SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
Gianluca Monaci, Oscar Divorra Escoda, Pierre Vandergheynst

ELD 241 (Bâtiment ELD)
Station 11
CH-1015 LAUSANNE

Tel: +41 21 6936874
Fax: +41 21 693 7600
e-mail: gianluca.monaci@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

MULTIMODAL ANALYSIS USING REDUNDANT PARAMETRIC DECOMPOSITIONS

Gianluca Monaci, Oscar Divorra Escoda and Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2004.24

October, 2004

Multimodal Analysis Using Redundant Parametric Decompositions

Gianluca Monaci, Oscar Divorra Escoda, Pierre Vandergheynst
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland
e-mail: {gianluca.monaci, oscar.divorra, pierre.vandergheynst}@epfl.ch

Abstract

In this work we explore the potentialities of a representational framework for the decomposition of audio-visual signals over redundant dictionaries, using Matching Pursuits [1] (MP). 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 from both 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 ignored. This is basically due to the complexity of the dependencies that exist between audio and video signals and to the signals representations that are considered when attempting to mix them in multimodal fusion systems. Redundant decompositions may describe audio-visual sequences in an extremely concise fashion, preserving good representational properties thanks to the use of redundant, well designed, dictionaries. We expect that this will help us to overcome two typical problems of multimodal fusion algorithms, that are the high dimensionality of the considered signals and the limitations of classical representation techniques, like pixel-based measures (for the video) or Fourier-like transforms (for the audio), that take into account only marginally the physics of the problem. The experimental results we obtain by making use of MP decompositions over redundant codebooks are encouraging and make us believe that such a research direction would allow to open a new way through multimodal signal representation.

Index Terms

Audiovisual fusion, multimodal data processing, sparse decompositions, Matching Pursuits, mutual information, Pearson correlation, Kendall correlation.

I. Introduction

Human perception of the world is essentially multimodal. We continuously combine different sensorial experiences to obtain an accurate and reliable representation of the surrounding environment, and this without any apparent effort. However, even our ability to correlate various input modalities is far from being perfect, and this because of the complex interaction between a number of factors that contribute to perception. In the audio-visual fusion context, for example, the *ventriloquism effect* emerges when sounds are erroneously located toward their apparent visual source. An even more impressive and curious effect is the McGurk illusion [2]: when dubbing an incongruous auditory syllable with a similar visual syllable the resulting percept is not either of the components, but a combination or a fusion of the auditory and visual components. It is thus obvious that automatic systems encounter enormous difficulties when trying to understand relationships between audio and video signals. Such a complex task requires the integration of a number of mathematical and statistical tools in order to represent the signals, extract meaningful features and fuse them. For this purpose most of the proposed algorithms make simplifying assumptions in order to model the audio-visual inputs and the complex relationships existing between them.

The problem we are study in this work is that of correlating audio tracks with video data to detect those regions in an image sequence from which the soundtrack originates. The topic was first faced by Hershey and Movellan [3], they measured the correlation between audio and video using an estimate of the mutual information between the

This work is founded by the Swiss NFS through the IM.2 National Center of Competence for Research.

Gianluca Monaci, Oscar Divorra Escoda and Prof. Pierre Vandergheynst are with the Signal Processing Institute (ITS) at the Swiss Federal Institute of Technology in Lausanne (EPFL). Web page: http://lts2www.epfl.ch.

energy of an audio track and the value of single pixels. Since a per-pixel measure was used, the hypothesis that pixels are independent of each other conditioned on the speech signal was introduced. In [3], the mutual information is derived from the Pearson correlation coefficient under the assumption that the joint statistics are Gaussian. Slaney and Covell [4] generalize this approach and look for a method able to measure the synchrony between audio signals and video facial images. In order to deduce a relationship between the cepstral representation of the audio and the video pixels, the authors use Canonical Correlation Analysis, which is equivalent to maximum mutual information projection in the jointly Gaussian case. Nock *et al*. [5] evaluate three different algorithms for assessing audio-visual synchrony in a speaker localization context, using different test sets up. Two of the considered methods are based on mutual information, one that assumes discrete distributions and the other one that considers multivariate Gaussian distributions. A third algorithm makes use of Hidden Markov Models trained on audio-visual data. Audio features are extracted from Mel-frequency cepstral coefficients, while different video features are tested, as the coefficients of the discrete cosine transform and the pixel intensity change. All three algorithms require training datasets in order to build *a priori* models, like the methods proposed in [3], [4].

Butz [6] proposes an approach based on Markov chains modeling audio and video signals. The audio-visual 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. For the audio, a linear combination of the power spectrum coefficients that has the biggest entropy is learnt from a dataset, while the video is represented by pixel intensity change. The audio and video joint densities are deduced by training the estimator on a set of audio-video sequences. A method that does not make use of any previous model training was first proposed by Fisher *et al.* [7] 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 general approach is used to solve a conversational audio-visual correspondence problem, obtaining encouraging results. A slightly different approach is used in [9], where is presented a methodology not only for correlating audio and video, but for extracting audio-visual independent components from video streams. 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-visual subspaces. However, this technique is not able to deal with dynamic scenes.

In this work, we explore a completely new representational framework for audio-visual fusion. This is based on the sparse decomposition of signals over atoms dictionaries using Matching Pursuits [1] (MP). There are several motivations for using such a description for our signals. The MP decomposition provides a very sparse representation of the information, allowing a considerable reduction of the dimensionality of the input signals. At the same time, an appropriate decomposition of a signal over a well designed redundant dictionary provides an interpretation of the information in terms of the most salient structures present in the signal. This should allow us to handle information in an easier and faster way, and thus to develop relatively simple and intuitive, but effective, fusion criteria. Moreover, we want to underline that, by representing the video using image structures (atoms) that evolve in time, we deal with dynamic features that have a true geometrical meaning, that is not the case when using pixel-based representations.

In order to combine audio and video representations, we are going to test three "classical" measures of correlation, mutual information, Pearson correlation and Kendall τ correlation. Surprisingly, we have obtained the most satisfactory results using the Pearson coefficients. Such results show that our technique allows the detection of the image zones that originate the audio signals.

The paper is structured as follows: in Section II, the representational framework for audio and video signals is introduced. Section III presents the different audio-visual fusion criteria that have been tested and Section IV shows the experimental results. In Section V the experimental results are discussed, and in Section VI conclusions are drawn and possible future extensions are depicted.

II. Audio and Video Representation

In this work, we consider multimodal sequences composed of an audio track together with its corresponding video component. Both audio and video signals are represented using a Matching Pursuit decomposition over redundant dictionaries. This kind of atomic decomposition seems, in fact, particularly suitable for representing and correlating audio-visual inputs. Indeed, we obtain a concise description of such signals that is explicitly related to the physical phenomena they are originated from. In Section II-A, the procedure used to represent 1-D signals using MP with

redundant codebooks of atoms is described in details, while in Section II-B we will present the techniques that have been developed to extend the MP algorithm to the complex case of video sequences.

A. Audio Decomposition

The audio signal $a(t)$ is decomposed using the MP algorithm over a redundant dictionary $\mathcal{D}_{\mathcal{A}}$, composed of unitary norm base functions called atoms. The family of atoms that compose $\mathcal{D}_{\mathcal{A}}$ is generated by scaling by s, translating in time by u and modulating in frequency by ξ a generating function $g(t) \in L^2(\mathbb{R})$. Indicating with the index γ the set of transformations (s, u, ξ), an atom can be expressed as

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

In our case, we consider a dictionary of Gabor atoms, that is, the generating function $g(t)$ is a normalized Gaussian window. The choice of a Gabor dictionary is due to the optimal time-frequency localization of the Gaussian window.

The first step of the MP algorithm decomposes a as

$$a = \langle a, g_{\gamma_0}\rangle g_{\gamma_0} + R^1 a\,, \tag{2}$$

where $R^1 a$ is the residual component after approximating a in the subspace described by g_{γ_0}. The function g_{γ_0} is chosen such that the projection $|\langle a, g_{\gamma_0}\rangle|$ is maximal. This procedure is recursively applied, and after N iterations the signal a is represented as

$$a = \sum_{n=0}^{N-1} \langle R^n a, g_{\gamma_n}\rangle g_{\gamma_n} + R^N a\,, \tag{3}$$

where $R^0 = a$ and $R^n a$ is the residual after n iterations. One of the analyzed signals together with its MP decomposition represented in the time-frequency plane are shown in Fig. 1.

B. Video Decomposition

The image sequence is represented using the algorithm proposed by Divorra and Vandergheynst [10]. This technique decomposes a sequence into a set of 2-D atoms evolving in time, allowing to represent salient geometric components present in the sequence and to track their temporal transformations.

An iteration on the MP algorithm decomposes the first frame of the sequence over a redundant dictionary $\mathcal{D}_{\mathcal{V}}$ of 2-D anisotropic atoms [11]:

$$I = \sum_{\gamma_i \in \Omega} c_{\gamma_i} g_{\gamma_i}\,, \tag{4}$$

where i is the summation index, c_γ corresponds to the projection coefficient for every atom g_γ and Ω is the subset of selected atom indexes from dictionary $\mathcal{D}_{\mathcal{V}}$. The changes suffered from a frame I_t to I_{t+1} are modelled as the application of an operator F_t to the image I_t such that $I_{t+1} = F_t(I_t)$ and $I_{t+1} = \sum_{\gamma_i \in \Gamma} F_t^{\gamma_i}(c_{\gamma_i}^t g_{\gamma_i}^t)$, where F_t represents the set of transformations F_t^γ of all atoms that approximate each frame. 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 parametric transformation F_t). However, at every greedy decomposition iteration some new criteria have to be considered in order to establish the relationship with the expansion of the reference frame. Only a subset of functions of the general dictionary is considered as candidate functions to represent each deformed atom. This subset is defined according to the past geometrical features of every particular atom in the previous frame, such that only a limited set of transformations (translation, scale and rotation) are possible.

The simple constraint of limiting possible atom transformations, and the simplicity of dictionary functions [11] turns into a lack of regularity (stability) of the atom motion. In order to include in the MP algorithm a regularity measure, a more flexible version of the selection criteria is considered (Weak Greedy Algorithm -WGA- [12]). Instead of selecting the function giving the biggest scalar product at every iteration, we select the most probable function with respect to a certain motion. The selection of the atom that gives the maximum scalar product is equivalent to select the most probable atom given that all transformations have equal *a priori* probability. However, in the case of smooth motion,

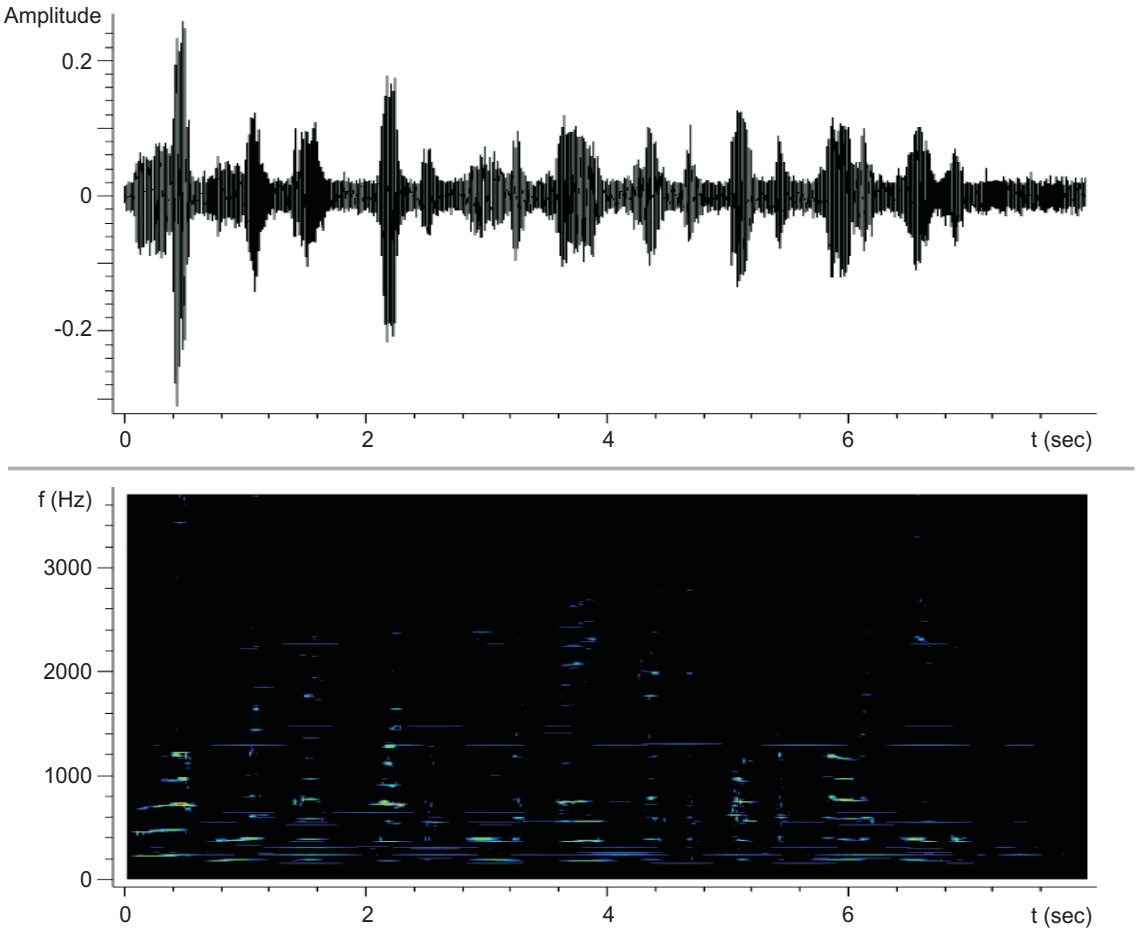

Fig. 1. Audio signal of a subject uttering the first ten digits in Italian (top) and its MP decomposition with 800 atoms represented in the time-frequency plane (bottom). The color map of the image goes from black to red, through blue, green and yellow, and the pixel intensity represents the value of the energy at each time-frequency location.

there will be a lot of transformations that are unlike and even impossible. A Bayesian modeling of the problem can be performed if some *a priori* information or knowledge about the parametric sequence description is available. The assumption that in the sequence approximation, neighboring atoms present regular motion is made, since several of them are needed to represent a region. In the greedy formulation, a Bayesian functional that maximizes the Maximum *a Posteriori* probability integrates the motion regularity assumption. A Markov Random Field (MRF) framework is considered, in order to define probabilistic relations among atoms. The formulation of the Bayesian approach to MP video representation is complex and is treated in detail in [10], [13], to which the interested readers are referred.

A cartoon example of the presented approach can be seen in Fig. 2, 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 reconstruction of the approximation. The bottom part of the figure shows the parametric representation of the sequence. We see the temporal evolution of the coefficient, the coordinates evolution of the translation parameters and the scales and angle evolution.

III. Audiovisual Fusion

The MP decompositions of audio tracks and video sequences represent some salient parametrization of those signals. Thus, a quite natural and, as it will be shown, effective way to relate audio and video sequences is that of comparing these parametric representations. The audio-visual features considered in the following of this work are presented in Sections III-A and III-B, while the criteria that are used to relate them are introduced and discussed in Section III-C.

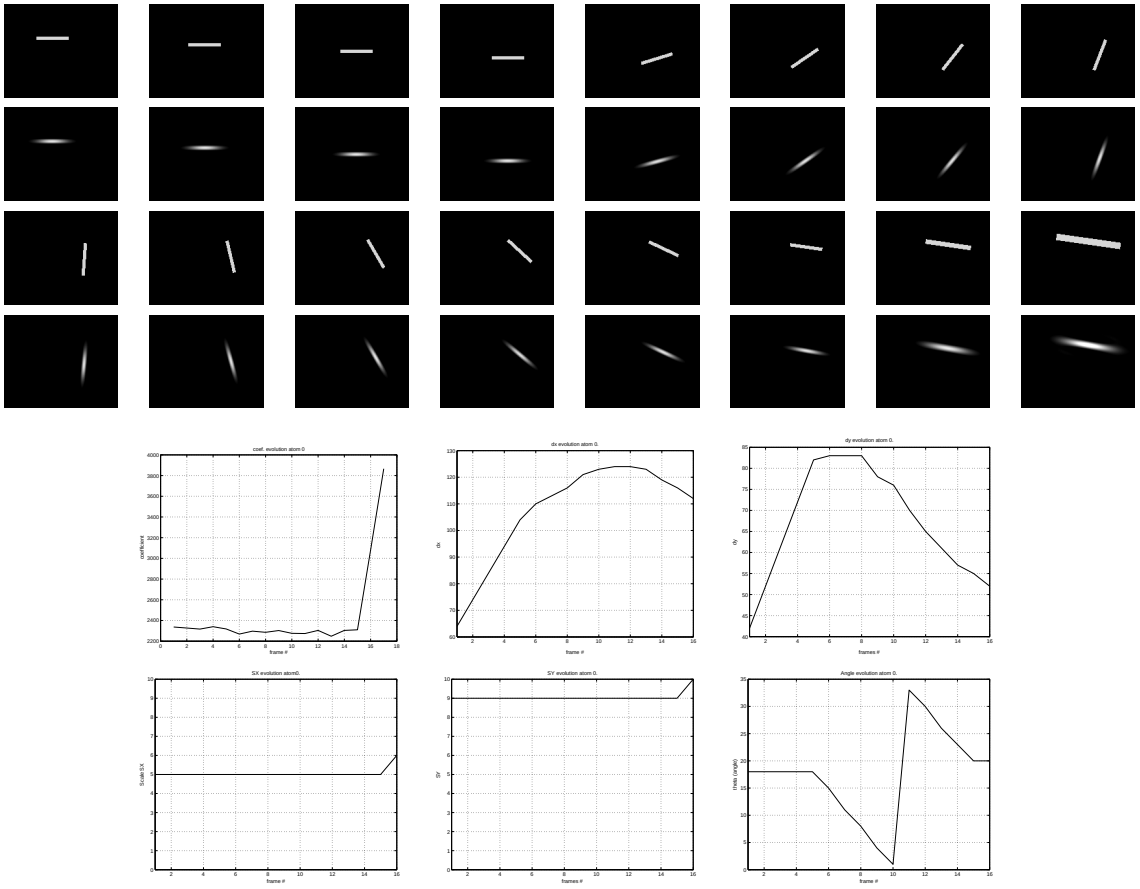

Fig. 2. Synthetic sequence approximated by 1 atom (top) and evolution over frames of the parameters describing such atom (bottom). From left to right and from up down, we find: amplitude of the coefficient, horizontal position, vertical position, x (short axis) scale, y (long axis) scale, rotation.

A. Audio Feature

The audio representation that we obtain from the MP decomposition it is not directly exploitable to our end and has to be further processed in order to obtain a function that is comparable with the evolution of the video parameters. We require a signal composed of the same number L of samples of the MP video parameters. Moreover, we would like to depict the audio signal with only one time-evolving feature, in order to speed-up the computation and to simplify the problem formulation. Typical features used to represent audio signals are the Mel-frequency cepstral coefficients (MFCCs) [14], that the dominant features used for speech recognition, and employed in [4], [5]. In [6] the audio feature is obtained from the spectrogram of the audio track by learning from a training dataset the linear combination of the power spectrum coefficients with the biggest entropy. Fisher and Darrell [8] propose a similar feature that maximizes the mutual information with the video. This uses an on-line procedure that does not require a training process. In both cases ([6] and [8]), the final feature is a 1-D function that is downsampled in order to obtain the same length for the audio and the video features.

In this work, we decided to use a much simpler approach exploiting the sparseness of the MP decomposition. Our audio feature is obtained by projecting over the time axis, the time-frequency representation obtained with MP (see Fig. 1). In fact, our feature is similar to those described in [6], [8], with the difference that we attribute to each frequency component the same weight. This can be seen as a representation of the frequency content that is present at each time instant. This is of course a very simple approach, and one of the first extensions of this work will be the conception of some more accurate criteria to select the audio feature. However, the sparseness and the fine time-frequency resolution of the dictionary decomposition allow us to obtain a description that captures nicely the evolution of the audio track. We show in Fig. 3 the audio feature obtained for the signal of Fig. 1. The values are normalized to a maximum of 1 and the mean value has been removed.

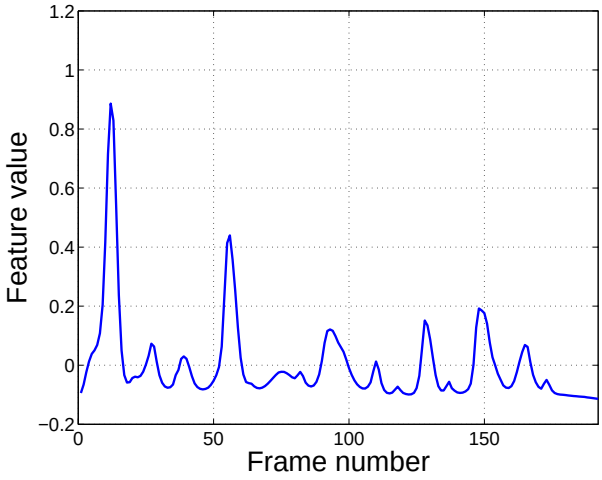

Fig. 3. Audio feature representing the signal of Fig. 1. The values are normalized to 1 and the mean has been removed.

B. *Video Features*

When considering the video signal, basically all the works we reviewed use pixels intensity values as video features, with the exceptions of [5] and [6], where the pixel intensity change is measured in a 3×3 averaging spatial window is considered. 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. We have decided thus to explore the possibilities offered by the MP video decomposition technique presented in Section II-B. 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 atoms 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. 4 shows the atom parameters evolution as a function of time. The described video sequence, in this case, is of length $L = 192$ frames.

The video features we consider, however, are not all the 6 video parameters. The scale parameters have been discarded, since they carry few information about the mouth movements. Moreover, the atom orientation appears to be a component that brings an unprecise description of the real geometric feature rotation and thus results to be unreliable. We have decided, therefore, to employ only the atoms coefficient and positions as video features, obtaining 3 descriptors per atom. Since a video sequence is represented with a fixed number N of time-evolving atoms, we end up with a list of $3 \times N$ functions composed of L samples. The video features that we compare with the audio feature are, in addition, normalized in amplitude and with zero mean.

C. *Fusion Criteria*

The way audio and visual features are correlated is a critical point in the processing. The effectiveness of the adopted criterion depends on how data is processed before this step and how the output of the fusion step is exploited. In this study, we want to explore the capabilities of MP decompositions to represent meaningful audio and visual structures, in order to correlate them and identify those visual primitives that originate the audio signal. Thus, starting from the atomic representations obtained using the procedures described in Section II, we want to detect those 2-D time-evolving atoms that are more correlated with the representation of the soundtrack. To do this, three different correlation measures have been tested. They are described in the next sections.

1) Pearson ρ correlation coefficient: The most common measure of correlation is the Pearson correlation coefficient [15]. The Pearson correlation is a parametric measure of correlation and reflects the degree of linear relationship between two variables that are on an interval or ratio scale. It ranges from $+1$ to -1. The observations for both variables should be approximately (bivariate) normally distributed. Given two observation vectors X and Y of length n,

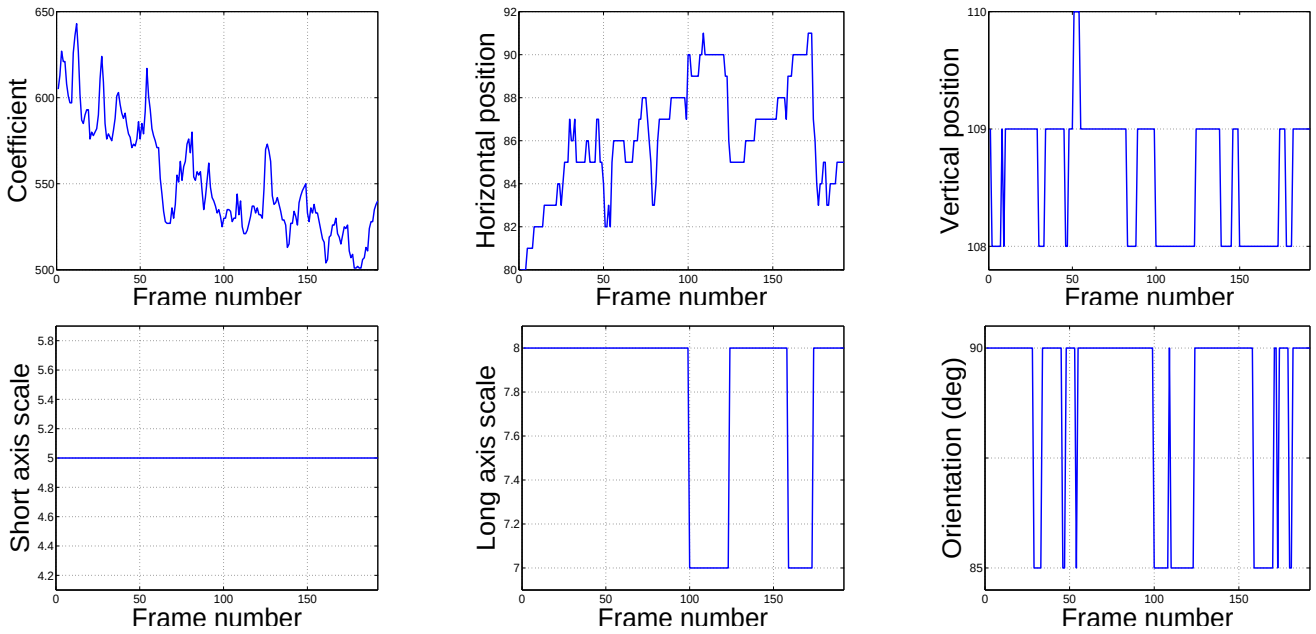

Fig. 4. Temporal evolution of the parameters of an atom used to decompose the video sequence corresponding to the audio track of Fig. 1. From left to right and from top to bottom: Amplitude of the coefficient, horizontal displacement, vertical displacement, short axis scale, long axis scale and rotation. Only the parameters depicted in the first row (coefficient and positions) are considered as video features and fused with audio.

the value of the Pearson correlation coefficient $\hat{\rho}$ between X and Y is computed as

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}\,, \tag{5}$$

where $\overline{X}$ and $\overline{Y}$ denote the mean values of X and Y respectively.

For each video feature, the value of the correlation $\hat{\rho}$ with the audio feature is computed. A probability p associated to the correlation coefficient is also computed, in order to assess the significance of the value of $\hat{\rho}$. When the true correlation is zero, the quantity

$$\hat{t} = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}\,, \tag{6}$$

belongs to a *Student's* t distribution with $n-2$ degrees of freedom, $t(n-2)$, with n number of available samples [15]. If the probability p that $\hat{t}$ belongs to $t(n-2)$ is small, then the correlation is significant. We want to remark that each atom is described by 3 quantities evolving in time. Hence, for each atom we have 3 correlation values. We select those video atoms that have all 3 correlation coefficients $\hat{\rho}$ with probability $p < 0.025$.

2) Kendall τ correlation coefficient: Kendall τ [16] is a non-parametric measure of correlation, which is intended to measure the strength of the relationship between two processes. Let (X_i, Y_i) and (X_j, Y_j) be a pair of (bivariate) observations. If $X_j - X_i$ and $Y_j - Y_i$ have the same sign, we shall say that the pair is concordant. On the other hand, if they have opposite signs, we shall say that the pair is discordant. In a sample containing n observations we can form $n(n-1)/2$ pairs corresponding to choices $1 \leq i < j \leq n$. Let C stand for the number of concordant pairs and D stand for the number of discordant pairs. Then, a simple way to measure strength of relationship is to compute $S = C - D$. A preponderance of concordant pairs, resulting in a large positive value of S, indicates a strong positive relationship between X and Y, while a preponderance of discordant pairs resulting in a large negative value of S, indicates a strong negative relationship between X and Y. As a measure of relationship strength, S has a disadvantage. Its range depends on the sample size n. But a simple normalization gets around this problem. Since S can vary between $-n(n-1)/2$ and $+n(n-1)/2$, we can therefore compute

$$\tau = \frac{2(C-D)}{n(n-1)}\,, \tag{7}$$

having always $-1 \leq \tau \leq 1$. Kendall correlation coefficient τ can measure linear and non-linear relationships, being at he same time robust to outliers.

We compute the value of τ between each video atom parameter and the audio feature, ending up with 3 values of τ per atom. A correlation index T is computed for every time-evolving atom, as the average of the absolute values of the 3 correlations τ. The bigger is the value of T, the more correlated is the time-evolving atom with the audio component. We consider as most correlated with the audio signal, the 10 video atoms that are characterized by the 10 biggest values of the index T.

3) Mutual information: The mutual information [17] is a general measure for statistical independence that quantifies the reduction in the uncertainty of one random variable given knowledge about another random variable. Considering two random variables X and Y with possible outcomes Ω_X and Ω_Y and with marginal probability densities $P(x)$, $P(y)$ and joint probability density $P(x, y)$, the mutual information, $MI(X, Y)$, of X and Y is defined as

$$MI(X,Y) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} P(x,y) log\Big(\frac{P(x,y)}{P(x)P(y)}\Big). \tag{8}$$

An estimate of the mutual information is computed by estimating the probability densities using a Parzen window technique [18], [19]. The Parzen density estimate is defined as

$$\widehat{P}(\nu) = \frac{1}{n} \sum_{i=1}^{n} K(\nu - \nu_i), \tag{9}$$

where $K(\cdot)$ is the Gaussian kernel and n is the available number of samples. The generalization of Eq. 9 to the multivariate case is obtained by simply substituting the 1-D Gaussian kernel with its multidimensional analogue.

Again, for each atom we have 3 mutual information values of the audio and the atom parameters. Thus, a mutual information index M is calculated for every primitive as the average of the mutual information coefficients. In the results we show in this report, the video atoms that are considered the most correlated with the soundtrack are the 10 with biggest information index M.

IV. EXPERIMENTS

The framework we have developed is used to detect the region in the video that contains the speaker that originates the corresponding audio signal. Such an application can be directly included in a conversational human-machine interface, on which one or more persons interact with a computer just by speaking in front of a camera, or in a smart video-conference system.

Experiments have been carried out on real-world video streams representing one or two persons speaking and moving in front of a camera. The video data was recorded at 25 frames per second at a resolution of 144×176 pixels. The input soundtrack was collected at 48 kHz and it was sub-sampled in order to obtain a signal at 8 kHz. All the test sequences last about 8 seconds, and are thus approximately 200 frames long.

The image sequence is represented using the procedure described in Section II-B. Thus, the video frames are high-pass filtered and decomposed using the MP algorithm of Divorra and Vandergheynst, obtaining a set of 2-D time-evolving atoms. The audio part is decomposed over a dictionary of Gabor atoms whose window lengths range from 512 to 16384 time samples, using the *LastWave* implementation of MP for 1-D signals [20]. Based on such decomposition, the audio feature is extracted as described in Section III-A. The number of basis functions used for the decomposition of the image and audio sequences is heuristically chosen for these experiments, in order to get convenient representations. However, note that a distortion criteria can be easily set, to automatically determine the required number of atoms. In Fig. 5, we show the results of the described procedure applied to the test video sequence *Elena 1*. The sequence shows a person repeating two simple words for 13 times and it lasts slightly less than 8 seconds, that is 192 frames. During the utterance of the phonemes, there is no significant movement on the scene. The subject speaking is close-up filmed; The frame number 15 of the video sequence is shown in Fig. 5(a). The video component is decomposed using 120 basis functions, while the audio track is represented with 500 Gabor atoms. In picture Fig. 5(b), the absolute value of the atoms obtained by correlating the video stream and the audio signal using Pearson correlations fusion criterion (Section III-C.1) are shown. At the top figure, results for the video sequence *Elena 1* correlated with

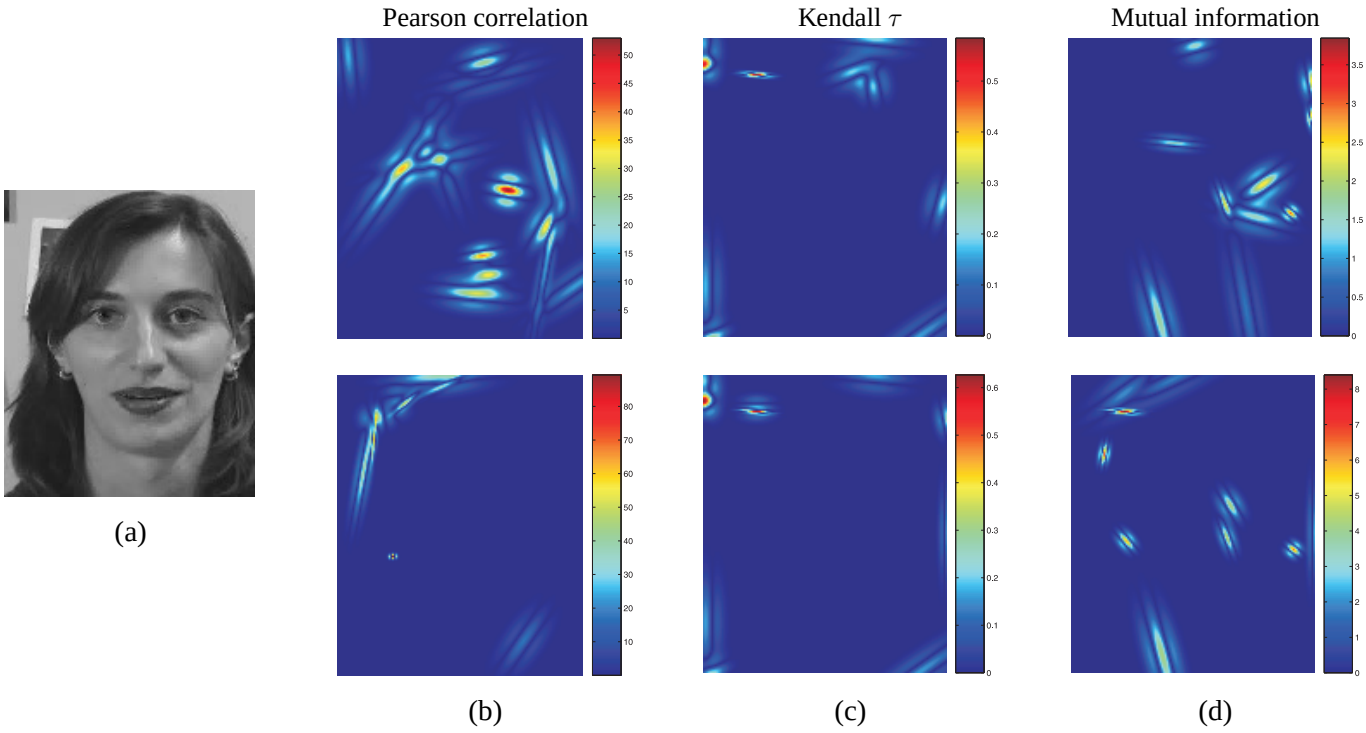

Fig. 5. Experiments on the sequence *Elena 1*; Frame 15 of the original sequence is shown in (a). In picture (b), results using Pearson correlations are shown. At the top figure, results for the video sequence *Elena 1* correlated with its audio component are depicted. At the bottom, we show the resulting atoms when the video sequence *Elena 1* is correlated with the audio signal of the sequence *Elena 2*. In (c) we show the results using the τ correlation coefficients as fusion criterion, when using the correct audio (top), and the one from sequence *Elena 2* (bottom). Finally, in (d) results for mutual information with correct (top) and incorrect (bottom) audio signals are depicted.

its audio component are depicted. At the bottom, we show the resulting atoms when the video sequence *Elena 1* is correlated with the audio track of the sequence *Elena 2*. The atoms are weighted by the coefficients computed by the MP algorithm. In Fig. 5(c) we show the results using the τ correlation coefficients as fusion criterion (Section III-C.2), when using the correct audio (top), and the one from sequence *Elena 2* (bottom). Each video atom is weighted with its correlation index T, therefore the image values are proportional to the strength of the correlation between the image areas and the soundtrack. Finally, in Fig. 5(d) results for mutual information fusion criterion (Section III-C.3) with correct (top) and incorrect (bottom) audio signals are depicted. The atoms are weighted with their mutual information index M, thus the picture intensity values are proportional to the strength of the correlation between the image zones and the audio. In this experiment, the Pearson correlation measure clearly outperforms the other fusion criteria, it is the only one that is able to clearly distinguish between correct and incorrect audio and it is capable of detecting the speaker's mouth (see Fig. 5(b) top).

The same type of results are depicted in Fig. 6 for the sequence *Elena 2*. The same person of the previous sequence utters the digits from 1 to 10 in Italian, and the stream is again 192 frames long. The sequence is static and the subject is filmed close to the camera. The frame number 10 of the video sequence is shown in Fig. 6(a). We represent the image sequence using again 120 video atoms, while the audio is represented with 800 Gabor functions, since the speech track in this case is more complex than the previous one. The audio component of this sequence, together with its MP decomposition, is depicted in Fig. 1. In picture 6(b), results using Pearson correlation for correct audio (top) and the audio track of *Elena 1* (bottom) are depicted, in (c) we can see the results using the τ correlation coefficients and in (d) using mutual information. Again, the best performances are achieved exploiting Pearson correlation coefficient.

A series of results on the sequence *Elisa* are shown in Fig. 7. This example is more challenging, since the subject moves the head while speaking and she pronounces a more articulated and complex phrase. In this case, we want to remark that using the mutual information criterion, we are able to detect the speaker's mouth, but the discrimination between correct and incorrect audio track is extremely poor. The original sequences, together with the resulting video sequences, are available on the author's web page [21].

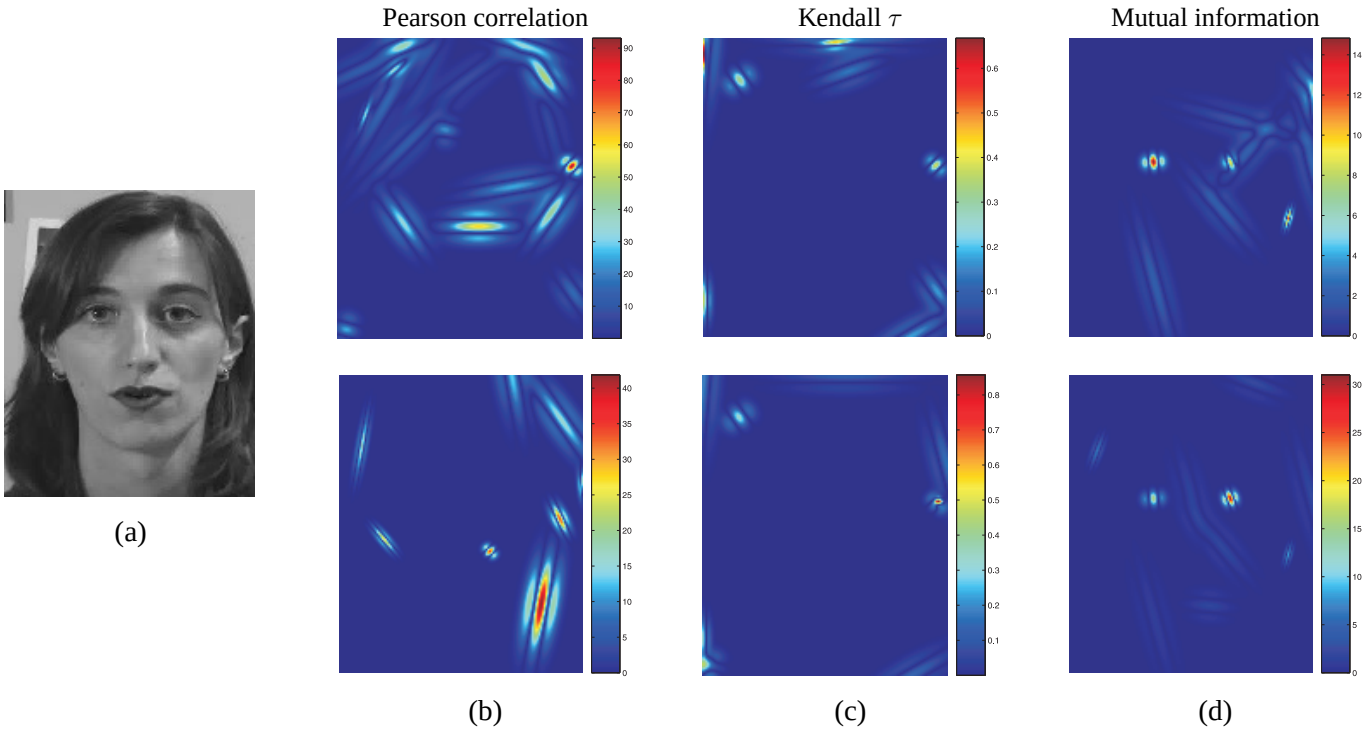

Fig. 6. Experiments on the sequence *Elena 2*; The frame number 10 of the video sequence is shown in (a). In picture (b), results obtained by correlating the video stream and the audio signal using Pearson correlations are shown. At the top figure, results for the video sequence *Elena 2* correlated with its audio component are depicted. At the bottom, we show the resulting atoms when the video sequence *Elena 2* is correlated with the audio signal of the sequence *Elena 1*. In (c) we show the results using the τ correlation coefficients as fusion criterion, when using the correct audio (top), and the one from sequence *Elena 1* (bottom). Finally, in (d) results for mutual information with correct (top) and incorrect (bottom) audio signals are depicted.

V. Discussion of the Results

The sequences we have used in this study are simple cases representing one person speaking in front of the camera without moving significantly. However, such a scenario is more realistic than one can think. In fact, one can assume that a face detector is available and thus, if more than one person are in front of a camera and only one is speaking, it seems possible to exploit our proposed approach to locate the speaker following a procedure like the one described in [8]. Moreover, videos have been filmed without any control on the illumination conditions and in addition some background noise is present in the audio streams.

The results we have obtained are somehow surprising, in the sense that the best performances are achieved by the simplest association measure, that is the Pearson ρ coefficients. The other two measures that we have tested, the Kendall τ coefficients and the mutual information, clearly fail in discriminating between correct and incorrect audio tracks, and in general are not able to locate the mouth of the person speaking. The τ coefficient is a correspondence measure that quantifies how much two variables tend to vary together. From our results, it seems that the video atoms that correlate most with the audio, according to this criterion, are those that vary the less in time. This is perhaps due to the fact that the visual primitives are not perfectly stable in time and thus the video features are affected by a certain noise. The correlation estimator based on Kendall τ seems to be not enough robust to such a noise.

The choice of mutual information as audio-visual fusion criterion is considered, often, the most natural one. This should be able to capture complex and general dependencies between variables. This is the choice that has been shown to be effective in [5], [6], [7], [8], [9]. In our case, to the contrary, mutual information performs poorly in relating audio and video features. In our opinion, this is primarily due to the fact that the joint distributions of audio-visual parameters are estimated on-line using a small number of samples, leading to non negligible errors in the mutual information estimate. One of the strong points of mutual information is its complete generality. However, this can turn into a weakness when dealing with small data samples.

This could be the reason of the fact that a simple correlation measure as the Pearson correlation coefficient is capable of detecting audio-visual dependencies much more better than more complex and general association criteria. We

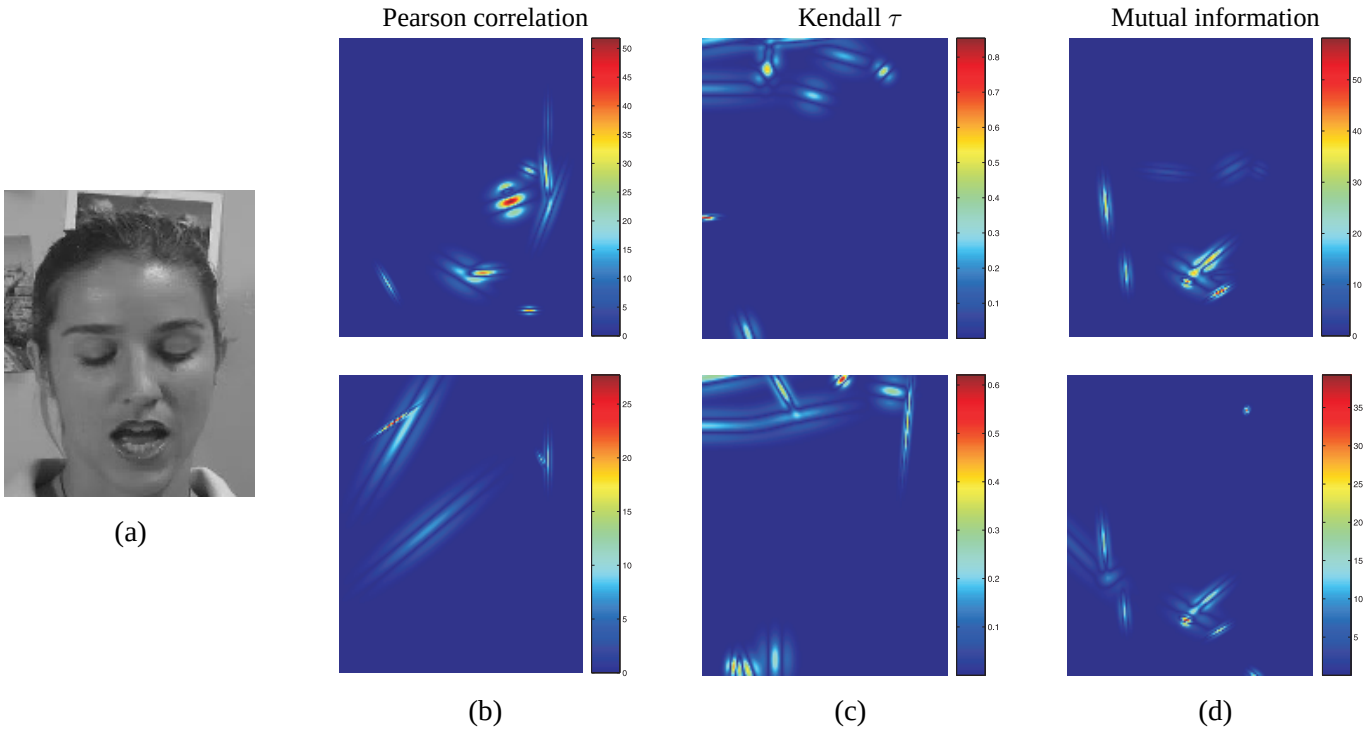

Fig. 7. Experiments on the sequence *Elisa*; The frame number 5 of the video sequence is shown in (a). In picture (b), results obtained by correlating the video stream and the audio signal using Pearson correlations are shown. At the top figure, results for the video sequence *Elisa* correlated with its audio component are depicted. At the bottom, we show the resulting atoms when the video sequence is correlated with a non coherent audio signal. In (c) we show the results using the τ correlation coefficients as fusion criterion, when using the correct audio (top), and an incorrect one (bottom). Finally, in (d) results for mutual information with correct (top) and incorrect (bottom) audio signals are depicted.

recall here that using the correlation coefficient ρ, we can detect linear relationships in the case of bivariate Gaussian distributions. If these hypotheses are satisfied, such statistical measure is more powerful than mutual information or Kendall τ. In other words, Pearson correlation is more precise than the other evaluated techniques, when the data sample has a reduced size. In our case such assumptions of linearity and normality seem to hold, probably because of the type of features we have used.

VI. Conclusions and Future Works

In the present work, we propose a dictionary approach to audio and video representation in the context of multimodal audio-visual fusion. The motivation for exploring this way is mainly the observation that image sequences are typically interpreted as huge pixel intensities matrices evolving in time. The fact of considering pixel-related quantities seems to us a remarkable limiting factor, since the pixel itself is a poor source of information. Video atoms, on the other hand, represent time-evolving image structures, and their parameters describe concisely how such structures move and change their characteristics in space and time. A very simple example can clarify this concept. If a person is moving back and forth while speaking in front of the camera, the pixel values on the mouth region change depending on the lips movements and on the person movement. These two cause the evolution of the pixel intensities in an undistinguishable way. On the other hand, if the mouth is represented using atoms that track the image structures and describe their intensity, position, scale and orientation variations, then, we are able to better interpret what is happening in the scene.

All the works in the field use very simple representations for the signals, that are processed involving huge computations and/or training of complex *a priori* models. One evident advantage of using redundant parametric decompositions, is that we obtain an extremely concise representation of information, that is at the same time accurate. In our case, for example, instead of processing $144 \times 176 = 25344$ time-evolving variables (pixel intensities) to deal with the video signal, we consider only $120 \times 3 = 360$ variables (atoms parameters). The price to pay, for the moment, is the high computational complexity of the MP algorithm, especially in what concerns the video signal. However, from our point of view this price is virtually zero, since the audio and video atoms we are using are exactly the same that the MP

decoders use to reconstruct the compressed audio-visual sequence. Moreover, recent results on signal approximation show that fast algorithms for the sparse representation of signals using redundant dictionaries are being achieved [22].

The results we show in this paper are purely explorative, we directly make use of algorithms that have been conceived for different purposes and the features and fusion criteria we consider are, at least, rough. However, the experimental results we have obtained encourage us in pursuing in this direction. In the future, we look forward to conceive an appropriate audio feature and studying more in details the relationship between audio and video, in order to define an accurate fusion strategy. Moreover, the stability of the video representation has to be considered, since for the moment the video MP algorithm is far from being perfect in tracking image features in complex scenes.

In this work, no *a priori* model on the audio-visual relationship has been established, except for the hypotheses underlying the Pearson correlation criterion (see Section III-C.1). We believe that thanks to the parametric representation we obtain using MP, it will be possible to build a general audio-visual model to allow for an improvement in the robustness of multimodal fusion of speech and images. For this purpose, we plan to study a database of sequences and learn a statistical model of the interdependencies between audio and video features.

VII. Acknowledgements

We thank Elena Salvador, Elisa Drelie Gelasca and Rosa Figueras i Ventura for the time they accorded to us when recording the test sequences. This work is supported by the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.

References

[1] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[2] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] J. Hershey and J. Movellan, "Audio-vision: using audio-visual synchrony to locate sounds," in *Proc. of Neural Information Processing Society*, vol. 13, 2000.

[4] M. Slaney and M. Covell, "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of Neural Information Processing Society*, vol. 12, 1999.

[5] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proc. of the 10^{th} ACM International Conference on Multimedia*, 2002.

[6] T. Butz, "From error probability to information theoretic signal and image processing," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 2003, [Online] Available: http://itswww.epfl.ch/~brain/.

[7] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. of Neural Information Processing Society*, vol. 13, 2000.

[8] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, June 2004.

[9] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. of International Symposium on Independant Component Analysis and Blind Source Separation (ICA)*, Nara (Japan), April 2003.

[10] O. Divorra Escoda and P. Vandergheynst, "A Bayesian approach to video expansions on parametric over-complete 2-D dictionaries," in *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Siena (Italy), September 2004.

[11] P. Vandergheynst and P. Frossard, "Efficient image representation by anisotropic refinement in matching pursuit," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Salt Lake City (USA), May 2001.

[12] V. N. Temlyakov, "Weak greedy algorithms," Department of Mathematics, University of South Carolina, Columbia, Tech. Rep., 1999.

[13] O. Divorra Escoda, P. Vandergheynst, and M. Bierlaire, "Video representation using greedy approximations over redundant parametric dictionaries," EPFL, 1015 Lausanne, Tech. Rep. TR-ITS 19/2004, September 2004, [Online] Available: http://lts2www.epfl.ch/publications.php.

[14] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2^{nd} ed. John Wiley & Sons, 1984.

[16] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 2^{nd} ed. Boca Raton: Chapman & Hall / CRC Press, 2000.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley & Sons, 1991.

[18] E. Parzen, "On the estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[19] R. V. Hogg, "Statistical robustness: One view of its use in applications today," *The American Statistician*, vol. 33, no. 3, pp. 108–115, 1979.

[20] http://www.cmap.polytechnique.fr/~bacry/LastWave/.

[21] http://lts2www.epfl.ch/~monaci/multimodal.html.

[22] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit," EPFL, 1015 Lausanne, Tech. Rep. TR-ITS 2004.13, July 2004, [Online] Available: http://lts2www.epfl.ch/publications.php.