BLIND AUDIOVISUAL SOURCE SEPARATION USING SPARSE REPRESENTATIONS

*Anna Llagostera Casanovas, Gianluca Monaci and Pierre Vandergheynst**

Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
E-mail : {anna.llagostera,gianluca.monaci,pierre.vandergheynst}@epfl.ch

ABSTRACT

In this work we present a method to jointly separate active audio and visual structures on a given mixture. Blind Audiovisual Source Separation is achieved exploiting the coherence between a video signal and a one-microphone audio track. The efficient representation of audio and video sequences allows to build relationships between correlated structures on both modalities. Video structures exhibiting strong correlations with the audio signal and that are spatially close are grouped using a robust clustering algorithm that can count and localize audiovisual sources. Using such information and exploiting audio-video correlation, audio sources are also localized and separated. To the best of our knowledge this is the first blind audiovisual source separation algorithm conceived to deal with a video sequence and the corresponding *mono* audio signal.

Index Terms— Audiovisual processing, blind source separation, sparse signal representation.

1. INTRODUCTION

Few methods exist that exploit audiovisual coherence to separate *stereo* audio mixtures [1, 2, 3, 4, 5]. 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 [6].

The approach we consider in this paper is very different from existing ones. It is inspired by [7], where audiovisual sources are localized using sparse geometric representations of video sequences. Here we first localize and separate the visual sources exploiting audiovisual synchrony. We create several clusters of video structures, each group corresponding to a detected source. Then, exploiting this information and the correlations established between audio and video entities we separate the audio mixture as well. We want to stress three important differences between the proposed approach and existing audiovisual separation methods :

1. State-of-the-art audiovisual separation algorithms exploit stereo audio signals, using classic BASS techniques helped by visual information. In contrast the audio signal we consider here comes from only *one microphone*;
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 [3, 4], either it is considered the case where one audiovisual source is mixed with an *audio-only* source [1, 2, 5]. Here, in contrast, we simultaneously separate audio-video sources building correlations between acoustic and visual entities;
3. Existing algorithms, except for [3], require off-line training to build the audiovisual source model. This is mainly due to the fact that the algorithms in [1, 2, 4, 5] 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). Here, in contrast, no training is required.

To summarize we want to solve a blind Single-Channel BASS problem [8], 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. In Sec. 2 we describe the audio and video features used to represent both modalities, while Sec. 3 details the *Blind Audiovisual Source Separation* (BAVSS) algorithm. In Sec. 4 we present the separation results obtained on real and synthesized audiovisual clips. Finally, in Sec. 5 achievements and future research directions are discussed.

2. AUDIO AND VIDEO REPRESENTATIONS

Audio Representation – The audio signal $a(t)$ is decomposed using the Matching Pursuit algorithm (MP) over a redundant dictionary of Gabor atoms $\mathcal{D}^{(a)}$ [7]. Thus, the signal $a(t)$ is approximated using K atoms as

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

where c_k are the coefficients for every atom $\phi_k^{(a)}(t)$.

Video Representation – The video signal is represented using the video MP algorithm adopted in [7]. The sequence is decomposed into a set of video atoms representing salient visual components and their temporal transformations. The video signal $V(x_1, x_2, t)$ is approximated using N video atoms $\phi_n^{(v)}$ as

$$V(x_1, x_2, t) \approx \sum_{n=0}^{N-1} c_{n(t)} \phi_n^{(v)}(x_1, x_2, t)\,, \tag{2}$$

where $c_{n(t)}$ are the coefficients. The atoms $\phi_n^{(v)}$ are edge-like functions that are tracked across time. Each function is represented by a set of parameters describing its shape and position and that evolve through time [7]. We compute a feature describing the displacement of each video atom, $d_n(t) = \sqrt{t_{1_n}^2(t) + t_{2_n}^2(t)}$, from its position parameters $(t_{1_n}(t), t_{2_n}(t))$.

3. BLIND AUDIOVISUAL SOURCE SEPARATION (BAVSS)

Figure 1 schematically illustrates the BAVSS process. First, video sources are localized using a clustering algorithm that spatially groups

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

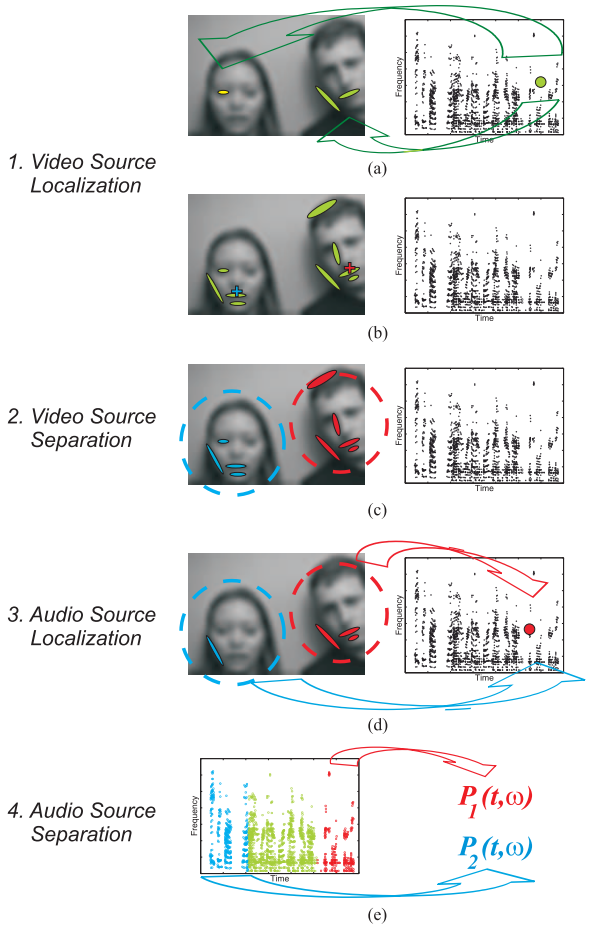


Fig. 1. Schema of the audiovisual source separation algorithm. *Phase 1* : in (a) audio entities (green dot on the right spectrogram) are correlated with video atoms (green and yellow footprints on the left image) and exploiting this information on picture (b) video sources are localized (blue and red crosses). *Phase 2* : video atoms are classified into the corresponding sources (c), as highlighted by the footprints colors. *Phase 3* : audio atoms (red dot on the right) are classified into the corresponding audio sources using the audiovisual association information (d), detecting periods with only one audiovisual active source. *Phase 4* : in temporal periods with a single active source (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).

the video structures that are correlated with the audio atoms. Second, a 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 active source. Finally, the sources frequency behavior is learned in time periods during which sources are active alone in order to separate them in the mixed periods. The interested reader should also refer to [10] for additional details about the proposed BAVSS procedure.

Two main assumptions are made on the type of analyzed sequences. First, for each detected video source there is one and only one associated source in the audio mixture. This means that audio "distractors" in the sequence (e.g. a person speaking out of the camera's field of view) are considered as noise and their contribution to the mixture is associated to the sources found in the video. 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 as it can be removed by analyzing the sequences using shifting time windows.

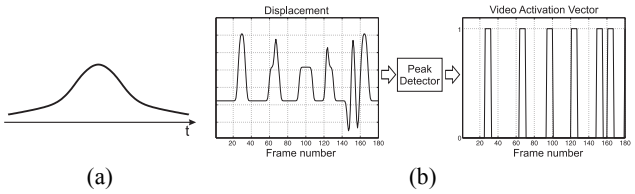


Fig. 2. Audio feature $\mathsf{f}_k(t)$ (a) and displacement function $d_n(t)$ with corresponding *Activation Vector* $\mathsf{y}_n(t)$ obtained for a video atom (b).

3.1. Video Source Localization

3.1.1. Audio and Video Atoms Association

Audio and video structures are associated computing the *correlation scores* $\chi_{k,n}$ between each audio atom $\phi_k^{(a)}$ and each video atom $\phi_n^{(v)}$. These scores measure the degree of synchrony between *relevant events* in both modalities : the presence of an audio atom (energy in the time-frequency plane) and a peak in the video atom displacement (oscillation from an equilibrium position).

Audio feature – The feature $\mathsf{f}_k(t)$ that we consider is the energy distribution of each audio atom projected over the time axis. In the case of Gabor atoms it is a Gaussian function whose position and variance depend on the atoms parameters (Fig.2(a)).

Video feature – An *Activation Vector* $\mathsf{y}_n(t)$ [7] is built for each atom displacement function $d_n(t)$ by detecting the peaks locations as shown in Fig. 2(b). The Activation Vector peaks are filtered by a window of width $W = 13$ samples in order to model delays and uncertainty.

Finally, a scalar product is computed between both features to obtain the *correlation scores*, $\chi_{k,n} = \langle \mathsf{f}_k(t), \mathsf{y}_n(t) \rangle, \forall\, k, n$. This value is high if the audio feature and the video displacement peak exhibit a big temporal overlap. Thus, a high correlation score means high probability for a video structure of having generated the sound.

3.1.2. Clustering

The idea, now, is to spatially group all the structures belonging to the same speaker in order to estimate its position on the image. We define the 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, $\kappa_n = \sum_k c_k$, with k such that $\chi_{k,n} \neq 0$. This value is a measure of the number of audio atoms related to this video structure 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. $((t_{1_n}, t_{2_n}), \kappa_n)$. We cluster all the video atoms correlated with the audio signal (i.e. with $\kappa_n \neq 0$) following these three main steps :

1 **Clusters Creation –** The algorithm creates Z clusters $\{C_i\}_{i=1}^{Z}$, by iteratively selecting the video atoms with highest confidence value and aggregating to them atoms closer than a *cluster size* D, whose value is set according to the video characteristics;
2 **Centroids Estimation –** The center of mass of each cluster is computed taking the confidence value of every atom as the mass.

The resulting centroids are the coordinates in the image where the algorithm locates the audio sources;

3 ***Unreliable Clusters*** **Elimination –** We define the *cluster confidence value* K_{C_i} as the sum of the 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 this measure, *unreliable clusters*, i.e. clusters with small confidence value K_{C_i} (here smaller than 0.2 times the maximum value of K_{C_i} found) are removed, obtaining the final set of $N_S \leq Z$ clusters, $\{C'_i\}_{i=1}^{N_S}$.

At this stage a good speaker localization is achieved. 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 parameters choice.

3.2. Video Source Separation

This step classifies *all* the video atoms closer than the cluster size D to a centroid into the corresponding source (in previous step only atoms with $\kappa_n \neq 0$ are considered). Each such group of video atoms, S_i, describes the video modality of an audiovisual source, achieving thus the Video Separation objective.

3.3. Audio Source Localization

The objective of this phase is to determine the temporal periods during which the sources are active. First, each audio atom $\phi_k^{(a)}$ is classified into its corresponding source in the following way :

1. Take all video atoms $\phi_n^{(v)}$ correlated with the audio atom $\phi_k^{(a)}$;
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 twice as big as any other value H_{S_h} for the other sources. If this condition is not fulfilled, this audio atom can belong to several sources and further processing is required.

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

3.4. Audio Source Separation

An audio atom $\phi_k^{(a)}$ is characterized by its position on the time-frequency plane, (u_k, ξ_k), and by a set of correlation scores $\{\chi_{k,n}\}_n$. Thus the set of points $A = \{(u_k, \xi_k), \{\chi_{k,n}\}_n\}_{k=0}^{K-1}$ collects the K audio atoms of the decomposition. Our aim is to associate all the points in A to one of the N_S detected sources. In Sec. 3.3 audio atoms in time slots with a single source present (red and blue markers in the spectrogram of Fig. 1(e)) have been assigned to a source. However, when several sources are present (green markers in Fig. 1(e)), temporal information alone is not sufficient to discriminate different audio sources in the mixture. The idea is to use the frequency characteristics of each source when it is active alone in order to classify the *ambiguous* atoms belonging to a mixture. These atoms are assigned according to their time-frequency coordinates in a *Map of Probabilities*, which is built computing the product between time and frequency probabilities of each source as :

$$P_{S_i}(\hat{t}, \hat{\omega}) = P_{S_i}^T(\hat{t}) \cdot P_{S_i}^\Omega(\hat{\omega}) \,, \tag{3}$$

where $P_{S_i}^T(\hat{t})$ is the probability of an audio atom with time index $\hat{t}$ to belong to source S_i, and $P_{S_i}^\Omega(\hat{\omega})$ is the probability for an audio atom with frequency index $\hat{\omega}$ to belong to source S_i.

Frequency probabilities $P_{S_i}^\Omega(\hat{\omega})$ are computed considering temporal slots during which the sources are active alone (red and blue markers in Fig. 1(e)), 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 as :

$$P_{S_i}^\Omega(\hat{\omega}) = \frac{\text{card}\left(A_{\hat{\omega},k \in S_i,n}\right)}{\text{card}\left(A_{\hat{\omega},k,n}\right)} \,. \tag{4}$$

where card$(\cdot)$ is the cardinality function. Not all the frequency values necessarily have a probability associated and, in this case, the closest frequency with a probability value associated is used in (3).

Temporal probabilities $P_{S_i}^T(\hat{t})$ are estimated in periods during which several sources are supposed to be active (green part in Fig. 1(e)). 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 as :

$$P_{S_i}^T(\hat{t}) = \frac{\sum_{k \in A_{\hat{t},k,n \in S_i}} \chi_{k,n}}{\sum_{k \in A_{\hat{t},k,n}} \chi_{k,n}} \,. \tag{5}$$

This probability acts like a mask : if it is 0 it means that no chance is given to source S_i to be active, since no correlation between the video source S_i and the audio signal is detected at this time instant.

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

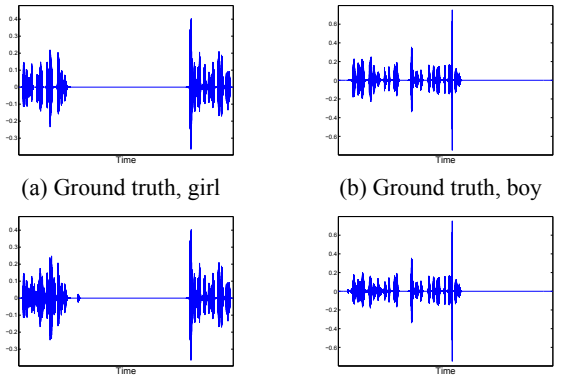
$$P_{S_i}(\hat{t}, \hat{\omega}) = \max\{P_{S_j}(\hat{t}, \hat{\omega})\} \text{ , with } j = 1, \ldots, N_S \,. \tag{6}$$

Reconstruction of the separated signals – Finally, the signal coming from the i-th audio source, $a_{S_i}(t)$, can be reconstructed by simply adding the audio atoms classified in this source as $\hat{a}_{S_i}(t) = \sum_{k \in S_i} c_k \, \phi_k^{(a)}(t)$, where c_k are the MP coefficients of $\phi_k^{(a)}(t)$ and S_i indexes the atoms attributed to the i-th source.

4. EXPERIMENTS

The proposed BAVSS method is evaluated on synthesized audiovisual mixtures, in order to have an objective evaluation of the algorithm's performances. Sequences are synthesized using clips taken from the *groups* partition of the CUAVE database [9] with one girl and one boy uttering sequences of digits alternatively. The video data is at 29.97 frames/sec with a resolution of 480×720 pixels, and the audio at 44 kHz. The video has been resized to a 120×176 pixels and the audio has been sub-sampled to 8 kHz. The video signal is decomposed into $N = 100$ video atoms and the soundtrack is decomposed into $K = 2000$ atoms. The video clustering algorithm uses a value of $D = 80$ pixels.

Ground truth mixtures are obtained by temporally shifting audio and video atoms of one speaker in order to obtain time slots with both speakers active simultaneously. For further details on the adopted procedure, please refer to [10]. Figures 3(a)-(b) show resulting synthetic clips generated by shifting by 150 frames the sequence part with the male speaker in clip g20 of the CUAVE database. At the beginning of the clip, both persons speak at the same time, then only the boy or the girl speak alone. Figures 3(c)-(d) show the sources extracted by the proposed algorithm. It is interesting to note that the



(a) Ground truth, girl (b) Ground truth, boy

(c) Estimated source, girl (d) Estimated source, boy

Fig. 3. Comparison between real (a)-(b) and estimated (c)-(d) soundtracks for a synthetic sequence generated by applying a shift of 150 frames to the male speaker in clip g20 of the CUAVE database.

separated signals present a perfect reconstruction when the boy or the girl speak alone, indicating a correct detection of these periods.

The quantities 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. For each source, this second measure 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 audio atoms to the incorrect speaker and represents the noise of the separated signal estimated by the algorithm.

Results obtained analyzing different synthesized sequences are summarized in Table 1. Classification results are satisfactory, except for sequence g12. The values obtained for the percentage of correct atoms and the percentage of energy that these atoms represent are similar, which means that the errors are distributed over audio atoms of all sizes. The obtained results seem to be independent of the shift introduced (sequence g21, last two lines of Table 1). Lower performances in sequence g12 are 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 source separation method based on probability maps. However, for all tested sequences the time periods during which the sources are active alone are correctly localized except for some minor errors in sequence g12.

5. CONCLUSION

In this paper we have introduced a novel algorithm to perform Blind Audiovisual Source Separation. We consider sequences made of a one-microphone soundtrack and the video signal associated. The method correlates acoustic and visual structures that are represented using atoms taken from redundant dictionaries. A robust clustering algorithm is proposed that can count and localize on the image plane audiovisual sources. Using such information and exploiting the coherence between audio and video patterns, audio sources are also localized and separated. The presented algorithm requires 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 sources are active at the same time.

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 1. Results obtained with synthetic sequences generated for different clips of CUAVE database.

Several tests are performed on real and synthetic sequences. The speaker spatial localization is successfully performed in challenging clips involving two persons speaking simultaneously. Concerning the audio source separation part, the audible quality of the separated audio signals is also reasonably good. However, the proposed method should be improved using more sophisticated audio source separation techniques in time slots presenting source mixtures. The framework developed in this paper seems to be appropriate to improve the proposed system by considering HMM-based models [11] or audio feature tracking techniques [8] at the audio separation stage.

6. REFERENCES

[1] 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.

[2] R. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," in *Proc. IEEE ICASSP*, 2004, pp. 645–648.

[3] S. Rajaram, A. V. Nefian, and T.S.; Huang, "Bayesian separation of audio-visual speech sources," in *Proc. IEEE ICASSP*, 2004, pp. 657–660.

[4] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *Proc. IEEE ICASSP*, 2005, pp. 425–428.

[5] B. Rivet, L. Girin, and C. Jutten, "Solving the indeterminations of blind source separation of convolutive speech mixtures," in *Proc. IEEE ICASSP*, 2005, pp. 533–536.

[6] 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, Queen Mary University of London, 2005.

[7] G. Monaci, Ò. Divorra, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, vol. 86, no. 12, pp. 3534–3548, 2006.

[8] 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 *Workshop on Statistical and Perceptual Audio Proc.*, 2004.

[9] 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 JASP*, , no. 11, pp. 1189–1201, 2002.

[10] A. Llagostera Casanovas, G. Monaci, and P. Vandergheynst, "Blind audiovisual source separation using sparse redundant representations," EPFL-ITS Technical Report 2007.01, 2007, [Online] Available: `http://lts2www.epfl.ch/`.

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