SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
Anna Llagostera Casanovas

ELE 227 (Bâtiment ELE)
Station 11
CH-1015 LAUSANNE

Tel: +41 21 693 68 74
Fax: +41 21 693 76 00
e-mail: anna.llagostera@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

BLIND AUDIOVISUAL SOURCE SEPARATION USING SPARSE REDUNDANT REPRESENTATIONS

Anna Llagostera Casanovas, Gianluca Monaci, Pierre Vandergheynst

École Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2007.03

January 25, 2007

Blind Audio-Visual Source Separation Using Sparse Redundant Representations

Anna Llagostera, Gianluca Monaci, Pierre Vandergheynst

Signal Processing Institute
École Polytech. Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

Abstract

In this work, we present a method that jointly separates active audio and visual structures on a given mixture. This new concept, the Blind Audiovisual Source Separation (BAVSS), is achieved by exploiting the coherence existing between the recorded signal of a video camera and only one microphone. An efficient representation of audio and video sequences allows to build robust audiovisual relationships between temporally correlated structures of both modalities or, what turns to be the same, two parts of the same audiovisual event. First, video sources are localized and separated on the image sequence exploiting the temporal occurrence of audiovisual events and using a spatial clustering algorithm, without necessity of any previous assumption about the number of sources in the mixture. Second, the same audiovisual relationships together with a time-frequency probabilistic analysis allow the separation of the audio sources in the soundtrack, and, consequently, the complete Audiovisual Separation.

Index Terms

Audiovisual processing, blind source separation, sparse signal representation.

I. Introduction

It is well known from every-day experience that visual information strongly contributes to the interpretation of acoustic stimuli. This is particularly evident if we think to speech signals : speaker's lips movements are correlated with the produced sound and the listener can exploit this correspondence to better understand speech, especially in adverse environments [1, 2]. The multi-modal nature of speech is exploited since at least two decades to design speech enhancement [3–5] and speech recognition algorithms [6, 7] in noisy environments. Lately, this paradigm has been adopted also in the speech separation field to increase the performances of audio-only methods.

Few methods exist that exploit audiovisual coherence to separate *stereo* audio mixtures [8–12]. All the existing algorithms consider the problem from an *audio source separation point of view*, i.e. they use the audio-video synchrony as side information to improve and overcome limitations of classical Blind Audio Source Separation (BASS) techniques. For a comprehensive survey of BASS terminology, methods and algorithms the reader is referred to an exhaustive report by Vincent and co-workers [13].

In [8] the authors propose to estimate the de-mixing process using a criterion based on audiovisual coherence: one speech source of interest is extracted using the visual information simultaneously recorded from the speakers face by video processing. The coherence between audio and video data is modeled by a joint audiovisual probability estimated as a mixture of Gaussian kernels whose parameters are learned from a large training set. Video information consists of geometric parameters describing the speaker's lips height and width that are extracted using a chroma-key process on lips under controlled head position and light conditions [14]. The system was shown to be able to estimate the un-mixing matrix in the case of instantaneous additive mixtures. A very similar approach, but for stationary convolutive mixtures, has been developed in [11]. Another method inspired by [8, 11] is presented in [12]. In this case video features are deduced using active appearance model [15] and the algorithm is tested on a limited set of 2×2 linear instantaneous mixtures.

The authors acknowledge the support of the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.

Dansereau [9] also propose an audiovisual speech source separation system plugging the visual information, representing again the speaker's lip height and width, in a decorrelation system with first-order filters. Visual cues are mapped to word structures with a continuous HMM that is trained on a corpus of visual speech. The method was tested simulating a 2×2 speech separation problem by mixing one audio source recorded with one microphone and one speaker captured with one camera and one microphone. Rajaram and colleagues [10] suggest instead a Bayesian framework for 2×2 linear mixtures of audio-video sources. In this case the video information is quite simple and it basically provides a binary weight that indicates the activation of a source, while the mixing model parameters are estimated on-line.

The approach we consider in this report is very different from existing ones. First, we localize and separate visual sources using audiovisual synchrony. Once located the video sources on the image sequence, we can reconstruct them by assuming that the structures close to a source belong to it. We obtain thus several groups of video structures, each group corresponding to a detected source. It is important to underline that sources in the video domain, e.g. people speaking in front of a camera, are typically well separated in space. This information will help us in separating the audio mixture as well, exploiting the correlations established between audio and video entities. Since only a one-microphone signal is considered, the separation of an unknown number of unknown sources is in fact extremely challenging.

We want to stress three important differences between our proposed approach and state-of-the-art audiovisual separation methods:

1) The BASS problem is solved for stereo audio signals, using separation techniques helped by visual information. In contrast the audio signal we consider here comes from only *one microphone*, which makes the source separation task considerably more challenging;
2) Existing methods simplify the task of associating audio and video information. Either the audio-video association is given *a priori*, i.e. it is known which audio signal corresponds to which video signal [10, 12], either it is considered the case where one single audiovisual source is mixed with an *audio-only* source [8, 9, 11]. In the latter case the separation problem basically turns into separate two mixed speech signals, one of which has a corresponding video counterpart. Here, in contrast, we simultaneously separate audio-video sources, automatically building correlations between acoustic and visual entities. The only hypothesis that we make is that each video source present in the scene has one and only one corresponding audio source in the audio mixture;
3) Existing audiovisual separation methods, with the only exception of [10], require an off-line training step to build the audiovisual source model. This is mainly due to the fact that the algorithms proposed in [8, 9, 11, 12] try to map video information into the audio feature space using techniques similar to lip-reading (requiring moreover accurate mouth parameters that are difficult to acquire). In contrast, in the proposed method no training will be required.

To summarize we essentially want to solve a blind Single-Channel BASS problem, but aided by the video. Since no hypothesis is made on the relationships between audio and video structures, video sources have to be localized and separated at the same time, exploiting the information contained in the audio channel. The approach we use is inspired by the previous work performed by by Monaci, Divorra and Vandergheynst [16], which already succesfully localized in the image the video sources of an audiovisual sequence. This method is based on sparse geometric representation of video sequences. They searched for the video structure more temporally correlated with a given audio feature, the average acoustic energy. Then, this structure was assumed to be the speaker mouth (body part whose movement is highly coherent with the speech energy [17, 18]) indicating, thus, the situation of the video source in the image.

The steps of our *Blind Audiovisual Source Separation* (BAVSS) algorithm will be detailed in the following of this report, while in the next section we describe the audio and video features that we use to represent both modalities.

II. Audio and Video Representations

The efficiency of the proposed algorithm is basically due to the representations used for describing the audio and video signals. These representations decompose the signals according to their reliant structures, whose variations in characteristics such as dimensions or position represent, at the same time, a relevant change in the whole signal. For example, a variation in one pixel value may mean movement or not, but a position change of one full structure will probably have this meaning. Next sections describe representation techniques used for both modalities.

A. Audio Representation

The previous work [16] used the average acoustic energy for the audio representation. With only this basic feature the Video Source Localization goal was achieved. However, more information is required in order to perform the Audio Separation task. In this research work, not only the distribution of the energy through time is considered, but also the information concerning the frequency components of the signal is included.

The audio signal in the time-frequency plane is decomposed using MP over a dictionary of Gabor atoms $\mathcal{D}^{(a)}$, where a single window function, $g^{(a)}$, generates all the atoms that compose the dictionary. Each atom $\phi_k^{(a)} = U_k g^{(a)}$, is built by applying a transformation U_k to the mother function $g^{(a)}$. The possible transformations are scaling by $s > 0$, translation in time by u and modulation in frequency by ξ. Then, indicating with an index k the set of transformations (s, u, ξ), an atom can be represented as

$$\phi_k^{(a)}(t) = \frac{1}{\sqrt{s}}\, g^{(a)}\left(\frac{t-u}{s}\right) e^{i\xi t}, \tag{1}$$

where the value $1/\sqrt{s}$ makes $\phi^{(a)}(t)$ unitary.

Thus, an audio signal $a(t)$ can be approximated using K atoms as

$$a(t) \approx \sum_{k=0}^{K-1} c_k \phi_k^{(a)}(t)\,,$$

where k is the summation index, c_k corresponds to the coefficient for every atom $\phi_k^{(a)}(t)$ from dictionary $\mathcal{D}^{(a)}$. In all the performed experiments the audio signals are approximated using $K = 2000$ Gabor atoms.

The main motivation behind the use of MP decomposition is that it provides a sparse representation of the audio energy distribution in the time-frequency plane, showing the frequency components evolution. Moreover, MP algorithm performs a denoising of the input signal, pointing out the most relevant structures [19].

B. Video Representation

The video signal is represented using the 3D-MP algorithm proposed by Divorra and Vandergheynst [20]. The image is decomposed into a set of video atoms representing salient video components and their temporal transformation is posteriorly tracked through time. A modified MP approach based on Bayesian decision criteria is used for the tracking.

The first frame of the video signal, $I_1(x_1, x_2)$, is approximated with a linear combination of atoms retrieved from a redundant dictionary $\mathcal{D}^{(v)}$ of 2-D atoms as

$$I_1(x_1, x_2) \approx \sum_{\gamma_i \in\ \Omega} c_{\gamma_i} g_{\gamma_i}^{(v)}(x_1, x_2), \tag{2}$$

where *n* is the summation index, c_{γ_i} corresponds to the coefficient for every 2-D video atom $g_{\gamma_i}^{(v)}(x_1, x_2)$ and Ω is the subset of selected atom indexes from dictionary $\mathcal{D}^{(v)}$. As in the audio case, the dictionary is built by varying the parameters of a mother function, an edge-detector atom with odd symmetry.

Then, this 2-D atoms are tracked from frame to frame. The possible transformations experienced by the atoms are: translations over the image plane, rotations to locally orient the function along the edge and scaling to adapt the atom to the considered image structure. Fig. 1 shows an schematic example of this procedure in a sequence of frames.

Thus, the video signal can be approximated using N 3-D video atoms $\phi_n^{(v)}$ as

$$\mathbf{V}(x_1, x_2, t) \approx \sum_{n=0}^{N-1} c_{n(t)} \phi_n^{(v)}(x_1, x_2, t)\,,$$

where n is the summation index and $c_{n(t)}$ are the coefficients corresponding to each video atom. In all experiments, sequences are represented using $N = 100$ video atoms, and each atom has an associated feature describing its displacement (considering spatial translations from frame to frame).

The interest in using this video decomposition is that, unlike the case of simple pixel-based representations, when considering image structures that evolve in time we deal with dynamic features that have a true geometrical

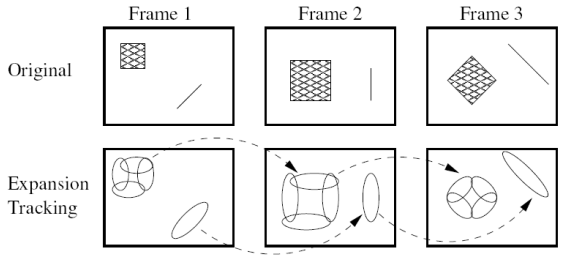


Fig. 1. Successive schematic updates of basis functions in a sequence of frames. In the second row, ellipses represent schematically the possible positioning of some 2D atoms.

meaning. Thus, the considered video features reflect the movement, from frame to frame, of the image relevant structures. Furthermore, geometric sparse video decompositions provide compact representations of information, allowing a considerable dimensionality reduction of the input signals.

III. Blind Audiovisual Source Separation (BAVSS)

Fig. 2 illustrates schematically the whole BAVSS process. First, the video sources are localized using a clustering algorithm that spatially groups the video structures in the image temporally correlated with the audio atoms of the soundtrack. Second, a purely spatial criterion is used to separate the sources. Then, the correlations between audio and video events are employed to identify temporal periods with only one source active (audio source localization). Finally, the sources frequency behavior is estimated in time periods during which only they are active alone in order to separate the sources in the mixed periods.

There are two main assumptions that we make on the type of sequences that we can analyze using the proposed algorithm. First, we assume that for each detected video source there is one and only one associated source in the audio mixture. This means that if there is an audio "distractor" in the sequence (e.g. a person speaking out of the camera's field of view), it is considered as noise and its contribution to the mixture is associated to the sources found in the video. This assumption clearly simplifies the analysis, since we know in advance that a one-to-one relationship between audio and video entities exists. Moreover, we consider the video sources approximately static, i.e. their positions over the image plane do not change too much. This assumption is less stringent in our opinion and it is formulated only not to have to worry about dynamic aspects of the scene. However it can be removed for example by analyzing the sequences using shifting time windows. One typical sequence that we consider in this work, taken from the *groups* section of the CUAVE database [21], is shown in Fig. 3. It involves two speakers arranged as in Fig. 3 [Left] that utter digits in English. As highlighted in Fig. 3 [Right], in the first part of the clip the girl on the left speaks alone, then the boy on the right starts to speak as well, and finally the girl stops speaking and the boy speaks alone.

A. Video Source Localization

This first phase of the Audiovisual Separation process consists in spatially locate the active video sources in the image. It is divided into two main parts: the temporal association between audio and video features with the correspondent measure of synchrony, and the spatial location of the video sources in the image.

1) Audio and Video Atoms Association: Correlation scores $\chi_{k,n}$ between each audio atom $\phi_k^{(a)}$ and each video atom $\phi_n^{(v)}$ are computed. These scores measure the degree of synchrony between *relevant events* in both modalities: more synchrony indicates higher possibility of belonging to the same audiovisual event. For the audio, a relevant event is the presence, at one particular moment, of an audio atom (audio energy concentration in the time-frequency plane), and, for the video, a peak in the video atom displacement, i.e., the uttering of a sound is caused by the movement of the lips, and both are relevant events in their modalities.

- **Audio feature** The feature $\mathsf{f}_k(t)$ that we consider is the projection over the temporal axis of the Wigner-Ville distribution of each audio atom [19], $\mathsf{f}_k(t) = W\phi_k^{(a)}(t,\ \omega = 0)$, which describes evolution of the atom energy through time. In the case of Gabor atoms is a 2D Gaussian function whose position and variance depend on the atoms parameters. An scheme of this feature is shown in Fig. 4. Thus, instead of considering only one

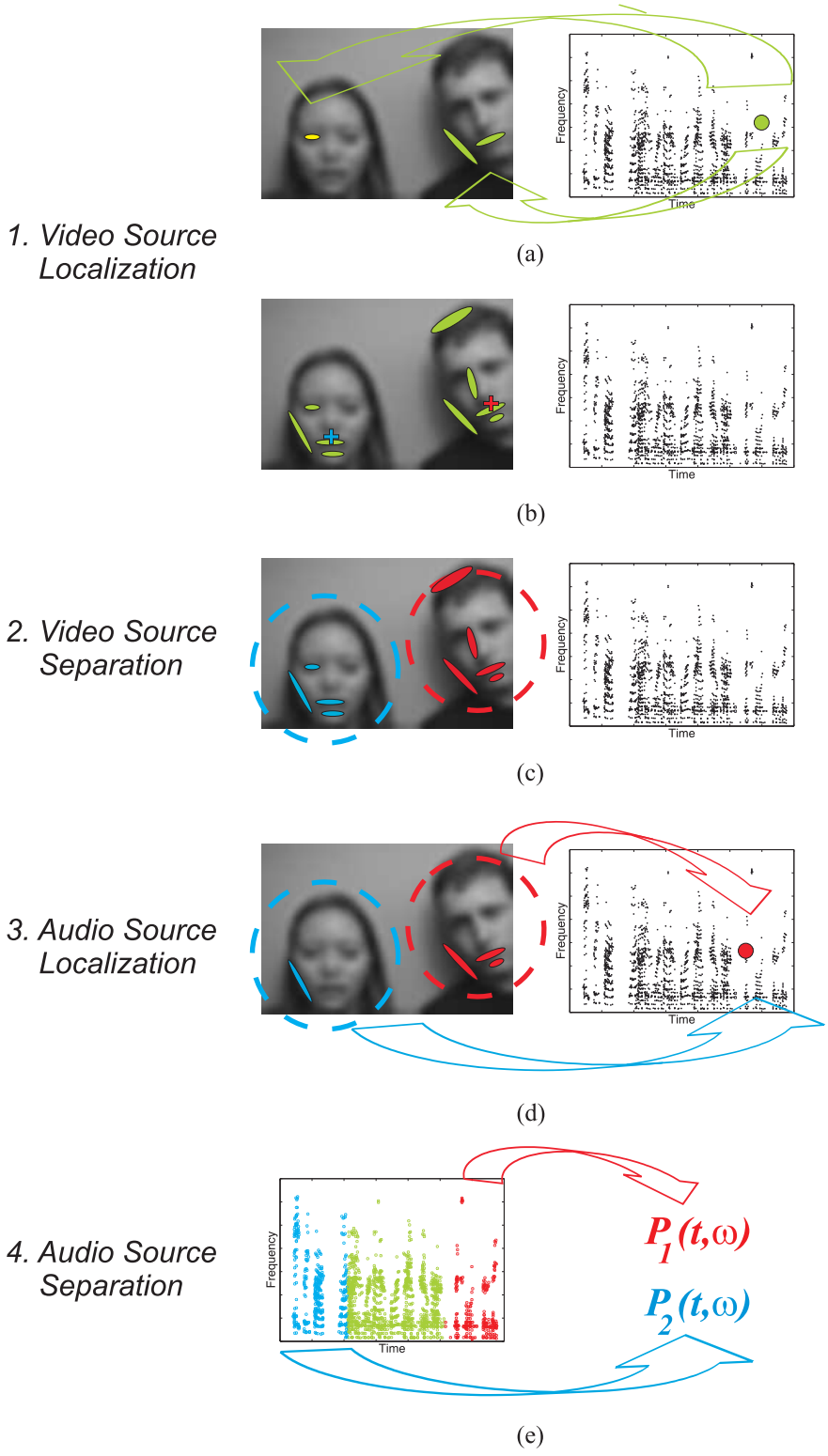


Fig. 2. Schema of the audiovisual source separation algorithm. *Phase 1* : in (a) audio entities (green dot on the spectrogram) are correlated with video atoms (green and yellow footprints are highlighted on the left image) and exploiting this information on picture (b) video sources are localized (blue and red crosses). *Phase 2* : video atoms are classified into the corresponding video sources (c), as highlighted by the footprints colors (blue for the left speaker and red for the right one). *Phase 3* : audio atoms (red dot on the right) are classified into the corresponding audio sources using the audiovisual association information (d). Periods with only one audiovisual active source are detected. *Phase 4* : in temporal periods when a single source is active (blue and red markers) the probability for each frequency to belong to one source is estimated (e). These probabilities are used to separate the sources in mixed periods (green markers).

Fig. 3. Example of a sequence analyzed with BAVSS algorithm. The sample frame [Left] shows the two speakers; as highlighted on the spectrogram of the audio [Right], in the first part of the clip the girl on the left speaks alone, then the boy on the right starts to speak as well, and finally the girl stops speaking and the boy speaks alone.

audio feature for all the soundtrack as in [16] (the average acoustic energy), a feature for each audio atom in the decomposition is used.

Fig. 4. Scheme of an audio feature.

- **Video feature** An Activation Vector $\mathsf{y}_n(t)$ is built for each video atom displacement function by detecting the peaks locations, a positive slope followed by a negative one, as shown in Fig. 5. The value of these Activation Vector peaks is 1 when the peaks in the displacement feature occur and 0 otherwise, and their duration is $W = 13$ samples. This length is chosen in order to model delays between audio and video signals, and it is big enough to associate each audio atom to at least one video atom (necessary condition to avoid energy losses on the reconstructed soundtrack).

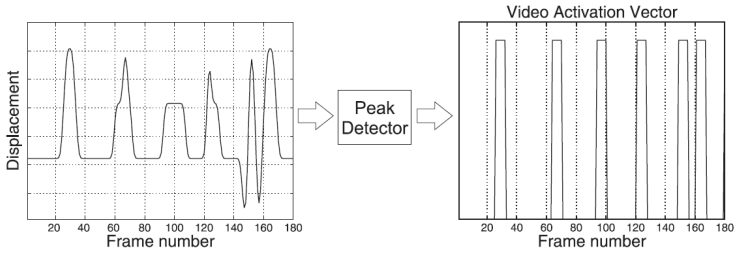


Fig. 5. Displacement function and Activation Vector obtained for a video atom.

Finally, a scalar product is computed between both features in order to obtain the *correlation scores*, $\chi_{k,n} = \langle \mathsf{f}_k(t), \mathsf{y}_n(t) \rangle$, $\forall\, k, n$.

2) Clustering: At this point, a list of correlations between audio and video atoms has been built. However, the goal is to locate the video sources on the image, and each one of these sources are composed of several video atoms. Therefore, the idea, now, is to spatially group all the structures belonging to the same speaker in order to estimate its location in the image.

In this section, we define an empirical *confidence value* κ_n of the n-th video atom as the sum of the MP coefficients c_k of all the audio atoms associated to it in the whole sequence:

$$\kappa_n = \sum_k c_k \quad \text{with } k \text{ s.t. } \chi_{k,n} \neq 0\,. \tag{3}$$

Thus, this confidence value is a measure of the number of audio atoms related to it and their weight in the MP decomposition of the audio track. Each video atom thus is characterized by its position over the image plane and

by its confidence value, i.e. $((t_{1_n}, t_{2_n}), \kappa_n)$. Looking at Fig. 6, the idea of a clustering is very intuitive. Atoms with higher confidence value form two different and well separated groups pointing out the sources, one at the left and the other at the right part of the image, while those lying far away from these regions have considerably smaller confidence values. Audio and video atoms association step has been successful, as it correlates atoms close to the source center much more often than the others.

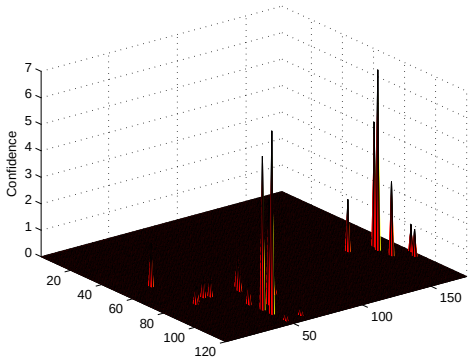


Fig. 6. Video atoms location in the image. Their confidence value is represented in the third dimension.

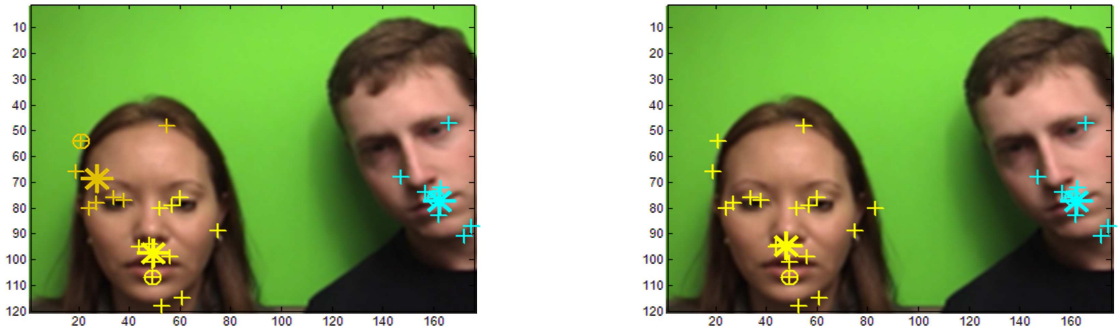
The proposed clustering algorithm is divided into three main steps:

- **Clusters Creation** First the algorithm creates Z clusters $C_i \subset P$ where $P = \{((t_{1_n}, t_{2_n}), \kappa_n)\}_n$ is the set of all points to be classified, i.e. all video atoms with confidence value different from zero. The clusters are created with the following iterative algorithm :
  1) Initialization : $Z = 0$, $P_Z = P_0 = P$;
  2) Find the point $((\tilde{t}_{1_n}, \tilde{t}_{2_n}), \tilde{\kappa}_n) \in P_Z$ with highest confidence value. It has the most important audio atoms associated, and consequently this video atom is the most probable to be the center of a source;
  3) Create a new cluster C_Z aggregating all the video atoms that are closer than a spatial maximum distance to $(\tilde{t}_{1_n}, \tilde{t}_{2_n})$ (*cluster size* defined in pixels);
  4) Remove all the video atoms assigned to this cluster from the set of points to be classified, i.e. $P_{Z+1} = P_Z \setminus C_Z$;
  5) Stop the algorithm if all the points with confidence over the mean are already classified, otherwise increment $Z \leftarrow Z + 1$ and go back to step 2. Only video atoms with significant confidence value can be the center of a new cluster.

  Considerations:
  - The *cluster size* used in step 3 determines the number of clusters created by the algorithm, and, consequently, the detected sources in first stage of the clustering. However, as we will see in the next paragraphs, the setting of this parameter does not affect significantly the final result.
  - It is basically impossible to remove real sources by the threshold applied in step 5, since most of the video atoms have a negligible confidence value, as we can see in Fig. 6.
- **Centroids Estimation** The center of mass of each cluster is computed. The confidence value of every atom is taken as the mass, and it ponders its contribution to the calculation of the centroid position over the image. Thus, for each created cluster, indexed by C_i, we calculate its centroid, $(\hat{t}_{1_i}, \hat{t}_{2_i})$, as :

$$(\hat{t}_{1_i}, \hat{t}_{2_i}) = \left(\frac{\sum_{j \in C_i} \kappa_j \cdot t_{1j}}{\sum_{j \in C_i} \kappa_j}, \frac{\sum_{j \in C_i} \kappa_j \cdot t_{2j}}{\sum_{j \in C_i} \kappa_j} \right), \tag{4}$$

  where (t_{1j}, t_{2j}) are the coordinates of the video atoms and κ_j their confidence values. These centroids are the coordinates in the image where the algorithm locates the audio sources. An example of the created clusters and their calculated centroids is shown in Fig. 7. Some of the clusters are, as expected, close to the speakers mouth, while others do not represent a source (*orange* cluster, the less important and the last one created, with cluster size 40 pixels). Next step goal is to remove these *unreliable clusters*.



(a) Clusters creation with radius 40 pixels

(b) Clusters creation with radius 60 pixels

Fig. 7. Clusters created using different cluster sizes in step 4 of the algorithm. The atom represented with a circle (○) is the one with highest *confidence value* and builds the cluster in step 2. Crosses (+) represent the coordinates of the video atoms aggregated to the cluster in step 3. Finally, the computed centroids of each cluster are indicated by an asterisk (∗). Each cluster is represented with a different color, from first to last created (descendent cluster importance) : yellow, cyan and the last one, orange, which is present only on picture (a).

- ***Unreliable Clusters* Elimination** We define the *cluster confidence value* K_{C_i} as the addition of the corresponding confidence values κ_j of the atoms belonging to the cluster, i.e. $K_{C_i} = \sum_{j \in C_i} \kappa_j$. Based on that measure, *unreliable clusters*, i.e. clusters with small confidence value K_{C_i}, are removed. A cluster is considered to be a *unreliable cluster* if its confidence value is 0.2 times the maximum value of K_{C_i} found.
  Considerations about the applied threshold:
  - High enough to eliminate the clusters that do not represent a speaker.
  - Not too high to avoid removing clusters indicating real sources. When one source is active much more time than the others, video atoms belonging to this speaker will have more correlated audio atoms making its cluster confidence value K_{C_i} considerably bigger.

At this point, a good speaker localization is achieved by means of the creation of audiovisual synchronous structures together with a robust clustering that spatially groups the video atoms forming these structures into sources. The number of sources does not have to be specified in advance since a confidence measure is introduced to automatically eliminate unreliable clusters. The algorithm is robust and the localization results do not critically depend on the choice of the cluster parameters.

B. Video Source Separation

Once the Video Source Location is achieved, each video atom is assigned to the speaker it belongs in order to posteriorly reconstruct the video sources. Regarding this objective, a *maximum distance* in pixels from the cluster centroid is defined. All the points that are closer than such distance from a centroid $(\hat{t}_{1_i}, \hat{t}_{2_i})$ are assigned to the corresponding source. With this procedure, we end up with a set of N_S clusters, $\{S_i\}_{i=1}^{N_S}$. Each group of video atoms S_i describes the video modality of an audiovisual source. To set the *maximum distance* parameter, we have to take into account several conditions:

- We do not want to assign one video atom to more than one source (no video separation).
- At the same time, the radius has to be big enough to contain the maximum number of atoms belonging to the source. If all the video atoms related to an audio atom are lost (not assigned), this audio atom cannot be assigned to one source, and severe energy losses could appear in the reconstruction of the audio signal.
- It is important not to assign to one source structures belonging to another one.

Figure 8 shows an example of the reconstruction of the current speaker detected by the algorithm. Only video atoms close to the sources estimated by the presented technique are considered. Thus, to carry out the reconstruction, the algorithm adds their energy and the effect is a highlight of the speaker's face. In both frames, the correct speaker is detected.

Fig. 8. Example of the video sources reconstruction. On the left picture the left person is speaking while on the right picture the right person is speaking.

C. Audio Source Localization

The objective of this phase is to determine the temporal periods where each source is active. This goal is achieved by means of the classification of each atom into its correspondent source using the information obtained in the last steps. For every audio atom we take into account all the video atoms related to it, their correlation scores and their classification into a source. According to this, the audio atom is assigned to the source with higher number of video atoms belonging to it, but also rewarding the temporal synchrony between these video atoms and the analyzed audio structure. Therefore, for each audio entity $\phi_k^{(a)}$ the assignation to a source can be done in the following way:

1) Take all the video atoms $\phi_n^{(v)}$ correlated with the audio atom $\phi_k^{(a)}$, i.e. for which $\chi_{k,n} \neq 0$;
2) Each of these video atoms is associated to an audiovisual source S_i ; for each source S_i compute a value H_{S_i} that is the sum of the correlation scores between the audio atom $\phi_k^{(a)}$ and the video atoms $\phi_j^{(v)}$ s.t. $j \in S_i$:

$$H_{S_i} = \sum_{j \in S_i} \chi_{k,j} \, ;$$

3) Classify the audio atom into the source S_i if the value H_{S_i} is "big enough" : here we require H_{S_i} to be twice as big as any other value H_{S_h} for the other sources. Thus we attribute $\phi_k^{(a)}$ to S_i if

$$H_{S_i} > 2 \cdot H_{S_h} \quad \text{with } h = 1, \ldots, N_S \, , \; h \neq i \, .$$

If this condition is not fulfilled, this audio atom can belong to several sources and further processing is required.

The decision bound in step 3 is introduced because, at this point of the processing, not all audio atoms can be clearly classified into one of the sources. Some of them are in an intermediate position and we cannot base the decision only on a small difference of the sources scores H_{S_i}. These atoms may belong to more than one source, or we could be making a mistake choosing one source instead of another one. This is typically the case when several speakers are simultaneously active. For these atoms additional processing is required, as it will be shown in the next section.

As an example, let us consider the situation shown in Table I. Here one audio atom has six video atoms associated (i.e. with correlation scores different from 0). Four of them belong to source S_1, and two to source S_2, with the correlation scores shown in in the table. Then, the sum of the scores are 13.88776 and 1.71717 for sources S_1 and S_2 respectively. The score for the first source is much bigger (approximately eight times bigger than the other) and thus this audio atom will be assigned to source S_1.

Using this labelling of audio atoms, time periods during which only one source is active are clearly determined. This is done using a very simple criterion: if in a continuous time slot longer than T seconds all audio atoms are assigned to source S_i, then during this period only source S_i is active. In the examples that we provide in this chapter, the value of T is set to 1 second.

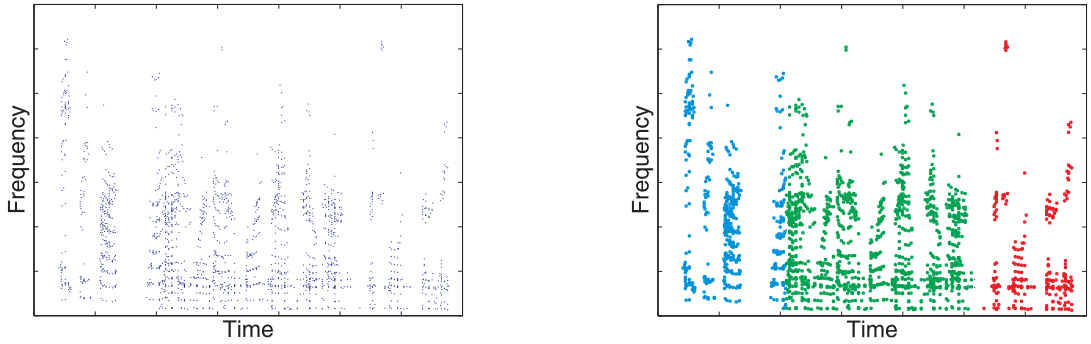
The classification of the audio atoms representing the test soundtrack shown in Fig. 3 is depicted in Fig. 9. The points in the pictures represent the position over the time-frequency plane of the audio atoms centers. The atoms locations in the original mixture are shown in picture (a), while the atoms classification is in (b). The sequence involves two speakers: at the beginning only the girl talks, then both persons speak together and finally the boy only talks. This partitioning of the signal is reflected by the proposed audio source classification method: atoms

Source S_1	**Source** S_2
6.9348	1.1146
5.8186	0.60257
0.809	
0.32536	
13.88776	1.71717

TABLE I

EXAMPLE OF THE LIST OF CORRELATION VALUES BETWEEN ONE AUDIO ATOM AND THE CORRELATED VIDEO ATOMS. FOUR OF THEM BELONG TO SOURCE 1 AND TWO TO SOURCE 2.

assigned to the girl and the boy are highlighted in blue and red respectively, while *ambiguous* atoms are indicated with green markers.



(a) Time-frequency representation (b) Audio segmentation

Fig. 9. Example of the classification of audio atoms into the correspondent sources. The points represent the time-frequency position of audio atoms. The atoms of the original mixture are in (a), while the atoms classification is in (b). The speech evolution on the sequence is reflected by the proposed classification method : at the beginning only speaker 1 is active (blue markers), then two persons are speaking (green markers) and finally only speaker 2 is active (red markers).

When several sources are present, temporal information alone is not sufficient to discriminate different audio sources in the mixture. To overcome this limitation, in these *ambiguous* time slots a time-frequency analysis is performed, which is presented in details in the next section.

D. Audio Source Separation

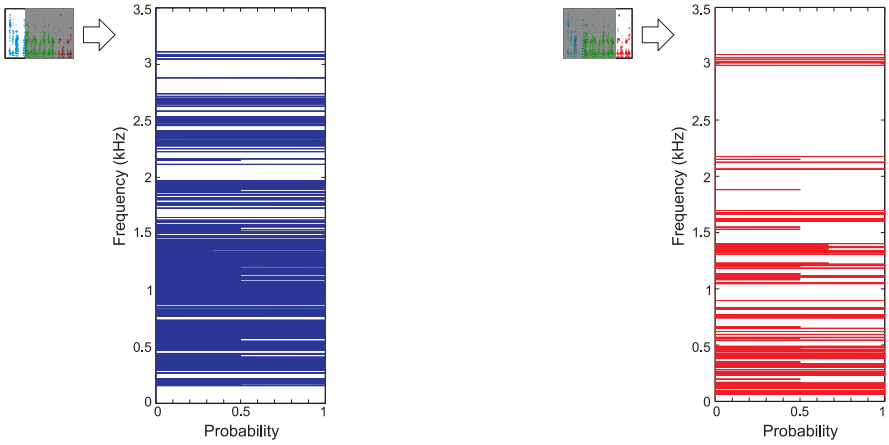
In this phase, the classification of the audio atoms in the correspondent source is performed in order to, posteriorly, reconstruct the separated soundtracks for each source. The idea is to use the frequency characteristics of each source when only this source is active in order to classify the *ambiguous* atoms of the previous phase. Thus, the audio atoms are assigned according to their time-frequency coordinates in a *Map of Probabilties*, which is built computing the product between time and frequency probabilities of each source as follows:

$$P_{S_i}(\hat{t},\hat{\omega}) = P^T_{S_i}(\hat{t}) \cdot P^\Omega_{S_i}(\hat{\omega}) \tag{5}$$

where $P^T_{S_i}(\hat{t})$ is the probability of an audio atom with time index $\hat{t}$ to belong to source S_i, and $P^\Omega_{S_i}(\hat{\omega})$ is the probability for an audio atom with frequency index $\hat{\omega}$ to belong to source S_i. This process, applied to the considered test sequence, is schematized in Figure 10. The steps for building this *Map of Probabilties* are the following:

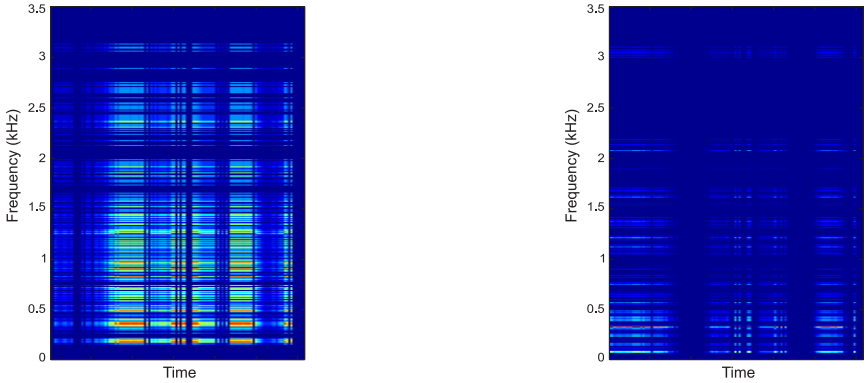
1) Frequency probabilities $P^\Omega_{S_i}(\hat{\omega})$ are computed considering temporal slots where the sources are active alone, so that a reliable association between audio atoms and sources can be established. For every value of $\hat{\omega}$ we keep the set of atoms $A_{\hat{\omega},k,n} = \{(u_k, \xi_k = \hat{\omega}), \{\chi_{k,n}\}_n\}_k$ and we estimate the frequency probability $P^\Omega_{S_i}(\hat{\omega})$ as:

$$P^\Omega_{S_i}(\hat{\omega}) = \frac{\text{card}\,(A_{\hat{\omega},k\in S_i,n})}{\text{card}\,(A_{\hat{\omega},k,n})}\,. \tag{6}$$



(a) Estimated frequency probabilities.

(b) Estimated temporal probabilities.



(c) Estimated time-frequency probabilities (Maps of Probabilities).

Fig. 10. First, frequency probabilities for the female [Left] and male [Right] speakers 10(a) are estimated on parts of the test sequence during which the subjects speak alone (blue and red dots in the spectrogram of Fig. 9(b) that is reproduced on the upper left corners of the figures). Then, temporal probabilities 10(b) are estimated using the part of the test sequence during which both persons speak together (indicated by green dots in the spectrogram of Fig. 9(b)). Finally, the Map of probabilities for the female [Left] and male [Right] speakers is build computing the product between both probabilities 10(c).

The probability of each frequency value is normalized to one, i.e. $\sum_{i=1}^{N_S} P_{S_i}^{\Omega}(\hat{\omega}) = 1$.

2) Temporal probabilities $P_{S_i}^{T}(\hat{t})$ instead, are estimated in the period where both sources are supposed to be active. These probabilities are estimated exploiting the correlation scores $\{\chi_{k,n}\}_n$ between audio atoms and video atoms classified into a source. For each time instant $\hat{t}$ we recover the set of atoms $A_{\hat{t},k,n} = \{(u_k = \hat{t}, \xi_k), \{\chi_{k,n}\}_n\}_k$ and we compute the temporal probabilities $P_{S_i}^{T}(\hat{t})$ as:

$$P_{S_i}^{T}(\hat{t}) = \frac{\sum_{k \in A_{\hat{t},k,n \in S_i}} \chi_{k,n}}{\sum_{k \in A_{\hat{t},k,n}} \chi_{k,n}} \,. \tag{7}$$

This probability basically acts like a mask: when it is 0 means that no chance is given to source S_i to be active, since no correlated event between the video source S_i and the audio signal is detected at this time instant. Again the probability of each temporal value is normalized to one, i.e. $\sum_{i=1}^{N_S} P_{S_i}^{T}(\hat{t}) = 1$.

3) For each time-frequency point $(\hat{t}, \hat{\omega})$ the probability $P_{S_i}(\hat{t}, \hat{\omega})$ in (5) is computed as a product between $P^T_{S_i}(\hat{t})$ and $P^{\Omega}_{S_i}(\hat{\omega})$ in order to penalize sources with low probability either in time or in frequency. One aspect has to be taken into account: not all the frequency values necessarily have a probability associated. In this case, the closest frequency with a probability value associated is used in (5).

Thus, according to this *Map of Probabilties* an audio atom centered in coordinates $(\hat{t}, \hat{\omega})$ will be classified into source S_i if

$$P_{S_i}(\hat{t}, \hat{\omega}) = \max\{P_{S_j}(\hat{t}, \hat{\omega})\} \text{ , with } j = 1, \ldots, N_S \,, \tag{8}$$

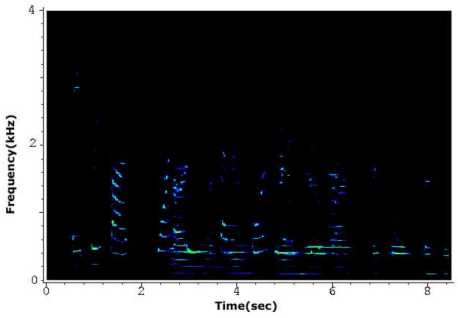
where N_S is the total number of detected sources.

Reconstruction of the Separated Signals

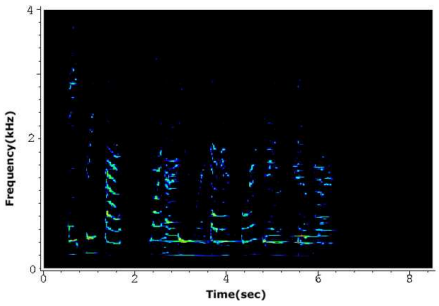
The audio signal coming from a source is reconstructed by simply adding the audio atoms classified in this source, weighted by their energy coefficients. Therefore the i-th audio source, $a_{S_i}(t)$, can be reconstructed as:

$$\hat{a}_{S_i}(t) = \sum_{k \in S_i} c_k \, \phi_k^{(a)}(t) \,, \tag{9}$$

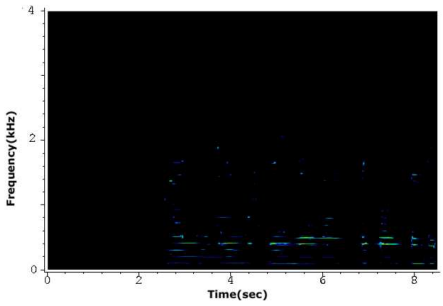
where c_k is the coefficient found by MP and corresponding to the Gabor atom $\phi_k^{(a)}(t)$ and S_i indexes the set of atoms attributed to the i-th source. The reconstructed sources $a_{S_i}(t)$ are time-evolving waveforms that can be listened using a media-player. The reconstructed sources shown in Fig. 11, for example, result well audible and the digits uttered by the two speakers can be clearly distinguished. However, quantitative measure of the quality of the source separation and reconstruction is required in order to asses the performances of the proposed algorithm.



(a) Original sequence



(b) Separated source 1 (girl)



(c) Separated source 2 (boy)

Fig. 11. Blind Source Separation of a real-world mixture representing a boy and a girl uttering digits simultaneously. The color map of the time-frequency plane images goes from black to red, through blue, green and yellow, and the pixel intensity represents the value of the energy at each time-frequency location.

IV. EXPERIMENTS

In this section the proposed BAVSS algorithm is evaluated on synthesized audiovisual mixtures. The interest of analyzing synthesized sequences resides in the fact that a ground truth can be assessed and thus an objective measure of the discrepancy between this ground truth and the reconstructed sources can be defined. The features used to

evaluate the algorithm are the percentage of correctly classified atoms for each audio source and the percentage of acoustic energy of the source that these correctly classified atoms represent.

Synthesized sequences are generated using clips taken from the *groups* partition of the CUAVE database [21] with one girl and one boy uttering sequences of digits alternatively. The video data is at 29.97 fps with a resolution of 480×720 pixels, and the audio at 44 kHz. The video data have been resized to a resolution of 120×176 pixels, while the audio signal has been sub-sampled to 8 kHz, with still a good audible quality. The video sequence is decomposed into 100 video atoms and the mixture soundtrack is decomposed into 1000 Gabor atoms. The audio and the video atoms of one speaker are then temporally shifted in order to obtain time slots with both speakers active. The steps carried out to synthesize the sequences employed in the experimental tests are the following:

1) Choose a clip of the *groups* section of the CUAVE database where two speakers (a boy and a girl) utter numbers in turns;
2) Shift the audio atoms of one speaker so that their voices are overlapped part of the time. The MP decomposition of the audio gives us the temporal position of the audio atoms belonging to each one of the speakers. Thus, we only need to take all the atoms of one speaker, which are temporally separated from those of the other one since they are speaking alternatively, and change their temporal index appropriately. The same quantity is added or subtracted from all the atoms;
3) The same procedure is applied to the video atoms. After their decomposition in 2D time-evolving atoms, the feature to analyze is the evolution of the video atoms displacement through time. In the CUAVE database, each speaker is located at one side of the image, so that video atoms belonging to one speaker have the abscissa value extracted from the decomposition between pixels 1 and 88, and the other one between 89 and 176 (the resolution of the video being 120×176). Thus, the procedure consists in temporally shifting the video atoms corresponding to one speaker by the same temporal value of the corresponding audio atoms. Notice that the shift in audio is in samples and we have to convert it in frames to apply the same temporal shift to the video.

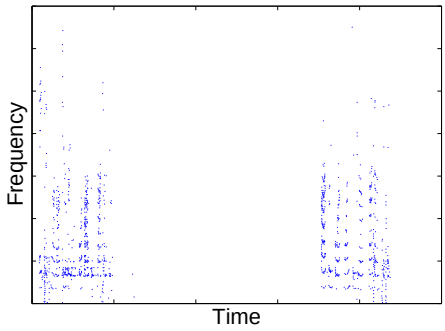
This procedure translates the whole part of the audiovisual sequence belonging to one speaker in order to have a synthetic mixture where both speakers are uttering different numbers at the same time. In the resultant synthetic clips, four cases are represented: both persons speak at the same time, only the boy or the girl speaks or silence.

First, the percentage of correct atoms is assessed. Figure 12 shows the sources extracted by the proposed algorithm [Top] and the real ones represented with 2000 Gabor atoms [Bottom], for a syntectic sequence generated by applying a shift of 150 frames to the sequence part with the male speaker in clip g20 of CUAVE database. For this synthetic sequence, on average our algorithm assigns 91% of the audio atoms to the correct source (Table II).

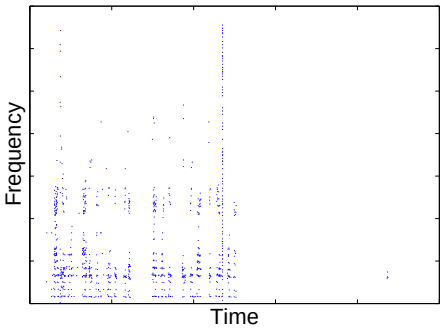
Another measure is employed in order to evaluate this method: the percentage of the original energy that these correct atoms represent. This value gives us the information relative to the difference of the original and estimated soundtracks for each speaker after the reconstruction step. This measure is performed in order to discard the very improbable fact that the 9% of audio atoms that are misclassified contribute to the separated soundtracks with the main part of the energy, i.e., this audio atoms are the first in the MP decomposition of the original mixture. For each source, this percentage is computed as the sum of the coefficients of all the atoms correctly assigned by the algorithm to the source divided by the sum of the coefficients of all the atoms belonging to this source. Therefore, this percentage can be seen as the part of the estimated signal belonging to the original one. The remaining energy is due to the assignation of the audio atoms to the incorrect speaker and constitutes the noise of the separated signal estimated by the algorithm. Figure 13 shows the original waveforms reconstructed with 2000 Gabor atoms on the bottom and those estimated by the proposed time-frequency analysis on the top.

Waveforms are very similar in the original and estimated sequences, and the percentages of the original energy that the correct atoms assigned to each source represent the 92% and 86% for the male and female speaker respectively. These percentages are high and similar to those obtained for the number of correct atoms assigned to each speaker (92% and 90%). It seems thus that correctly assigned audio atoms represent most of the energy of the speakers separated signals. Results obtained analyzing different sequences are summarized in Table II.

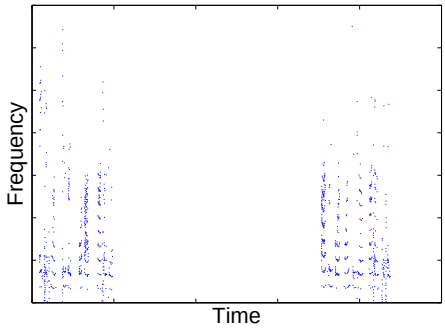
The values obtained for the percentage of correct atoms and the percentage of energy that these atoms represent are similar. We can thus argue that the algorithm distributes the errors over audio atoms of all sizes, and the percentage of correct atoms is already a good measure of the algorithm performance. Results are satisfactory, around 80–90% except for sequence g12 of CUAVE database, with a worse performance for the boy. Table II also shows that the results obtained are linked with the sequence to analyze and they are independent of the shift



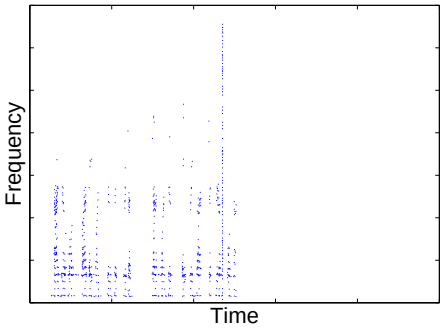
(a) Time-frequency analysis, speaker 1 (girl)



(b) Time-frequency analysis, speaker 2 (boy)



(c) Real audio atoms, speaker 1 (girl)



(d) Real audio atoms, speaker 2 (boy)

Fig. 12. Comparison between audio atoms resulting of time-frequency analysis in a synthetic mixture [Top] and the original ones [Bottom]. The points are the centers of the audio atoms over the time-frequency plane. The sequence is generated by applying a shift of 150 frames to the male speaker in clip g20 of CUAVE database.

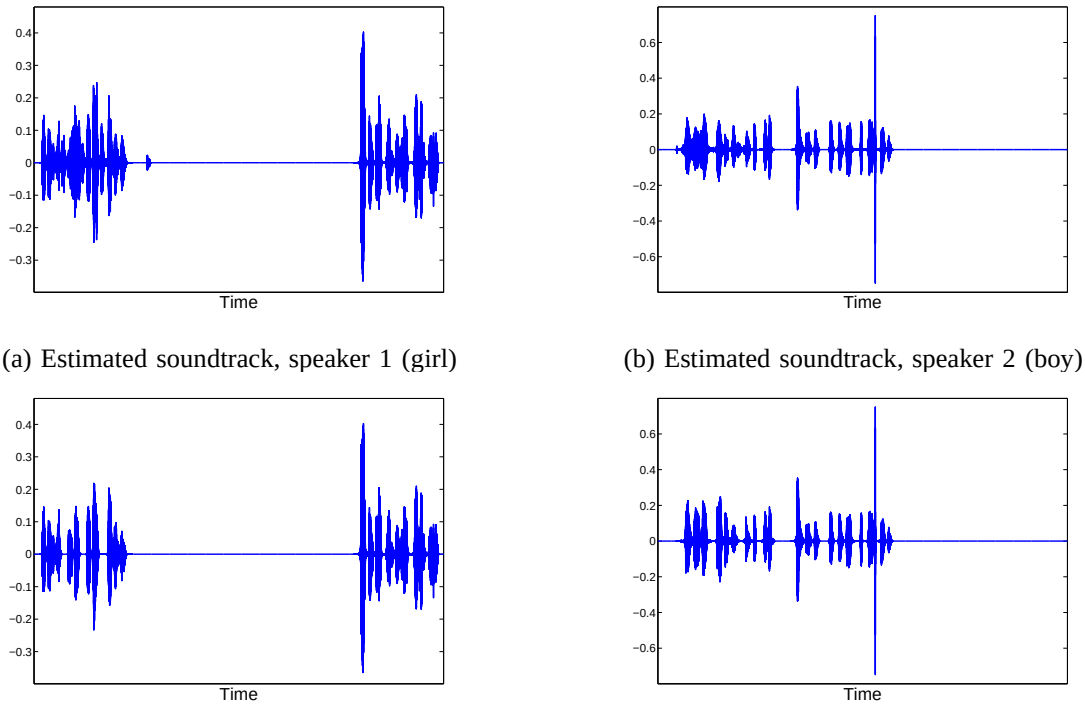
Sequence	% correct atoms		% correct energy	
	girl	boy	girl	boy
g12 shift 100 frames	86	54	73	42
g20 shift 150 frames	92	90	92	86
g21 shift 130 frames	83	81	81	75
g21 shift 169 frames	82	78	84	73

TABLE II

RESULTS OBTAINED WITH SYNTHETIC SEQUENCES GENERATED FOR DIFFERENT CLIPS OF CUAVE DATABASE.

introduced. The performance for sequence g21 is around 80% with shifts of 130 or 169 frames, with a small difference in favor of the first case.

It is important to underline that lower performances in sequence g12 are mostly due to errors done in the sequence part during which both speakers are active and they are caused by the low discriminative power of the simple model based on the probability maps of the speakers. Actually, for all tested sequences the time periods during which the sources are active alone are correctly localized except for some minor error in sequence g12. The signals in these time slots are essentially perfectly reconstructed, with a Signal to Noise Ratio (SNR) between the ground-truth MP reconstructions and the separated sources of about 50 dB. In contrast, performances are much lower in mixed periods. Although the separated speech signals are still audible and the uttered digits can be clearly distinguished most of the time, we have measured SNR values ranging from 3 dB (for the first part of the signals shown in Fig. 13(b),(d)), down to -1dB. This shows that while the proposed framework is able to localize the sources on the video and to detect time slots during which a speaker alone is present, improvements are needed in the time-frequency separation of audio mixtures. This can be done using more complex one-microphone source



(a) Estimated soundtrack, speaker 1 (girl)

(b) Estimated soundtrack, speaker 2 (boy)

(c) Real soundtrack, speaker 1 (girl)

(d) Real soundtrack, speaker 2 (boy)

Fig. 13. Comparison between estimated [Top] and real [Bottom] soundtracks for a synthetic sequence generated by applying a shift of 150 frames to the male speaker in clip g20 of CUAVE database.

separation techniques. An HMM-based generative model like the one proposed in [22] would probably match well our considered scenario, since we could still keep a completely blind setting and we could think of learning a model of the sources in time slots during which they are active alone. However this type of techniques typically require large training audio portions that can be unavailable in the presented scenario. Another interesting option could be then the use of a blind method to track the evolution of harmonics and resonances, like the one proposed in [23], but aided here by the information available in time periods presenting audiovisual sources active alone.

As a final remark, we have noticed that the quality of the reconstructed signals is considerably better for synthetic sequences than for real ones. This effect is caused by the change in the speakers fundamental frequency, and, consequently, spectral harmonics, when they speak simultaneously in real sequences. Humans tend to change their speech characteristics in order to differ more from the other speakers and to be, thus, more easily heard. This change in the sources frequency behavior causes a worse performance of the algorithm, since the speakers models are learned in temporal periods during which they are alone.

V. Discussion

In this report we have introduced a new algorithm to perform a Blind Audiovisual Source Separation task. We consider sequences made of one soundtrack and the video signal associated, without the stereo audio signal usually employed for the BASS task. The method builds correlation between acoustic and visual structures that are represented using atoms taken from redundant dictionaries. Video atoms that exhibit strong correlations with the audio track and that are spatially close are grouped together using a robust clustering algorithm that can confidently count and localize on the image plane audiovisual sources. Then, using such information and exploiting the coherence between audio and video signals, audio sources are localized as well and separated. The presented algorithm needs time periods with sources active alone to predict their behavior in the mixture. This condition is however not very restrictive, since it is rare that in real-world mixtures all the sources are active all the time.

Several tests are performed in real-world and synthetic sequences, and encouraging results are obtained for both of them. The speaker spatial localization is successfully performed in challenging sequences where two persons speak simultaneously. Concerning the audio source separation part, the audible quality of the separated audio signals

is also reasonably good, with reconstructed waveforms close to the original ones. However, we believe that the proposed method can be improved using more sophisticated techniques for the separation of audio sources in time slots that present source mixtures. To this end, HMM-based models [22] or audio feature tracking techniques [23] could be plugged in the proposed framework.

REFERENCES

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds., pp. 3–51. Lawrence Erlbaum Associates, 1987.

[3] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[4] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (Audiovisual Codebook Dependent Cepstral Normalization)," in *Proc. Int. Conf. Spoken Language Proc. (ICSLP)*, 2002, pp. 1449–1452.

[5] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2002, pp. 2025–2028.

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

[7] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 495–506, 2005.

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

[9] R. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2004, vol. 5, pp. 645–648.

[10] S. Rajaram, A. V. Nefian, and T.S.; Huang, "Bayesian separation of audio-visual speech sources," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2004, vol. 5, pp. 657–660.

[11] B. Rivet, L. Girin, and C. Jutten, "Solving the indeterminations of blind source separation of convolutive speech mixtures," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2005, vol. 5, pp. 533–536.

[12] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2005, vol. 5, pp. 425–428.

[13] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Blind audio source separation," Tech. Rep. C4DM-TR-05-01, Centre for Digital Music, Queen Mary University of London, 2005.

[14] M.T. Lallouache, *Un poste visage-parole couleur. Acquisition et traitement automatique des lèvres*, Ph.D. thesis, Institut National Polytechnique, Grenoble, France, 1991.

[15] G. Edwards, C. Taylor, and T. Cootes, "Interpreting face images using active appearance models," in *Proc. 3^{rd} Int. Conf. on Automatic Face and Gesture Recognition*, 1998, pp. 300–305.

[16] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *International Conference on Image Processing*, 2005, pp. III: 145–148.

[17] P. Bertelson, J. Vroomen, G. Wiegeraad, and B. de Gelder, "Exploring the relation between mcgurk interference and ventriloquism," *In Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 2, pp. 559–562, 1994.

[18] J. Driver, *Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading*, chapter 381, pp. 66–68, Nature, 1996.

[19] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[20] O. Divorra Escoda and P. Vandergheynst, "A bayesian approach to video expansions on parametric over-complete 2-d dictionaries," in *International Workshop on Multimedia Signal Processing*. IEEE, September 2004, IEEE.

[21] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1189, Nov. 2002.

[22] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Subband audio modeling for single-channel acoustic source separation," in *ICASSP*, Montreal, 2004.

[23] M. Reyes-Gomez, N. Jojic, and D. Ellis, "Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation/tracking model," in *ResearchWorkshop on Statistical and Perceptual Audio Processing*, Korea, October 2004, SAPA04.