threshold to be 0.1 and compute the precision and recall for each frame. The final criteria are averaged over all frames. The F_1-measure is reported to show the overall performance. That is,

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Moreover, we define an L_c term similar to [3] to evaluate the localization results from an energy perspective. The resulted image W_v can have a number of positive and negative values. The energy of the pixels is defined as

$$e(\vec{x}) = |W_v(\vec{x})|^2$$

where $\vec{x}$ is the pixel coordinate. A satisfactory localization is obtained if most of the energy $e(\vec{x})$ is concentrated in the region of the ground truth. The localization criterion is defined as [3]

$$L_c = \frac{\sum_{\vec{x}\in\mathcal{D}_c} e(\vec{x})}{\sum_{\vec{x}} e(\vec{x})} \times \frac{R_t + R_u}{R_t \cap R_d}$$

where R_t is the ground truth and R_u is the remaining uncorrelated region, so $R_t + R_u$ represents the whole frame. R_d stands for the region detected by the proposed method. $\mathcal{D}_c$ represents the set of correctly detected pixels: $\mathcal{D}_c \doteq \{\vec{x} : \vec{x} \in R_t \cap R_d\}$. So $\sum_{\vec{x}\in\mathcal{D}_c} e(\vec{x})$ represents the sum of energy in the correctly detected region $R_t \cap R_d$, and $\sum_{\vec{x}} e(\vec{x})$ is the sum of energy in the whole frame $R_t + R_u$.

Because of the accumulation of energy, the detected pixels with negligible values make no difference to L_c. So the L_c criterion mainly focus on the regions with the major energy distribution. This property makes L_c well suited to pixel-level approaches and more robust to outliers in results.

2) Audio Separation Evaluation: Following the evaluation framework in [10] and [18], we examine the audio separation results by BSS-EVAL metrics [19]. Specifically, the source to distortion ratio (SDR) is often used to represent the overall performance of audio separation. We define the normalized SDR (NSDR), which only measures the improvement of the SDR from the mixed noisy audio $\hat{s}$ to the reconstructed separated sound $\hat{v}$ (namely, $\mathbf{P}_a$). That is [18]

$$\text{NSDR}(\hat{v}, v, \hat{s}) = \text{SDR}(\hat{v}, v) - \text{SDR}(\hat{s}, v)$$

where $\hat{v}$ is the reconstructed audio signal, v is the original sound, and $\hat{s}$ is the mixture of the original sound and artificial noises. In the second set of experiments (visually assisted audio separation), the artificial noises could be either white noise or irrelevant background music.

[3] https://db.sewaproject.eu/

3) Speaker Detection Evaluation: The diarization error rate (DER) is the metric used to quantitatively measure the performance: the smaller the DER value, the better the performance. DER was introduced by the NIST-RT[4] as the fraction of speaking time which is not attributed to the correct speaker, or to none of them in the case of a silent frame. To compute the DER on speech segments, three error types have to be defined.

1) *Confusion error*, when the output of the speaker label does not match the ground truth.
2) *Miss error*, when speech is present, but the method fails to detect the speech activity.
3) *False alarm error*, when speech is incorrectly detected in the case of silence.

The DER contains the composition of the three error measurements

$$\text{DER} = \frac{\text{confusion} + \text{miss} + \text{false alarm}}{\text{total speech time}}.$$

C. Experimental Setup

Parameters fall into two categories: 1) $\{\mu, \rho, \delta\}$ to control the iteration process of inexact ALM and 2) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, K\}$ to control the rank and sparsity of the output matrices. It is worth mentioning that only the $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, K\}$ are model parameters of the proposed method, while $\{\mu, \rho, \delta\}$ exist for any ALM process. Details are summarized as follows.

1) *Penalty Parameter μ:* The penalty parameter μ increases in each iteration via the multiplicative update factor $\rho > 1$. It is common to initialize μ with a small value for ALM, and we pick $\mu_0 = (1.25/\|\mathbf{V}\|_2)$ consistent with [16]. To avoid ill-conditioning, the algorithm requires an upper bound for μ, in which we choose $\mu_{\max} = 10^{18}$.
2) *Multiplicative Update Factor ρ (Algorithm 1):* The factor $1 < \rho < 2$ can be viewed as a tradeoff between the precision and speed of the iteration process. In other words, for ρ close to 1, the algorithm converges slowly but precisely, while for ρ close to 2, it works the other way around. Due to the high dimensionality and volume of the datasets, we choose $\rho = 1.5$ to reduce the computational load, as well as guarantee decent results in terms of precision.
3) *Convergence Threshold δ:* In general, a smaller value of δ tends to produce more accurate results but takes more iterations to converge. In our experiments, we choose $\delta = 10^{-5}$.
4) *Regularization Parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$:* They are used to balance the significance of minimization in the objective function (6). Take λ_1 as an example: a larger value of λ_1 tends to obtain a sparser matrix $\mathbf{P}_v$, while a smaller λ_1 would obtain a denser one. In all experiments, we set $\lambda_1 = 1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.2$, $\lambda_4 = 0.4$, and $\lambda_5 = 0.4$. These values of λ are obtained by optimizing the performance on V7 and then applied to all other videos. To further investigate the robustness of each parameters, we have conducted a sensitivity analysis in Appendix B in the supplementary material, where the performances of different values of λs are plotted. The results of sensitivity analysis show the stability of the proposed method, and there is a large set of values where the proposed method achieves comparable results.
5) *Rank Parameter K:* The K denotes the number of rows in matrix $\mathbf{C}$, which is the upper bound of the rank of $\mathbf{B}_v, \mathbf{B}_a$. In our experiments, we use $K = 50$ for all videos.

[4]www.itl.nist.gov/iad/mig/tests/rt/

D. Visual Localization of Sound Source

1) Comparison With the Pixel-Level Approach: The state-of-the-art method in the pixel level was proposed in [21]. It uses SCCA to identify the dynamic pixels which are most correlated to the audio signal. To gain a fair comparison with this pixel-level approach, we report the L_c criterion which shows the energy distribution of localization results.

Both the proposed method and the method in [21] were tested and evaluated on all 20 videos in the SOP dataset. The sample frames of localization results are shown in Fig. 2, as well as the manually labeled ground truth for a visual comparison. As you can see, the proposed method has successfully localized the sound sources in test videos. Thanks to the sparsity of $\mathbf{P}_v$, the localization results remain in one or two concentrated areas, which automatically construct the physical boundary of the sounding object. On the other hand, although the method in [21] is able to find several pixels that are most correlated to the audio signal, its localization result is severely influenced by distracting moving objects (e.g., its detection on pedestrians).

The quantitative evaluation for each algorithm is shown in Table II. It is clear that the proposed method outperforms the method [21] in terms of L_c for all but one video, which means its energy distribution is more concentrated in the region of ground truth. Besides, one may find that the value of L_c in different videos varies a lot, and this is because the size of the sound-producing region in videos is quite different, which leads to a different concentration level of energy distribution.

To further demonstrate the contribution of this paper, the proposed method is compared with our preliminary work [36]. The model in [36] is actually a special case of the proposed method, where it decomposes each modality into only two components (either low-rank background or sparse correlated component) and does not consider any uncorrelated information (e.g., distracting movements in visual modality and noise in audio modality). By inspecting Table II, it is easy to see that the consideration of uncorrelated components in the proposed method helps a lot and results in better performances in almost all test videos.

2) Comparison With the Object-Level Approach: The object-level approach usually first decomposes the video signals into a number of video regions and then performs correlation analysis to identify one or two video regions which are most correlated to the audio signal. The state-of-the-art method in the object level was proposed by Li *et al.* [8]. Their method not only segments video frames into spatial regions but

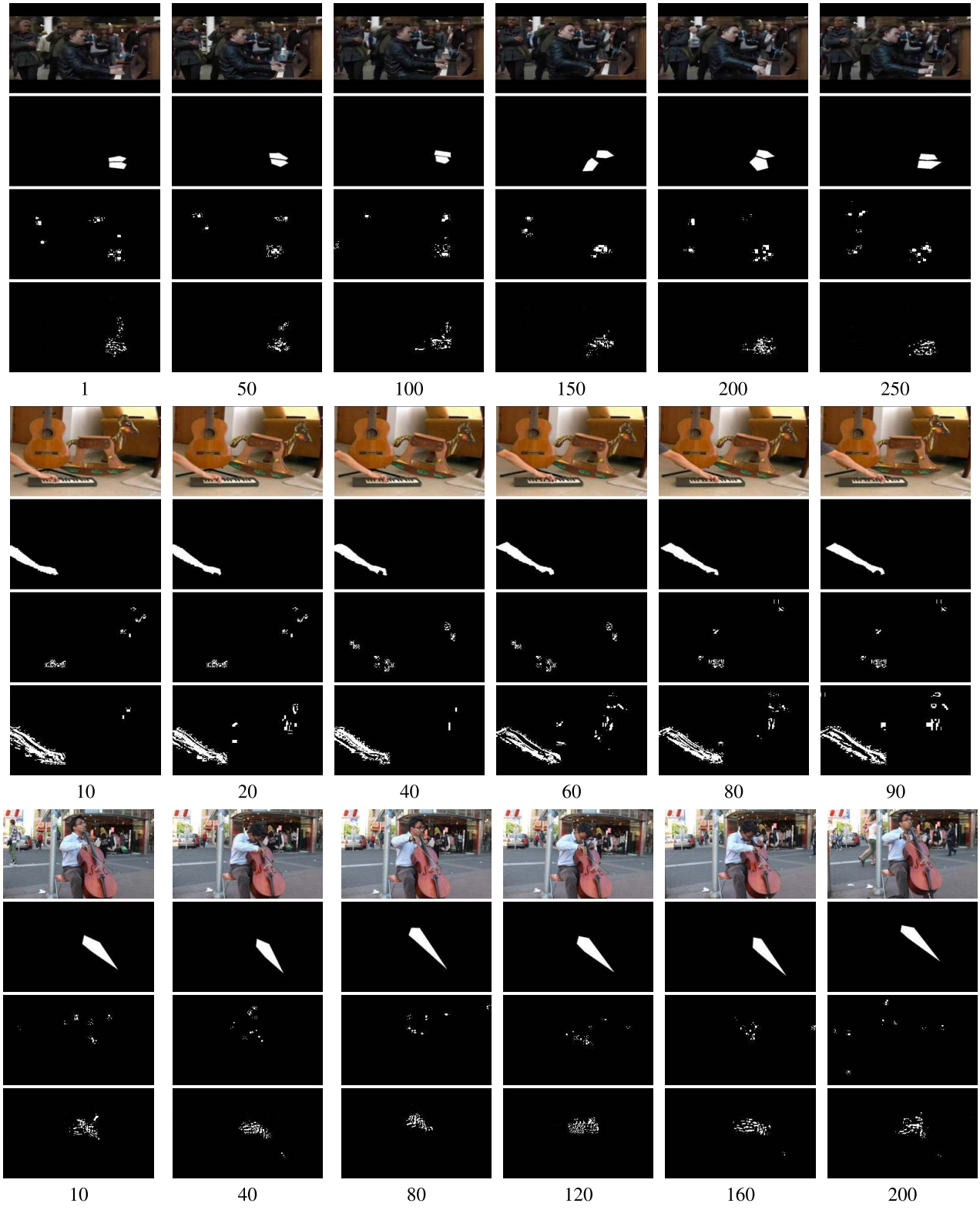


Fig. 2. Sample frames of the localization results. The frame number is marked at the bottom of the sample frames. These groups of figures are for video sequences V2, V8, and V19. Within each group, each row from top to bottom corresponds to the original video frames, the manually labeled ground truth, and results produced by the method [21] and by our algorithm (from the sparse component $\mathbf{P}_v$). To facilitate the visual comparison, 5% of pixels with most energy are plotted in white, while others remain in black.

also applies a region tracking algorithm to record the temporal evolution of one region.

As a result of Li's segmentation and region tracking algorithm, the video is decomposed into a number of region trackers and each tracker represents the temporal evolution of one region. The visual features (acceleration) are extracted from these region trackers and used in their correlation analysis. Although Li's method produced good localization results

TABLE II
QUANTITATIVE COMPARISON FOR VISUAL LOCALIZATION OF THE SOUND SOURCE, IN TERM OF THE L_c CRITERION. THE BEST PERFORMANCE IN EACH ROW IS IN BOLD

Index	Pixel-level [21]	Preliminary [36]	Proposed method
V1	17.7147	25.0812	**35.9697**
V2	38.4121	18.9275	**53.4280**
V3	27.8205	24.4916	**41.1083**
V4	16.6262	25.9166	**37.3420**
V5	37.5824	23.7763	**69.3898**
V6	32.8769	18.7008	**36.7466**
V7	10.1565	**21.5093**	11.8834
V8	20.1201	**24.2709**	18.7321
V9	15.6017	24.3466	**48.2306**
V10	36.4695	16.2149	**48.0794**
V11	125.2610	17.8773	**148.8214**
V12	49.9097	21.4786	**67.4180**
V13	18.8748	16.3823	**33.9825**
V14	17.4748	47.4792	**75.9254**
V15	7.7405	4.0114	**9.8328**
V16	10.5244	12.9377	**19.2947**
V17	27.5647	10.2106	**41.1681**
V18	8.3142	5.6821	**18.2406**
V19	6.2431	8.1276	**17.0999**
V20	4.3793	6.0100	**7.7506**

on their six test videos, it has two vital disadvantages which greatly influence its localization results on other unseen videos. The two disadvantages are also common in other object-level approaches, so we conduct a detailed investigation here.

1) *"Binary" Result:* Since Li's method keeps 15 region trackers as candidates and their correlation measure only picks one tracker out of the 15, the localization task becomes a binary problem. In other words, the algorithm can either pick the "right" region or a completely uncorrelated one.
2) *Sensitivity to Parameters:* The 15 region trackers are produced by two segmentation processes and two clustering processes, in which each process is laborious and determined by its own parameters (e.g., shape of segmented regions and the number of clusters). These parameters play a crucial role in feature extraction and consequently influence the localization results. A detailed investigation about how these parameters influence the localization results is given in Table III.

As shown in Table III, the optimal parameters for V7, V8, V16 are [15, 0.2], [25, 0.05], and [20, 0.15], respectively. Although their results under optimal parameters are reasonably good, the performances under suboptimal parameters are unsatisfactory. Hence, their localization results are very sensitive to parameters. The reason is that the construction of the 15 region trackers is sensitive to these parameters and any suboptimal parameters will produce undesirable visual features, which eventually undermines the localization results. Moreover, these optimal parameters can be different for different videos and there is no principled way to find them.

To gain a fair comparison, we consider the parameter sensitivity of different methods. For Li's method, we take the three sets of optimal parameters for V7, V8, V16 in Table III and apply them on other videos of our SOP dataset. Each video was analyzed three times under the three sets of parameters. For the proposed method, its parameters are found by optimizing performance on V7, which is described in detail in Section IV-C. Then, the found parameters are used for all other videos.

TABLE III
PARAMETER INVESTIGATION OF LI'S METHOD [8]. THE THREE VIDEOS HERE ARE THE ORIGINAL VIDEOS USED IN THEIR PAPER, WHICH ARE ALSO IN OUR SOP DATASET. "PARAMETER1" IS THE NUMBER OF REMAINING REGIONS AFTER ITS INTRAFRAME CLUSTERING PROCESS. "PARAMETER2" IS A SIMILARITY THRESHOLD TO DECIDE WHETHER TO CREATE A NEW REGION TRACKER DURING ITS INTERFRAME CLUSTERING PROCESS

Video name	Parameter1	Parameter2	F1
Violin Yanni (V7)	15	0.3	0.0287
	15	0.25	0.0243
	15	**0.2**	**0.6074**
	15	0.15	0.0601
	15	0.1	0.0601
Wooden Horse (V8)	25	0.25	0.0103
	25	0.2	0.0103
	25	0.15	0.0103
	25	0.1	0.0103
	25	**0.05**	**0.4357**
Guitar Lessons (V16)	20	0.3	0.1718
	20	0.25	0.0648
	20	0.2	0.0648
	20	**0.15**	**0.4541**
	20	0.1	0.0648

TABLE IV
QUANTITATIVE COMPARISON OF THE PROPOSED METHOD WITH LI'S METHOD [8], IN TERMS OF THE F1-CRITERION. FOR LI'S METHOD, THE PERFORMANCE IS REPORTED IN THREE COLUMNS AND EACH COLUMN HAS DIFFERENT PARAMETER CONFIGURATIONS. IN THE FIRST COLUMN, PARAMETERS ARE FOUND BY OPTIMIZING PERFORMANCE ON THE VIDEO V7, I.E., [15, 0.2]. THE SECOND COLUMN IS FOR THE OPTIMAL PARAMETERS OF V8 AND THE THIRD COLUMN IS FOR V16. FOR THE PROPOSED METHOD, PARAMETERS ARE FOUND BY OPTIMIZING PERFORMANCE ONLY ON THE VIDEO V7

Index	Compared method [8]			Proposed method
	Parameters optimized on			Parameters optimized on
	V7	V8	V16	V7
V1	0.1129	0.1224	0.0748	**0.4457**
V2	0.0888	0.1550	0.0688	**0.4598**
V3	0.0367	0.0772	0.0891	**0.4470**
V4	0.0725	0.1575	0.1673	**0.4857**
V5	0.1699	0.0313	0.0207	**0.4812**
V6	**0.3775**	0.0	0.0	0.3158
V7	**0.6074**	0.0601	0.0601	0.3393
V8	0.0103	0.4357	0.0103	**0.5219**
V9	0.0549	0.0768	0.1645	**0.3477**
V10	0.0148	0.0225	0.0149	**0.0758**
V11	0.0013	0.0084	0.0182	**0.1889**
V12	0.0	0.0	0.0	**0.1545**
V13	0.0730	0.0872	0.0821	**0.1172**
V14	0.0246	0.0098	0.0017	**0.2251**
V15	0.0344	0.3783	**0.4721**	0.1837
V16	0.0648	0.0648	**0.4541**	0.3385
V17	0.0025	0.0	0.2021	**0.2093**
V18	**0.3295**	0.2237	0.2837	0.1391
V19	0.0793	0.1020	0.1699	**0.2147**
V20	0.1913	0.0970	0.1923	**0.4311**

The quantitative comparison for the proposed method and Li's method is shown in Table IV. As you can see, Li's method is quite sensitive to the parameters from the comparison of its own three columns. It can achieve better results on five videos (V6, V7, V15, V16, and V18) when the tested parameters are desirable for its feature extraction. By contrast, our method performs better on the remaining 15 videos and it produces

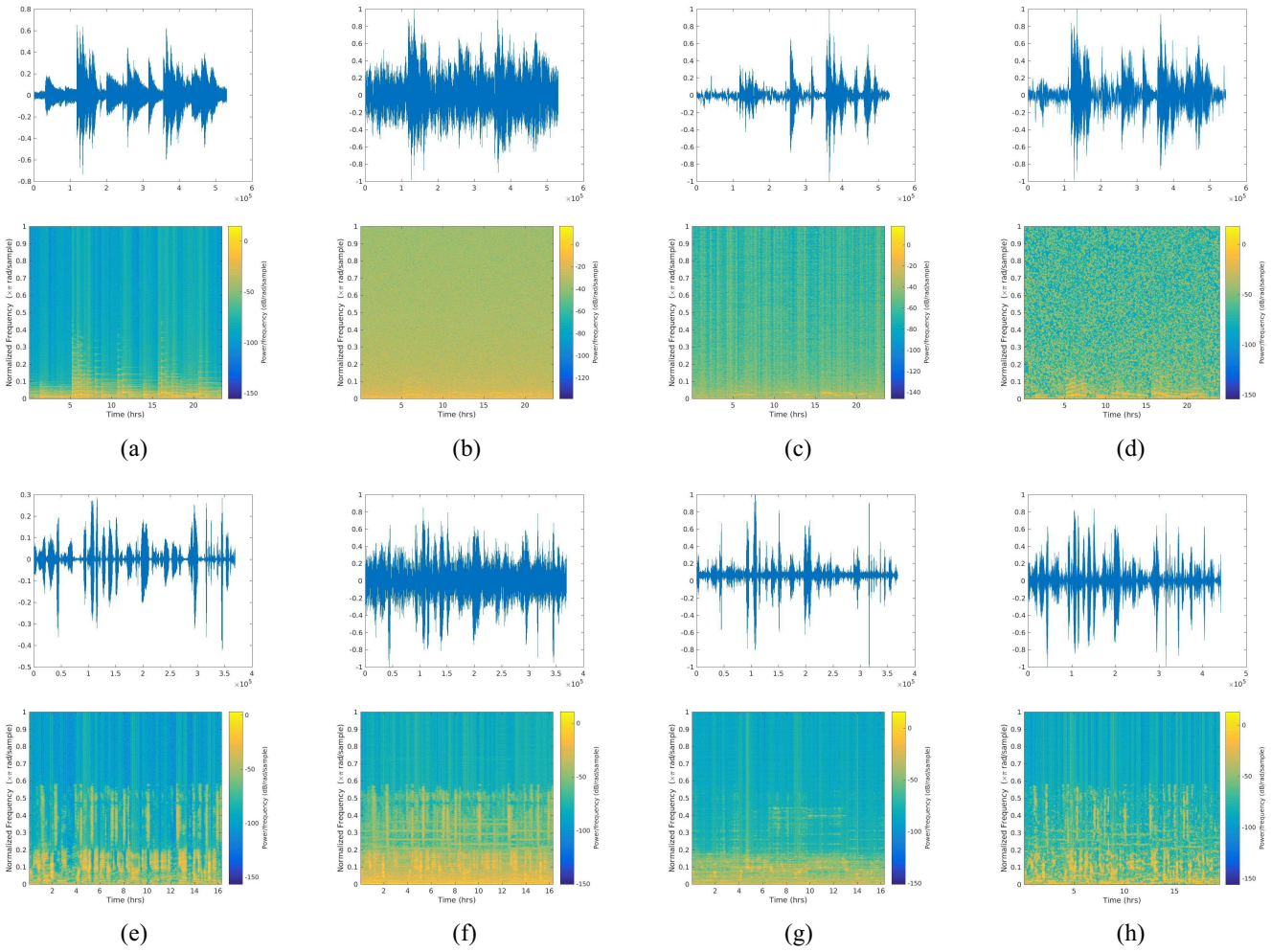


Fig. 3. Audio separation results. The group of figures (a)–(d) is for video sequence V8 and (e)–(h) for V9. Within each group, the top row is the plain plot of audio signals, while the second row is the spectrogram of audio. (a), (e) Original audio. (b) Audio + white noise. (f) Audio + background music. (c), (g) Separated audio of the proposed method. (d), (h) Separated audio of [10].

more stable localization results without relying on parameter tuning.

It is worth mentioning that the localization results of the proposed method in Tables II and IV are all obtained under the same parameter configuration. Besides, one may find for different videos that the value of F1 in Table IV varies a lot, and it is because some videos are more challenging than others.

E. Visually Assisted Audio Separation

In this section, we investigate the capability of the proposed method in audio separation with the assistance of visual information. The experiment is conducted on 12 videos (V1, V4–V10, V15, V16, V18, V20), where their original audio signals are relatively clean and can be kept as the ground truth of audio separation. To synthesize noisy audio inputs, these audio signals are corrupted by either white noise or irrelevant background music, which are from the Audio Degradation Toolbox [37]. The first six videos (V1 and V4–V8) are corrupted by white noise, while the other 6 (V9, V10, V15, V16, V18, and V20) are corrupted by background music.

The corrupting noises and the original audio signal were mixed at 0 dB signal to noise ratio, that is, they are at the same energy level. In this scenario, the proposed method will output an audio sparse component $\mathbf{P}_a$ that corresponds to the separated audio signal. The samples of the audio separation results are plotted in Fig. 3. As you can see, in the time-frequency domain (spectrogram), the separated audio obtained by the proposed method successfully captures the most prominent part of the original audio signal. The comparison is made against the state-of-the-art unsupervised monaural audio separation method [10], which also performs reasonably well.

To make the comparison more clear, the quantitative result of the audio separation is shown in Table V, along with the comparison to our preliminary work [36]. It is easy to see that the proposed method is able to obtain decent separation results and achieve competitive performance, outperforming [10] in 8 out of 12 videos (V1, V4, V5, V6, V7, V8, V16, and V20).

F. Active Speaker Detection

The capability of the proposed method in speaker detection is assessed on ten videos (V21–V30). These videos are

TABLE V
QUANTITATIVE COMPARISON FOR AUDIO SEPARATION IN TERMS OF THE NSDR CRITERION. THE BEST PERFORMANCE IN EACH ROW IS IN BOLD

Index	State-of-the-art [10]	Preliminary [36]	Proposed method
V1	-1.0809	2.5542	**4.3106**
V4	-1.8427	0.4389	**2.1210**
V5	4.3240	4.1101	**4.7042**
V6	-0.6818	2.3165	**3.1029**
V7	-3.0257	5.8963	**7.5347**
V8	3.4864	8.8156	**8.8810**
V9	**3.8541**	0.7112	1.6892
V10	**7.1564**	0.4902	1.0707
V15	**1.2723**	0.8769	1.2661
V16	-2.8458	-1.2459	**-0.5897**
V18	**7.0333**	3.9239	4.0539
V20	-3.4429	-0.5301	**0.2835**

TABLE VI
QUANTITATIVE COMPARISON FOR ACTIVE SPEAKER DETECTION, IN TERMS OF THE DER SCORES(%). THE BEST PERFORMANCE IN EACH ROW IS IN BOLD

Index	LIUM [22]	Preliminary [36]	Proposed method
V21	33.89	41.87	**32.81**
V22	**41.94**	45.63	45.94
V23	54.72	56.25	**47.19**
V24	49.72	42.19	**39.69**
V25	51.11	40.63	**25.31**
V26	50.31	47.81	**46.56**
V27	59.37	52.19	**49.06**
V28	49.69	47.50	**44.06**
V29	35.94	25.94	**23.75**
V30	51.25	49.38	**35.62**
Average	47.79	44.94	**38.99**

created from the SEWA database and described in detail in Section IV-A. Since the dialogues in videos contain only two persons (sitting on the left and right side of images, respectively), we exploit the results of visual localization and compare the magnitude of energy between the left and right side of generated images. The active speaker is identified as the one with the highest energy magnitude.

The experimental comparison is made against the widely used toolbox LIUM [22]. LIUM uses acoustic features that are composed of 13 mel-frequency cepstral coefficients. The quantitative result of each method (LIUM [22], the proposed method, and our preliminary work [36]) is shown in Table VI. It is easy to see that the proposed method outperforms the LIUM toolbox, with a smaller average score of DER, for all but one video (V22). However, there is still space to improve the performance since both the LIUM toolbox and the proposed method find difficult to detect silent frames and simultaneous speech.

The purpose of the experiment is to demonstrate the decent localization result of the proposed method, which can be further used for speaker detection. Besides, the idea of exploiting the visual localization to help the speaker detection task could also be promising to improve the multimodal speaker detection in future.

V. CONCLUSION

In this paper, we proposed a low-rank and sparse model to handle the audio–visual localization and separation problem, with one additional application on active speaker detection. The proposed CLS method is formalized as a constrained optimization problem and solved by applying inexact ALM algorithm. Experiments are conducted on 30 videos to evaluate the capability of the proposed method, with comparison against other state-of-the-art methods in the field. Specifically, we conducted three sets of experiments: 1) visual localization of sound source; 2) visually assisted audio separation; and 3) active speaker detection. In these experiments, the proposed method is able to correctly localize the sound source and separate the associated audio with competitive performance and also successfully performs speaker detection.

Currently, the proposed method cannot be applied to handle discrete audio signals due to the lack of audio information during the silent intervals. In addition, it cannot detect the nonlinear correlation between visual modality and audio modality due to the linear shared matrix $\mathbf{C}$ between the sparse components $\mathbf{P}_v$ and $\mathbf{P}_a$. So a kernel version of the proposed method and its extension will be investigated in future.

REFERENCES

[1] 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.

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

[3] E. Kidron, Y. Y. Schechner, and M. Elad, "Pixels that sound," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2005, pp. 88–95.

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

[5] A. L. Casanovas and P. Vandergheynst, "Unsupervised extraction of audio-visual objects," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2011, pp. 2284–2287.

[6] A. L. Casanovas and P. Vandergheynst, "Audio-based nonlinear video diffusion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2010, pp. 2486–2489.

[7] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 378–390, Feb. 2013.

[8] K. Li, J. Ye, and K. A. Hua, "What's making that sound?" in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 147–156.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, 2011.

[10] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 57–60.

[11] G. Monaci *et al.*, "Learning multimodal dictionaries," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2272–2283, Sep. 2007.

[12] G. Liu and S. Yan, "Active subspace: Toward scalable low-rank learning," *Neural Comput.*, vol. 24, no. 12, pp. 3371–3394, Dec. 2012.

[13] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, 1996.

[14] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

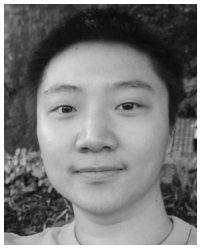
[15] S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4453–4461.

[16] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1665–1678, Aug. 2015.

[17] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[18] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2011, pp. 221–224.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[20] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2402–2415, Dec. 2016.
[21] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, Apr. 2007.
[22] M. Rouvier *et al.*, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Interspeech*, 2013.
[23] X. Anguera *et al.*, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.
[24] A. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal speaker diarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 79–93, Jan. 2012.
[25] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1218–1225.
[26] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.
[27] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 895–910, Oct. 2010.
[28] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.
[29] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2011, pp. 209–212.
[30] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Belmont, MA, USA: Academic, 2014.
[31] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.
[32] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 241–245.
[33] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016, pp. 31–35.
[34] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2013, pp. 141–145.
[35] D. E. Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 256–260.
[36] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 2901–2905.
[37] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. 14th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Curitiba, Brazil, 2013, pp. 83–88.
[38] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.
[39] F. Sedighin, M. Babaie-Zadeh, B. Rivet, and C. Jutten, "Multimodal soft nonnegative matrix co-factorization for convolutive source separation," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3179–3190, Jun. 2017.
[40] Q. Liu *et al.*, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5520–5535, Nov. 2013.
[41] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[42] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Comput. Sci. Rev.*, vol. 23, pp. 1–71, Feb. 2017.
[43] O. Oreifej, X. Li, and M. Shah, "Simultaneous video stabilization and moving object detection in turbulence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 450–462, Feb. 2013.
[44] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 609-Ű617.
[45] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.
[46] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 892–900.
[47] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, 2002.
[48] M. Chen, Z. Lin, Y. Ma, and L. Wu, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Coordinated Sci. Lab., Urbana, IL, USA, Rep. UILU-ENG-09-2215, 2009.
[49] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.



Jie Pu (S'18) received the M.Sc. degree in computing from Imperial College London, London, U.K., in 2016, where he is currently pursuing the Ph.D. degree in computer vision and machine learning.

He is a member of the iBUG Group with the Department of Computing, Imperial College London. His current research interests include machine learning, mathematical optimization and its applications in computer vision, audio-visual analysis, and human behavior analysis.

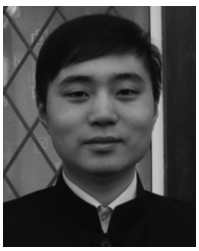
Yannis Panagakis received the B.Sc. degree in informatics and telecommunication from the University of Athens, Athens, Greece, and the M.Sc. and Ph.D. degrees in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece.

He is a Senior Lecturer (Associate Professor equivalent) of computer science with Middlesex University London, London, U.K., and a Research Faculty with the Department of Computing, Imperial College London, London. His research has been published in leading journals and conferences proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, CVPR, and ICCV. His current research interests include machine learning and its interface with signal processing, high-dimensional statistics, computational optimization, as well as statistical models and algorithms for robust and efficient learning from high-dimensional data and signals conveying audio, visual, behavioral, medical, and social information.

Dr. Panagakis was a recipient of the Prestigious Marie-Curie Fellowship, among various scholarships and awards for his studies and research. He currently serves as the Managing Editor-in-Chief of *Image and Vision Computing* Journal. He co-organized the British Machine Vision Conference in 2017 and numerous workshops in conjunction with international conferences.

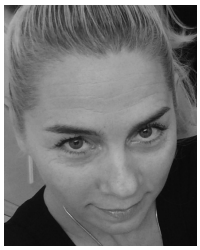
Stavros Petridis (S'07–M'10) received the B.Sc. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004 and the M.Sc. degree in advanced computing and the Ph.D. degree in computing from Imperial College London, London, U.K., in 2005 and 2012, respectively.

He is a Research Fellow with the Department of Computing, Imperial College London. He has been a Research Intern with the Image Processing Group, University College London, London, and the Field Robotics Centre, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, and a Visiting Researcher with the Affect Analysis Group, University of Pittsburgh, Pittsburgh. His current research interests include deep neural networks and machine learning and their application to multimodal recognition of human nonverbal behavior and audio–visual speech recognition.



Jie Shen received the B.Eng. degree in information engineering from Zhejiang University, Hangzhou, China, in 2005 and the M.Sc. degree in advanced computing and the Ph.D. degree in computing from Imperial College London, London, U.K., in 2008 and 2014, respectively.

He was a Research Assistant with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, until 2007. He is a Research Fellow with Imperial College London. His current research interests include affect-sensitive human–computer interaction, social robotics, system engineering for HCI/HRI systems, and face tracking and segmentation.



Maja Pantic (M'98–SM'06–F'12) received the M.Sc. and Ph.D. degrees in computer science from the Delft University of Technology, Delft, The Netherlands, in 1997 and 2001, respectively.

She is a Professor of affective and behavioral computing with the Department of Computing, Imperial College London, London, U.K., and the Department of Computer Science, University of Twente, Enschede, The Netherlands.

Prof. Pantic was a recipient of the European Research Council Starting Grant Fellowship in 2008, the Roger Needham Award in 2011, and various awards for her research on automatic analysis of human behavior. She currently serves as the Editor-in-Chief of *Image and Vision Computing* Journal and as an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.