Learning Multi-Modal Dictionaries

Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhe, Sylvain Lesage, Rémi Gribonval

***Abstract*— Real-world phenomena involve complex interactions between multiple signal modalities. As a consequence, humans are used to integrate at each instant perceptions from all their senses in order to enrich their understanding of the surrounding world. This paradigm can be also extremely useful in many signal processing and computer vision problems involving mutually related signals. The simultaneous processing of multi-modal data can in fact reveal information that is otherwise hidden when considering the signals independently. However, in natural multi-modal signals, the statistical dependencies between modalities are in general not obvious. Learning fundamental multi-modal patterns could offer a deep insight into the structure of such signals. Typically, such recurrent patterns are shift invariant, thus the learning should try to find the best matching filters. In this paper we present an algorithm for iteratively learning multi-modal generating functions that can be shifted at all positions in the signal. The learning is defined in such a way that it can be accomplished by iteratively solving a generalized eigenvector problem, which makes the algorithm fast, flexible and free of user-defined parameters. The proposed algorithm is applied to audiovisual sequences and we show that it is able to discover underlying structures in the data.**

I. Introduction

Multi-modal signal analysis has received an increased interest in the last years. Multi-modal signals are sets of heterogeneous signals originating from the same phenomenon but captured using different sensors. Each modality typically brings some information about the others and their simultaneous processing can uncover relationships that are otherwise unavailable when considering the signals separately. Multi-modal signal processing is widely employed in medical imaging, where the spatial correlation between different modalities (e.g. magnetic resonance and computed tomography) is exploited for segmentation [1] or registration [2], [3]. In this work we analyze a broad class of multi-modal signals exhibiting correlations along time. In many different research fields, the temporal correlation between multi-modal data is studied : in neuroscience, electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) data are jointly analyzed to study brain activation patterns [4]. In environmental science, connections between local and global climatic phenomena are discovered by correlating different spatio-temporal measurements [5]. Many multimedia signal processing problems involve the simultaneous analysis of audio and video data, e.g. speech-speaker recognition [6], [7], talking heads creation and animation [8] or sound source localization [9]–[14]. Interestingly, humans as well are used to integrate acoustic and visual inputs [15]–[17] or tactile and visual stimuli [18], [19] to enhance their perception of the world.

The temporal correlation across modalities is exploited by seeking for patterns showing a certain degree of synchrony. Research efforts typically focus on the statistical modelling of the dependencies between modalities. In [4], EEG and fMRI structures having maximal temporal covariance are extracted. In [9] the correlation between audio and video is assessed measuring the correlation coefficient between acoustic energy and the evolution of single pixel values. In [10], audio-video correlations are discovered using Canonical Correlation Analysis (CCA) for the cepstral representation of the audio and the video pixels. Smaragdis and Casey [11] find projections onto maximally independent audiovisual subspaces performing Independent Component Analysis simultaneously on audio and video features that are respectively the magnitude of the audio spectrum and the pixel intensities. In [12] the video components correlated with the audio are detected by maximizing Mutual Information between audio energy and single pixel values. In [13] the wavelet components of difference images are correlated with the audio signal applying a modified CCA algorithm which is regularized using a sparsity criterion.

While research efforts appear to be concentrated in the development of multi-modal fusion strategies, it seems that the features employed to represent the different modalities are often basic and barely connected with the physics of the observed phenomena (e.g. video sequences are typically represented using time series of pixel intensities). This can be a limitation of existing approaches : multi-modal features having low structural content can be difficult to extract and manipulate. Moreover, the interpretation of the results can be problematic without an accurate modelling of the observed phenomenon. To cope with these limitations, the cross-modal correlation problem can be attacked from a different point of view, by focusing on the modelling of the modalities, so that *meaningful* signal structures can be extracted and synchronous patterns easily detected.

In this paper we propose an algorithm that allows to learn dictionaries of basis functions representing recurrent multi-modal structures. Such patterns are learned using a recursive algorithm that enforces synchrony between the different modalities and de-correlation between the dictionary elements. The learned multi-modal functions are translation invariant, i.e. they are *generating functions* defining a set of structures corresponding to all their translations. The proposed algorithm is applied to real audiovisual sequences and the learned audio-visual dictionaries seem to capture well underlying structures in the data. The dictionary functions are used to analyze com-

This work was supported by the Swiss National Science Foundation through the IM.2 National Center of Competence for Research and by the EU HASSIP network HPRN-CT-2002-00285 .

Gianluca Monaci, Philippe Jost and Pierre Vandergheynst are with the Signal Processing Institute - Ecole Polytechnique Federale de Lausanne (EPFL), 1015 Lausanne, Switzerland. E-mail: {gianluca.monaci,philippe.jost,pierre.vandergheynst}@epfl.ch. Boris Mailhe, Sylvain Lesage and Rémi Gribonval are with IRISA-INRIA, Campus de Beaulieu, 35042 Rennes CEDEX, France. E-mail: {boris.mailhe,sylvain.lesage,remi.gribonval}@irisa.fr

plex multimedia clips, showing the ability to detect meaningful correlated audio-video structures and to localize the sound source in the video sequence.

The structure of the paper is the following : Section II describes the proposed model for multi-modal signals. Section III constitutes the central part of this work, presenting the learning algorithm for multi-modal signals. In Sec. IV experimental results based on real audiovisual signals are shown. Section V concludes the paper with a discussion of the achieved results and of the possible developments of this research.

II. Modelling and Understanding

A. Sparse approximations of multi-modal signals

Multi-modal data are made up of M different modalities and represented as M-tuples $s = (s^{(1)}, \dots, s^{(M)})$ which are not necessarily homogenous in dimensionality : for example, audiovisual data consist of an audio signal $s^{(1)}(t)$ and a video sequence $s^{(2)}(\vec{x}, t)$ with $\vec{x} \in \mathbb{R}^2$ the pixel position. Other multi-modal data such as hyperspectral images or biomedical sequences could be made of images, time-series and video sequences at various resolutions.

To date, methods dealing with multi-modal fusion problems basically attempt to build general and complex statistical models to capture the relationships between the different signal modalities $s^{(m)}$. However, as underlined in the previous section, the employed features are typically simple and barely connected with the physics of the problem. Efficient signal modelling and representation require the use of methods able to capture particular characteristics of each signal. Therefore, the idea is basically that of defining a proper model for signals, instead of defining a complex statistical fusion model that has to find correspondences between barely meaningful features.

Applications of this paradigm to audiovisual signals can be found in [14], [20]. A sound is assumed to be generated through the synchronous motion of important visual elements like edges. Audio and video signals are thus represented in terms of their most salient structures using redundant dictionaries of functions, making it possible to define acoustic and visual *events*. An audio event is the presence of an audio signal with high energy and a visual event is the motion of an important image edge. The synchrony between these events reflects the presence of a common source, which is effectively localized. The key idea of this approach is to use high-level features to represent signals, which are introduced by making use of codebooks of functions. The audio signal is approximated as a sparse sum $s^{(1)} \approx \sum_{k \in I_1} c_k^{(1)} \phi_k^{(1)}$ of Gabor atoms from a Gabor dictionary $\{\phi_k^{(1)}\}_k$, while the video sequence is expressed as a sparse combination $s^{(2)} \approx \sum_{k \in I_2} c_k^{(2)} \phi_k^{(2)}$ of edge-like functions $\{\phi_k^{(2)}\}_k$ that are tracked through time. Such audio and video representations are still quite general, and can be employed to represent any audiovisual sequence.

One of the main advantage of dictionary-based techniques is the freedom in designing the dictionary, which can be efficiently tailored to closely match signal structures. For multi-modal data, distinct dictionaries $\mathcal{D}^{(m)} = \{\phi_k^{(m)}\}_k$ for each modality do not necessarily reflect well the interplay between events in the different modalities, since the sets of salient features I_m involved in the models of each modality are not necessarily related to one another. An interesting alternative consists in capturing truly multi-modal events by the means of an intrinsically *multi-modal dictionary* $\mathcal{D} = \{\phi_k\}$ made of *multi-modal atoms* $\phi_k = (\phi_k^{(1)}, \dots, \phi_k^{(M)})$, yielding a multi-modal sparse signal model

$$s \approx \sum_{k \in I} \left(c_k^{(1)} \phi_k^{(1)}, \dots, c_k^{(M)} \phi_k^{(M)} \right). \tag{1}$$

Here, a common set I of salient multi-modal features forces *at the model level* some correlation between the different modalities.

Given the multi-modal dictionary $\mathcal{D} = \{\phi_k\}$ and the multi-modal signal s, the inference of the model parameters I and $\{c_k^{(m)}\}_{k,m}$ is not completely trivial : on the one hand, since the dictionary is often redundant, the are infinitely many possible representations of any signal; on the other hand, choosing the best approximation with a given number of atoms is known to be an NP-hard problem. Fortunately, several suboptimal algorithms such as multichannel Matching Pursuit [21], [22], can provide generally good sparse approximations.

B. Synchrony and shift invariance in multi-modal signals

Very often, the various modalities in a multi-modal signal will share synchrony of some sort. By synchrony, we usually refer to time-synchrony, i.e. events occurring in the same time slot. When multi-modal signals share a common time-dimension, synchrony is a very important feature, usually tightly linked to the physics of the problem. As explained above, synchrony is of particular importance in audio-visual sequences. Sound in the audio time series is usually linked to the occurrence of events in the video *at the same moment*. If for example the sequence contains a character talking, sound is synchronized with lips movements. More generally though, multi-modal signal could share higher-dimensions, and the notion of synchrony could refer to spatial co-localization, for example in multi-spectral images where localized features appear in several frequency bands at the same spatial position.

For the sake of simplicity, we will focus our discussion on time-synchrony and we now formalize this concept further. Let

$$\phi = \left(\phi^{(1)}(\vec{x}_1, t), \dots, \phi^{(M)}(\vec{x}_M, t) \right), \ \vec{x}_m \in \mathbb{R}^{d_m}$$

be a multi-modal function whose modalities $\phi^{(m)}$, $m = 1, \dots, M$ share a common temporal dimension $t \in \mathbb{R}$. A modality is temporally localized in the interval $\Delta \subset \mathbb{R}$ if $\phi^{(m)}(\vec{x}_m, t) = 0$, $\forall t \notin \Delta$. We will say that the modalities are synchronous whenever all $\phi^{(m)}$ are localized in the same time interval Δ.

Most natural signals exhibit characteristics that are time-invariant, meaning that they can occur at any instant in time. Think once again of an audio track : any particular frequency pattern can be repeated at arbitrary time instants. In order to account for this natural shift-invariance, we need to be able to shift patterns on modalities. Let ϕ be a multi-modal function localized in an interval centered on $t = 0$. The operator T_p shifts ϕ to time $p \in \mathbb{R}$ in a straightforward way :

$$T_p \phi = \Big(\phi^{(1)}(\vec{x}_1, t-p), \ldots, \phi^{(M)}(\vec{x}_M, t-p) \Big) \, . \quad (2)$$

This temporal translation is homogeneous across channels and thus preserves synchrony. With these definitions, it becomes easy to express a signal as a superposition of synchronous multi-modal patterns $\phi_k,\ k \in I$ occurring at various time instants $t_1, \ldots, t_k$:

$$s \approx \sum_{k \in I} c_k T_{t_k} \phi_k ,$$

where the sum and weighting coefficients are understood as in (1). We often construct a large subset of a dictionary by applying such synchronous translations to a single multi-modal function. In that case, we will often refer to this function as a *generating function* and we will indicate it with g_k.

In complex situations, it is sometimes difficult to manually design good dictionaries because there is no good a priori knowledge about the generating functions g. In these cases, one typically would want to learn a good dictionary from training data. Successful algorithms to learn dictionaries of basis functions have been proposed in the last years and applied to diverse classes of signal, including audio data [23]–[25], natural images [25]–[29] and video sequences [30]. In the next section, we propose a learning strategy adapted to synchronous multi-modal signals.

III. Learning Multi-Modal Dictionaries

Our goal is to design an algorithm capable of learning sets of multi-modal synchronous functions adapted to particular classes of multi-modal signals. However, the design of an algorithm for learning dictionaries of multi-modal atoms is non-trivial and an extended literature survey showed that it has never been attempted so far. Two major challenges have to be considered:

- Learning algorithms are inherently time and memory consuming. When considering sets of multi-modal signals that involve huge arrays of data, the computational complexity of the algorithm becomes a challenging issue.
- Natural multi-modal signals often exhibit complex underlying structures that are difficult to explicitly define. Moreover, modalities have heterogeneous dimensions, which makes them complicated to handle. Audiovisual signals perfectly illustrate this challenge: the audio track is a 1-D signal typically sampled at high frequency rate ($\mathcal{O}(10^4)$ samples/sec), while the video clip is a 3-D signal sampled with considerably lower temporal resolution ($\mathcal{O}(10^1)$ frames/sec).

We will design a novel learning algorithm that captures the underlying structures of multi-modal signals overcoming both of these difficulties. We propose to learn *synchronous multi-modal generating functions* as introduced in the previous section using a generalization of the MoTIF algorithm [25]. Each such function defines a set of atoms corresponding to all its translations. This is notably motivated by the fact that natural signals typically exhibit statistical properties invariant to translation, and the use of generating functions allows to generate huge dictionaries while using only few parameters. In order to make the computation feasible, the proposed algorithm learns the generating functions by alternatively localizing and learning interesting signal structures on the different signal components. As detailed in the following, this allows moreover to enforce synchrony between modal structures in an easy and intuitive fashion. Generating functions are learned successively and the procedure can be stopped when a sufficient number of atoms have been found. A constraint that imposes low correlation between the learned functions is also considered, such that no function is picked several times.

The goal of the learning algorithm is to build a set $\mathcal{G} = \{g_k\}_{k=1}^K$ of multi-modal generating functions g_k such that a very redundant dictionary $\mathcal{D}$ adapted to a class of signals can be created by applying all possible translations to the generating functions of $\mathcal{G}$. The function g_k can consist of an arbitrary number M of modalities. For simplicity, we will treat here the bimodal case $M = 2$; however, the extension to $M > 2$ is straightforward. To make it more concrete, we will write a bimodal function as $g_k = (g_k^{(a)}, g_k^{(v)})$ where one can think of $g_k^{(a)}$ as an audio modality and $g_k^{(v)}$ as a video modality of audiovisual data. More generally, the components do not have to be homogeneous in dimensionality; however, they have to share a common temporal dimension.

For the rest of the paper, we denote discrete signals of infinite size by lower case letters. Real-world finite signals are made infinite by padding their borders with zeros. Finite size vectors and matrices are denoted with bold characters. We need to define the time-discrete version $\mathcal{T}_p,\ p \in \mathbb{R}$ of the synchronous translation operator (2). Since different modalities are in general sampled at different rates over time the operator $\mathcal{T}_p$ must shift the signals on the two modalities by a different integer number of samples, in order to preserve their temporal proximity. We define it as $\mathcal{T}_p = (\mathcal{T}_p^{(a)}, \mathcal{T}_p^{(v)}) := (T_{q^{(a)}}, T_{q^{(v)}})$, where $T_{q^{(a)}}$ translates an infinite (audio) signal by $q^{(a)} \in \mathbb{Z}$ samples and $T_{q^{(v)}}$ translates an infinite (video) signal by $q^{(v)}$ samples. In the experiments that we will conduct at the end of this paper, typical values of the sampling rates are $\nu^{(a)} = 1/8000$ for audio signals sampled at 8 kHz and $\nu^{(v)} = 1/29.97$ for videos at 29.97 frames per second. Therefore the discrete-time version of the synchronous translation operator $\mathcal{T}_p$ with translation $p \in \mathbb{R}$ is defined with discrete translations $q^{(a)} := \mathrm{nint}(p/\nu^{(a)}) \in \mathbb{Z}$ and $q^{(v)} := \mathrm{nint}(p/\nu^{(v)}) \in \mathbb{Z}$ where $\mathrm{nint}(\cdot)$ is the nearest integer function. Without loss of generality we may assume that $\nu^{(v)} \geq \nu^{(a)}$ and define a *re-sampling factor* $\mathrm{RF} = \nu^{(v)}/\nu^{(a)}$.

For a given generating function g_k, the set $\{\mathcal{T}_p g_k\}_{p \in \mathbb{R}}$ contains all possible atoms generated by applying the translation operator to g_k. The dictionary generated by $\mathcal{G}$ is then $\mathcal{D} = \{\{\mathcal{T}_p g_k\}_p, k = 1 \ldots K\}$. Learning is performed using a training set of N bimodal signals $\{(f_n^{(a)}, f_n^{(v)})\}_{n=1}^N$, where $f_n^{(a)}$ and $f_n^{(v)}$ are the components of the signal on the two modalities. The signals are assumed to be of infinite size but they are non zero only on their support of size $(S_f^{(a)}, S_f^{(v)})$. Similarly, the size of the support of the generating functions to learn is $(S_g^{(a)}, S_g^{(v)})$ such that $S_g^{(a)} < S_f^{(a)}$ and $S_g^{(v)} < S_f^{(v)}$. The proposed algorithm iteratively learns translation invariant filters. For the first one, the aim is to find $g_1 = (g_1^{(a)}, g_1^{(v)})$ such that the dictionary $\{(\mathcal{T}_p^{(a)} g_1^{(a)}, \mathcal{T}_p^{(v)} g_1^{(v)})\}_p$ is the most correlated in mean with the signals in the training set. Hence, it is equivalent to the following optimization problem :

$$\text{UP} : g_1 = \underset{\|g^{(a)}\|_2 = \|g^{(v)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} \sum_i \mid \langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \rangle \mid^2 , \tag{3}$$

which has to be solved simultaneously for the two modalities ($i = a, v$), i.e. we want to find a pair of synchronous filters $(g^{(a)}, g^{(v)})$ that minimize (3). There are two main differences with respect to classical learning methods, which make the present problem extremely challenging. First of all, we do not only want the learned function g_1 to represent well in average the training set (as expressed by the first maximization over g), but we want g_1 to be the best representing function up to an arbitrary time-translation on each training signal (as indicated by the second maximization over p_n) in order to achieve shift-invariance. In addition, we require these characteristics to hold for both modalities simultaneously, which implies an additional constraint on the synchrony of the couple of functions $(g_1^{(a)}, g_1^{(v)})$. Note that solving problem UP requires to compute simultaneous correlations across channels. In the audio-visual case, the dimension of the video channel makes this numerically prohibitive. To avoid this problem, we first solve UP restricted to the audio channel :

$$\text{UP'} : g_1^{(i)} = \underset{\|g^{(i)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} \mid \langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \rangle \mid^2 , \tag{4}$$

where $i = a$. We can then solve (4) for $i = v$ but limit the search for best translations around the time-shifts already obtained on the audio channel, thus avoiding the burden of long correlations between video streams.

For learning the successive generating functions, the problem can be slightly modified to include a constraint penalizing a generating function if a similar one has already been found. Assuming that $k-1$ generating functions have been learnt, the optimization problem to find g_k can be written as :

$$\text{CP} : g_k^{(i)} = \underset{\|g^{(i)}\|_2 = 1}{\arg\max} \frac{\sum_{n=1}^{N} \max_{p_n} \mid \langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \rangle \mid^2}{\sum_{l=0}^{k-1} \sum_{q \in \mathbb{Z}} \mid \langle g_l^{(i)}, T_q g^{(i)} \rangle \mid^2} , \tag{5}$$

which again has to be solved simultaneously for the two modalities ($i = a, v$). In this case the optimization problem is similar to the unconstrained one in (4), with the only difference that a de-correlation constraint between the actual function $g_k^{(i)}$ and the previously learned ones is added. The constraint is introduced as a term at the denominator that accounts for the correlation between the previously learned generating functions (the first summation over l) and the actual target function shifted at all possible positions (the second sum over q). By maximizing the fraction in (5) with respect to g, the algorithm has to find a balance between the goodness of the representation of the training set, which has to be maximized being expressed by the numerator, and the correlation between g_k and g_l ($l = 1, \dots, k-1$), which has at the same time to be minimized, being represented by the denominator.

Finding the best solution to the unconstrained problem (UP') or the constrained problem (CP) is indeed hard. However, the problem can be split into several simpler steps following a *localize and learn* paradigm [25]. Such a strategy is particularly suitable for this scenario, since we want to learn synchronous patterns that are localized in time and that represent well the signals. Thus, we propose to perform the learning by iteratively solving the following four steps:

1. ***Localize:*** for a given generating function $g_k^{(a)}[j-1]$ at iteration j, find the best translations $p_n^{(a)}[j] := \nu^{(a)} \cdot q_n^{(a)}[j]$ with
$$q_n^{(a)}[j] := \underset{q \in \mathbb{Z}}{\arg\max} \mid \langle f_n^{(a)}, T_q g_k^{(a)}[j-1] \rangle \mid ;$$
2. ***Learn:*** update $g_k^{(v)}[j]$ by solving UP' (4) or CP (5) only for modality (v), with the translations fixed to the values $p_n = p_n^{(a)}[j]$ found at step 1, i.e. $q_n^{(v)} := \text{nint}(\text{RF} \times q_n^{(a)}[j])$;
3. ***Localize:*** find the best translations $p_n^{(v)}[j] := \nu^{(v)} \cdot q_n^{(v)}[j]$ using the function $g_k^{(v)}[j]$;
$$q_n^{(v)}[j] := \underset{q \in \mathbb{Z}}{\arg\max} \mid \langle f_n^{(v)}, T_q g_k^{(v)}[j] \rangle \mid$$
4. ***Learn:*** update $g_k^{(a)}[j]$ by solving UP' (4) or CP (5) only for modality (a), with the translations fixed to the values $p_n = p_n^{(v)}[j]$ found at step 3 i.e. using $q_n^{(a)} = \text{nint}(q_n^{(v)}[j]/\text{RF})$.

The first and third steps consist in finding the location of the maximum correlation between one modality of each training signal $f_n^{(i)}$ and the corresponding generating function $g^{(i)}$. The temporal synchrony between generating functions on the two modalities is enforced at the learning steps (2 and 4), where the optimal translation p_n found for one modality is also kept for the other one.

We now consider in detail the second and fourth steps. We define $\mathbf{g}_k^{(i)} \in \mathbb{R}^{S_g^{(i)}}$ the restriction of the infinite size signal $g_k^{(i)}$ to its support. We will use the easily checked fact that for any translation p, any signal $f^{(i)}$ and any filter $g^{(i)}$ we have the equality $\langle f^{(i)}, \mathcal{T}_p^{(i)} g^{(i)} \rangle = \langle \mathcal{T}_{-p}^{(i)} f^{(i)}, g^{(i)} \rangle$, in other words the adjoint of the discrete translation operator $\mathcal{T}_p^{(i)}$ is $\mathcal{T}_{-p}^{(i)}$. Let $\mathbf{F}^{(i)}[j]$ be the matrix (with $S_f^{(i)}$ rows and N columns), whose columns are made of the signals $f_n^{(i)}$ shifted by $-p_n[j]$. More precisely, the n^{th} column of $\mathbf{F}^{(i)}[j]$ is $\mathbf{f}_{n,-p_n[j]}^{(i)}$, the restriction of $\mathcal{T}_{-p_n[j]}^{(i)} f_n^{(i)}$ to the support of $g_k^{(i)}$, of size $S_g^{(i)}$. We also denote $\mathbf{A}^{(i)}[j] = \mathbf{F}^{(i)}[j] \cdot \mathbf{F}^{(i)}[j]^T$, where $.^T$ indicates the transposition.

With these notations, the second step (respectively fourth step) of the *unconstrained* problem can be written as :

$$\mathbf{g}_k^{(i)}[j] = \underset{\|\mathbf{g}^{(i)}\|_2 = 1}{\arg\max} \mathbf{g}^{(i)T} \mathbf{A}^{(i)}[j] \, \mathbf{g}^{(i)} . \tag{6}$$

with $i = v$ (respectively $i = a$). The best generating function $\mathbf{g}_k^{(i)}[j]$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}^{(i)}[j]$. Let us underline that in this case it is possible to easily solve the learning problem because of the particular form of the function to optimize. In fact, it is only because the objective function in (4) can be expressed as the quadratic form (6), given the translations p_n, that it is possible to turn the learning problem into an eigenvector problem.

For the *constrained* problem, we want to force $g_k^{(i)}[j]$ to be as de-correlated as possible from all the atoms in $\mathcal{D}_{k-1}$. This corresponds to minimizing

$$\sum_{l=1}^{k-1}\sum_{q\in\mathbb{Z}} \mid \langle T_{-q} g_l^{(i)}, g^{(i)} \rangle \mid^2 \tag{7}$$

or, denoting

$$\mathbf{B}_k^{(i)} = \sum_{l=1}^{k-1}\sum_{q\in\mathbb{Z}} \mathbf{g}_{l,-q}^{(i)}\, \mathbf{g}_{l,-q}^{(i)\,T}, \tag{8}$$

to minimizing $\mathbf{g}^{(i)T}\mathbf{B}_k^{(i)}\mathbf{g}^{(i)}$. With these notations, the constrained problem can be written as :

$$\mathbf{g}_k^{(i)}[j] = \underset{\|\mathbf{g}^{(i)}\|_2=1}{\arg\max} \frac{\mathbf{g}^{(i)T}\mathbf{A}^{(i)}[j]\,\mathbf{g}^{(i)}}{\mathbf{g}^{(i)T}\mathbf{B}_k^{(i)}\mathbf{g}^{(i)}}\,. \tag{9}$$

The best generating function $\mathbf{g}_k^{(i)}[j]$ is the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem defined in (9). Defining $\mathbf{B}_1^{(i)} = \mathbf{Id}$, we can use CP for learning the first generating function $\mathbf{g}_1$. Note again that the complex learning problem in (5) can be solved as the generalized eigenvector problem (9) because of the particular quadratic form imposed to the objective function to optimize, when the translations p_n are fixed.

The proposed multi-modal learning algorithm is summarized in **Algorithm 1**.

It is easy to demonstrate that the unconstrained single-modality algorithm converges in a finite number of iterations to a generating function locally maximizing the unconstrained problem. It has been observed on numerous experiments that the constrained algorithm [25] and the multi-modal constrained algorithm typically converge in few steps to a stable solution independently of the initialization.

Algorithm 1 Principle of the multi-modal learning algorithm

```
k = 0, training set {(f_n^(a), f_n^(v))};
for k = 1 to K do
    j ← 0;
    random initialization of {(g_k^(a)[j], g_k^(v)[j])};
    compute constraint matrices B_k^(a) and B_k^(v) as in (8);
    while no convergence reached do
        j ← j + 1;
        localize in modality (a):
            for each f_n^(a), find the translation
            p_n^(a)[j] ← ν^(a) · argmax_q | ⟨f_n^(a), T_q g^(a)[j − 1]⟩ |,
            maximally correlating f_n^(a) and g^(a)[j − 1];
        learn modality (v):
            set A^(v)[j] ← Σ_{n=1}^N f^(v)_{n,−p_n^(a)[j]} f^(v)_{n,−p_n^(a)[j]}^T;
        find g_k^(v)[j], the eigenvector associated to the biggest
            eigenvalue of the generalized eigenvalue problem
            A^(v)[j] g = λ B_k^(v) g, using (9);
        localize in modality (v):
            for each f_n^(v), find the translation
            p_n^(v)[j] ← ν^(v) · argmax_q | ⟨f_n^(v), T_q g^(v)[j]⟩ |,
            maximally correlating f_n^(v) and g^(v)[j];
        learn modality (a):
            set A^(a)[j] ← Σ_{n=1}^N f^(a)_{n,−p_n^(v)[j]} f^(a)_{n,−p_n^(v)[j]}^T;
        find g_k^(a)[j], the eigenvector associated to the biggest
            eigenvalue of the generalized eigenvalue problem
            A^(a)[j] g = λ B_k^(a) g, using (9);
    end while
end for
```

IV. Experiments

A. Audiovisual Dictionaries

The first experiment demonstrates the capability of the proposed learning algorithm to recover meaningful synchronous patterns from audiovisual signals. In this case the two modalities are audio and video, which share a common temporal axis, and the learned dictionaries are composed of generating functions $g_k = (g_k^{(a)}, g_k^{(v)})$, with $g_k^{(a)}$ and $g_k^{(v)}$ respectively audio and video component of g_k. Two joint audiovisual dictionaries are learned on two training sets. The first audiovisual dictionary, that we call *Dictionary 1* ($\mathcal{D}_1$), is learned on a set consisting of four audiovisual sequences representing the mouth of the same speaker uttering the digits from zero to nine in English. *Dictionary 2* ($\mathcal{D}_2$) is learned on a training set of four clips representing the mouth of four different persons pronouncing the digits from zero to nine in English. *Dictionary 1* should represent a collection of basis functions adapted to a particular speaker, while *Dictionary 2* aims at being a more "general" set of audio-video atoms.

For all sequences, the audio was recorded at 44 kHz and sub-sampled to 8 kHz, while the gray-scale video was recorded at 29.97 frames/second (fps) and at a resolution of 70×110 pixels. The total length of the training sequences is 1060 video frames, i.e. approximately 35 seconds, for $\mathcal{D}_1$, and 1140 video frames, i.e. approximately 38 seconds, for $\mathcal{D}_2$. Note that the sampling frequencies along the time axis for the two modalities are different, thus when passing from one modality to the other a re-sampling factor RF equal to the ratio between the two frequencies has to be applied. In this case the value of the re-sampling factor is RF $= 8000/29.97 \approx 267$. Video sequences are filtered following the procedure suggested in [30], in order to speed up the training. The video component is thus "whitened" using a filter that equalizes the variance of the input sequences in all directions. Since the spatio-temporal amplitude spectrum of video signals roughly falls as $1/f$ along all directions [27], [31], whitening can be obtained applying a spherically symmetric filter $W(f) = f$ that produces an approximately flat amplitude spectrum at all spatio-temporal frequencies. The obtained whitened sequences are then low-pass filtered to remove the high-frequency artifacts typical of digital video signals. We use a spherically symmetric low-pass filter $L(f) = e^{-(f/f_0)^4}$ with cut-off frequency f_0 at 80% of the Nyquist frequency in space and time.

The learning is performed on audio-video patches $(f_n^{(a)}, f_n^{(v)})$ extracted from the original signals. The size of the audio patches $f_n^{(a)}$ is 6407 audio samples, while the size of the video patches $f_n^{(v)}$ is 31×31 pixels in space and 23 frames in time. We learn 20 generating functions g_k consisting of an audio component $g_k^{(a)}$ of 3204 samples and a video component $g_k^{(v)}$ of size 16×16 pixels in space and 12 frames in time. The 20 elements of $\mathcal{D}_2$ are shown in Fig. 1. The dictionary $\mathcal{D}_1$ has similar characteristics. The video component $g_k^{(v)}$ of each function is shown on the left, with time proceeding left to right, while the audio part $g_k^{(a)}$ is on the right, with time on the horizontal axis.

Concerning the video components, they are spatially localized and oriented edge detector functions that shift smoothly from frame to frame, describing typical movements of different parts of the mouth during the utterances. The audio parts of the generating functions contain almost all the numbers present in the training sequences. In particular, when listening to the waveforms, one can distinguish the words *zero* (functions #11, #13, #16), *one* (#7, #9), *two* (#5, #6), *four* (#3), *five* (#1), *six* (#4), *seven* (#8, #18), *eight* (#10). Functions #12, #14, #15, #17, #19, #20 express the first two phonemes of the word *five* (i.e. /f/,/ay/), and they are also very similar to the word *nine* (i.e. /n/,/ay/). Typically, different instances of the same number have either different audio characteristics, like length or frequency content (e.g. compare audio functions #7 and #9), or different associated video components (e.g. functions #12, #14, #15, #17, #19, #20). As already observed in [25], both components of generating function #2 are mainly high frequency due to the de-correlation constraint with the first atom.

The learning algorithm captures well high-level signal structures representing the synchronous presence of meaningful acoustic and visual patterns. All the learned multi-modal functions consist in couples of temporally close signals : a waveform expressing one digit when played, and a moving edge (horizontal, diagonal or curved) that follows the contour of the mouth during the utterances.

B. Audiovisual Speaker Localization

In this experiment we want to test if the learned dictionaries are able to recover meaningful audiovisual patterns in real multimedia sequences. The dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$ are used to detect synchronous audio-video patterns revealing the presence of a meaningful event (the utterance of a sound) that we want to localize. We consider three test clips, Movie 1, Movie 2 and Movie 3, consisting in two persons placed in front of the camera arranged as in Fig. 2. One of the subjects is uttering digits in English, while the other one is mouthing *exactly the same words*. Test sequences consist in an audio track at 8 kHz and a video part at 29.97 fps and at a resolution of 480×720 pixels[1]. In all three sequences, the speaker is the same subject whose mouth was used to train $\mathcal{D}_1$; however, the training sequences are different from the test sequences. In contrast, none of the four speaking mouths used to train $\mathcal{D}_2$ belongs to the speaker in the test data set. We want to underline that the test sequences are particularly challenging to analyze, since both persons are mouthing the same words at the same time. The task of associating the sound with the "real" speaker is thus definitely non-trivial. The clips can be downloaded through `http://lts2www.epfl.ch/~monaci/avlearn.html`.

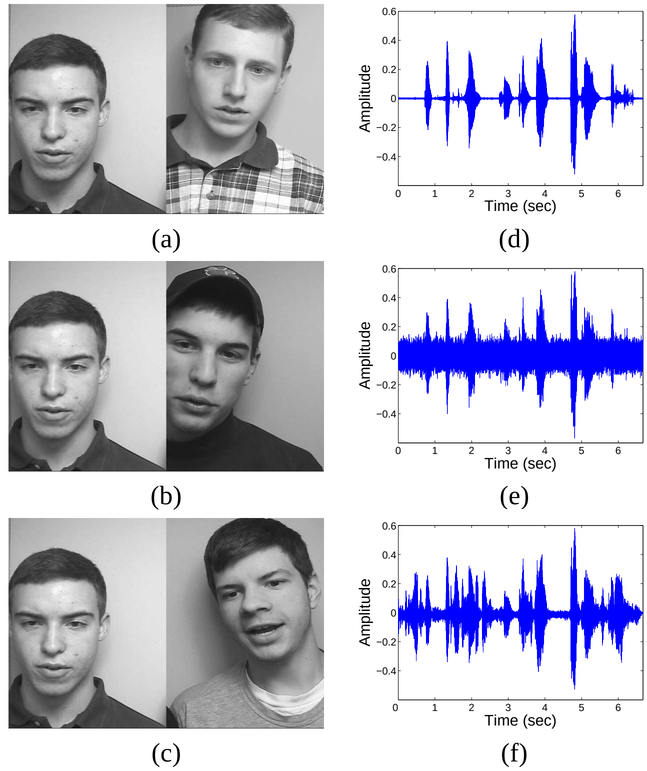

Fig. 2. Test sequences. Sample frames of Movie 1 (a), Movie 2 (b) and Movie 3 (c) are shown on the left. The original audio track a (d), together with its noisy versions with additive gaussian noise a+AWGN (e) and added distracting speech and music a+speech (f) are plotted on the right. All test clips can be downloaded through `http://lts2www.epfl.ch/~monaci/avlearn.html`.

With the experimental results that we will show in the following we want to demonstrate that:

- For both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$, the positions of maximal projection between the dictionary atoms ϕ_k and the test sequences are localized on the actual location of the audiovisual source.
- The detection of the actual speaker using both $\mathcal{D}_1$ and $\mathcal{D}_2$ is robust to severe visual noise (the person mouthing the same words of the real speaker) as well as to acoustic noise. The mouth of the correct speaker is effectively localized also when strong acoustic noise (SNR=1dB) is summed to the audio track in the form of additive white gaussian noise or out-of-view talking people.
- The detection of the speaker's mouth is more robust and accurate using dictionary $\mathcal{D}_1$, which is adapted to the speaker, than using the general dictionary $\mathcal{D}_2$.

[1]Only the luminance component is considered, while the chromatic channels are discarded.

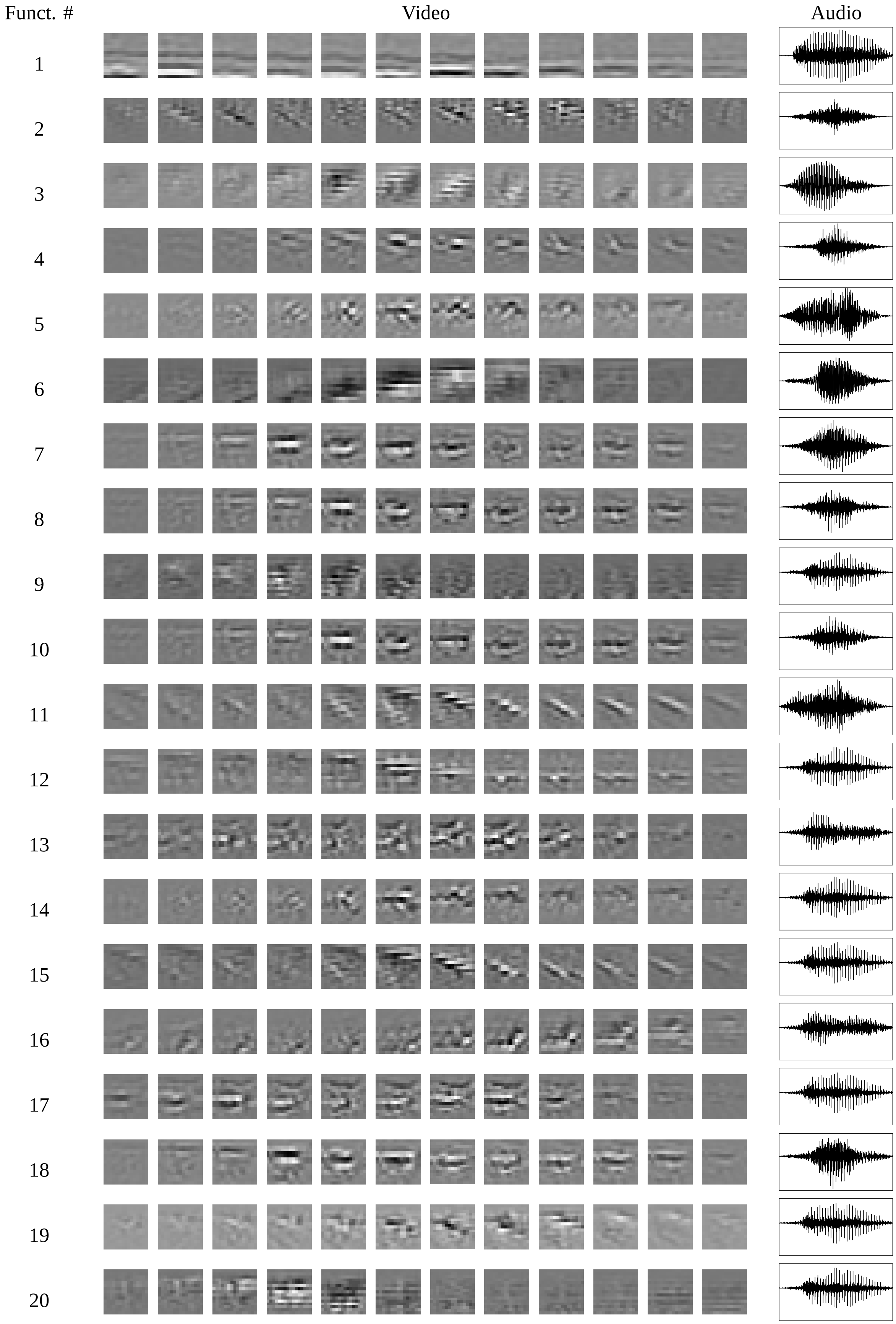

Fig. 1. Audio-video generating functions of *Dictionary 2*. Shown are the 20 learned functions, each consisting on an audio and a video component. Video components are on the left, with time proceeding left to right. Audio components are on the right, with time on the horizontal axis.

Fig. 3. Sample frames of Movie 1 [Left], Movie 2 [Center] and Movie 3 [Right]. The left person is the real speaker, the right subject mouths the same words pronounced by the speaker but his audio track has been removed. The white cross highlights the estimated position of the sound source, which is correctly placed over the speaker's mouth.

The audio tracks of the test clips are correlated with all time-shifted version of each audio component $g_k^{(a)}$ of the 20 learned generating functions g_k, which is efficiently done by filtering. For each audio function we find the the time position of maximum correlation, $\hat{p}_k^{(a)}$, and thus the audio atom $\phi_k^{(a)}$ with highest correlation. We consider a window of 31 frames around the time position in the video corresponding to $\hat{p}_k^{(a)}$, which is computed as $\tilde{p}_k^{(v)} = \mathrm{nint}(\hat{p}_k^{(a)}/\mathrm{RF})$. This restricted video patch consists of frames in the interval $[\tilde{p}_k^{(v)} - 15; \tilde{p}_k^{(v)} + 15]$ and we compute its correlation with all spatial and temporal shifts of the video component $g_k^{(v)}$ of g_k. The spatio-temporal position $\left(\widehat{\vec{x}}_k, \hat{p}_k^{(v)}\right)$ of maximum correlation between the restricted video patch and the learned video generating function yields the video atom $\phi_k^{(v)}$ with highest correlation. The positions of maximal projection of the learned atoms over the image plane $\widehat{\vec{x}}_k$, $k = 1, \ldots, 20$, are grouped into clusters using a hierarchical clustering algorithm[2]. The centroid of the cluster containing the largest number of points is kept as the estimated location of the sound source. We expect the estimated sound source position to be close to the speaker's mouth.

In Fig. 3 sample frames of the test sequences are shown. The white marker over each image indicates the estimated position of the sound source over the image plane, which coincides with the mouth of the actual speaker. The sound source location is correctly detected for all the tested sequences and using both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$. Results are accurate when the original sound track a is used (signal in Fig. 2 (d)), as well as when considerable acoustic noise (SNR=1dB) is present (signals a+AWGN and a+speech in Fig. 2 (e-f)).

In order to assess the goodness of the estimation of the sound source position, a simple measure can be designed. We define the *reliability* of the source position estimation, r, as the ratio between the number of elements belonging to the biggest cluster, which is the one used to estimate the sound source location, and the total number of elements considered, N (i.e. the total number of functions used for the analysis of the sequence, in this case 20). The value of r ranges from $1/N$, when each point constitutes a one-element cluster, to 1, when all points belong to the same group. Clearly, if most of the maxima of the projections between the video basis functions and the sequence lie close to one another, and are thus clustered together, it is highly probable that such cluster indicates the real position of the sound source and the value of r is high in this case. On the other hand, if maxima locations are placed all over the image plane forming small clusters, even the biggest cluster will include a small fraction of the whole data. In this situation it seems reasonable to deduce that the estimated source position is less reliable, which is reflected by the value of r being smaller in this case.

[2]The MATLAB function `clusterdata.m` was used. Clusters are formed when the distance between groups of points is larger than 50 pixels. According to several tests, the choice of the clustering threshold is non-critical.

As we have already observed, for all the test sequences the sound source position is correctly localized. Moreover, it is interesting to remark that in all cases, the detection of the speaker's mouth is more *reliable* using dictionary $\mathcal{D}_1$, which is adapted to the speaker, than using the general dictionary $\mathcal{D}_2$. An example of the described situation is depicted in Fig. 4. The images show sample frames of Movie 3. The positions of maximal projection between video functions belonging to dictionaries $\mathcal{D}_1$ (Left) and $\mathcal{D}_2$ (Right) and the test sequence are plotted on the image plane. Points belonging to the same cluster are indicated with the same marker. In both cases *Cluster 1* is the group containing the largest number of points and it is thus the one used to estimate the sound source position. When using dictionary $\mathcal{D}_1$ (Left), the biggest cluster has 17 elements and thus the reliability of the source position is $r = 17/20 = 0.85$, while when using $\mathcal{D}_2$ (Right), the biggest cluster groups only 13 points and the reliability equals $r = 13/20 = 0.65$. This behavior is indeed interesting, since it suggests that the learning algorithm actually succeeds in its task. The algorithm appears to be able to learn general meaningful synchronous patterns in the data. Moreover, the fact that more reliable localization results are achieved using the dictionary adapted to the speaker ($\mathcal{D}_1$) suggests that the proposed method allows to capture important signal structures typical of the considered training set.

The experimental results for all tested sequences and both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$ are summarized in Table I. The first column indicates the video clip used, the second one the audio track used and the third one the dictionary employed for the analysis. The fourth column shows the source localization result and the fifth column indicates the reliability r of the localization. In all cases the audio source is correctly localized on the image plane.

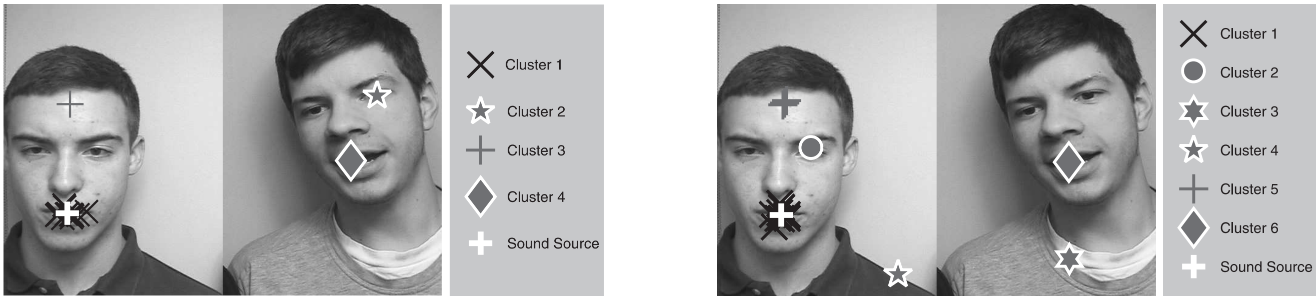

Fig. 4. Sample frames of Movie 3. The positions of maximal projection between video functions and test sequence are plotted on the image plane. Points belonging to the same cluster are indicated with the same marker. The biggest cluster is in both cases *Cluster 1*; it contains 17 elements when $\mathcal{D}_1$ is used [Left] and 13 when $\mathcal{D}_2$ is used [Right].

Video	Audio	Dictionary	Localization correct	r
Movie 1	a	$\mathcal{D}_1$	YES	0.75
		$\mathcal{D}_2$	YES	0.40
	a+AWGN	$\mathcal{D}_1$	YES	0.75
		$\mathcal{D}_2$	YES	0.45
	a+speech	$\mathcal{D}_1$	YES	0.75
		$\mathcal{D}_2$	YES	0.40
Movie 2	a	$\mathcal{D}_1$	YES	0.65
		$\mathcal{D}_2$	YES	0.45
	a+AWGN	$\mathcal{D}_1$	YES	0.65
		$\mathcal{D}_2$	YES	0.45
	a+speech	$\mathcal{D}_1$	YES	0.65
		$\mathcal{D}_2$	YES	0.45
Movie 3	a	$\mathcal{D}_1$	YES	0.85
		$\mathcal{D}_2$	YES	0.65
	a+AWGN	$\mathcal{D}_1$	YES	0.80
		$\mathcal{D}_2$	YES	0.65
	a+speech	$\mathcal{D}_1$	YES	0.85
		$\mathcal{D}_2$	YES	0.70

TABLE I

SUMMARY OF THE SOURCE LOCALIZATION RESULTS FOR ALL THE TESTED SEQUENCES. IN ALL CASES THE AUDIO SOURCE IS CORRECTLY LOCALIZED ON THE IMAGE PLANE.

V. CONCLUSIONS

In this paper we present a new method to learn translation invariant multi-modal functions adapted to a class of multi-component signals. Generating functions are iteratively found using a *localize and learn* paradigm which enforces temporal synchrony between modalities. Thanks to the particular formulation of the objective function, the learning problem can be turned into a generalized eigenvector problem, which makes the algorithm fast and free of parameters to tune. A constraint in the objective function forces the learned waveforms to have low correlation, such that no function is picked several times. The main drawback of this method is that the few generating functions following the first one are mainly due to the de-correlation constraint, more than to the correspondence with the signal. Despite that, the algorithm seems to capture well the underlying structures in the data. The learned dictionaries include elements that describe typical audiovisual features present in the training signals. The learned functions have been used to analyze complex multi-modal sequences, obtaining encouraging results in localizing the sound source in the video sequence.

One extension of the proposed method, based on the properties of the inner product, is to add to the translation invariance the invariance to other transformations that admit a well defined adjoint (e.g. translations *plus* rotations for images). Moreover, the application of this technique to other types of multi-modal signals, like climatologic or EEG-fMRI data, are foreseen.

REFERENCES

[1] A. A. Farag, A. S. El-Baz, and G. Gimel'farb, "Precise segmentation of multimodal images," *IEEE Trans. Imag. Proc.*, vol. 15, no. 4, pp. 952–968, 2006.

[2] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, 1997.

[3] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, vol. 85, no. 5, pp. 875–902, 2005.

[4] E. Martínez-Montes, P. A. Valdés-Sosa, F. Miwakeichi, R. I. Goldman, and M. S. Cohen, "Concurrent EEG/fMRI analysis by multiway partial least squares," *NeuroImage*, vol. 22, pp. 1023–1034, 2004.

[5] C. Carmona-Moreno, A. Belward, J. Malingreau, M. Garcia-Alegre, A. Hartley, M. Antonovskiy, V. Buchshtaber, and V. Pivovarov, "Characterizing inter-annual variations in global fire calendar using data from earth observing satellites," *Global Change Biology*, vol. 11, no. 9, pp. 1537–1555, 2005.

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

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

[8] E. Cosatto, J. Ostermann, H. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *Proc. IEEE*, vol. 91, no. 9, pp. 1406–1429, 2003.

[9] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. of NIPS*, vol. 12, 1999.

[10] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of NIPS*, vol. 13, 2000.

[11] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. of ICA*, April 2003, pp. 709–714.

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

[13] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in *Proc. of IEEE CVPR*, 2005, pp. 88–95.

[14] G. Monaci, O. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, 2006, in press. [Online] Available: http://lts2www.epfl.ch/.

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

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

[17] S. Watkins, L. Shams, S. Tanaka, J.-D. Haynes, and G. Rees, "Sound alters activity in human V1 in association with illusory visual perception," *NeuroImage*, vol. 31, no. 3, pp. 1247–1256, 2006.

[18] A. Violentyev, S. Shimojo, and L. Shams, "Touch-induced visual illusion," *Neuroreport*, vol. 10, no. 16, pp. 1107–1110, 2005.

[19] J.-P. Bresciani, F. Dammeier, and M. Ernst, "Vision and touch are automatically integrated for the perception of sequences of events," *Journal of Vision*, vol. 6, no. 5, pp. 554–564, 2006.

[20] G. Monaci, O. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proc. of IEEE ICIP*, vol. 3, 2005, pp. 46–49.

[21] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," in *Proc. of IEEE ICASSP*, vol. 3, 2002, pp. 3057–3060.

[22] J. Tropp, A. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proc. of IEEE ICASSP*, vol. 5, 2005, pp. 721–724.

[23] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[24] S. Abdallah and M. Plumbley, "If edges are the independent components of natural images, what are the independent components of natural sounds?" in *Proc. of ICA*, 2001, pp. 534–539.

[25] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "MoTIF: an efficient algorithm for learning translation invariant dictionaries," in *Proc. of IEEE ICASSP*, vol. 5, 2006, pp. 857–860.

[26] A. Bell and T. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[27] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Resesarch*, vol. 37, pp. 3311–3327, 1997.

[28] M. Lewicki and B. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America*, 1999.

[29] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, pp. 349–396, 2003.

[30] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. of IEEE ICIP*, vol. 1, 2003, pp. 41–44.

[31] D. Dong and J. Atick, "Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Network: Computation in Neural Systems*, vol. 6, pp. 159–178, 1995.