BLIND AUDIO-VISUAL SOURCE SEPARATION USING SPARSE REDUNDANT REPRESENTATIONS

DIPLOMA PROJECT

ANNA LLAGOSTERA CASANOVAS

Assistant:
GIANLUCA MONACI
Supervisor:
Prof. PIERRE VANDERGHEYNST

Lausanne, EPFL
July 2006

Table of Contents

List of Figures

List of Tables

Abstract

This report presents a new method to confront the Blind Audio Source Separation (BASS) problem, by means of audio and visual information. In a given mixture, we are able to locate the video sources first and, posteriorly, recover each source signal, only with one microphone and the associated video.

The proposed model is based on the Matching Pursuit (MP) [18] decomposition of both audio and video signals into meaningful structures. Frequency components are extracted from the soundtrack, with the consequent information about energy content in the time-frequency plane of a sound. Moreover, the MP decomposition of the audio is robust in front of noise, because of its plain characteristic in this plane. Concerning the video, the temporal displacement of geometric features means movement in the image. If temporally close to an audio event, this feature points out the video structure which has generated this sound.

The method we present links audio and visual structures (atoms) according to their temporal proximity, building audiovisual relationships. Video sources are identified and located in the image exploiting these connections, using a clustering algorithm that rewards video features most frequently related to audio in the whole sequence.

The goal of BASS is also achieved considering the audiovisual relationships. First, the video structures close to a source are classified as belonging to it. Then, our method assigns the audio atoms according to the source of the video features related.

At this point, the separation performed with the audio reconstruction is still limited, with problems when sources are active exactly at the same time. This procedure allows us to discover temporal periods of activity of each source. However, with a temporal analysis alone it is not possible to separate audio features of different sources precisely synchronous.

The goal, now, is to learn the sources frequency behavior when only each one of them is active to predict those moments when they overlap. Applying a simple frequency association, results improve considerably with separated soundtracks of a better audible quality.

In this report, we will analyze in depth all the steps of the proposed approach, remarking the motivation of each one of them.

Chapter 1

Introduction

1.1 Problem Statement

It is relatively easy for a human to correctly interpret a scene consisting on a combination of acoustic and visual stimuli and to take profit of both the information to experience a richer perception of the world. On the contrary, computer systems have considerable difficulties when having to deal with multimodal signals, and the information that each component contains about the others is usually discarded.

Audio and video modalities experiment synchronous changes when these variations are caused by the same physical phenomenon. This observation is the basis of the proposed model. Assessing the temporal correspondence between audio and video parts of the audiovisual sequence, we have the key to discover the relationships between both modalities and to obtain the maximum of information.

The previous research work performed by Monaci, Divorra and Vandergheynst [19] explored the capabilities of redundant parametric decompositions to describe audiovisual sequences. These techniques allow to interpret signals in terms of their most salient structures, preserving good representational properties thanks to the use of redundant, well designed, dictionaries. In this way, it is possible to combine audio and video representations using simple and intuitive, but effective, criteria.

Our project starts from this point. Only with the received signal of one single microphone and the video associated we are already able to spatially locate the active speaker of an audiovisual sequence.

The objective of our research is to perform a kind of Audiovisual Separation. First, we want to separate and reconstruct the video sources combining information of both modalities. Once located the video sources on the image, we only have to reconstruct them by assuming that the structures close to a source belong to it. Second, the goal is to perform a Blind Audio Source Separation (BASS) aided by video. This kind of Audiovisual

separation was already studied by Smaragdis in [29], but in this case the objective of BASS is much more ambitious for the complexity of speech signals.

The previous approach [19] lost a lot of information, since it only used the energy feature in the audio part. Thus, we think it is possible to perform a good BASS by incorporating audio processing. Then, in order to achieve this ambitious second objective, we characterize the audio signal with a redundant parametric decomposition (the Matching Pursuit technique [18]).

1.2 Proposed model

The proposed model is based on temporal synchrony between audio and video *relevant events*. We assume that one observed physical phenomenon is generated by changes temporally close in both features. For example, in our case it is clear that speech is caused by a moving mouth, both events are synchronous and clearly related.

Therefore, the correspondence between both features can be evaluated with a temporal analysis. The proposed model looks for related audiovisual features that characterize the sequence to posteriorly analyze them and extract the maximum information to achieve our purpose.

1.2.1 Audio and video representation

We have discussed before about the synchrony between audio and visual *relevant events*, but what are these events we want analyze? Why have we chosen these events and not others? What makes them *relevant*?

Concerning the **video signal**, we use redundant parametric decompositions as in the previous research work [19]. Therefore, the Matching Pursuit (MP) algorithm proposed by Divorra and Vandergheynst [10] is used to represent the video signal.

The sequence is decomposed into a set of 2-D atoms evolving in time. Relevant geometric video components are represented and their temporal transformations tracked. The video sequence is represented by meaningful unities, while other pixel-based analysis have no visual global sense, representing a poor source of information.

This MP tracking algorithm characterizes the whole video signal with 6 parameters describing the temporal evolution of the atoms. They are characterized by: a coefficient, x position, y position, x scale, y scale and the angle of rotation.

The most important feature concerning our analysis is the variation of x and y position parameters (modulus of displacement of atoms). This feature represents the modification of the video atom situation in the image, and quantifies the movement present in the scene.

As a result, the *relevant events* in the video part are the peaks in the video atoms displacement.

Regarding the **audio**, we also perform the Matching Pursuit decomposition of the signal, obtaining the resultant features in the time-frequency plane with the LastWave software [1]. In this case, the elements coincide with the audio formants or the signal frequency components.

Another advantage of this representation is the robustness in front of noise. The plain characteristic of this element causes the apparition of audio atoms spread across the time-frequency plane. However, the structures presents in the original signal are still recognizable, due to the higher concentration of their energy in the time-frequency domain. Thus, while important structures in the original signal are always represented (they are chosen in the first iterations of the algorithm), part of the noise is filtered out.

This audio decomposition has a lot of output parameters, but the proposed model uses only the time and frequency index, the windowSize used and the coefficient (energy of the audio atom). However, the most important is the temporal situation, as our analysis is based in the audiovisual correlation in this domain.

In the previous research work only the energy of the audio signal was used in the analysis. This is the reason why this model has much more possibilities of discovering the relationship between audio and video and extract the joint information they contain.

1.2.2 Procedure

The first goal to achieve is, like in the previous work of Monaci, Divorra and Vandergheynst, to **detect the speakers** in an audiovisual sequence. As in [19], we will perform the speaker detection task exploiting the temporal synchrony between audio and visual events, but now the concept of event changes for the acoustic signal. As a result, it is necessary to change the analysis principles.

In the previous work, there was only one feature that characterized all the audio signal, and the algorithm looked for the video feature more correlated and decided wether it was the cause of the audio event or not.

Now, each element of the audio decomposition has a feature associated. So, our method looks for the video features temporally correlated to each one of the audio ones (audio and video atoms assignation). Then, the video atoms related to higher number of important (high energy) audio atoms have more possibilities to belong to one speaker's mouth. A clustering algorithm groups these video atoms and calculates their centroid, estimating, thus, the spatial position of the video sources.

Then, the most important part and the next objective of this research work is to perform the **Blind Audio Source Separation** aided by video information. In our sequence, we want to extract separately the signals that form the audio mixture.

Our method confronts these problem with the same information extracted from the previous temporal analysis. At the end of the last part, we already know: the relationship between audio and video atoms, with a *correlation score* measuring the synchrony between them, the 2-D position of the sources over the image plane, and the video atoms belonging to each speaker (defining a maximum distance in pixels from the mouth).

The method classifies each audio geometric feature in one of the sources, according to which source the video atoms related to it belong. Therefore, if the majority of these video elements with highest *correlation score* belong to one source, logically the acoustic atom will belong to the same source.

The last part lies in reconstructing the audio signal of each source by adding the energy distribution of all the audio atoms belonging to this source.

The implemented method allows us to separate quite well the mixture. The main problem at this step of the processing is that the performed analysis is only temporal and, as a result, our method is incapable of distinguish between two speakers when they are speaking exactly at the same time (audio atoms with the same time index). However, at this point we have already temporally separated the mixture, detecting clear periods with only one active source. The goal is to learn the frequency behavior of each one of the speakers in these periods in order to, posteriorly, predict them when they are mixed.

Therefore, a complementary *frequency analysis* is performed in order to improve the present results. The audio features are situated in the time-frequency domain, and we were not exploiting the information present in this second dimension to perform the BASS. Our algorithm assigns the frequencies to the sources when they are alone (known cases) to later classify the audio atoms into one of them when they are mixed. The past experience gives us the key for future and more difficult situations.

Results show an important improvement in the Blind Audio Source Separation with the addition of this frequency analysis. Now, the algorithm is able to separate both in time and frequency, and a good BASS is performed.

The investigated approach is tested on a set of real-world sequences taken from the CUAVE database [21], in order to test the proposed model in a multi-modal context.

To summarize, obtained results demonstrate that is possible to perform a good *Audiovisual* Source Separation through a multimodal analysis, extracting the mutual information between audio and video features. However, an adequate representation pointing out *relevant events* is necessary.

1.3 Report organization

This report is organized as follows. In **Chapter 2**, the previous work in the two main aims of the proposed model are reviewed. Several methods to perform speaker localization based on audio-visual synchrony, and Blind Audio Source Separation are explained. **Chapter 3** describes the pre-processing performed to the audio and video part of the sequence, jointly with the motivations of the MP decomposition. The phases of the temporal analysis are carefully detailed and illustrated in **Chapter 4**: first, locate the sources, then, separate and reconstruct them, and, finally, carry out the BASS aided by video. **Chapter 5** describes the frequency analysis applied to complement the temporal one when several sources are active exactly at the same time. In **Chapter 6** the CUAVE database and the methodology of the analysis are presented. The obtained results are evaluated and discussed. Finally, the conclusions and future work are analyzed in **Chapter 7**.

Chapter 2

Related Work

This research work is based on the previous one by Monaci, Divorra and Vandergheynst [19], which explored the capabilities of redundant parametric decompositions to describe audiovisual sequences. These representations show the temporal evolution of the most salient structures in both audio and video modalities, and, through an analysis to detect multimodal synchronous relevant events, they allow to locate in the image the sources of a given soundtrack.

This report presents an extension of this algorithm in order to aspire a higher objectives. Now we introduce a new concept, the **Audiovisual Separation**, which is achieved by exploiting the information contained in audio and video to perform a kind of separation in each one of these modalities. First, as in the previous work [19], we use multimodal relevant events, that is, the audio and video events generated by the same physical phenomenon, in order to locate the video sources in the image. The **visual separation** goal, which is the first part of the process, is achieved using audiovisual synchrony. In this research work, the information relative to the speaker situation is employed to determine which are the visual structures that generate the soundtrack (those structures that are close to the detected speakers mouth) and reconstruct them separately. Then, the innovation introduced by the present model consists on the analysis of these multimodal relevant events in order to perform a **Blind Audio Source Separation**. Therefore, the algorithm exploits the video information to separate the audio signals corresponding to each source that are present in the mixture.

The proposed model achieves a satisfactory Blind *Audiovisual* Source Separation analyzing the multimodal information extracted from one microphone soundtrack and the video signal associated. Unlike most of the techniques for achieving *Blind Audio Source Separation*, we do not use arrays of microphones. Therefore, we can consider, concerning the audio analysis, that we are facing a Single Channel Source Separation, but employing also the video information.

As a result, taking into account the different parts of the **Audiovisual Separation** process and the limited literature on this multimodal separation field, the state-of-the-art is divided in three sections. In the first one, we show the evolution of the research works on *source localization* using audio and visual information. The second one contains the main motivations to perform the speaker detection task in *a different way*, that is, we focus our work in the audio and video representations and not in building complex pixel-wise models. This section also provides the summary of the main characteristics of the previous research work [19], and the differences between this approach and the model proposed in this report. Finally, an overview of the related work performed in the *Single Channel Source Separation* field is exposed.

Next, we describe the related work performed in these three fields, with a brief description of the problem, a general overview over the proposed algorithms, and, finally, the description of some representative approaches.

2.1 Source localization using multimodal signal analysis

The problem we confront here is that of detecting those audiovisual events generated by the same physical phenomenon by means of the information present in both modalities. Thus, a temporal analysis has to be performed in order to assess the synchrony between audio and video events, build these multimodal structures and localize the present sources in the image. However, the construction of these audiovisual pairs is a difficult task due to the different and complex nature of the signals to analyze.

In this section we will first do a fast chronological overview of the related work in the source localization field to later explain in detail four of the most representative approaches already mentioned.

Hershey and Movellan [13] were the first to propose a method to locate the sound sources in an image. This work was inspired by psychophysical and physiological evidence [3, 9] that the human spatial localization of acoustic signals is strongly influenced by their synchrony with visual stimuli (*ventriloquism effect*). In [13], this synchrony is defined as the degree of mutual information between the energy of a soundtrack and the intensity of single pixels, which introduces the hypothesis that pixels are independent conditioned on the speech signal. Assuming that the joint statistics of audio and video are Gaussian, the mutual information between both features (pixels intensity in video and energy evolution in audio) is computed by means of the Pearson correlation coefficient.

In [28], Slaney and Covell generalize this approach and look for a method that is able to measure the synchrony between audio signals and video facial images. Canonical Correlation Analysis, which is equivalent to maximum mutual information projection in the

jointly Gaussian case, is used to deduce the multimodal relationship between the cepstral coefficients of the audio signal and the value of single pixels for the video.

Nock et *al.* consider in [20] two mutual information approaches, one assumes discrete distributions and the other one multivariate Gaussian distributions, and one third algorithm that uses Hidden Markov Models trained on audiovisual data. The objective is evaluate the consistency of these three different approaches. Audio features are extracted from Mel-frequency cepstral coefficients, while different video features are tested: the coefficients of the discrete cosine transform (DCT) and the pixel intensity changes. All three methods utilize training corpus in order to build a priori models, like the methods proposed in [13, 28].

Recently, more general algorithms based on information theoretic features optimization have been introduced.

Butz and Thiran [7] propose an approach based on Markov chains modeling audio and video signals. The audiovisual consistency is assessed by maximizing the mutual information between audio and video features, where the distributions of such features are estimated using nonparametric density estimators. The video is represented by pixel intensity change and the audio feature is the linear combination of the power spectrum coefficients that brings biggest entropy. The framework developed in [7] is used in [4], to extract optimal audio features with respect to video features. These audiovisual features are then correlated by maximizing their mutual information, in order to locate the active speaker among several candidates.

A similar multi-modal fusion framework was proposed by Fisher et al. [15] and has been extended in their latest work [8]. The algorithm is based on a probabilistic generation model that is used to define projection rules on maximally informative subspaces. The learnt densities are used to define the relationship between different signal modalities using a nonparametric density estimator. This approach is used to solve a conversational audiovisual correspondence problem, obtaining encouraging results.

In [29], a slightly different approach is used to find, in a joint manner, an optimal modeling and fusion criteria of data. Principal Components Analysis and Independent Component Analysis are performed on audio and video features at the same time, in order to find the maximally independent audio-video subspaces, and thus extract audiovisual independent components. However, this technique is not able to deal with dynamic scenes.

2.1.1 Hershey and Movellan approach

The method introduced by Hershey and Movellan [13] in 1999 is based on the observation of the *ventriloquism effect*. This effect causes a mislocation of sounds toward their apparent visual source and depends strongly on the degree of synchrony between both signals [3, 9].

The *synchrony* is measured as the mutual information (MI) between the average acoustic energy and the intensity of single pixels.

Modeling audiovisual signal as a non-stationary Gaussian process, the MI between audio and video features ($a(t)$ and $v(x, y, t)$ respectively) can be written as follows:

$$I(A(t_k); V(x, y, t_k)) = \frac{1}{2} \log(1 - \rho^2(x, y, t_k)) \tag{2.1}$$

where $\rho(x, y, t_k)$ is the Pearson correlation coefficient between $A(t_k)$ and $V(x, y, t_k)$

This approach chooses, for each frame, the pixels with maximum MI, which correspond to different parts of the current speaker face as shown in Fig. 2.1. As a result, a centroid weighted by the estimated MI is computed in order to estimate the speaker position.

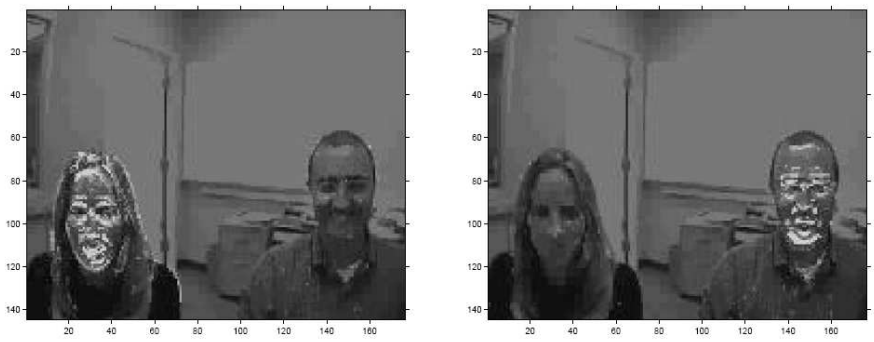


Figure 2.1: Estimated mutual information(MI) between pixel intensity and audio intensity (bright areas indicate greater MI) overlaid on stills from the video where one person is in mid-utterance, girl (left picture) and boy (right). These images are taken from [13].

To calculate the next frame centroid, a Gaussian *influence function* [11] is employed, that is, this function reduces the weight given to MI from pixels far from the current frame centroid. A threshold is introduced to reduce the effects of background noise, such as camera and microphone jitter. Thus, the speaker estimated position is computed with:

$$\hat{S}_x(t) = \frac{\sum_x \sum_y x\ \theta(log(1 - \hat{\rho}^2(x, y, t)))\psi(x, \hat{S}_x(t-1))}{\sum_x \sum_y \theta(log(1 - \hat{\rho}^2(x, y, t)))\psi(x, \hat{S}_x(t-1))} \tag{2.2}$$

where $\hat{S}_x(t)$ represents the estimate of the x coordinate for the position of the speaker at time t. $\theta(.)$ is the thresholding function, and $\psi(x, \hat{S}_x(t-1))$ the influence function. Finally, $\hat{\rho}^2(x, y, t)$ is the estimate of the correlation when using the past 16 video frames.

The main limitation of this method concerning the video representation is that changes in fundamental frequency are not captured if they do not affect the average energy.

2.1.2 Nock approach

Nock et *al.* evaluated in [20] (2003) the performances of three different algorithms for assessing face and speech consistency using audiovisual synchrony:

Gaussian MI Assumes continuous multivariate Gaussian distributions.

Discrete MI Assumes discrete distributions.

Audio-visual Likelihood ("AV-LL") Uses Hidden Markov Models (HMMs) trained on joint sequences of audiovisual data. Unlike the other two *generic* algorithms, this one is an *specific* measure and requires that audio correspond to speech and images contain faces.

Like [13,28] methods, all three algorithms require training datasets to build *a priori* models. Existence of good face and speech detection is assumed. Audio features are Mel-frequency cepstral coefficients (MFCCs). Concerning video, a normalized mouth region-of-interest (ROI) is extracted of each frame and compressed using a discrete cosine transform(DCT).

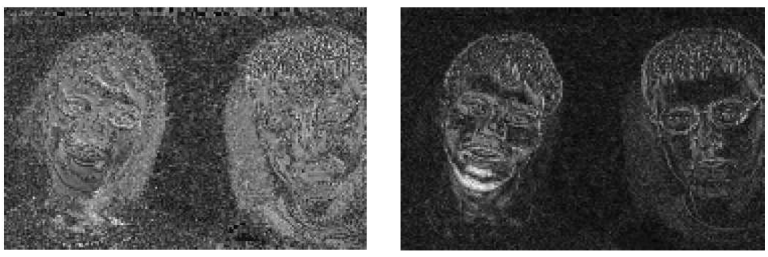
This approach concludes that Gaussian MI obtains significantly better results than the other two algorithms in identifying the active speaker, but not in determining the degree of synchrony (can not distinguish between voiceovers and monologues).

Nock et *al.* also compare also the performances of Gaussian MI if the video feature is related to pixel intensities or pixel intensity *changes*, the second one defined as:

$$F^{IC}(x,y,t) = \sum_{l,m=-1}^{1} F(x+l, y+m, t+1) - F(x+l, y+m, t-1) \tag{2.3}$$

where $F(x,y,t)$ is the original image.

Results evaluated in one sequence of CUAVE database are shown in Fig. 2.2. As desired, mutual information in 2.2(b) is highest around speaker's mouth and jawline.



(a) (b)

Figure 2.2: Mutual Information Images: (a) Pixel Intensities (b) Pixel Intensity Changes.

Finally, under the assumptions of known number of speakers in known regions without background motion, pixel-wise Gaussian MI performance is close to two video only techniques: *Intensity Change Image Sums* that compare the total pixel intensity changes on the left and right halves, and *Intensity Change Image X-Projection Peak* that chooses the column with maximum sum of intensity changes.

2.1.3 Fisher and Darrell approach

Fisher and Darrell proposed in [8] a general information-theoretic approach to identify cross-modal correspondences without any assumption about the content of the audio and video signals. It presents a probabilistic model for audiovisual signal generation, and demonstrate that correspondences in both modalities can be discovered by applying lineal projections that maximize the mutual information of the derived measurements.

This technique uses the **probabilistic models** of audiovisual fusion described in Fig. 2.3. B represents the joint source (audio and video), while A and C represent the background interferences in each single modality. $\{X^a, X^v\}$ are the audio and video observations respectively.

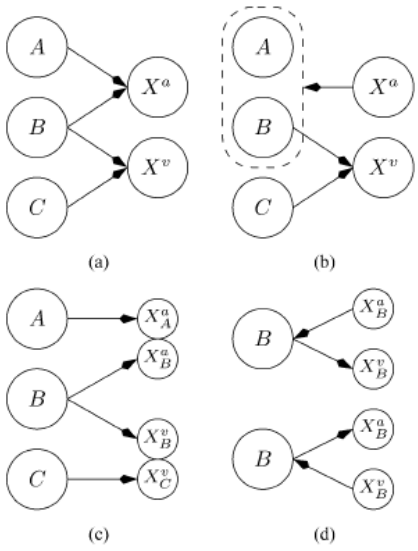


Figure 2.3: Graphs illustrating the various statistical models exploited by the algorithm: (a) the independent cause model– X^a and X^v are independent of each other conditioned on $\{A, B, C\}$; (b) information about X^a contained in X^v is conveyed through *joint* statistics of A and B, (c) the graph implied by the existence of a separating function, and (d) two equivalent Markov chains which can be extracted from the graphs *if* the separating functions can be found.

Supposing that decompositions of the measurement X^a and X^b *exist* such as the model can be represented as in Fig. 2.3(c), then there is no influence due to A or C, the interferences. Thus,

$$I(X_B^a; I(X_B^v)) \leq I(X_B^a; B) I(X_B^a; I(X_B^v)) \leq I(X_B^v; B) \tag{2.4}$$

and, processing this inequality, the following conclusion can be extracted: maximizing separate projections of the audio-video measurements with high mutual information we are also maximizing a lower bound on $I(X_B^a; I(X_B^v))$.

After this demonstration, **projections** that maximize the mutual information of the derived measurements have to be estimated, and can be parameterized as:

$$y_t^v = h_v^T x_t^v \tag{2.5}$$

$$y_t^a = h_a^T x_t^a \tag{2.6}$$

where x_t^v and x_t^a are lexicographic samples of images and periodograms, respectively, of an A/V sequence. The linear projections h_v^T and h_a^T map A/V samples to low dimensional features y_t^v and y_t^a.

Computing h_v can be divided in three stages:

1. Prewhiten the images using the inverse of the average spectrum **once** followed by iterations of
2. Updating the feature values y_t^v
3. Solving for the projection coefficients using least squares and an L_2 penalty in order to introduce a capacity control mechanism.

The projection coefficients related to audio h_a are solved in the same form but without the prewhitening step.

Results are shown in Fig. 2.4. This approach can detect which speaker is active when several are facing a device and distracting motion is present, and without making any assumptions about acoustic or visual models.

2.1.4 Smaragdis approach

Smaragdis and Casey proposed in [29] (2003) a statistical method that operates in a fused dataset, without distinguishing between the auditory and visual data. Thus, audiovisual features corresponding to relevant events are extracted jointly from the stream, while other approaches as those presented in [4, 7, 8, 13, 20, 28] only correlate the auditory with visual cues.

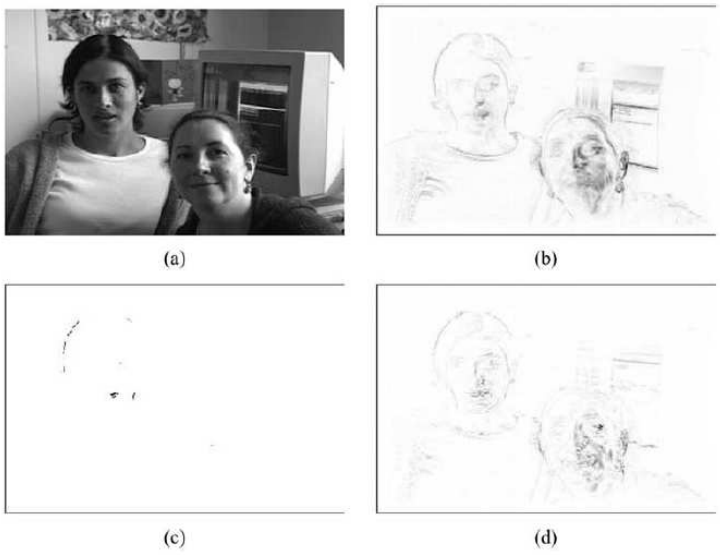


Figure 2.4: Video sequence containing one speaker (person on left) and one person who is randomly moving their mouth/head (but not speaking): (a) one image from the sequence, (b) pixel-wise image of standard deviations taken over the entire sequence, (c) image of the learned projection, h_v, and (d) image of h_v for incorrect audio.

The soundtrack is processed by a short term Fourier transform, obtaining a time-frequency representation $\mathbf{f}(t)$, and the video frames are reshaped as vectors $\mathbf{m}(t)$. Audio and visual streams are treated as one set of data and combined into one vector:

$$x(t) = \begin{bmatrix} \alpha \cdot \mathbf{f}(t) \\ \beta \cdot \mathbf{m}(t) \end{bmatrix} \tag{2.7}$$

where the two scalars α and β are used for variance equalization (results influenced more by video or audio components depending on their values).

The procedure is divided in two steps:

Dimensionality reduction Is performed by principal components analysis (PCA) over $\mathbf{x}(t)$ with zero mean, a linear transform $\mathbf{W}_o$ that projects the input to make it orthonormal:

$$\mathbf{x}_o(t) = \mathbf{W}_o \cdot \mathbf{x}(t), \tag{2.8}$$

so that $E\{\mathbf{x}_o \cdot \mathbf{x}_o^T\} = \mathbf{I}$. Then, the algorithm keeps the first few dimensions with maximal variance, so that $\mathbf{x}_r(t) = \mathbf{x}_o^{1\ldots m}(t)$ and $\mathbf{W}_r(t) = \mathbf{W}_o^{1\ldots m}(t)$.

Independence transform This step employs independent component Analysis (ICA) [14], which ensures that the the output will be maximally statistically independent. The linear transform is represented as

$$\mathbf{x}_i(t) = \mathbf{W}_i \cdot \mathbf{r}(t), \tag{2.9}$$

where $\mathbf{W}_i$ is estimated using a natural gradient algorithm.

The two steps can be described by the linear transformation $\mathbf{W} = \mathbf{W}_i \cdot \mathbf{W}_r$. And, to get a better idea of what the component bases $\mathbf{W}$ mean, we can rewrite the equation to show the audio and video parts:

$$\mathbf{x}_i(t) = \mathbf{W} \cdot \mathbf{x}(t) \Rightarrow \begin{bmatrix} \mathbf{f}_i(t) \\ \mathbf{m}_i(t) \end{bmatrix} = [\mathbf{W}_a, \mathbf{W}_v] \cdot \begin{bmatrix} \mathbf{f}(t) \\ \mathbf{m}(t) \end{bmatrix} \tag{2.10}$$

We can visualize the results of a simple video example in Fig. 2.5. For the audio, the rows of $\mathbf{W}_a$ representing spectral profiles are plotted, and, to visualize de video component bases, each row of $\mathbf{W}_v$ reshaped to the size of input frames is represented. The component weights $\mathbf{x}_i(t)$ indicate the presence of each audiovisual component through time.

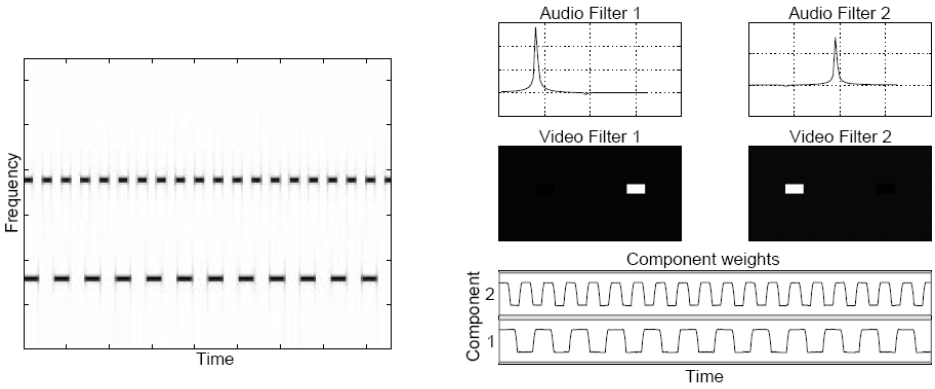


Figure 2.5: Simple video example. The left plot is a spectrogram of the soundtrack, which consists of two periodically gated sine waves. The audio segment of the component bases $\mathbf{W}_a$ is shown at the top right plots, and video segment $\mathbf{W}_v$ at the middle right. The component weights $\mathbf{x}_i(t)$ are shown on the bottom right.

This approach has good results in static scenes, but it has problems with dynamic ones. An object moving through the image cannot be tracked by only one component and, as a result, it will be distributed among many visual bases.

2.2 Why we perform the source separation in a different way?

As we have explained before, this research work is based on the previous one performed by Monaci, Divorra and Vandergheynst [19]. This work was also placed in the field of source localization using multimodal signal analysis, but now with a different sight.

In contrast to the previous works in this field, in [19] the attention is focused on modeling the audiovisual sequence. This is because of the fact that features employed in the other approaches (pixels) are poorly connected with the physics of the problem, so that a huge and barely meaningful amount of data has to be analized. In this algorithm, the scene content is described concisely, with a video decomposition into geometric 2-D features evolving from frame to frame. These features describe the visual content of the scene: when they experiment a movement what is moving is one structure representing a part of a real object, while a change in the intensity of one pixel have no global visual sense alone.

Such a representation allows the design of a intuitive audiovisual fusion criteria that do not require the formulation of any complex statistical model describing the relationships between both modalities. Thus, unlike the previous works in this field, the computational cost is not very high as the dimensions of the problem have been significantly reduced.

In this research work, we use the same video features than [19], but now we exploit the time-frequency characteristics of the audio sequence instead of using only the soundtrack energy. This change affects both the audio representation and the fusion criteria applied posteriorly, as we are adding information and dimensions in order to achieve more ambitious objectives. Before, there was only one feature for the audio, its energy. Now, the energy distribution of the audio in the time-frequency plane is captured by several audio features (atoms). Therefore, the new fusion criteria has more dimensions to analyze, since it evaluate the synchrony between each video and audio feature.

In the following subsection we scarcely explain the research work of Monaci, Divorra and Vandergheynst [19] as it is on this algorithm that the proposed model is based. Moreover, the importance of this approach consists in confronting the source localization problem in a innovative way, and with successful results.

2.2.1 Monaci, Divorra and Vandergheynst Approach

In [19], Monaci, Divorra and Vandergheynst model audiovisual data in order to reduce dimensionality and using only relevant signal information for the speaker localization objective.

For the **video representation**, this approach employs the MP algorithm proposed by Divorra and Vandergheynst in [10]. The video sequence is decomposed into a set of 2-D atoms evolving in time, tracking, thus, temporal transformations of salient geometric video components, that is, transformations that represent changes in the scene. The procedure is the following:

1. First frame of the sequence is decomposed over a redundant dictionary of 2-D atoms

$$I = \sum_{\gamma_i \in \ \Omega} c_{\gamma_i} g_{\gamma_i}, \tag{2.11}$$

where c_γ corresponds to the projection coefficient for every atom g_γ and Ω is the subset of selected atom indexes from the dictionary.

2. For each frame we decompose it as for the first one, and the algorithm tries to track the video atoms in the image through time, that is, it tries to find in the new frame the video atoms presents in the last one. Thus, changes can be modeled as $I_{t+1} = F_t(I_t)$, where F_t represents the set of transformations experimented by the video atoms from frame to frame.

Each video atom is characterized by 6 parameters, but this approach uses only the x and y positions as a measure of the video atoms displacement, which means movement in the scene.

Concerning the **audio representation**, the MP algorithm for 1-D signals [18] is used. The audio signal $a(t)$ is decomposed over a dictionary of Gabor atoms D_A. In each iteration, the MP algorithm chooses the atom in the dictionary $g_{\gamma i}$ in D_A that maximizes the projection $|\langle a, g_{\gamma i}\rangle|$ in order to minimize the residual part after the approximation of $a(t)$ in the subspace described by $g_{\gamma i}$. Thus, after N iterations

$$a = \sum_{n=0}^{N-1} < R^N a, g_{\gamma n} > g_{\gamma n} + R^N a, \tag{2.12}$$

where $R^0 = a$ and $R^N a$ is the residual part after n iterations. The audio feature analyzed in the fusion criteria is the average of the energy present at each time instant after perform the MP decomposition.

The **fusion criteria** is shown in Fig. 2.6. The algorithm looks for audio energy peaks temporally correlated to video displacement peaks in order to build audiovisual meaningful structures. Scalar product is employed to measure the correlation between features in both modalities: one audio feature representing the energy of the soundtrack along time, and *several* video features showing the evolution of each video atom from frame to frame. The algorithm select the video atoms with highest *Synchronization Score* (output value of the scalar product).

Results involving two active speakers are shown in Fig. 2.7. This approach is able to locate the mouth of the active speaker in an audiovisual sequence, and the accuracy results are better than the obtained in the previous research works.

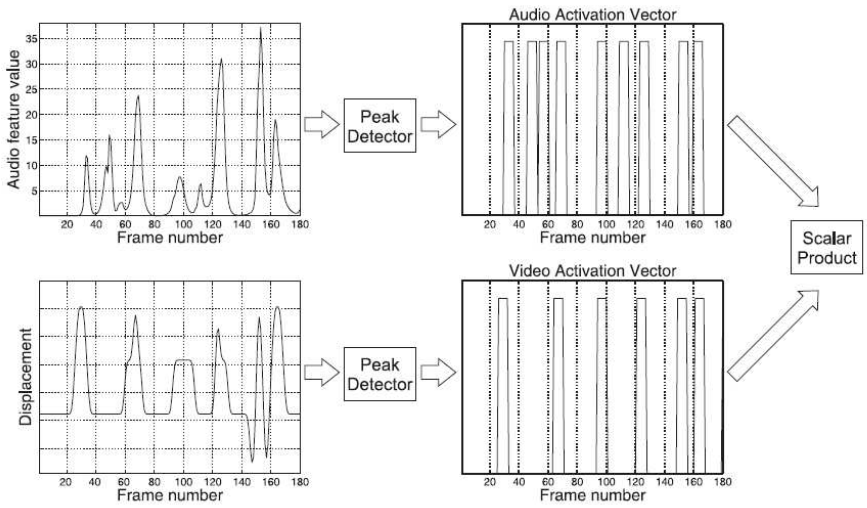


Figure 2.6: Scheme of the audiovisual fusion criterion proposed by Monaci, Divorra and Vandergheynst. The audio energy peaks and the displacement peaks for each video atom are extracted and activation vectors are build. The Synchronization Scores between the audio activation vector and the video activation vectors are computed as the scalar product between those signals.

Figure 2.7: Four frames taken from a clip with two speakers in front of the camera taking turn in reading digit string. In the first two frames the right person is speaking, while in the last two the left one is speaking. The footprint of the most correlated atom is highlighted in white. The mouth of the correct speaker is detected.

2.3 Blind Audio Source Separation

The most difficult step of this research work is the Blind Audio Source Separation (BASS). The problem of the BASS consists in separating the audio signals present in a given mixture. Existing solutions to this problem generally require microphone arrays and prior knowledge of the sources, but that is not our case. We want to perform the BASS only with one microphone and the video signal associated.

There are not a lot of studies in the field of BASS using audiovisual information, and therefore, we focus on the related work performed in the case of BASS with only one microphone. The temporal analysis of the proposed model determines the time instants

with only one source is active. Thus, when both of them are active at the same time we can consider we are facing a Single-Channel Source Separation with the prior knowledge achieved in the temporal periods with only one active speaker.

This is a new, hard an still open problem, first time faced by Roweis in [26] (2000). When only the input signal of one microphone is available, simple generic assumptions do not suffice. For the Single-Channel Source Separation task is necessary to model different characteristics of the speech signal, such as the spectral envelope, the fundamental frequency or the temporal continuity. These known cues for the speech separation [5, 6] have to be taken into account in order to build models that face this problem.

The main limitation of the Single-Channel Source Separation methods is the necessity of noticeable differences between the sources spectrum. Their energy must be distributed in a different way in the time-frequency plane, as it is in this domain that this methods operate. This is the reason why in this field, and in the BASS field in general, the performances of the developed approaches are tested with mixtures of one boy and one girl speech.

The existent research works relative to the Single-Channel Source Separation can be divided into two main groups according to their blindness:

Generative: Approaches in this group build their models according to the present speakers in the mixture, that is, for each mixed speaker the algorithm is trained in sequences where only he or she is speaking. Thus, these works are situated in a *non blind* context. First approaches in this field belong to this group [16, 24, 26, 27]. The problem is that if the model is too simple, it does not separate, while, on the other hand, if the algorithm is too complex we are facing an intractable inference of huge dimensionality.

Discriminative: These approaches focus in the spectral separation task instead of building complex models for each speaker. They try to exploit the sparsity of speech signals in the time-frequency domain, and do not have any prior knowledge about the present speakers in the mixture. Examples of algorithms in this *blind* group are [2,25].

Next, we describe the *chronological* evolution of the research works in the Single-Channel Source Separation field.

The first approach to face the problem was proposed by Roweis in [26]. This technique use a factorial Hidden Markov Model (FHMM) trained in sequences where the speakers present in the mixture are recorded alone. Through HMM, *binary mask functions* are computed for each frequency sub-band and applied to the mixture in order to extract the original signal of each speaker. In [27], Roweis introduce to his previous model a factorial-max vector quantization (MAXVQ), which combines the outputs from the various causes and adds non-negative noise.

Jang and Lee proposed in [16] a new technique that utilizes the time-domain ICA basis functions previously learned from a training database which consists in sequences where the speakers in the mixture speak alone. This method recovers original signals through gradient-ascendent adaptation steps to found the maximum likelihood estimate.

In [24], Reyes-Gomez reduce the dimensions of the problem raised in [26] by dividing the spectral representation of the source signals into *multiple sub-bands*, that is, multiple parallel horizontal sections of the speactrogram. Then, this approach computes a separate HMM to model each band, requiring many few states per model and, for comparable computation expense, can achieve more accurate signal separations than *full-band* models. This model also presents an interesting basis for learning source models directly from mixed signals, since there are more opportunities to find a time-frequency space with the energy of only one speaker. This observed characteristic is exploited by the same authors in [25]. This approach captures local deformations of the time-frequency energy distribution and describes each frame with a lineal transformation applied to its predecessor. The spectrum is analyzed as the addition of harmonics and formant structure and no prior models about the present speakers are necessary.

A different approach is proposed by Bach and Jordan in [2]. This algorithm builds *affinity* matrices combining classical cues for from speech psychophysics [5, 6]. These matrices are employed to define a spectral segmenter, which, applied to the mixture, performs the speech separation with one signal channel and without prior knowledge about the speakers.

A selection of the most representative approaches in the Single-Channel Source Separation field are described in the next subsections.

2.3.1 Roweis Approach

The method introduced by Roweis in [26] (2000) has an important relevance as it was the first work in Single-Channel Source Separation. The proposed algorithm employs a Factorial Hidden Markov Model (FHMM) system which is trained in one speaker sequences to later separate mixtures of known speakers using only one audio signal. That is achieved by computing binary masking functions of frequency sub-bands through HMM and then *refiltering* the mixed signal.

The concept of **refiltering** is introduced. *Unmixing* algorithms try to reweight output signals from the microphones to estimate the original source, using coefficients constant over time for the weights. On the other hand, *refiltering* techniques reweight multiband signals with *masking signals*, that is, the coefficients are no longer constant over time. Thus, the estimated source can be represented as

$$s(t) = \sum_{i=1}^{K} \alpha_i(t) b_i(t) \tag{2.13}$$

where $b_i(t)$ are the multiband signals and $\alpha_i(t)$ the masking signals for each sub-band. This *refiltering* concept is illustrated in Fig. 2.8.

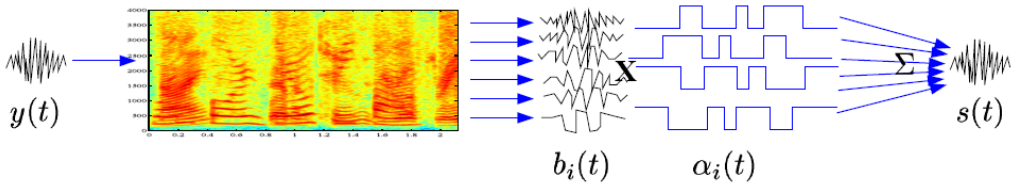


Figure 2.8: The refiltering approach to one microphone source separation. Multiband analysis of the original signal $y(t)$ gives sub-bands signals $b_i(t)$ which are modulated by masking signals $\alpha_i(t)$ (binary or real valued between 0 and 1) and recombined to give the estimated source or object $s(t)$

This unsupervised model first trains speaker dependent hidden Markov models (HMM) on sequences with only one talker, and fits, thus, a simple HMM for each present speaker in the mixture. Then, these models are combined into the factorial hidden Markov model (FHMM) shown in Fig. 2.9 in order to perform the separation task. A FHMM consists on two or more Markov chains (this approach uses only two) that evolve independently. At each time,each chain proposes an output vector, $\mathbf{a}_{x_t}$ and $\mathbf{b}_{z_t}$, and the algorithm choses the *elementwise maximum* of the proposals.

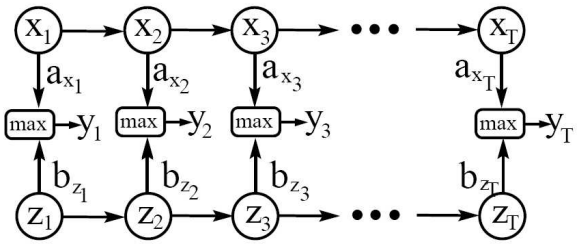


Figure 2.9: Factorial HMM with max output semantics. Two Markov chains x_t and z_t evolve independently. Observations $\mathbf{y}_t$ are the elementwise max of the individual emission vectors $\max[\mathbf{a}_{x_t}, \mathbf{b}_{z_t}]$ plus Gaussian noise.

Therefore, the binary **masking signal** is computed as

$$\alpha_t(i) = 1 \quad if \quad \mathbf{a}_{\hat{x}_t}(i) > \mathbf{b}_{\hat{z}_t}(i) \quad and \quad 0 \quad if \quad \mathbf{a}_{\hat{x}_t}(i) \leq \mathbf{b}_{\hat{z}_t}(i) \tag{2.14}$$

This masking signal is later applied to the mixture to obtain the separated signals in the *refiltering* step.

2.3.2 Bach and Jordan Approach

Bach and Jordan faced in [2] (2004) the problem of the single-channel speech separation by segmenting the spectrogram of the signal into two or more disjoint sets, depending on the number of speakers in the mixture, and without modeling individual speakers as done in the previous works in this field [16, 26]. The proposed algorithm combines classical cues from speech psychophysics [5, 6] into parameterized affinity matrices, which are employed to define a spectral segmenter. To adjust the parameters of the affinity matrices a learning algorithm trained with any speaker, in other words, it is not necessary to train the algorithm with the present speakers in the mixture, and this can be considered a really *blind* separation algorithm.

This approach takes into account the classical cues for speech separation [5,6] in order to build the parameterized affinity matrices. These cues are associated to different time scales depending on their duration in frames: *small*, *medium* and *large*. The speech cues are divided into two groups

Non-harmonic cues Similar to vision cues.

- *Continuity* If two time-frequency points are close in one of both modalities, they are likely to belong to the same segment. Features: time and frequency. Time scale: small.
- *Common fate cues* Elements that exhibit the same variation (both in time or in frequency) are likely to belong to the same source. The features are captured using oriented filters. Time scale: medium.

Harmonic cues They act at all time scales: small, medium and large.

- *Pitch estimation* Several pitches are estimated
- *Timbre* Spectral envelope of the signal. Feature: reduced dimensionality of spectral envelope using principal component analysis.
- *Building feature maps from pitch information* A set of features is built from the pitch information.

From each one of this features, a basis affinity matrix is defined as $W_j(\beta_j)$, where β_j is a parameter. For *non-harmonic features*

$$W_{ab} = exp(-\|f_a - f_b\|^{\beta}) \tag{2.15}$$

where f_a is the value of the feature for data point a and $\beta > 1$. On the other hand, for *harmonic features* the proposed method takes into account the strength of the feature

$$W_{ab} = exp(-|g(y_a, y_b) + \beta_3|^{\beta_4} \|f_a - f_b\|^{\beta_2}) \tag{2.16}$$

where y_a is the strength of the feature in the data point a and the function $g(u,v) = (ue^{\beta_5 u} + ve^{\beta_5 v})/(e^{\beta_5 u} + e^{\beta_5 v})$ ranges from the minimum of u and v for $\beta_5 = -\infty$ to their maximum for $\beta_5 = +\infty$.

The **spectral clustering** algorithm used to segment the spectrogram of the mixed signal is composed of the following steps

1. Build *affinity/similarity* matrix $W \in \Re^{P \times P}$. Given m basis matrices, the following parametrization of W is used: $W = \sum_{k=1}^{K} \gamma_k W_1^{\alpha_{k1}} \times ... \times W_m^{\alpha_{km}}$. A sum of $K = 3$ matrices is used in this algorithm, one matrix for each time scale.

2. Normalize the affinity matrix: $\tilde{W} = D^{-1/2} W D^{-1/2}$ where D is diagonal with sums of rows of W.

3. Compute the R largest eigenvectors $U(W) \in \Re^{P \times R}$ of $\tilde{W}$

4. Considering $U(W)$ as P points in $\Re^R$, cluster U using weighted K-means.

Due to the huge dimensionality of the affinity matrices W_i, some approximations depending on the time scale are done.

Results obtained by analyzing a mixture with this method are shown in Fig. 2.10. The spectral segmenter has been trained with different speakers of those that contribute to the mixture.

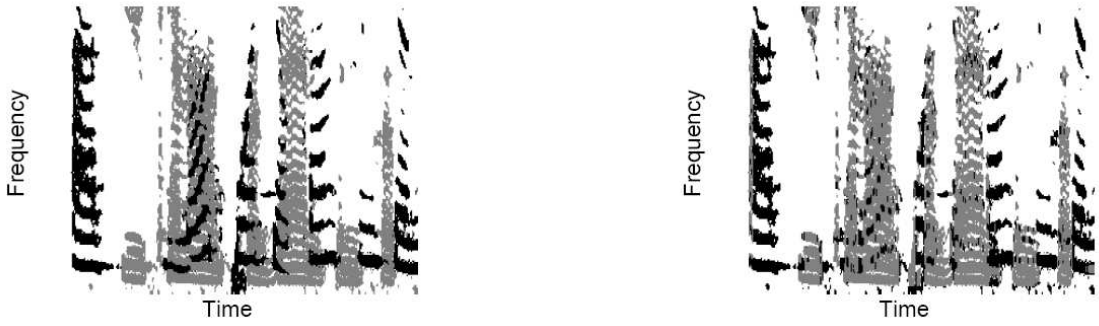


Figure 2.10: (Left) Optimal segmentation for the spectrogram in Figure 1 (right), where the two speakers are black and grey; this segmentation is obtained from the known separated signals. (Right) The blind segmentation obtained with Bach and Jordan algorithm.

The proposed approach obtains audible signals of reasonable quality, and the only requirement is that the speakers in the mixture have distinct and far enough pitches most of the time. The main problem is the long duration of the process and the high memory requirements.

2.3.3 Reyes-Gomez, Jojic and Ellis Approach

Reyes-Gomez, Jojic and Ellis introduced in [25] (2004) a new method to perform single-channel source separation based on capturing the detailed dynamic of the speech describing each spectral frame as a deformation of its predecessor, and without any pre-trained speech or speaker models. The spectrum of the audio sequence is decomposed into two additive layers, which describe separately the harmonics and formant structure. This approach have very few parameters, only a limited set of initial states to cover the full spectra variety through transformations, and these parameters can be learned from mixed data without supervision.

The **spectral deformation model** captures the variations experimented by the speech harmonics in the time-frequency plane. It can be represented as

$$\mathbf{X}_t^{[k-n_c,k+n_c]} \approx \mathbf{T}_t^k \cdot \mathbf{X}_{t-1}^{[k-n_p,k+n_p]} \tag{2.17}$$

where $\mathbf{T}_t^k$ is the *transformation matrix*, and N_C and N_P are the number of frequency bins of,respectively, frames t and $t-1$ centered at the k^{th} frequency bin, with $N_C < N_P$ to permit both upward and downward motions. Fig. 2.11 illustrates an example of this procedure.

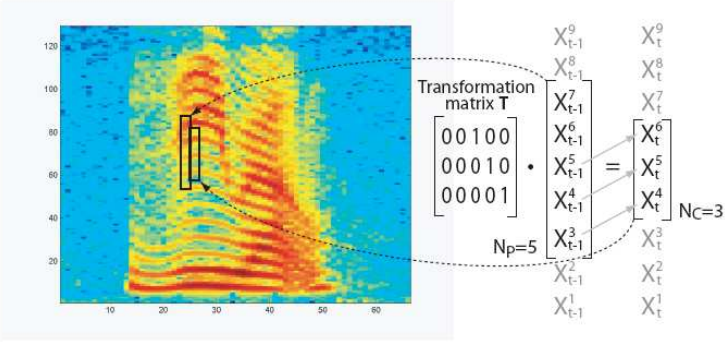


Figure 2.11: The $N_C = 3$ patch of time-frequency bins outlined in the spectrogram can be seen as an upward version of the marked $N_P = 5$ patch in the previous frame. This relationship can be described using the matrix shown.

The inference consists in finding probabilities for each transformation index at each time-frequency bin, and this is approximated using Loopy Belief Propagation [32, 33].

The **two layer source-filter transformations** separately model the deformation fields for harmonics and formant resonance components. Thus, the mixed spectra X can be modeled as the sum of variables F and H, harmonics and formants decomposition, as shown in Fig. 2.12.

A **Matching-tracking model** is also implemented. A small set of *initial states* is introduced to model translations between silence and speech or between speakers. Thus,

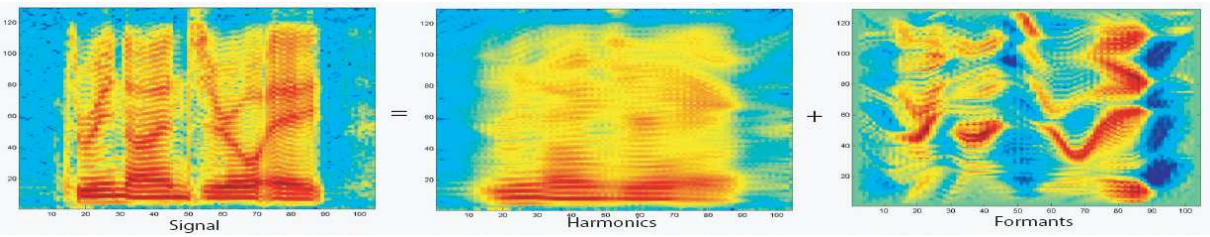


Figure 2.12: Harmonics/Formants decomposition.

in this discontinuities the algorithm keeps on tracking the "old" speaker at the same time as estimating the initial state of the "new" speaker.

This approach has an interactive model that implements a lot of applications: formant and harmonics tracking, missing data interpolation, formant/ harmonics decomposition, and semi-supervised source separation of two speakers. The dynamics of the speech are captured with only a few parameters.

Chapter 3

Audio and Video Representations

3.1 Introduction

The retrieval of correlation between audio and video signals is a problem with a very high dimensionality. The goal is that of locating those spatio-temporal video regions that are interrelated with a certain audio track. In order to make this problem feasible, audiovisual data need to be modeled such that dimensionality gets reduced and only relevant signal information is used. Data modeling is thus supposed to capture the main characteristics of each signal modality that may contain information about the other modality. However, existing approaches to multimodal processing typically focus on the modeling of relationships between audio and video data, rather than on modeling the data itself.

To date, methods dealing with audiovisual fusion problems basically attempt to build complete, general and complex statistical models to capture the relationships between audio and video features. But surprisingly, the employed features are extremely simple and poorly connected with the physics of the problem, in particular for what concerns visual information. Efficient signal modeling and representation require the use of methods able to capture particular characteristics of each signal kind. A question that arises at this point is: Why should we use a representation of video based on a basis of *deltas* (*i.e.* pixel wise features), if video is made of moving regions surrounded by contours with high geometrical content? Pixel-related quantities seem to us a relatively poor source of information that has a huge dimensionality, it is quite sensitive to noise and does not exploit structures in images. A very simple example can clarify this concept. If a person is moving back and forth while speaking in front of a camera, the pixel values on the mouth region change depending on the lips movements *and* on the person movement. The result is that pixel intensities evolve in an undistinguishable way.

Therefore, the idea is basically that of defining a proper model for visual signals, instead of defining a complex statistical fusion model that has, however, to find

correspondences between barely meaningful features. If an accurate description of the scene is available, we can actually think of detecting consistent audiovisual pairs generated by the same phenomenon (in this case, a speaker uttering a sound), by simply observing the co-occurrence of interesting audio and video events (*i.e.* the presence of sound and the movement of the mouth). For particular applications, one may consider the use of adapted template based approaches for video representation (in order to model particular objects and their trajectories: Lips, faces, etc...). However, for generic non-application constrained approaches, the answer seems to be that we should, indeed, use a signal model capable of exploiting video structural properties while keeping generic and flexible enough.

As we have already discussed, the proposed model is based on temporal synchrony between audio and video features, looking for *relevant events* temporally close in both modalities. This chapter explains the audio and video representations that extract these meaningful events necessary for the later analysis.

In this research work, both audio and video signals are represented using the Matching Pursuit (MP) algorithm [18], which decomposes them into a linear expansion of waveforms chosen from a redundant dictionary of functions called *atoms*.

The decompositions used in this research work are described in the next sections. MP algorithm for 1-D signals is used to represent the audio track. Concerning the video, the complexity of the analysis causes the utilization of more complex techniques. MP algorithm proposed by Divorra and Vandergheynst [10] decomposes the image into a set of video atoms which represent salient video components and it tracks their temporal transformation from frame to frame.

In this chapter, the features used for audio and video signals after their decomposition are also presented. Those features determine the later analysis our method is capable to perform. Thus, we detail which parameters of the decomposition in each modality are more important for our algorithm.

Finally, the reasons of this choice concerning both audio and video representation are detailed. The main reason is that these features are strongly related to the physics of the problem. Video features indicate the movement of image structures (composed by atoms) through time. Concerning the audio, the MP decomposition describes the energy distribution of the signal in the time-frequency plane.

3.2 Audio representation

Audio signals have a rich variety of components that human auditive system is able to perceive (Fig. 3.1). Correlations of the wide diversity of sounds with the also large variety of geometric configurations of the visual stimulus of a mouth are possible. Indeed, this is

the main basis for *lip reading*. A positional model of lips may be assigned to each sound and transitional models between sounds can be established.

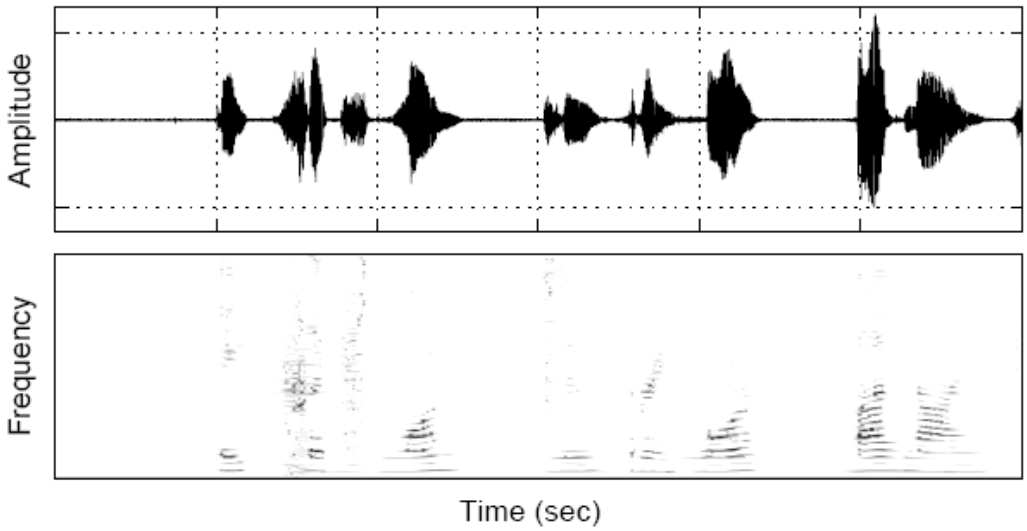


Figure 3.1: Audio signal of a subject uttering eight digits in English (top), its time-frequency energy distribution $E_a(t, \omega)$ (bottom). The color map of the time-frequency plane image goes from white to black, and the darkness of a pixel represents the value of the energy at each time-frequency location.

We consider here a much simpler and generic approach. As already stated, we look for synchrony between audio-video events. However, in contrast to the previous work [19], we need more characteristics of the audio signal than only the temporal situation of a sound. The proposed model tries to exploit all the information present in the audio signal spectrum, that is, the situation of each relevant structure in the time-frequency plane and its energy coefficient or importance in the whole sequence decomposition.

Typical features used to represent audio signals are the Mel-frequency cepstral coefficients (MFCC) [23], mainly used in the speech recognition field, and employed in [4,20,28]. In [7] the audio feature is obtained from the spectrogram of the audio track as the linear combination of the power spectrum coefficients exhibiting the biggest entropy. Fisher and Darrell [8] propose a similar feature that maximizes the mutual information with the video.

Monaci, Divorra and Vandergheynst [19] give a different approach estimating the audio energy contained par frame. To compute this estimation the MP decomposition [18] of the audio signal over redundant dictionaries is used. The estimated energy present at each time instant is the addition of the energy of all the 2-D Gaussian functions (atoms) in the MP decomposition. The sparse decomposition of the audio track performs a denoising of the signal, pointing out its most relevant structures.

In all cases, the final feature is a 1-D function that is downsampled in order to obtain the same length for audio and video features.

In this research work, we exploit widely the properties the audio signal representation over redundant dictionaries using also the MP algorithm. This decomposition in the time-frequency plane gives us a complete description of the spectrum of the audio signal. As well as the distribution of the energy along time, the information concerning the frequency components of the signal is also included.

3.2.1 Signal decomposition

The **Matching Pursuit** algorithm [18] is used to decompose the audio signal *a(t)* into a linear expansion of waveforms, that are chosen from a redundant dictionary D_A of functions in order to best match the signal structures. The unit norm functions in D_A are called time-frequency atoms, as is in this plane where they are situated.

Therefore, selecting an appropriate countable subset of atoms, any function $a(t)$ can be represented as

$$a(t) = \sum_{n=0}^{+\infty} c_n g_{\gamma n}(t), \tag{3.1}$$

where the expansion coefficients c_n give us information about the function $a(t)$.

A single window function, $g(t) \in L^2(R)$, generates all the atoms that compose the dictionary D_A. Each atom $g_\gamma = U_\gamma g$, is built by applying a geometrical transformation U_γ to the mother function g. The possible transformations applied to the function are the following: scaling by $s > 0$, translating in time by u and modulating in frequency by ξ.

Then, indicating with an index γ the set of transformations (s, u, ξ), an atom can be represented as

$$g_\gamma(t) = \frac{1}{\sqrt{s}}\, g\left(\frac{t-u}{s}\right) e^{i\xi t}, \tag{3.2}$$

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

In this research work, a dictionary of Gabor functions is used. As a result, the generating function $g(t)$ is the normalized Gaussian window showed in figure 3.2, with different possible values for the window size. The optimal time-frequency situation of the function is the cause of its election [12].

To represent the audio signal with a set of waveforms selected from D_A, MP algorithm makes successive approximations of *a(t)* with orthogonal projections on elements of the dictionary.

Thus, in the first step, MP algorithm decomposes the audio signal as

$$a =< a, g_{\gamma 0} > g_{\gamma 0} + R^1 a, \tag{3.3}$$

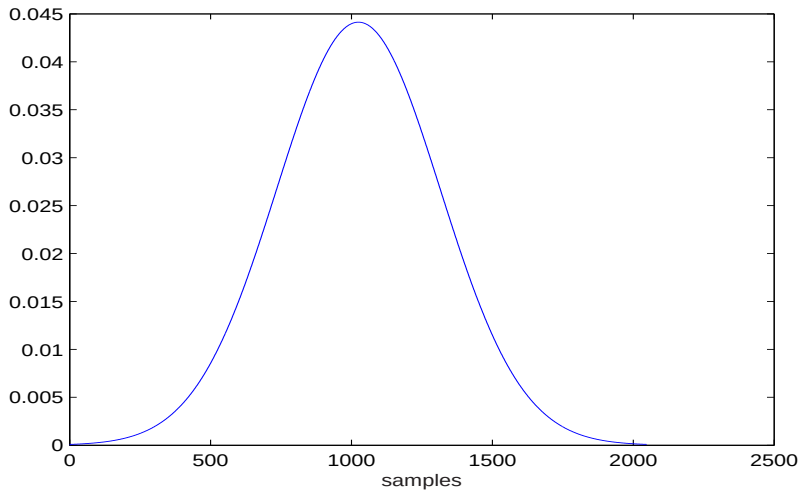


Figure 3.2: Gaussian distribution of the energy of one audio atom. In this case, the window size is 2024 samples.

where R^1a is the residual part after approximating a in the subspace described by the atom $g_{\gamma 0}$. This atom is orthogonal to the residual part R^1a and as a result the squared norm of the audio signal is expressed as

$$\|a\|^2 = | < a, g_{\gamma 0} > |^2 + \|R^1 a\|^2. \tag{3.4}$$

Therefore, the algorithm chooses $g_{\gamma 0} \in D_A$ such that the projection $| < a, g_{\gamma 0} > |$ is maximal, in order to minimize the residual component $\|R^1 a\|$.

Applying recursively the same procedure N times, the signal a is then represented as

$$a = \sum_{n=0}^{N-1} < R^N a, g_{\gamma n} > g_{\gamma n} + R^N a, \tag{3.5}$$

where $R^0 = a$ and $R^N a$ is the residual part after n iterations.

3.2.2 Processed Features

In the previous research work of Monaci, Divorra and Vandergheynst [19], the feature used in the analysis was the average of the energy present at each time instant. After the MP decomposition of the audio sequence, this value was computed as the addition of all the MP audio atoms energy.

With this audio feature we are loosing a lot of information present in the MP decomposition. The proposed model tries to take advantage of all the output parameters instead of using only the energy feature. These additional information allow us to aspire

to the Audiovisual Separation objective. First, we perform the video separation aided by audio as in the previous research work, with the resulting speaker detection. After, a new concept of Audio Source Separation is introduced using the information present in the video sequence. We will show that a good Audiovisual Separation is possible combining the features of both modalities.

The most relevant parameters obtained by the Lastwave [1] MP audio decomposition are the following:

timeId Temporal situation of the audio atom. One of the most important parameters, as the analysis we perform is basically in the temporal domain.

freqId Localization in frequency of the atom. Used in the later frequency analysis.

coeff2 Coefficient of the audio atom. Measure of relevance of this audio atom in the decomposition of the whole sequence. A big coefficient means that the atom is one of the first ones obtained by the MP iterative decomposition and, as a result, one of the most important in the representation of the audio signal energy in the time-frequency plane.

windowSize Size of the window used for this atom. As we can see in figure 3.2, this parameter gives us information about the energy distribution of the audio atom, as it determines its duration. The smaller is the windowSize, the more concentrated is the energy around the audio atom temporal center.

The proposed model uses all these Lastwave parameters. In the next chapter, the time index and the coefficient are used to determine the relationship between audio and video atoms, looking for synchronous *relevant events* in both features. Then, the audio reconstruction takes into account the window size to compute the energy feature of each atom. And, finally, the frequency index is utilized in the frequency analysis introduced after the temporal one.

3.2.3 Motivation

The main motivation of the use of MP algorithm is the sparse representation of the information present in the audio signal. These decomposition into atoms provides a clear representation of the audio energy distribution in the time-frequency plane, showing the frequency components evolution.

Another reason to choose this representation is that MP algorithm performs a denoising of the input signal, pointing out the most relevant structures. This characteristic of the decomposition is widely explained in [18].

The white characteristic of the noise causes the apparition of time-frequency atoms spread across the time-frequency plane. However, the structures presents in the original

signal are still recognizable, due to the higher concentration of their energy in the time-frequency domain.

Therefore, the algorithm chooses the most relevant structure in each iteration. As a result,the atoms created by the presence of noise are selected later than the structures in the original signal, which have a bigger coefficient and importance in the decomposition. The number of iterations determines how much of these noise atoms are created by the algorithm. While important structures in the original signal are always represented (in the first iterations), part of the noise disappears.

Moreover, the MP representation over redundant dictionaries in the time-frequency plane provide a complete description of the spectrum of the audio signal. This decomposition describes, for each relevant structure, its situation in the time-frequency plane and its temporal energy distribution. All this information is exploited in this research work in order to obtain a good Audiovisual Separation.

3.3 Video representation

Natural image sequences are composed of successive projected snapshots of 3-D objects. Considering these objects to describe smooth trajectories through time as shown in Fig. 3.3, one usually assumes that image sequences are well modeled by smooth transformations of a reference frame [31].

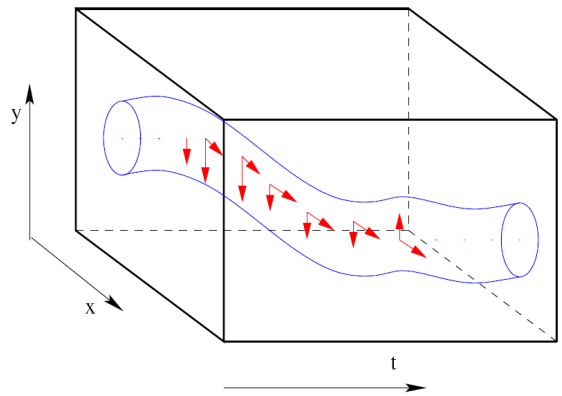


Figure 3.3: Schematic smooth evolution of an object through time.

A video sequence can thus be considered as a series of frames represented by a mixture of homogeneous regions and regular contours, where the motion is represented by smooth local deformations of those regions. Coping with regular geometric deformations necessitates the use of flexible visual primitives. In order to achieve this, we advocate the use of parametric over-complete dictionaries of basic waveforms, referred to as atoms. Local

deformations are then propagated along the sequence by updating the atoms parameter field in order to approximate the succession of frames.

3.3.1 Signal decomposition

In the proposed model, the video signal is represented using the **Matching Pursuit algorithm proposed by Divorra and Vandergheynst** [10]. In this technique, the image is decomposed into a set of video atoms which represent salient video components tracking their temporal transformation through time. A modified MP approach based on Bayesian decision criteria is used for the tracking.

Assuming that an image $I(x_1, x_2)$ can be approximated with a linear combination of atoms retrieved from a redundant dictionary D_V of 2-D atoms, we can write:

$$I(x_1, x_2) \approx \sum_{\gamma_n \in \ \Omega} c_{\gamma_n} G_{\gamma_n}(x_1, x_2), \tag{3.6}$$

where n is the summation index, c_γ corresponds to the coefficient for every atom $G_\gamma(x_1, x_2)$ and Ω is the subset of selected atom indexes from dictionary D_V. We also require that the representation is *sparse*, *i.e.* the cardinality of Ω is much smaller than the dimension of the signal. The decomposition of $I(x_1, x_2)$ on an overcomplete dictionary is not unique, and several decomposition approaches have been proposed, like the method of frames [10], Matching Pursuit [18] or Basis Pursuit [12]. Here we consider Matching Pursuit (MP), an iterative greedy algorithm that selects the element of the dictionary that best matches the signal at each iteration.

Each video frame is decomposed into a low-pass part, that takes into account the smooth components of images, and a high-pass part, where most of the energy of edge discontinuities lays. The low frequency component is obtained by lowpass filtering and downsampling the images in the sequence, using the Laplacian pyramid scheme [13]. We employ here the FIR low-pass filter proposed in [14]. The high-pass frames are obtained by subtracting the low frequency parts from the original frames. These high frequency residual frames which contain the geometric structures of images, are represented using MP. The approach we consider here consists of decomposing a reference frame in terms of geometric 2-D primitives and tracking them through time. Thus, starting from the first frame of the sequence, I_1, MP iteratively picks up the function belonging to D_V that best approximates the image I_1.

The first step of the MP algorithm decomposes I_1 as

$$I_1 = < I_1, G_{\gamma 0} > G_{\gamma 0} + R^1 I_1, \tag{3.7}$$

where $R^1 I_1$ is the residual component after approximating I_1 in the subspace described by $G_{\gamma 0}$. The function $G_{\gamma 0}$ is chosen such that the projection $< I_1, G_{\gamma 0} >$ is maximal. At the

next step, we simply apply the same procedure to $R^1 I_1$, which yields:

$$R^1 I_1 =< R^1 I_1, G_{\gamma 1} > G_{\gamma 1} + R^2 I_1. \tag{3.8}$$

This procedure is recursively applied, and after N iterations, we can approximate I_1 as $\hat{I}_1$:

$$\hat{I}_1 = \sum_{n=0}^{N-1} c_{\gamma_n} G_{\gamma_n}, \tag{3.9}$$

where $c_{\gamma_n} =< R^n I_1, G_{\gamma_n} >$.

As in the audio case, the dictionary D_V is built by varying the parameters of a mother function, in such a way that it generates an overcomplete set of functions spanning the input image space. The choice of the generating function $G(x_1, x_2)$ is driven by the observation that it should be able to represent well edges on the 2-D plane. Thus, it should behave like a smooth scaling function in one direction and should approximate the edge along the orthogonal one. We use here an edge-detector atom with odd symmetry, that is a Gaussian along one axis and the first derivative of a Gaussian along the perpendicular one (see Fig. 3.4). The generating function $G(x_1, x_2)$ is thus expressed as:

$$G(x_1, x_2) = 2x_1 \cdot e^{-(x_1^2 + x_2^2)} \tag{3.10}$$

The codebook of functions D_V can be defined as $D_V = \{G_\gamma : \gamma \in \Gamma\}$. Each atom $G_\gamma = U_\gamma g$

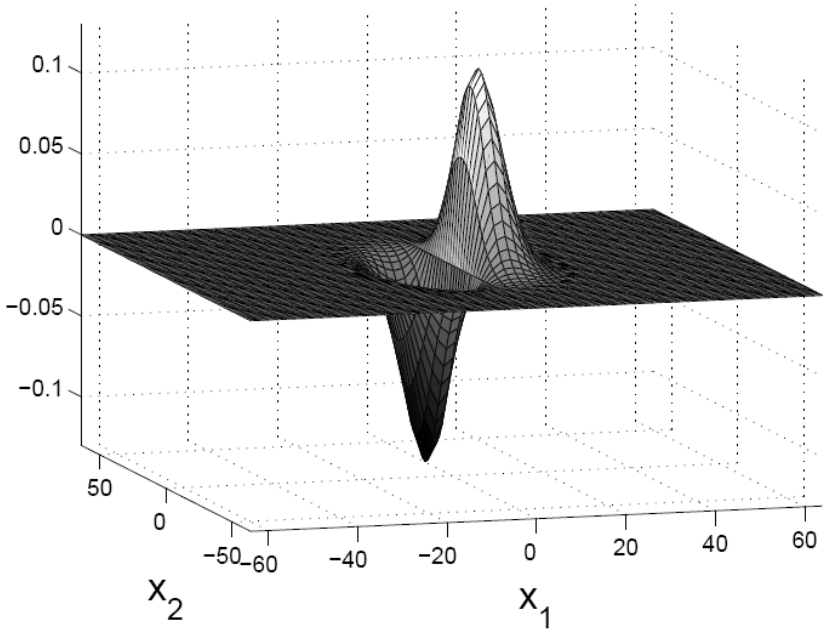


Figure 3.4: The generating function $G(x_1, x_2)$ described by Eq. 3.10

is built by applying a set of geometrical transformation U_γ to the mother function $G(x_1, x_2)$. Basically, this set has to contain three transformations:

- Translations $\vec{t} = (t_1, t_2)$ all over the image plane.
- Rotations θ to locally orient the function along the edge.
- Anisotropic scaling $\vec{s} = (s_1, s_2)$ to adapt the atom to the considered image structure.

Any atom G_γ in the dictionary rotated by θ, translated by t_1 and t_2, and anisotropically scaled by s_1 and s_2 can thus be written as:

$$G(x_1, x_2) = \frac{C}{\sqrt{s_1 s_2}} \cdot 2u \cdot e^{-(u^2+v^2)} \tag{3.11}$$

where C is a normalization constant and

$$u = \frac{\cos\theta(x_1 - t_1) + \sin\theta(x_2 - t_2)}{s_1}, \tag{3.12}$$

and

$$v = \frac{-\sin\theta(x_1 - t_1) + \cos\theta(x_2 - t_2)}{s_1}. \tag{3.13}$$

We consider an approach where 2-D spatial primitives obtained in the expansion of a reference frame of the form of Eq. 3.9 are tracked from frame to frame. Given a set of images belonging to a sequence, the changes suffered from a frame I_t to I_{t+1} are modeled as the application of an operator F_t to the image I_t such that

$$\begin{aligned} I_{t+1} &= F_t(I_t) \\ I_{t+2} &= F_{t+1}(I_{t+1}) = F_{t+1}(F_t(I_t)) \\ I_{t+3} &= ... \end{aligned} \tag{3.14}$$

where t is the time index.

Fig. 3.5 shows an example of the application of the operator F_t in a sequence of frames. The possible situation in the image of some 2-D video atoms is schematically represented. The transformations they experiment in each frame are symbolized by arrows, modeling with the operator F_t their translations, rotations and changes in the scale.

From the model of Eq. 3.9 and 3.14, follows that

$$\hat{I}_1 = F_t \left(\sum_{n=0}^{N-1} c_{\gamma_n^t} G_{\gamma_n^t} \right), \tag{3.15}$$

Making the hypothesis that F_t represents the set of transformations F_t^γ of all individual atoms that approximate each frame, we obtain:

$$\hat{I}_{t+1} = \sum_{i=0}^{N-1} F_t^{\gamma_i}(c_{\gamma_i}^t G_{\gamma_i}^t) \tag{3.16}$$

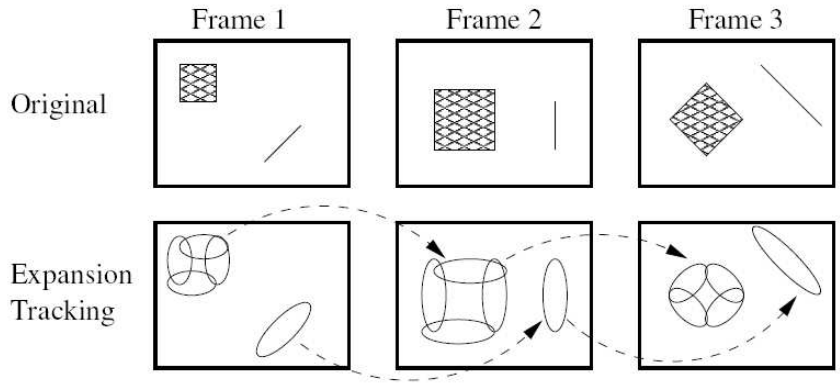


Figure 3.5: Successive schematic updates of basis functions in a sequence of frames. In the second row, ellipses represent schematically the possible positioning of some 2D atoms.

A MP-like approach similar to that used for the first frame is applied to retrieve the new set of g_γ^{t+1} (and the associated transformation F_t). At every greedy decomposition iteration only a subset of functions of the general dictionary is considered to represent each deformed atom. This subset is defined according to the past geometrical features of every atom in the previous frame, such that only a limited set of transformations (translation, scale and rotation) are possible. This imposes smoothness on the set of deformed primitives, following the assumption of smooth transformation.

Due to dictionary coherence, and the fact that normally more than one atom is necessary to represent a signal structure, direct MP full search in a frame at $t+1$ does not have any guarantee to recover the corresponding deformed atoms from frame t.

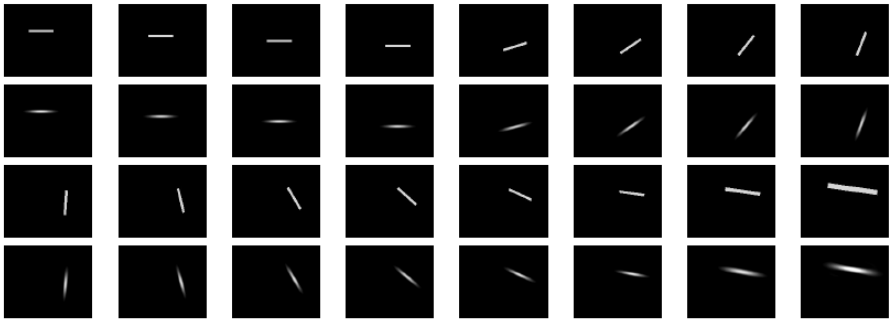
Thus, a Bayesian framework is introduced to regularize motion among expansion terms of frame representations. This method considers a modified MP approach based on a Bayesian decision criteria to deform the atoms in a predictive fashion from frame to frame.

Motion is assumed to be uniform over the support of an atom, and consequently, a local search on a reduced subspace is performed for each video atom, starting from higher energy atoms (first in the MP expansion) to weaker ones.

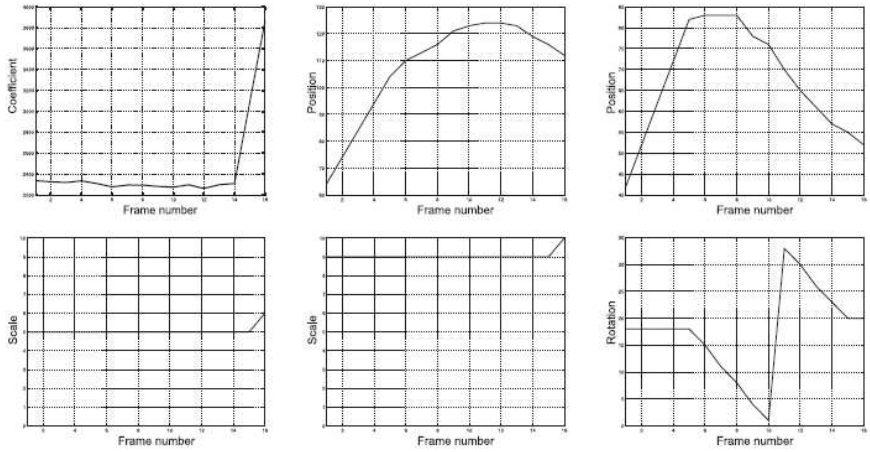
In other words, the use of some *a priori* knowledge in the selection criteria give us the most probable transformation, considering the constrained motion (translation, rotation, scale and projection coefficient c_γ).

The formulation of the MP approach to geometric video representation is complex and is studied in detail in [10], to which the interested readers are referred.

A cartoon example of the used approach can be seen in Fig. 3.6(a), where the approximation of a simple synthetic object by means of a single atom is performed. The first and third row of pictures show the original sequence and the second and fourth rows provide the approximation composed of a single geometric term. Fig. 3.6(b) shows the parametric



(a) Synthetic sequence approximated by 1 atom: First and third row show the original sequence made by a simple moving object. Second and fourth row depict the different slices that form a 3-D geometric atom.



(b) Parameter evolution of the approximated object; from left to right and from up down, we find: Coefficient c_γ, horizontal position t_1, vertical position t_2, short axis scale s_1, long axis scale s_2, rotation θ.

Figure 3.6: Approximation of a synthetic scene by means of a 2-D time-evolving atom.

representation of the sequence. We see the temporal evolution of the coefficient c_γ^t, and of the position, scale and orientation parameters. The MP decomposition of the video sequence provides a parametrization of the signal which represents the image geometrical structures *and* their evolution through time. In this way we can track the movements of relevant image features, getting an accurate description of the scene content. Besides, it is important to underline that the stream of video atoms that we consider is absolutely generic. It could be generated using different approximation techniques and it can be used to encode video sequences, as it is shown in [10].

3.3.2 Processed Features

Clearly, video features need to capture temporal variations. To date, video features used for multimodal audiovisual fusion are often based on pixel-wise intensity difference measures. In [20] and [4], the pixel intensity change measured in a 3×3 averaging spatial window is considered. The approach in [5] looks forward exploiting local motion information by means of optical flow measures. In any case, none of the actual approaches try to exploit the real structural nature of video signals.

We have decided thus to explore the possibilities offered by the MP video decomposition technique presented before. In this way, we hope to be able to track important geometric features over time and to effectively parameterize those transformations that represent changes in the scene. The output of the MP algorithm is a set of atom parameters that describe the temporal evolution of 3-D video features. Each atom is characterized by a coefficient, 2 position parameters, 2 scale parameters and a rotation, *i.e.* 6 parameters. Fig. 3.6(b) shows the atom parameters evolution as a function of time.

The video features we consider, however, are not all these 6 video parameters. The scale and orientation parameters have been discarded, since they carry few information about the mouth movements. Clearly, they can be used if needed in a more complex application, but in this context the natural choice seems that of considering a feature that takes into account the movement of image structures. Therefore, for each video atom we compute the absolute value of the displacement as

$$d = \sqrt{t_1^2 + t_2^2}, \tag{3.17}$$

where t_1 and t_2 are the horizontal and vertical position parameters of the atom. The quantity d is used as video feature, and indicates a sort of activation of the video structure that it represents.

Modulus of the displacement of each video atom is computed as a measure of movement in the image, from one frame to next one.

The video *relevant events* that the implemented method temporally compares to the audio ones are the peaks in the video atoms displacement. These peaks mean movement in the audio atom situation and, consequently, movement in the image. Thus, one peak in the displacement of one video atom temporally close to the apparition of one sound is studied as possible cause of this sound.

3.3.3 Motivation

An image sequence is decomposed in 3-D video components intended to capture geometric features (like oriented edges) and their temporal evolution. In order to represent the large variety of geometric characteristics of video features, redundant codebooks of functions

have to be considered. The use of geometric video decomposition has at least two main advantages:

- 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 meaning. Thus, the considered video features reflect the movement, from frame to frame, of the image structures associated with the corresponding geometric primitives. A peak in the displacement suggests the presence of an event, that is, a possible movement with respect to a certain equilibrium position (*i.e.* movement of the lips in the speaker mouth).
- Geometric sparse video decompositions provide compact representations of information, allowing a considerable dimensionality reduction of the input signals. This property is particularly appealing in this context, since we have to process video sequences, which have a very high dimensionality.

3.4 Discussion

We have now a generic representation of the video that describes synthetically how the scene is composed and how image components evolve. Using such a parametric representation, we try to follow the temporal evolution of relevant image features, like those constituting the speakers mouth or chin. The whole set of 3-D geometric primitives used to represent the video are considered, and they are sorted by correlation with the audio. Such correspondence between acoustic and visual signals is assessed by comparing the evolution of visual structures with that of some audio track descriptors.

Audiovisual pairs are considered to be correlated when we observe a temporal synchrony between events present in both audio and video signals, that are thus supposed to be caused by the same physical phenomenon. Events will be defined as local perturbations of an equilibrium situation, exploiting the motion information of the geometric primitives describing the scene and the energy distribution of the audio track in the time-frequency plane.

The approach we propose is derived directly from the physics of the phenomenon. On the one hand, the considered video features reflect the movement, from frame to frame, of the image structures associated with the corresponding geometric primitives. On the other hand, the audio feature provides a complete description of the soundtrack energy distribution. Peaks in such signals suggest the presence of an event. In the video case, it can be the movement with respect to a certain equilibrium position (*i.e.* lips opening or closing). For the audio, the presence of an audio atom indicates the utterance of a sound. If those audio and video peaks occur at time instants that are temporally close, we can expect that they reflect the presence of two expressions (acoustic and visual signals) of the same physical phenomenon (utterance of a sound).

Chapter 4

Phases of the Temporal Analysis

According to psychophysical experiments, *temporal synchrony* strongly contributes to integrate cross-modal information in humans [9, 30]. These observations motivated several research works that used audio-video synchrony in order to locate the sources of a multimodal sequence [8, 13, 20, 29]. Just like all these methods, the proposed model is based on a temporal analysis that looks for audiovisual pairs generated by the same physical phenomenon. Therefore, as in the previous research work performed by Monaci, Divorra and Vandergheynst [19], the proposed model uses video 2-D geometric features instead of pixel-related features.

Therefore, the detection of this audiovisual pairs is performed combining the features extracted in Chapter 3. We want to associate temporally proximal *relevant* audio-video events and extract in this way meaningful audiovisual structures. The determination of this relationship is the most important step of the proposed algorithm, and provides all the information necessary to perform a satisfactory Audiovisual Separation.

The temporal analysis that the proposed method uses can be divided in three different parts. To achieve the Audiovisual Separation, first we situate the sources in the image using the information present in the soundtrack, then we reconstruct them separately, and, finally, the relationships established between features in both modalities are used for the Blind Audio Source Separation objective. Thus, this chapter is structured as follows:

- In the first part, we use the information contained into the audio signal to situate the video sources in the image, that is, the objective already achieved in the previous research work performed by Monaci, Divorra and Vandergheynst [19], but now correlating audio atoms instead of the soundtrack energy. The localization of the video sources is performed with a temporal analysis of the video and audio atoms, the most important step that provides us a measure of the synchrony between both features, the *correlation score*.

- The objective of the second part is to classify the video atoms into the detected video sources. These assignation is carried out with a spatial proximity criterium, atoms closer to the estimated source center in step 1 than a maximum distance defined in pixels belong to this source.

- The last and more ambitious objective (the Blind Audio Source Separation) is performed using the already extracted audiovisual features. At the end of step 2, we have the video atoms classified into the sources, and their respective *correlation scores* with the audio features. All what we need to do is to classify the audio atoms into one of the sources using these scores and the source of each video atom associated. Finally, the reconstruction of the audio signals is carried out by adding the energy of the audio atoms belonging to each one of the sources.

4.1 Locate video sources of an acoustic signal

4.1.1 Introduction

The correct localization of the video sources in the spatial domain, speaker detection problem, is the first part of the Audiovisual Separation and provides the relationship between atoms in both features, a necessary step in order to face the Blind Audio Source Separation (BASS). In this research work, the first of our objectives will be achieved by characterizing the video through the audio information.

This problem was firstly faced by Hershey and Movellan [13], who design a simple algorithm to locate sounds using audio-video synchrony. The correlation between audio and video was measured using the correlation coefficient between the energy of an audio track and the value of single pixels. Successive studies in the field [8, 17, 20, 28, 29] focused on the statistical modeling of relationships between audio and video features, proposing audiovisual fusion strategies based on Canonical Correlation Analysis [17, 28], Independent Subspace Projections [29] and Mutual Information maximization [8, 20]. Pixel-related features typically used for video representations and employed in all these works were and barely connected with the physics of the problem. This makes it difficult to deal with dynamic scenes, since the variables that are observed (pixel values or related quantities) are static. Moreover, pixel-related values have low semantic content, which makes it impractical to extract and manipulate correlated audiovisual structures.

In order to understand more in detail audio-video structures and to improve the performances of audiovisual fusion algorithms, an effort should be done to model the observed physical phenomenon. As in the work [19], we introduce a new framework for detecting meaningful events in audiovisual signals. The main difference is the dimension of the problem. Now, we have one feature for each audio atom, whereas in the previous work one unique feature represented all the soundtrack. Consequently, we confront a problem with bigger number of dimensions now, but there is also more information to

exploit. A new method to correlate audio and visual features is implemented in order to confront the multi-dimensionality of the audio feature.

The problem we are studying in this work is that of correlating audio tracks with visual data to detect those regions in an image sequence from which the sound is originated. This problem can be divided into two different parts.

In the first one, the changes in the image which are related to the present acoustic signals are detected. These resultant audiovisual events are represented, in our case, by changes in the video atoms displacement temporally correlated to the apparition of the audio atoms.

The second part consists on a spatial clustering which groups into possible sources all the video atoms related to any audio atom in the whole sequence. At the end of the process we will obtain the estimated spatial coordinates of the video sources.

4.1.2 Audio and Video Atoms Association

The association between visual features and acoustic signals is the most important part in the process, which extracts all the relevant information used in the next steps of the algorithm. All this resultant information is posteriorly processed in order to obtain the spatial situation of the video sources. In the next sections we will show that a correct audio-video association allows a good source localization, pointing out the video atoms closest to the speaker mouth in our case.

As already discussed in Chapter 3, the features to analyze here have a true geometrical meaning. Thus, the considered video features (displacement as modification of the position with regard to the last frame) reflect the movement of the image structures associated with the corresponding geometric primitives. A peak in the displacement suggests the presence of an event, that is, a possible movement with respect to a certain equilibrium position (*i.e.* movement of the lips in the speaker mouth). Concerning the audio signal, the decomposition into atoms provides a clear representation of its energy distribution in the time-frequency plane, showing the frequency components evolution. Thus, the temporal situation of an audio atom indicates the presence of a sound in this period.

More concretely, considering the *relevant events* extracted in last chapter, the goal is to discover all the video atoms with a peak in the displacement temporally close to the center of the audio atom. In this way, the algorithm creates audiovisual relationships, which are probably generated by the same physical phenomenon.

At this moment we know which ones are the relevant events in each modality, but what we have to do to discover the relationships between them?

The proposed model performs a temporal analysis taking into account the temporal index of *relevant events* in both modalities: the time center of the audio atoms and the

moment when the peaks in the video displacement occur. Thus, for each video atom we obtain its displacement function with several peaks as shown in Fig. 4.1, and, for each audio atom we know their temporal center. As a result, the assignation consists in compare for each audio atom (with one peak in its time index) all the video atoms features (with several peaks). Then, for each audio atom, we select the video atoms with a peak close to the audio atom temporal center in frames.

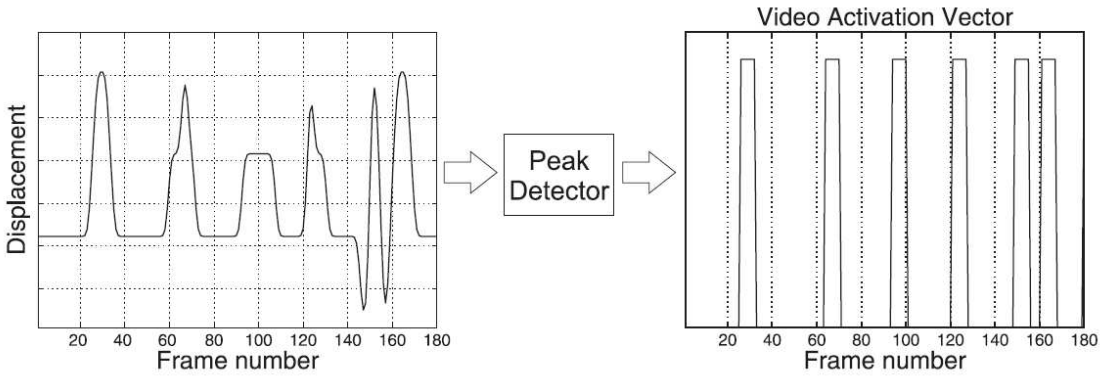


Figure 4.1: Displacement function and resulting peaks of a video atom.

This method uses a scalar product between audio and video features to obtain the existent relationships between them, but many questions appear at this moment about how to compute this scalar product: should we use deltas in both features (one for the audio, as much as peaks for the video) or we would be too much restrictive? Should we consider a possible delay between them with temporal windows? What would be the appropriate length of those windows?

This part has gone through a lot of changes from the first stage of the investigation until the last one, with important improvements in the results. The most representative stages are explained in the following subsections.

Delta in audio, window in video

Concerning the audio part, we put a delta in the central frame of the audio atom, in other words, we process it like a punctual event with duration one frame. Then, for the video, we put a window of W=2 frames to all its displacement peaks, taking into account a possible delay between audio and video features.

Fig. 4.2 illustrates the features analyzed in the scalar product, where W=2 implies a maximum delay of 5 frames.

According to these characteristics, when we compute the scalar product the output value is 1 or 0, respectively for a video atom synchronous or not. For each audio atom,

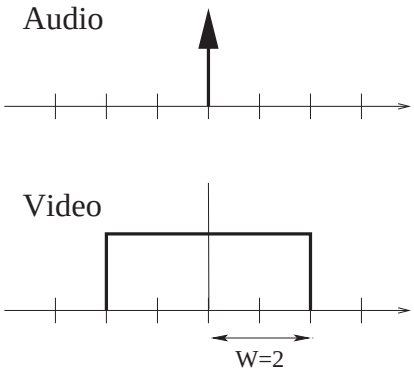


Figure 4.2: Scalar product between audio and video features, first stage (time index in frames)

the proposed algorithm chooses the video atoms with this *correlation score* different from zero.

The main problem is that we are not pondering the temporal superposition between both features in this form, all the video atoms are equally associated to the audio atom, *correlation score* 1. Therefore, a video atom with a peak in the displacement situated exactly in the same frame than the audio atom center has the same correlation score than another one shifted two frames away. We are not giving more importance to the more synchronous video atom.

In addition, the real distribution of one video atom energy is not a delta but a gaussian as shown in Fig. 3.2, so in the next stage we will change this representation to ponder the audiovisual synchrony more accurately.

Gaussian in audio, window in video

For the audio, we consider a gaussian function in the central frame of the audio atom, according to its characteristics of length and energy coefficient. For the video, we use the same window of W=2 frames or 534 samples (1 video frame is equivalent to 267 audio samples), 1335 samples of maximum delay. There is a representation of this scalar product in figure 4.3.

This first significative change regarding the energy distribution of the audio atoms, involves more variability in the result of the scalar product between both features.

Thus, every video atom has associated a *correlation score*, the result of the scalar product. This value is higher if the audio atom and the peak of displacement of this video atom have a bigger temporal overlap. In other words, a high *correlation score* means high probability of being the structure that has generated the sound.

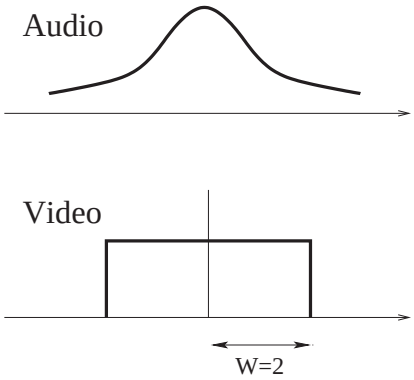


Figure 4.3: Scalar product between audio and video features, second stage (time index in samples)

Gaussian in audio, window in video with peak constraint

The most relevant improvement in this stage has been to introduce a constraint relative to the video atoms displacement. The gaussian for the audio and the window in video are kept.

In this part, we define a *peak* in the modulus of the displacement of a video atom as a positive slope followed by a negative one in the next frames. The peak detector used in the other versions detected as a peak a positive slope, but it didn't look what occurred afterwards. Figure 4.4 shows the different peaks definition, the first one with considerable improvements in the results.

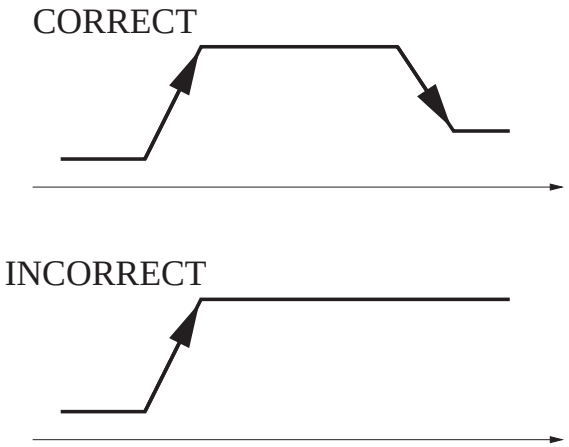


Figure 4.4: *Peak* current and former definition in the displacement evolution

The *errors* are video atoms temporally synchronous with an audio atom and not belonging to the related source or belonging to it but localized far away form the center of the source. Thus, possible errors could be the detection of the hair of the speaker, the

eyes, etc. And still more dangerous in our case, the association with an atom in the mouth of the other speaker.

Most of the *errors* in the other associations had this temporal comportment, only a positive slope, and with this simple redefinition are now solved.

In this stage, the window length is bigger (W=6 frames) because we want to assign all the audio atoms to almost one video atom. With the constraint we are removing a lot of video atoms, and then some important audio atoms would be related to no video atom. As a result we could not decide to which source they belong and these audio atoms would be lost in the reconstruction.

Fig. 4.5 shows the considerable **improvements** in the audiovisual association when we introduce the constraint relative to the form of the peaks in the video atoms displacement. The frames are reconstructed by summing to the low-pass images those video atoms that are associated to some audio atom temporally situated in this frame, that is, with the time index in samples equivalent to the represented frame.

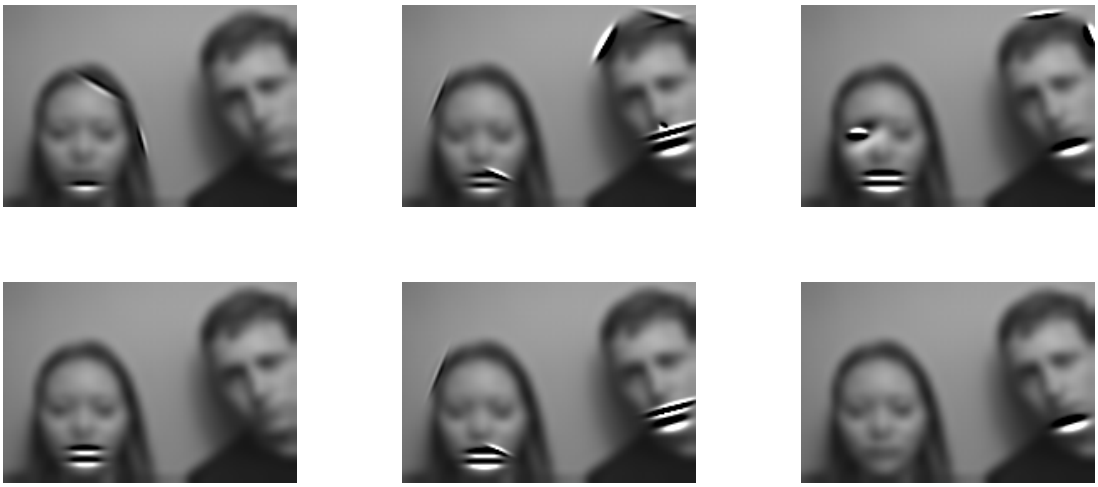


Figure 4.5: Audiovisual association in frames 30, 100 and 240 of the clip `g20` of CUAVE database. The video atoms related to some audio atom in this frame are highlighted. The first row corresponds to the association performed in 4.1.2 while the second row corresponds to this new procedure, which incorporates the constraint for the peaks in video atoms displacement.

Thus, analyzing more accurately this figure, the principal improvements of this method are represented. First column shows a frame where only the girl is speaking. The selection of some video atoms far away from her mouth is removed with this new definition of *peak* 4.4. This improvement is also shown in the central column, since, now, the video atoms in the boy head are not reconstructed. However, the most important change regarding our objective of detecting the current speaker can be visualized in third column. Despite

only the boy is speaking in this frame, if we perform the association according to 4.1.2, that is, without constraint, several atoms are highlighted in the mouth of the incorrect speaker. This would be a big problem for the proposed method, since this wrong association would involve the detection by the proposed approach of a speaker that is not active in this moment. Thus, this figure shows as most of the errors are removed only with this redefinition of the concept of peak relative to the video atoms displacement.

4.1.3 Clustering

Our objective is to detect and locate the sources of an audio signal into the image. At this moment we know the temporal relationship of audio and video atoms in this sequence, but how we can localize the signals? The video atoms more frequently related to the audio atoms are chosen in the proposed model as sources of this sounds, locating them in the image.

The problem is that one visual structure is composed of several video atoms, each one of them selected by a set of audio atoms. If one of the sources were considerably more time active than the other one, selecting the video atoms with more associations the algorithm would detect several sources in the same speaker and none in the other one. In addition, some other structures such as the eyes are also usually correlated with the soundtrack. Thus, the implemented clustering allows us to group spatially the most important video atoms belonging to the same source.

In this section, we define the *confidence value* of a video atom as the addition of the Matching Pursuit (MP) coefficients of all the audio atoms associated in the whole sequence. 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 sequence.

Looking at the figure 4.6, the idea of a clustering is very intuitive. We can see the remarked video atoms (with confidence different from zero) grouped around the speakers mouth, one at the left and the other at the right of the image. Atoms with higher *confidence value* form two differentiated groups pointing out the sources, while the more separated have a considerably smaller confidence. We can conclude that the audio and video atoms association has been successful, detecting features close to the source center much more usually than the others.

We have to point out one of the main advantages of this spatial clustering: it groups the video atoms without any assumption about the number of sources present in the sequence. This characteristic makes the proposed model robust to analyze any audiovisual sequence without previous external information.

The algorithm is divided in three main steps: create clusters, calculate centroids and eliminate bad clusters. Next, we explain them more accurately.

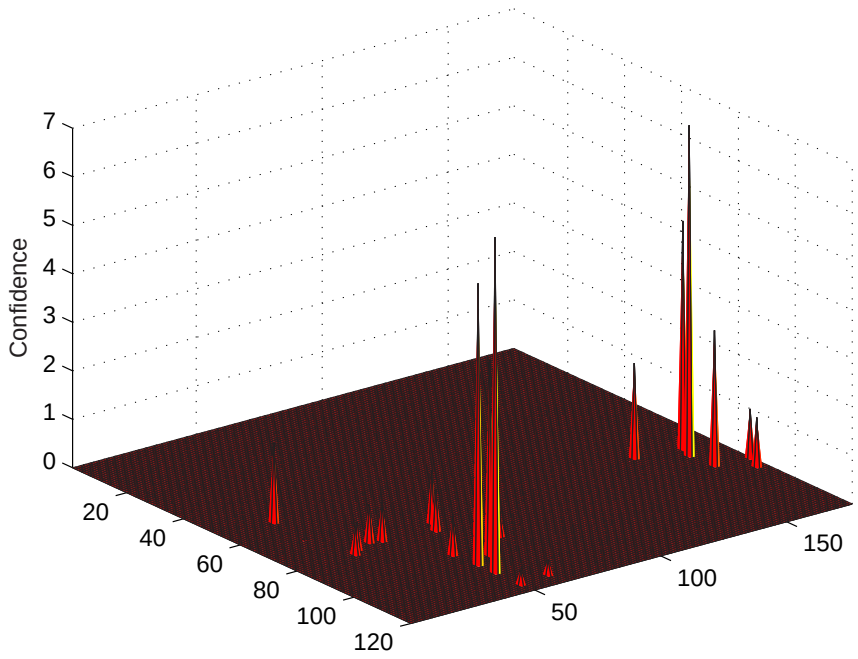


Figure 4.6: Video atoms situation in the image. Their confidence value is represented in the third dimension.

Create clusters

The clusters are created with the following iterative algorithm:

1. Initialization: the video atoms to be classified are all those related to almost one audio atom in the whole sequence (confidence value different from zero).
2. The video atom with highest confidence value builds the first cluster. 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. Aggregate closest video atoms, spatial maximum distance from the center of the most important atom (*cluster size* defined in pixels).
4. Remove the video atoms assigned to this cluster from the total of those to be classified.
5. Stop the algorithm when all the points with confidence over the mean are already classified, otherwise go back to step 2. Only video atoms with significant confidence value can be the center of a new cluster (possibly a source).

We have to take into account some considerations about this algorithm.

Concerning the clusters creation, the most important parameter to fix is the cluster size. This characteristic determines the number of clusters created by the algorithm, and, consequently, the number of sources detected in a first stage of the clustering. However, as we will prove after in the discussion, this characteristic does not affect significantly the final result.

Thus, in the third step of the algorithm, a radius around the main video atom between 30 and 60 pixels (wide of the image: 176 pix) is appropriated to the case we are analyzing. The database we are using contains sequences with 2 speakers significantly separated. However, this algorithm has no problems with bigger cluster sizes (radius until 90 pixels).

As we can see in figure 4.6, most of the video atoms selected have a negligible *confidence value* (related sometimes to only one audio atom). As a result, the threshold applied in step 5 is not very high, and it is basically impossible to remove real sources.

Calculate centroids

This step computes the center of mass of the video atoms belonging to the clusters. In order to perform it, the confidence value of every atom is taken as the mass, and ponders its position in the image.

Thus, for each created cluster we calculate its centroid as:

$$Centroid = \frac{\sum position_i \times confidence_i}{\sum confidence_i} \tag{4.1}$$

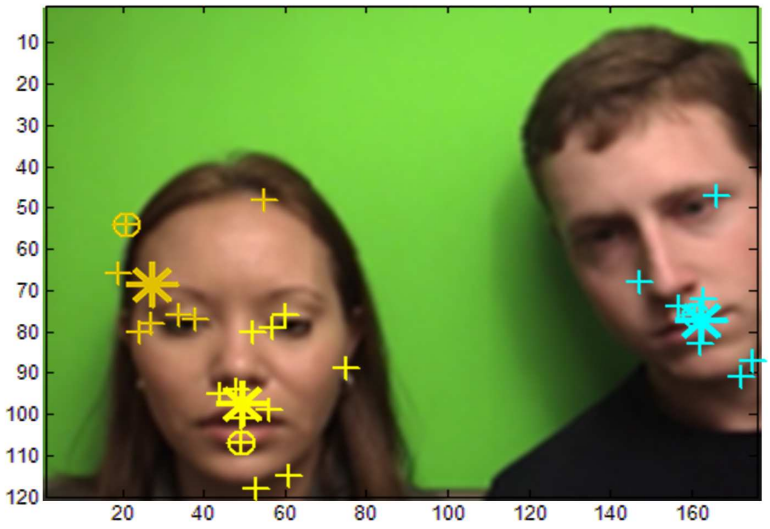
where $position_i$ are the coordinates of the video atoms belonging to the cluster and $confidence_i$ their confidence values.

This centroids are the coordinates in the image where the algorithm locates the sources of the audio feature. In this kind of sequences with several speakers, the centroid should be close to their mouths.

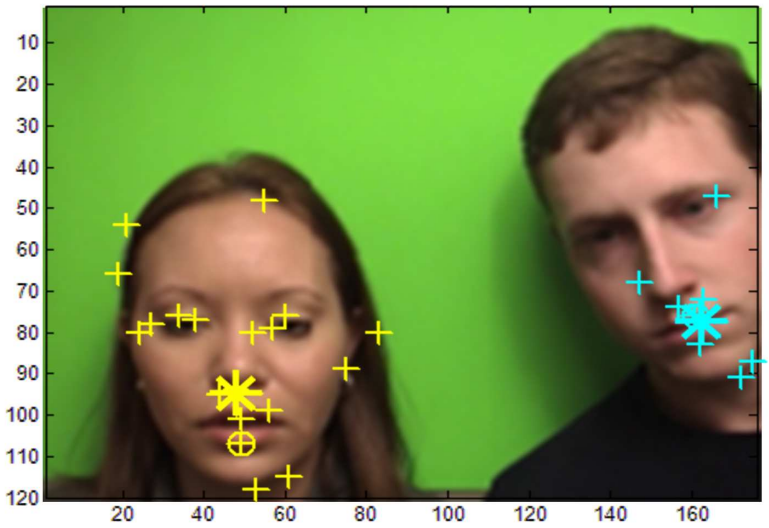
An example of the created clusters and their calculated centroids is shown in Fig. 4.7. We can see that some of the clusters are, as expected, close to the speakers mouth, while others do not represent a source (*orange* cluster, the less important and created the last one, with cluster size 40 pixels). The next step the proposed clustering algorithm takes into account these *bad clusters* and eliminates them.

Eliminate bad clusters

We define the *cluster confidence value* as the addition of the corresponding confidence values of the atoms that belong to this cluster.



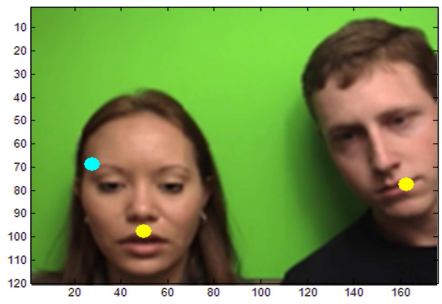
(a) Clusters creation with radius 40 pixels



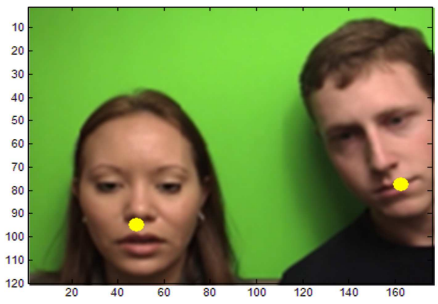
(b) Clusters creation with radius 60 pixels

Figure 4.7: Clusters created using different cluster sizes in the step 4 of the algorithm. The atom represented with a **o** is the one with higher *confidence value* that builds the cluster in step 2 of the algorithm. Then, the + are the coordinates of the video atoms aggregated to the cluster in step 3. Finally, the * are the calculated centroids of the cluster. Each cluster is represented in a different color, from first to last created (descendent importance of the cluster): yellow, cyan and, the last one, orange.

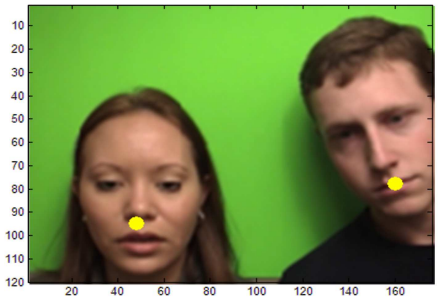
Then, *bad clusters*, that is, clusters with slight confidence value are removed. Fig. 4.8 shows the centroids of the clusters created by the algorithm with different cluster sizes. There are represented in cyan those clusters eliminated by the algorithm in this step.



(a) Clusters with 40 pixels of radius



(b) Clusters with 60 pixels of radius



(c) Clusters with 90 pixels of radius

Figure 4.8: Bad clusters (*cyan*) and good clusters (*yellow*) created by the algorithm using different cluster sizes. As shown in picture (a), a cluster size too small according to the scene causes the apparition of undesired centroids. The last step of the algorithm successfully removes them and selects only the centroids corresponding to the speakers mouths.

The threshold applied is 0.2 times the maximum value. There are two main factors that influence this choice.

On the one hand, this threshold has to be high enough to eliminate the created clusters that do not represent a speaker. Sometimes, a small cluster size involves the appearance of more than one cluster per source

On the other hand, if this value is very high the algorithm can remove clusters indicating real sources. This would be the case if one of the sources is active much more time than the other ones. As a result, the video atoms belonging to these speakers would have much more audio atoms related and their *cluster confidence value* will be considerably bigger.

Empirically, we observed than the threshold value we have applied is adequate to the explained requisites.

4.1.4 Discussion

In the proposed model, a good speaker localization is achieved by means of the creation of audiovisual synchronous structures combined with a robust clustering, which spatially groups the video atoms that form these structures into sources. This is not an easy task, and without these results we could not go on with the aspiration of the Blind Audio Source Separation.

An important point to clarify is the setting of the cluster size parameter. Figure 4.8 shows the calculated coordinates of the sources, with a cyan point where the algorithm has detected a bad source. Therefore, we can visualize the slight variations caused by different choices concerning the cluster size.

Despite a small radius, taking into account our database (two speakers close to the camera), last step 4.1.3 removes the clusters not belonging to an active mouth in the sequence.

Thus, this algorithm is robust enough to not depending strongly on the parameters we choose. Even if some of them are not adequate, we can perform a successful situation of the video sources in the image.

4.2 Separation and reconstruction of video sources

Once the sources location is calculated, the next step to carry out is to extract the whole visual structures, spatially separate them and then reconstruct the video sources.

The main characteristic to use in this point is the spatial distance between elements. We can use pixels or video atoms close to the estimated source coordinates to carry out the video reconstruction.

However, the most important function of this step regarding to the audio separation objective is to classify the video atoms into the detected sources.

Thus, we define a *maximum distance* in pixels from the centroid. All the video atoms located inside this region belong to this source.

To set this parameter, we have to take into account several conditions.

- Not to assign one vide atom to more than one source. In this case, we would not be separating, and there will be errors in this classification and the posterior audio separation.

- At the same time, the radius has to be big enough to contain the maximum number of atoms of each source. It is a very important aspect not to loose all the atoms related to an audio atom (in that case it would not be possible to posteriorly assign it to a source and reconstruct completely the sequence without energy losses).

- Not to classify into one source points belonging to another one. The activation of these atoms will cause mistakes and remark the wrong source.

Empirically, a radius around the centroids of 60 pixels (width of the image: 176 pix) is appropriated to the case we are analyzing. CUAVE database consists of different sequences with two speakers significantly separated.

At the end of this phase, video sources are already detected and easily reconstructed with a method lying in considering only the atoms closest to the source. Thus, the video separation is satisfactorily performed.

Fig. 4.9 shows an example of the reconstruction of the current speaker detected by the algorithm. For each frame, 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 face. Therefore, we can see as, in both frames, the correct speaker is detected.

4.3 Blind Audio Source Separation using video

4.3.1 Introduction

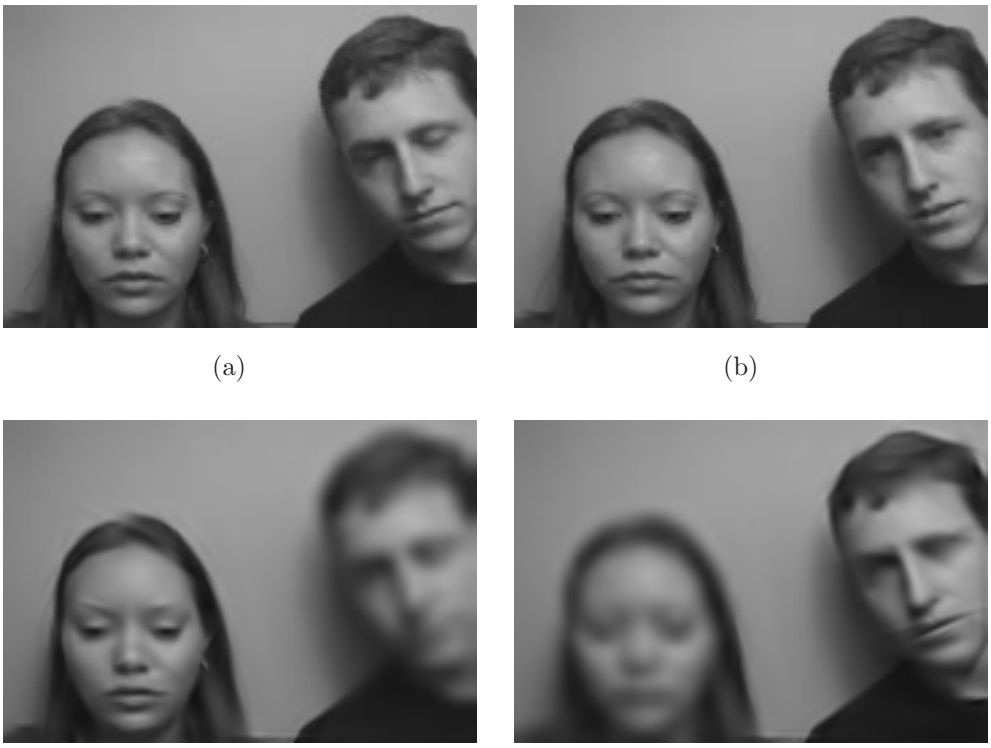
We confront now the most difficult part of the whole process, the problem of the Blind Audio Source Separation.

The information of the other parts of the process will aid us in this task. We already know the location of the video sources, the video atoms belonging to each one of this sources and, finally, the temporal relation between audio and video atoms with their *correlation score*.

From here we will combine all this elements to achieve the goal of the audio source separation.

4.3.2 Procedure

The BASS objective is to extract separately the signals that form an audio mixture. What we have to do, first, is to decide which audio atoms belong to each one of the sources. Then, we will reconstruct every one of these separated audio signals by adding the energy values of the atoms belonging to it.



(a) (b)

(c) (d)

Figure 4.9: Example of the video sources reconstruction. Pictures (a) i (b) show two original frames of an audiovisual sequence, where the speakers are the girl and the boy respectively. (c) and (d) correspond to the frame reconstruction using only video atoms close to the estimated source.

Assign every audio atom to one source

At this point, we know the video atoms belonging to each source and also the *correlation score*, the measure of synchrony between audio and video atoms. What we are going to do to assign the audio atoms to the sources?

The proposed model is the following. First, for every audio atom we have to take into account the total of 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 of these video atoms with the analyzed audio atom.

Therefore, here is the detailed description of the implementation steps for each audio atom:

1. Take all the video atoms related to this audio atom with its *correlation score*.

2. Add the *correlation score* of each video atom to the corresponding source. At the end of this step the sources have a value associated, the sum of the scores of the video atoms related.

3. Classify the audio atom into one source if its *total score* is big enough (more than two times the value of the other sources). Otherwise, this audio atom belongs to both sources.

For example, one audio atom has six video atoms associated (score different from 0). Four of them belong to speaker 1, and two to speaker 2, with the *correlation scores* shown in table 4.1. Then, the sum of the scores are 13.88776 and 1.71717 for sources 1 and 2 respectively. The score for the first source is much bigger (approximately eight times the other) and, as a result, this audio atom will be assigned to source 1.

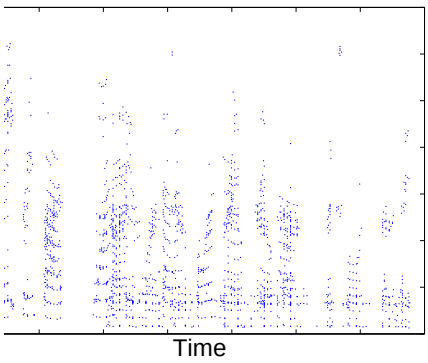
Source 1	**Source 2**
6.9348	1.1146
5.8186	0.60257
0.809	
0.32536	

Table 4.1: Example of one video situation atom before its assignation to one of the sources. A description of the correlation values of each video atom related and their assigned source is made. Four of them belong to source 1 and two to source 2.

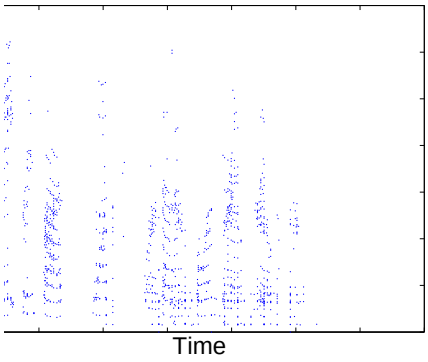
Repeating this steps, all the audio atoms will be classified into one or more of the sources, and can be used to reconstruct the corresponding one.

Fig. 4.10 shows an example of the audio atoms assignation into a source according to the presented temporal analysis. The centers of the audio atoms in the time-frequency plane are represented for the original sequence (mixture of two speakers, one boy and one girl) and for the separated soundtracks. This temporal analysis determines the time periods where only one source is active. The first part of the sequence corresponds to speaker 1 and the last part to speaker 2, and, correctly, audio atoms in this periods are assigned only to them. Additionally, when both sources are active at the same time some audio atoms are related to one of them and some to the other.

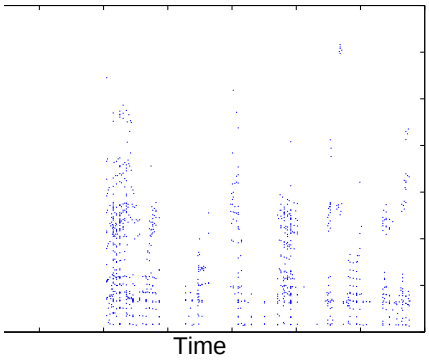
Before continuing, it is necessary to do the following explanation about the decision bound in step 3. Not all the audio atoms in this point are clearly classified into one of the sources. Some of them are in an intermediate position and we cannot decide only with a small difference of the *source total score*. This atom may belong to more than one source, or we could be making a mistake choosing one speaker and not the other. Thus, with this *source total score* bound, we can be more sure about the fact that the analyzed audio atom belongs to this video source.



(a) Original sequence decomposed into atoms.



(b) Atoms assigned to speaker 1.



(c) Atoms assigned to speaker 2.

Figure 4.10: Example of the classification of the audio atoms into the correspondent source. The points in the pictures represent the situation in the time-frequency plane of the audio atoms centers. Therefore, (a) corresponds to the centers of the atoms in the original mixture, and (b) and (c) to speakers 1 and 2 respectively.

Reconstruct the separated signals

The audio signal that comes from each source is reconstructed by adding the energy of the audio atoms classified in this source. If the atom belongs to several sources, its energy is equally distributed among all of them.

Therefore, to reconstruct the audio signal, we simply sum the selected atoms a(t) together. Each atom a(t) is computed as:

$$a(t) = \text{coeff} \cdot g(t - \tilde{t}) \cdot \cos(2\pi(\varphi + f(t - \tilde{t}))) \tag{4.2}$$

where $\tilde{t}$ and f are, respectively, the time and frequency coordinates of the audio atom $a(t)$ in the time-frequency plane. φ is the phase of the atom in the dictionary that best matches the analyzed signal and g is a gaussian function, the energy distribution of $a(t)$. The gaussian

form depends on its windowSize parameter in the Matching Pursuit decomposition. There are 6 possible sizes in this decomposition, from 512 samples to 16384 depending on the audio atom duration.

In the MP decomposition of Lastwave, *coeff* is the energy coefficient of the audio atom, *timeID* and *freqID* are the central coordinates in the time-frequency domain.

Fig. 4.11 shows the reconstruction with LastWave software [1] of the separated sequences for a boy and a girl obtained with the explained temporal analysis.

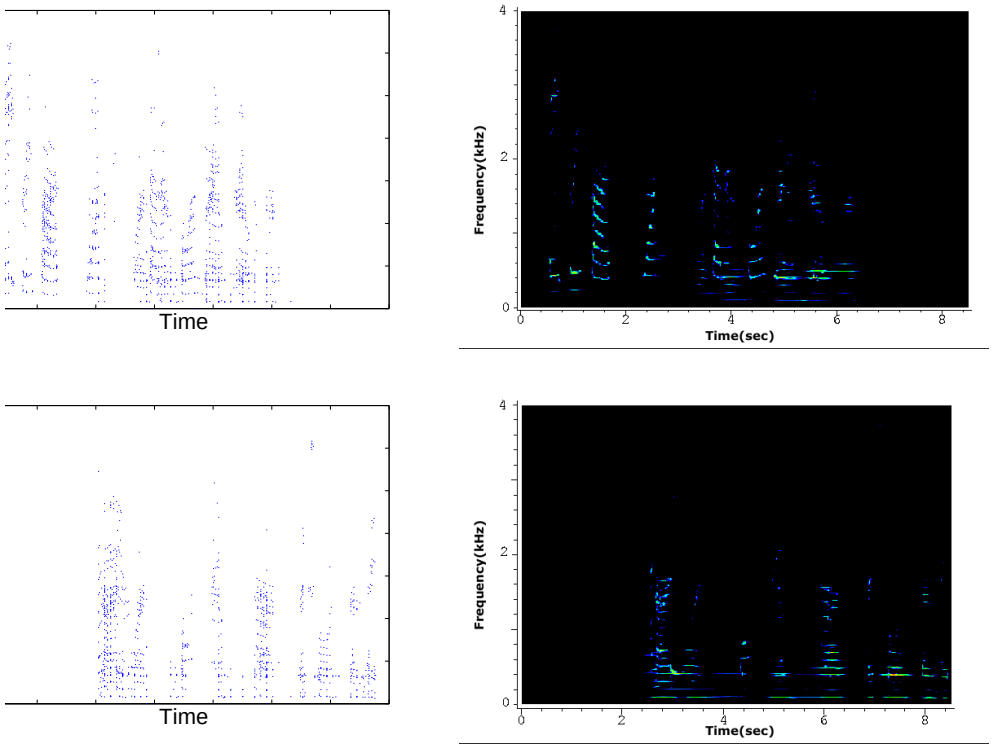


Figure 4.11: Reconstruction with LastWave software of the separated sequences obtained with the explained temporal analysis. Left column shows the centers of the audio atoms that are classified into the girl [up] and the boy [down]. Right column represents their respective reconstructions using the Lastwave software

4.3.3 Discussion

In this form, the temporal Blind Audio Source Separation of the sources is performed. The *Blind* concept is assured by the fact that the proposed model is not using any external information, such as the number of sources in the mixture.

Combining the temporal information present in the audio and video features, this method is capable to determine the number of speakers and their spatial situation, as

well as perform an acceptable source separation. And the best part, all these results are obtained with only a temporal analysis, without using the other information present in the Matching Pursuit decomposition of the audio.

Nevertheless, difficulties appear when the sources are active exactly at the same time. In these temporal periods we are not able to decide because we are not selective in frequency, all the information that we use is relative to temporal events.

Therefore, it is logical to think that the results obtained with the implemented method would be improved by adding some kind of frequency analysis.

The temporal separation of the audio signal we have already performed is the main advantage in our case. We can use the periods where we know there is only one source active to predict the sources behavior when they are mixed.

Chapter 5

Time-Frequency Analysis

The main achievement of the implemented method until now is to determine clearly the temporal periods where each one of the present sources is active. Therefore, the proposed algorithm automatically detects when the sources are alone and when superposed to another one.

The goal of the complementary analysis presented in this chapter lies in using the temporal periods where a single source is active to learn its frequency behavior. Then, we will try to predict this source evolution in those instants when more than one speaker contributes to the mixture. And, as a result, this frequency analysis will make possible to perform a better Blind Audio Source Separation.

Therefore, the idea is to establish, in temporal periods where only one speaker is active, a probability for each frequency f to belong to one or the other speaker. In other words, we **assign the frequencies** to a speaker, the other, or both of them taking into account the frequency values of the audio atoms in temporal periods with only one active source, and, as a result, a probability is obtained for each frequency f. Thus, combining the probability of one speaker in this domain with his probability in time, that is, the number of atoms of this speaker divided by the total, we can build a **map of probabilities**, where each point in the time-frequency plane has a probability of belonging to one or the other speaker. In temporal periods where both speakers are active at the same time, the proposed method decides to what speaker belong the audio atom according to this probability values.

5.1 Motivation

At this point, we have exploited the temporal information present in the sequence in order to have a first BASS approach. What is necessary to do now is to use the additional information that the Matching Pursuit decomposition of Lastwave provides.

Each audio atom has several parameters, but the most relevant are *timeID*, *coeff2* and *freqID*.

The most important parameter we have used until now is the *timeID* or temporal center of the audio atom. The implemented method is based on the temporal synchrony between audio and video events, and consequently this term has an important role. We use this information to determine the relationship between atoms of both modalities (audiovisual pairs), the most determinant phase in the whole process. Our method looks for peaks in video atoms displacement temporally close to this audio atom parameter.

Another relevant parameter for the proposed model is the *coeff2*, which provide us the information about the audio atom energy or, in other words, the importance of this atom in the decomposition of the whole sequence. We use this parameter to calculate the video atom *confidence value*, addition of the Matching Pursuit coefficient of the audio atoms related. Then, we build the cluster around the most important ones to estimate the coordinates of the video sources. So, this coefficient is mainly relevant concerning the determination of the source position in the image.

So far, we have not used the last parameter. Just like the first one, the *freqID* is the audio atom center, but now in the frequency domain. The audio atoms are situated in the time-frequency plane, and so these two coordinates are very relevant to characterize them. Therefore, the importance of this parameter consists in the addition of a second dimension in the analysis and, as a result, a new possibility of separation.

Thus, in order to perform the Audio Source Separation task, the goal is to separate the audio atoms of the sequence both in time and in frequency combining the time-frequency information. Then, this second dimension will aid us to obtain better separation results when the sources are temporally overlapped.

5.2 Frequency assignation

This step consists in using the audio atoms present in temporal periods where only one speaker is active to characterize the frequency axis, so that, using these observations, the proposed model can assign a probability to each frequency f of belonging to one or the other speaker.

This method is based on the hypothesis that the speakers have pitches differentiated, since, otherwise, their frequency assignation would be similar, and this second analysis vain. Another characteristic of the speech signal is that frequency components tend to move, they are not fixed in the pitch multiples. So, the frequency range of each component is quite big with this assignation, involving a big improvement on the performance when their pitches are well separated.

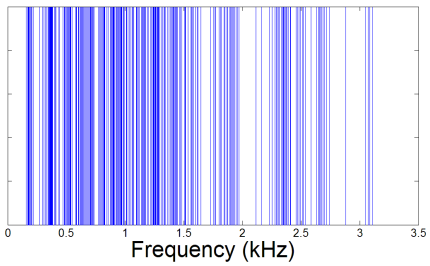
The proposed model consists of the following steps:

1. Extract the audio atoms in temporal periods with a single active source.

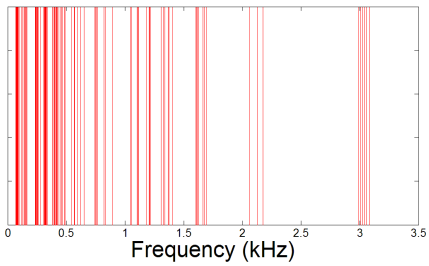
2. The central frequencies (*freqID*) corresponding to these audio atoms are assigned to the active source. Each frequency f has a number of atoms belonging to one speaker and another value for the other one.

3. The probability of each frequency f of belonging to each one of the speakers is computed as the number of atoms in this frequency that belong to one speaker divided by the total number of atoms.

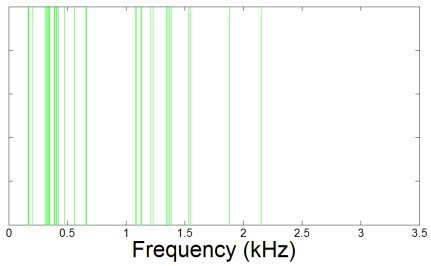
Figure 5.1 shows the frequency classification for a sequence with two speakers, one boy and one girl. The figure represents the frequencies with probability 1 for the boy and the girl in 5.1(a) and 5.1(b) respectively. Then, figure 5.1(c) shows the frequencies with value different from 1, that is, frequencies with audio atoms of both speakers in the temporal periods with only one source active.



(a) Speaker 1 (girl)



(b) Speaker 2 (boy)



(c) Audio atoms of both speakers

Figure 5.1: Example of frequencies assigned by the algorithm only to speakers 1 and 2 or with audio atoms of both of them in 5.1(c).

As expected, lower frequencies are assigned to the boy while higher ones are associated with the girl. This characteristic remarks the pitch frequency for both speakers. Another observation about the figure is that almost all the highest frequencies belong to the girl. It is also possible to see the formant periodicity for both speakers, with their frequency components situated in pitch multiples.

All these observations reinforce the theory that the temporal separation is well performed, correctly discriminating, thus, the temporal periods with only one source active. That allows to characterize clearly every speaker in the frequency domain.

Furthermore, in the third picture of 5.1, we can notice that frequencies belonging to both speakers are not very usual. Sequences with speakers with close pitches would mean more overlapping in the frequency classification and, consequently, the impossibility to perform a good separation in this domain using such a simple approach.

We have to take into account the following point: one frequency may belong to several sources. Even though different speakers have different pitches, frequency components tend to move in speech, and they may be overlapping fundamental frequencies of other speakers. That is the main reason of the probabilistic time-frequency analysis, since, if every frequency had belong only to one speaker, the probabilities would have been always 1 and the atoms assigned clearly into one or the other source.

5.3 Map of probabilities

Concerning the last step of the proposed model, the assignation of the audio atoms to one of the sources, first we have to consider in which of the following cases we are situated:

Time period with only one source active We use the temporal analysis result to classify this atom. We already know what source is active at this moment and so it is not necessary to use the frequency information.

Time period with several sources actives There is a mixture in this period, and also frequency analysis is required. Each audio atom in this period is classified into a source according to the probabilities of its coordinates in the time-frequency plane. Therefore, an audio atom with center in coordinates (t, f) is classified into source 1 if

$$P_1(t, f) > P_2(t, f) \tag{5.1}$$

otherwise it is classified into source 2.

This map of probabilities is built computing the product between time and frequency probabilities as following:

$$P_i(t, f) = P_i^T(t) \cdot P_i^F(f) \tag{5.2}$$

where $P_i^T(t)$ is the probability for an audio atom situated in the time index t of belonging to source i, and $P_i^F(f)$ is the probability for an audio atom situated in the frequency index f of belonging to source i. The probabilities are computed as:

$$P_i^T(t) = \frac{\#\ atoms_i\ in\ t}{\#\ atoms\ in\ t} \tag{5.3}$$

and

$$P_i^F(f) = \frac{\#\ atoms_i\ in\ f}{\#\ atoms\ in\ f} \tag{5.4}$$

A product is employed in equation 5.2 in order to penalize sources with low probabilities to belong to one source in time or in frequency. Therefore, if the analysis in the temporal domain or in the frequency one are sure, independently of the analysis in the other domain, the product favors this probability in front of others with less capacity of decision (more close to probability 0.5).

One consideration has to be taken into account. Not all the frequency values have a probability associated, only the values with some audio atoms in temporal periods where only one source is active. The closest frequency with probability is used in equation 5.2, in other words, the frequency probability of the audio atom is the same than that one calculated for the closest point in the frequency assignation.

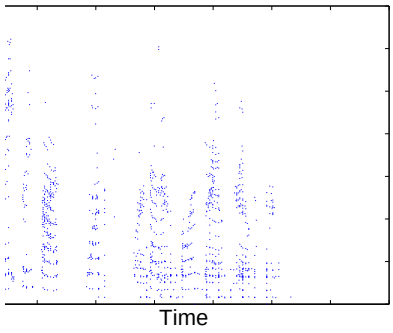
Fig. 6.5 show the comparison between video atoms resulting of temporal and frequency analysis in a real-world mixture with one boy and one girl speaking at the same time. Fig. 5.3 makes the same comparison, but now with the separated signals already reconstructed with the LastWave software [1].

Looking at this figures we can see the **improvements** introduced by the analysis in this second dimension of the spectrogram. The temporal analysis limitations are visible in the time period when the two sources are active exactly at the same time. This analysis relative to the synchrony between audio and video *relevant events* discover the periods were only one speaker is active with relative facility, but, logically, when there are audio atoms of both speakers in the same time instant, this technique can not make a decision, or it decides the speaker with more audio atoms in this moment. Thus, if we introduce the analysis in the second dimension of the problem, these atoms can be separated according to each speaker energy distribution in the frequency axe.

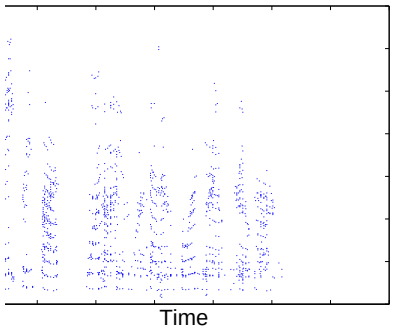
Therefore, comparing the temporal analysis alone, and after adding the frequency analysis, considerable differences are illustrated.

- Temporal analysis can not decide most of the time when the two sources are active at the same time, and, consequently atoms are assigned to both sources and their energy split between them. This classification is performed with the analysis in the second dimension, and consequently, the number of atoms classified into both speakers descend considerably.

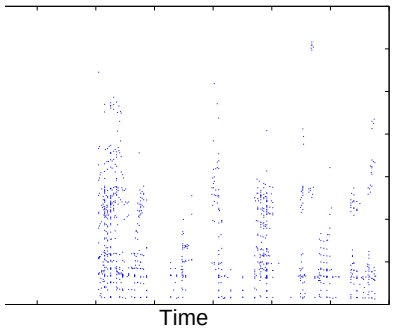
- Characteristic energy distribution in the time-frequency plane of each speaker is extracted correctly when we analyze the two dimensions of the problem, but not when only the temporal analysis is used. For example, in Fig. 6.4(b) we can see the separated signal for a girl. The first part of the soundtrack only contains her speech, so that it is possible to observe clearly her characteristic evolution of the frequency components. The same evolution is repeated in the period were the two are speaking at the same time with this time-frequency analysis(center of the spectrogram), but not in the temporal alone as shown in Fig. 5.3(a). Another aspect to remark is that low frequencies characteristics of boys are removed in the separation with time-



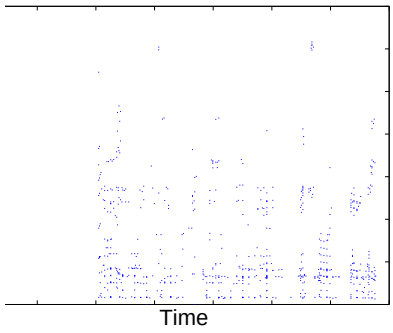
(a) Temporal analysis, speaker 1.



(b) Frequency analysis, speaker 1.



(c) Temporal analysis, speaker 2.



(d) Frequency analysis, speaker 2.

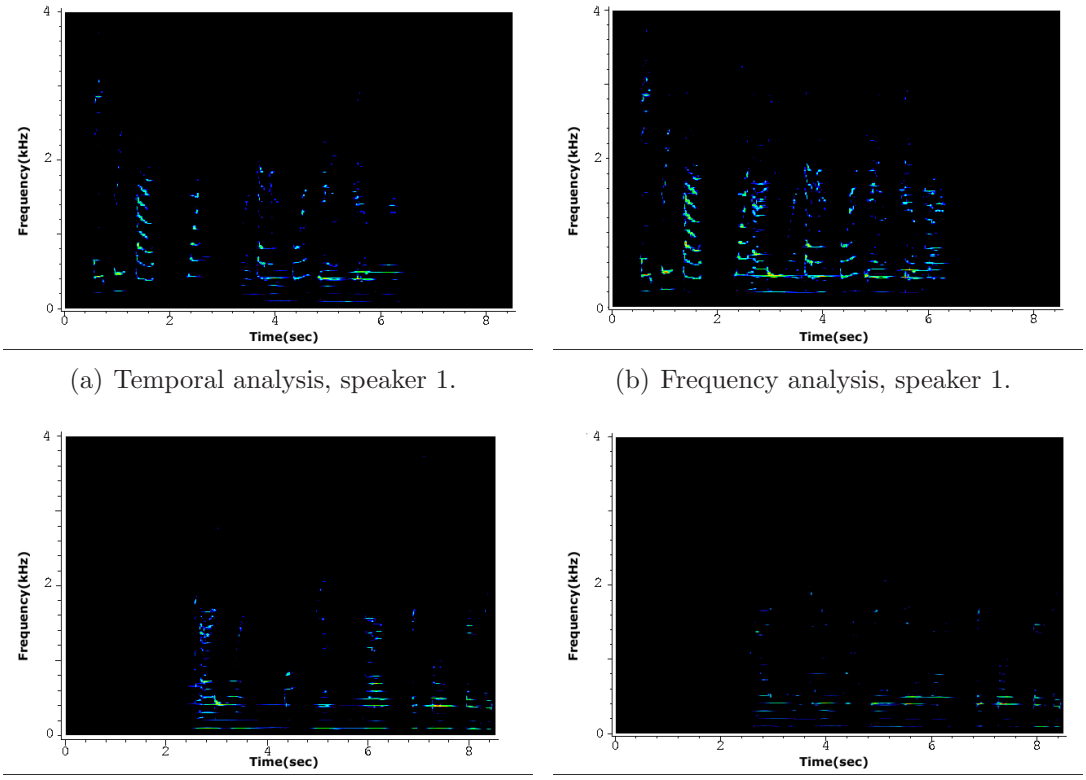
Figure 5.2: Comparison between video atoms resulting of temporal and frequency analysis in a real-world mixture with one boy and one girl speaking at the same time. The points are the centers of the audio atoms, which the algorithm estimates that they belong to this speaker.

frequency analysis, with and important improvement in the audible quality of the resultant soundtrack.

All this improvements demonstrate the necessity of this second analysis in the frequency domain, and its relevance in the achievement of a correct Blind Audio Source Separation, with reasonable auditory quality in the case of a mixture of a boy and a girl (more disjoint distribution of the energy in the time-frequency plane).

5.4 Discussion

This combined method is more robust than the temporal one alone. With a simple probabilistic time-frequency analysis we obtain better audible results and a good Blind Audio Source Separation.



(a) Temporal analysis, speaker 1.

(b) Frequency analysis, speaker 1.

(c) Temporal analysis, speaker 2.

(d) Frequency analysis, speaker 2.

Figure 5.3: Comparison between LastWave reconstructed sequences resulting of temporal and frequency analysis in a real-world mixture with one boy and one girl speaking at the same time.

Concerning the conditions to perform this frequency separation, clear temporal periods with only one source active are necessary to characterize the speaker. That is not a too restrictive requirement, since one long sequence with both sources active all the time simultaneously is a very unreal case. In this case, even if the observer were a human it would be necessary a lip reading mechanism.

The source separation that we perform is totally *blind* because, while other methods need previous knowledge as the number of sources, spatial situation of them, or training datasets to characterize the speaker, we obtain all this information combining audio and video features. Therefore, the temporal analysis already gives us these training sequences, without previous knowledge or suppositions about the speakers present in the mixture. We are performing a Blind Audio Source Separation aided by the visual component.

Some other difficulties appear when there is a frequency overlapping between the sources. As a result, sequences with boy and girl are preferred for the analysis, as usual in the Blind Source Separation research works.

More elaborated methods to separate in frequency could be:

- Apply Hidden Markov Models, then create masks and refiltering as in [26].
- Represent speakers by dynamic models and then refiltering as in [2]. Method CUES: Continuity of the speech signal, synchrony of movement of the frequency components (oriented filters), pitch, etc.
- Follow the temporal evolution of harmonics and resonances as in [25]

All these models could obtain better results, but it is necessary to remark that applying only a simple algorithm, it is already possible to obtain a good Blind Audio Source Separation. Implementing one of these laborious methods is not the object of this research work, the proposal is only to show the possible way to improve it.

Chapter 6

Analysis and Results

6.1 Introduction

The **Audiovisual Separation** task is divided into two main steps in the proposed model. The first objective is to *locate the speakers* present in an audiovisual sequence into the image, that is, the video separation part. This first goal is achieved by means of the creation of multimodal relevant structures, which are audio and video features synchronous and probably caused by the same physical event. The same structures that allow these localization provide us the information to perform the *Blind Audio Source Separation*, which is the second objective of this research work.

Results concerning these two main aims of the algorithm are presented separately in this chapter. Performance of the proposed approach is evaluated in CUAVE database clips [22] where two speakers, one boy and one girl, are active at the same time. Concerning the BASS, results are also evaluated in synthesized sequences generated from the same database, which allows the quantification of the results. The output sequences of the algorithm are compared with the original ones, where each speaker is isolated, and the percentage of correct atoms is computed. Another measure to evaluate the presented algorithm is the percentage of the original energy that these correct atoms represent, that is, the energy of the correct atoms assigned by the algorithm divided by the energy of all the atoms of the speaker.

In the next sections we describe, first, the characteristics of the sequences to analyze, and, next, the results obtained by this method concerning the two aims of the proposed model.

6.2 Test dataset: CUAVE database

Experiments have been carried out on clips taken from groups sections of the CUAVE database [22] where two persons in the scene, one boy and one girl, are speaking at the same time. First tests are performed in real-world sequences, extracted directly from this database, while the others are artificially synthesized from sequences where the the two talkers are speaking alternatively.

CUAVE database consists of sequences where one or two speakers (individuals and groups sections, respectively) are uttering different sequences of numbers in front of a camera. The video data was recorded at 29.97fps with a resolution of 480×720 pixels, and the audio at 44kHz.

The sizes of audio and video data have been reduced to allow a quicker processing. Therefore, concerning the **video** the dimensions of the images employed for this algorithm are 120×176 pixels. Applying the procedure described in section 3.3.1 the decomposition of the video signal into 2-D time-evolving atoms representing the scene evolution is obtained. For the **audio**, the signal is sub-sampled to 8kHz, with still a good audible quality. Using the Lastwave software [1], the mixture soundtrack is decomposed into 1-D atoms, which are situated in the time-frequency plane and represent the energy distribution of the audio sequence in this domain.

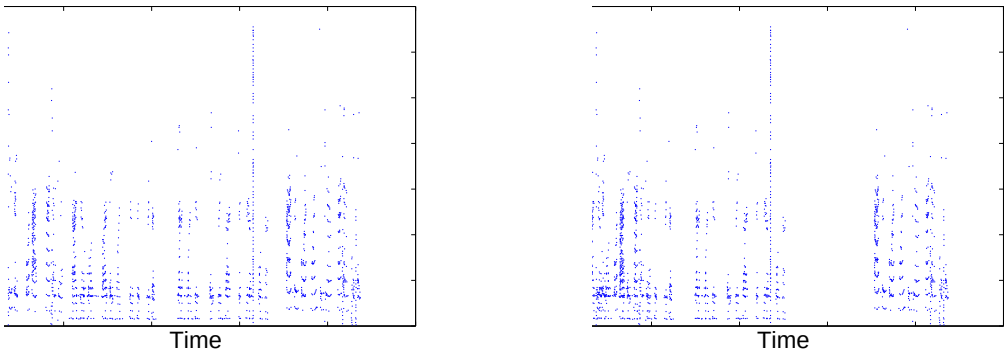
The input data of test 1 (real world clips) is obtained by choosing CUAVE audiovisual sequences where one boy and one girl speak simultaneously. The steps carried out to **generate the synthesized sequences** employed in test 2 are the following:

1. Choose a clip of the groups section of the CUAVE database where two speakers (boy and girl) utter numbers in turns.

2. Shift the audio atoms of one speaker so that their voices are overlapped part of the time. The Matching Pursuit decomposition of the audio [18] gives us the temporal situation 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. Fig. 6.1 illustrates an example of this procedure.

3. The same procedure is applied to the video atoms. After their decomposition in 2-D time-evolving atoms [10], the feature to analyze is the evolution of the video atoms displacement through time. In the CUAVE database, each speaker is situated 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 is 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 shift to the video.



(a) Clip g20 before shift. (b) Clip g20 after shift.

Figure 6.1: Shift applied to clip g20 of CUAVE database. The central part corresponding to a boy has been shifted 150 frames to the left, that is, by subtracting the same quantity in samples to the time index of all the audio atoms belonging to the boy.

This procedure shifts 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. Therefore, the performance of the proposed model can be compared with the original sequence, with the two speakers temporally separated.

6.3 Results concerning the Speaker detection task

Results are evaluated in clips of CUAVE database, with one girl and one boy uttering sequences of numbers. First, the girl speaks alone, then both of them at the same time, and, finally, the boy keeps on speaking while the girl has already stopped.

The first point to check is the performance of the proposed algorithm in **locating the sources** of the soundtrack into the image. Fig. 6.2 shows an example of the obtained results for sequence $g20$. The detected sources, that is, yellow points in the image, are close to the speakers mouth, and, as a result, this first objective is satisfactorily achieved.

Another goal is to **detect the *current* speaker**, that is, the speaker active in each moment. Results for clip $g20$ are illustrated in Fig. 6.3, with frames where only the girl or the boy are speaking. The correct speaker is reconstructed in both cases.

As in the previous research work [19] the speaker detection task is satisfactorily performed, detecting both, where are situated the sources in the image and in what moment

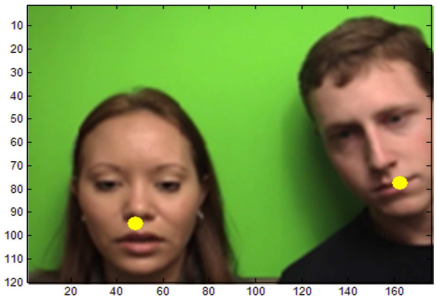
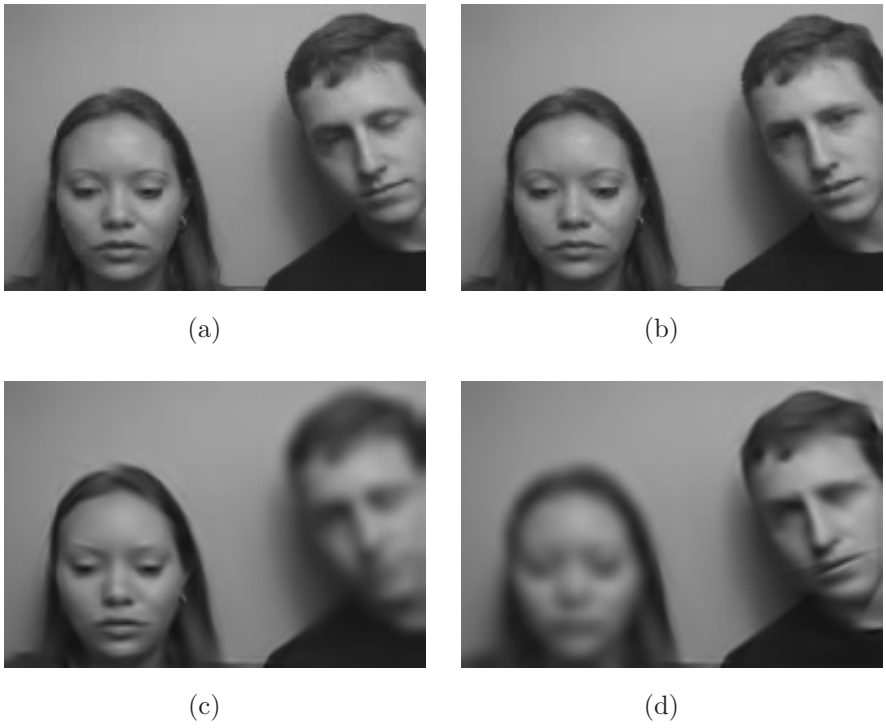


Figure 6.2: Results concerning the speaker localization obtained by analyzing clip $g20$ of CUAVE database.



(a) (b)

(c) (d)

Figure 6.3: Results concerning the current speaker detection for clip $g20$ of CUAVE database. Pictures (a) i (b) show two original frames of the sequence, where the current speakers are the girl and the boy respectively. (c) and (d) correspond to the frame reconstruction using only video atoms close to the estimated source.

they are active in an audiovisual sequence. This challenge is more difficult in our case, since in the analyzed mixture boy and girl are speaking at the same time. Therefore, we can conclude that this procedure is more robust thanks to the applied clustering, and also more precise since we associate video with audio atoms instead of audio energy feature.

6.4 Results concerning the Blind Audio Source Separation task

The performance of the proposed model in the BASS task is evaluated through two kind of tests:

- **Test 1** analyzes real-wold sequences extracted from the groups section of CUAVE database.
- **Test 2** is performed over synthetic sequences generated by following the steps described in section 6.2. The main advantage in this case is the possibility of quantify the results by comparing them with a known grown truth.

Results of both tests are presented in the next subsections, showing, for each one of them, the performance of the presented technique, with quantitative and qualitative evaluations. The percentage of correct atoms and the percentage of the original energy that these correct atoms represent are assessed in Test 2.

6.4.1 Test 1: Real world mixtures

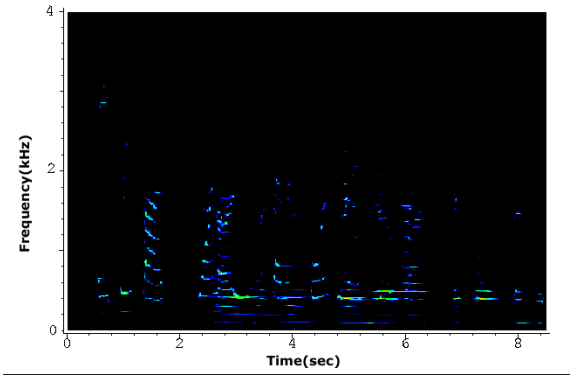
For this first test we clips of CUAVE database with one girl and one boy uttering sequences of numbers. In this sequences, there are the three possible situations represented: girl and boy speak alone and also both at the same time.

Qualitative results of analyzing a real-world mixture with the presented technique are shown in Fig. 6.4. This sequence corresponds to clip $g20$ of CUAVE database. Figures show the LastWave [1] reconstruction of the original sequence with all the atoms in the decomposition in 6.4(a), first, and then of the audio atoms assigned to each one of the speakers in the mixture, 6.4(b) and 6.4(c) for girl and boy respectively.

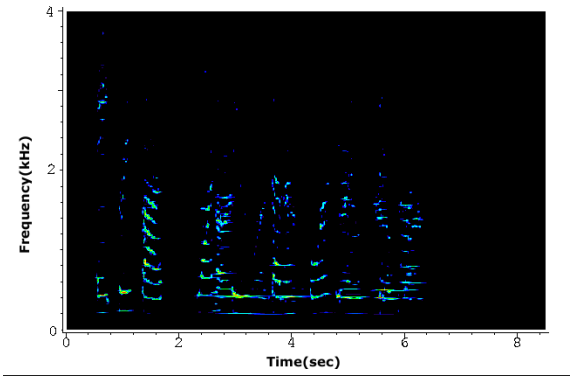
First **consideration** to do is that all the three situations are correctly interpreted and delimited by the proposed approach, with the energy assigned to the correct speaker if he is uttering the numbers alone or to both of them if their speech is overlapped.

Another point to remark is that the inherent structures in the time-frequency plane for both speakers are correctly detected. Therefore, it is possible to see the same structures repeated for the girl when she is speaking alone (2 first seconds in the spectrogram) and when the boy is also speaking: her pitch and frequency harmonics evolution are clearly represented. Concerning the boy, his structures are not so visible, neither in the period when he speaks alone, due to his lower energy. However, as expected, the lowest frequencies in the mixture are assigned to him and there are no presence of them in the girl spectrogram.

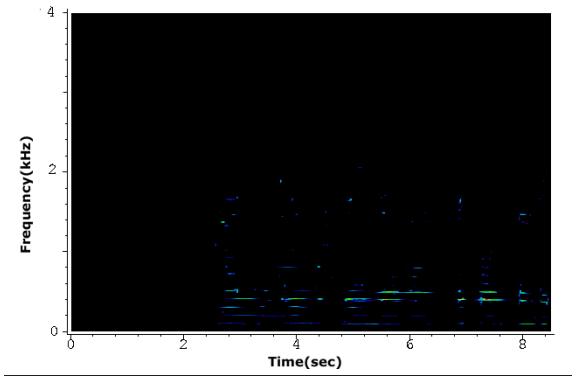
Thus, we can conclude that a satisfactory Blind Audio Source Separation is achieved by the algorithm in real-world mixtures, and with a good audible quality.



(a) Original sequence.



(b) Results speaker 1 (girl).



(c) Results speaker 2 (boy).

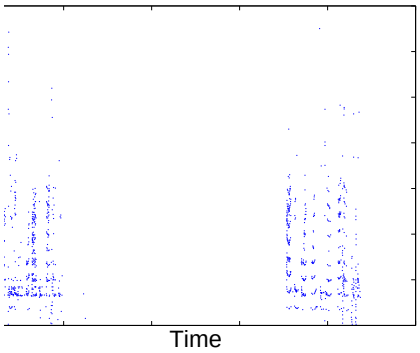
Figure 6.4: Results obtained for clip $g20$ of CUAVE database with the explained time-frequency analysis.

6.4.2 Test 2: Synthesized mixtures

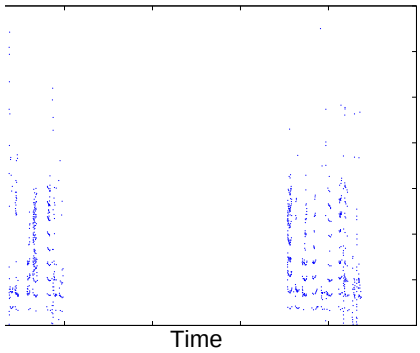
Synthesized sequences are generated using clips of CUAVE database, with one girl and one boy uttering sequences of numbers alternatively. The audio and the video atoms of one speaker are temporally shifted as previously explained in 6.2. In the resultant synthetic sequence, four cases are represented: both are speaking at the same time, only the boy or the girl and, the last possibility, a period of silence.

The interest of analyzing synthesized mixtures resides in the fact that **quantitative results** can be extracted. The features used to evaluate this technique are the percentage of correct atoms for each one of the speakers and the percentage of the original energy that these correct atoms represent.

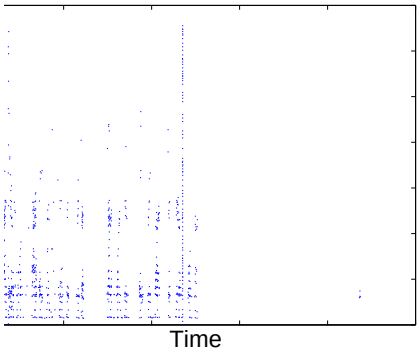
First, the **percentage of correct atoms** is assessed. Fig. 6.5 shows, for a syntectic sequence generated by shifting the boy 150 frames in clip $g20$ of CUAVE database, the speakers video atoms estimated by the algorithm at left and the real ones at right.



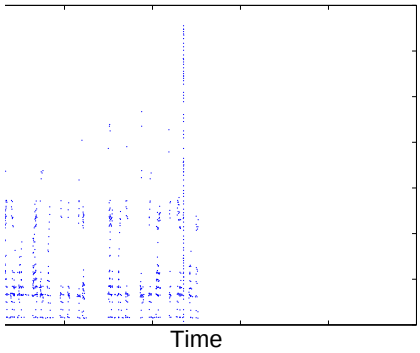
(a) Time-frequency analysis, speaker 1.



(b) Real video atoms belonging to speaker 1.



(c) Time-frequency analysis, speaker 1.

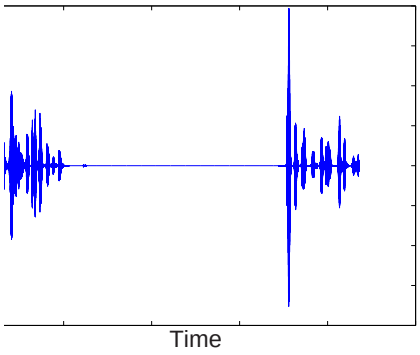


(d) Real video atoms belonging to speaker 2.

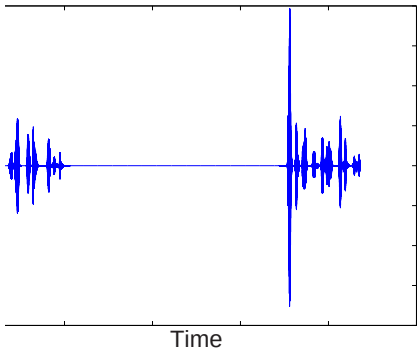
Figure 6.5: Comparison between video atoms resulting of time-frequency analysis in a synthetic mixture with the original ones. The points are the centers of the audio atoms, which the algorithm estimates that they belong to this speaker.

Results obtained by the proposed technique in this sequence are: 92% of correct atoms for the girl and 90% for the boy. This one is a good result taking into account that is at the atoms level that our algorithm is performed. Thus, in global, our algorithm classifies 91% of the audio atoms to the correct source.

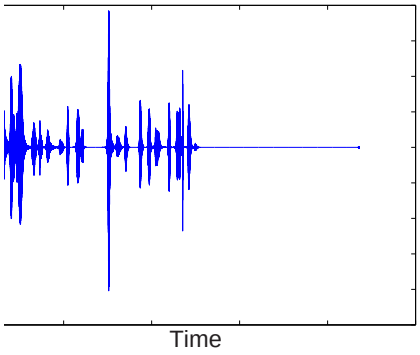
However, another measure is employed in order to evaluate this method: the **percentage of the original energy that these correct atoms represent**. This value gives us the information relative to the difference of the original and estimated soundtracks for each speaker after the reconstruction step. This measure is performed in order to discard the very improbable fact that the 9% of audio atoms that the algorithm classifies into the wrong source contribute to the separated soundtracks with the main part of the energy, that is, this video atoms are the first in the 1-D MP decomposition [18] of the original mixture.



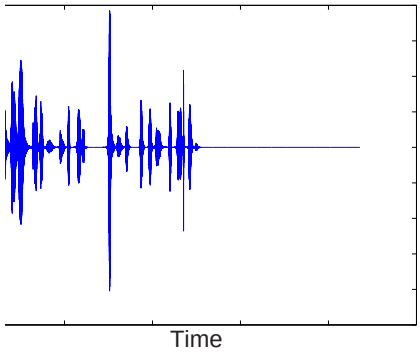
(a) Estimated soundtrack for speaker 1 (girl).



(b) Real soundtrack for speaker 1 (girl).



(c) Estimated soundtrack for speaker 2 (boy).



(d) Real soundtrack for speaker 2 (boy).

Figure 6.6: Comparison between estimated [left] and real[right] soundtracks for a synthetic sequence generated by shifting the boy 150 frames in clip $g20$ of CUAVE database

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.

Fig. 6.6 shows an example of reconstructed waveforms, the originals at right and the estimated by the proposed time-frequency analysis at left.

Waveforms are very similar in original and estimated sequences, and the percentage of the original energy that the correct atoms assigned to each source represent are of 92% and 86% for boy and girl respectively. These percentages are considerably high and similar of those obtained for the number of correct atoms assigned to each speaker (92% and 90%). Thus, we can discard the fact that incorrect audio atoms represent most of the

energy of each speaker separated signal.

Results obtained analyzing different sequences are summarized in table 6.1

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 6.1: Results obtained with syntethic sequences generated for different clips of CUAVE database.

Values obtained for the percentage of correct atoms and the percentage of energy that these atoms represent are similar. As a result, we can conclude that the algorithm has the errors distributed in audio atoms of all sizes, and the percentage of correct atoms is already a good measure of the algorithm performance.

Results are quite good, around 80-90% except for sequence $g12$ of CUAVE database, with a worse performance for the boy. Table 6.1 shows also that the results obtained are linked with the sequence to analyze and they are independent of the shift we introduce. The performance for sequence $g21$ is around 80% with a shift of 130 frames or one of 169, with a small difference in favor of the first one.

Therefore, we can conclude that the proposed method presents a new and successful approach for the Blind Audio Source Separation problem, with most of the atoms and energy assigned to the correct source.

Chapter 7

Conclusions

7.1 Conclusions

In this report, a new algorithm to perform Blind Audio-Visual Source Separation is presented. This method is based on the representation of both, audio and video signals, using sparse redundant dictionaries of functions. For the audio 1-D Matching Pursuit decomposition [18] is employed, representing the soundtrack as the addition of Gabor functions in the time-frequency plane. This representation provides, thus, the information relative to the energy distribution of the soundtrack in this plane, denoising also the input signal. Concerning the video, its most relevant structures are represented by a set of 2-D atoms and their changes tracked from frame to frame using the video Matching Pursuit algorithm proposed by Divorra and Vandergheynst [10].

The presented method uses the idea introduced in [19], that is, to look for the synchronous *relevant events* in both modalities in order to build relationships between them. The **innovation** now consists in assessing the synchrony between each audio and video atom instead of using only one audio feature (the energy) for all the soundtrack. Therefore, several video structures are related to the audio and selected by the algorithm, and the introduction of a spatial clustering is required in order to locate the sources in the image. The function of this clustering is to group the video atoms related to the audio into bigger structures representing the speakers and compute the center of this sources, the mouths in the speakers case. Another important innovation is the use of this audiovisual relationships to confront the Blind Audio Source Separation problem. Information present in video signal is employed to separate an audio mixture in the time-frequency plane. One soundtrack and the video signal associated are the only features used in this procedure, without the microphone arrays usually employed for the BASS task.

Several tests are performed in **real-world and synthetic sequences**, with encouraging results obtained for both of them. As in the previous research work [19],

the speaker localization is now successfully performed, with a more difficult challenge here, since the two speakers are active at the same time. Therefore, the presented model is more robust due to the clustering and more precise since we associate video with audio atoms. Concerning the second part of this audiovisual separation, and also the most difficult one, the audible quality of the separated audio signals is also quite good, with reconstructed waveforms close to the original ones.

Synthetic mixtures are generated in order to obtain quantitative results for the proposed model, since the analysis of real-world sequences gives us only a qualitative idea of the algorithm performance. However, we have noticed that performance of the algorithm in synthetic sequences is considerably better than in real ones. This effect is caused by the change in the speakers fundamental frequency, and, consequently, spectral harmonics, when they speak at the same time in the reality. Humans tend to change this speech parameters in order to differ more from the others speakers and to be, thus, easily heard. This change in their frequency behavior causes a worse performance of the algorithm, since the speakers model is learned in temporal periods where they are alone.

Despite of the good performance of the proposed approach, some **considerations** have to be taken into account.

First, the need of clear periods with only one source active to predict its behavior in the mixture. This requisite is not a very restrictive one, since it is not usual that in real-world mixtures the sources are active all the time, and, as a result, there are temporal periods with only one speaker where the learning of frequency characteristics is possible.

Another point to considerate is the characteristic frequencies of the speakers. In a mixture, if the two speakers are temporally overlapped, that is, they are speaking at the same time, and also in frequency, with very similar pitches, the separation of the mixture is almost impossible. For this reason, sequences with one girl and one boy are chosen, since their characteristic frequencies and their harmonics are more separated. Then, the frequency assignation performed in chapter 5 is more discriminative than in the case with two boys or two girls.

Finally, the last consideration to do is that frequency components of the speakers tend to move, taking thus a quite big band in the frequency axis for each harmonic. This characteristic causes that, even though the speakers pitches are different, frequencies are sometimes assigned to both speakers. However, this effect is solved with the introduction of frequency probabilities, that combined with the temporal ones provide a more robust assignation of the audio atoms to the correct speaker.

7.2 Future Work

The Blind Audio-Visual Source Separation presented in this report can be improved mostly in the Audio Separation field. The use of temporal and frequency probabilities to assign the atoms to one of the sources represents only a first and simple approach to all the possible ways to explore for the BASS goal. As already explained in the related work in this field, we can consider that we are performing a Single-Channel Source Separation, but aided with video information. Therefore, a possible future work is to apply some of the methods in this field in order to separate the mixture in the time-frequency plane. Some ideas to improve the proposed model in the future could be:

- To train factorial Hidden Markov Models (FHMM) in the temporal periods where one source is active in order to, later, compute binary mask functions through HMM and apply them to separate the mixture. This procedure is proposed in [26], but in our case it would be *blind* because the proposed model automatically detects this periods with only one speaker.
- To represent speakers by dynamic models and then apply them to the mixture as in [2]. Classical cues from speech psychophysics [5, 6] are used to define these models: pitch, continuity of the spectrum, synchrony of movement of the frequency components, etc.
- Another possible work in this field would be to track the evolution of harmonics and resonances through time due to their continuity in this domain. This idea is introduced in [25].
- Refined processing of mixed parts using audio dictionaries adapted to the speakers.

Bibliography

[1] J. Abadia, E. Bacry, and R. Gribonval. Matching pursuit software and documentation. `http://www.cmap.polytechnique.fr/bacry/LastWave/packages/mp/mp.html`.

[2] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In *NIPS*, 2004.

[3] 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*, 2:559–562, 1994.

[4] P. Besson, M. Kunt, T. Butz, and J. Thiran. A multimodal approach to extract optimized audio features for speaker detection. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, September 2005.

[5] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990.

[6] G. J. Brown and M. P. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–333, 1994.

[7] T. Butz and J. Thiran. From error probability to information theoretic (multi-modal) signal processing. *Signal Processing*, 85(5):875–902, 2005.

[8] T. Darrell and J. W. Fisher III. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.

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

[10] 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, IEEE, September 2004.

[11] C. Goodall. *M-Estimators of Location: an outline of the Theory*. Wiley series in probability and mathematical statistics. Applied probability and statistics, 1983.

[12] R. Gribonval, E. Bacry, and S. Mallat. Analysis of sound signals with high resolution matching pursuit, July 04 1996.

[13] J. Hershey and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *NIPS*, pages 813–819. The MIT Press, 1999.

[14] A. Hyvrinen. Survey on independent component analysis, June 04 1999.

[15] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NIPS*, pages 772–778, 2000.

[16] G.-J. Jang and T.-W. Lee. A probabilistic approach to single channel blind signal separation. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 1173–1180. MIT Press, 2002.

[17] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *CVPR*, pages 88–95. IEEE Computer Society, 2005.

[18] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

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

[20] H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. In Erwin M. Bakker, Thomas S. Huang, Michael S. Lew, Nicu Sebe, and Xiang Sean Zhou, editors, *CIVR*, volume 2728 of *Lecture Notes in Computer Science*, pages 488–499. Springer, 2003.

[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*, 2002(11):1189, November 2002.

[22] 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*, 2002(11):1189, November 2002.

[23] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

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

[25] 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.

[26] S. T. Roweis. One microphone source separation. In *NIPS*, pages 793–799, 2000.

[27] S. T. Roweis. Factorial models and refiltering for speech separation and denoising, September 10 2003.

[28] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *NIPS*, pages 814–820, 2000.

[29] P. Smaragdis and M. Casey. Audio/visual independent components. Technical report, Mitsubishi Electric Research Laboratories, April 2003. "Audio/Visual Independent Components", International Symposium on Independent Component Analysis and Blind Source Separation (ICA), pp. 709-714, April 2003, ICA 2003 (http://ica2003.jp/),.

[30] M. T. Wallace, W. D. Hairston G. E. Roberson, B. E. Stein, J. W. Vaughan, , and J. A. Schirillo. *Unifying multisensory signals across time and space*, chapter 158, pages 252–258. Experimental Brain Research, 2004.

[31] Y. Wang, J. Ostermann, and Y. Zhang. *Video processing and communications*. Prentice-Hall signal processing series. Prentice-Hall, pub-PH:adr, 2002.

[32] Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

[33] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical report, Mitsubishi Electric Research Laboratories, November 2001. "Understanding Belief Propagation and Its Generalizations", Exploring Artificial Intelligence in the New Millennium, ISBN 1558608117, Chap. 8, pp. 239-236, January 2003, Science & Technology Books ,.