# Learning Multimodal Dictionaries

Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, *Member, IEEE*, Boris Mailhé, Sylvain Lesage, and 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 multimodal data can, in fact, reveal information that is otherwise hidden when considering the signals independently. However, in natural multimodal signals, the statistical dependencies between modalities are in general not obvious. Learning fundamental multimodal patterns could offer deep insight into the structure of such signals. In this paper, we present a novel model of multimodal signals based on their sparse decomposition over a dictionary of multimodal structures. An algorithm for iteratively learning multimodal generating functions that can be shifted at all positions in the signal is proposed, as well. 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 it is able to discover underlying structures in the data. The detection of such audio-video patterns in audiovisual clips allows to effectively localize the sound source on the video in presence of substantial acoustic and visual distractors, outperforming state-of-the-art audiovisual localization algorithms.**

***Index Terms*—Audiovisual source localization, dictionary learning, multimodal data processing, sparse representation.**

## I. Introduction

### A. Motivation

MULTIMODAL signal analysis has received an increased interest in the last years. Multimodal 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. Multimodal 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 registration [1], [2]. In remote sensing, multispectral satellite images are jointly segmented using measurements from visible, infra-red and radar sensors [3] or ice charts are built combining information from satellite images captured with very high resolution radiometer, synthetic aperture radar, operational line scanner and sensor microwave/imager [4].

In this work, we analyze a broad class of multimodal signals exhibiting correlations along time. In many different research fields, the temporal correlation between multimodal data is studied: in neuroscience, electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) data are jointly analyzed to study brain activation patterns [5]. In environmental science, connections between local and global climatic phenomena are discovered by correlating different spatio-temporal measurements [6]. Many multimedia signal processing problems involve the simultaneous analysis of audio and video data, e.g., speech–speaker recognition [7], [8], talking heads creation and animation [9], or sound source localization [10]–[17]. Interestingly, humans, as well, integrate acoustic and visual inputs [18]–[20] or tactile and visual stimuli [21], [22] to enhance their perception of the world.

### B. Related Work

The temporal correlation across modalities is typically exploited by seeking for patterns showing a certain degree of synchrony. Research efforts generally focus on the statistical modeling of the dependencies between modalities. In [5], EEG and fMRI structures having maximal temporal covariance are extracted. In [10], the correlation between audio and video is assessed measuring the correlation coefficient between acoustic energy and the evolution of single pixel values. In [11], audio-video correlations are discovered using canonical correlation analysis (CCA) for the cepstral representation of the audio and the video pixels. Nock and co-workers [15] evaluate three audiovisual synchrony measures and several video representations (coefficients of the DCT, pixel intensities and pixel intensity changes) in a speaker localization context. Two measures are based on mutual information (MI) maximization: one assumes discrete distributions and the other one considers multivariate Gaussian distributions as in [10]. The third correlation method defines the likelihood of audio-video configurations using hidden Markov models (HMMs) trained on audiovisual data. Tests are performed on a large database of audiovisual sequences, the CUAVE data set [23]. Smaragdis and Casey [12] 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

Manuscript received August 30, 2006; revised April 12, 2007. This work was supported in part by the Swiss National Science Foundation through the IM.2 National Center of Competence for Research and in part by the EU HASSIP network HPRN-CT-2002-00285. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ercan E. Kuruoglu.

G. Monaci, P. Jost, and P. Vandergheynst are with the Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: gianluca.monaci@epfl.ch; philippe.jost@epfl.ch; pierre.vandergheynst@epfl.ch).

B. Mailhé, S. Lesage, and R. Gribonval are with IRISA-INRIA, 35042 Rennes Cedex, France (e-mail: boris.mailhe@irisa.fr; sylvain.lesage@irisa.fr; remi.gribonval@irisa.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.901813

intensities. In [13], the video components correlated with the audio are detected by maximizing MI, estimated using Parzen windows, between audio energy and pixel values. In [14], the wavelet components of difference images are correlated with the audio signal applying a modified CCA algorithm which is regularized using a sparsity criterion.

Reviewed methods dealing with multimodal fusion problems basically attempt to build statistical models to capture the relationships between the different data streams. Surprisingly enough, however, the features employed to represent the modalities are basic and barely representative of the structural properties of the observed phenomena: we refer, in particular, to pixel-related features typically used for video representations. Such an approach can hide several drawbacks. Basic signal features convey few information about the structure of the data. This flaw has to be compensated by a statistical modeling step that attempts to capture the complex interplay between the different modalities. Modality features are, thus, considered as *random variables* whose degree of correlation is estimated using statistical measures under more or less restrictive assumptions. First, it seems improbable that the complex relationships between multimodal stimuli could be effectively modeled by statistical quantities. Besides, the estimation of cross-modal correlations forces one to consider an uncomfortable tradeoff. Either the statistical relationships between different modalities are supposed to be very simple, assuming, for example, linearity [10], [15], independence [12], or mutual Gaussianity [11], [14]. Either complex models involving the estimation of MI [13] or HMMs parameters [15] have to be conceived if no strong assumption is made, incurring in problems of parameter sensitivity and lack of data. In contrast to previous research works, in this paper, we suggest to avoid a complex statistical modeling of cross-modal interactions and we propose a generative structural model of multimodal data.

### C. Our Approach

We attack the cross-modal correlation problem taking a completely new point of view, by focusing on the modeling of the modalities, so that *meaningful* signal structures can be extracted and synchronous patterns easily detected. To this end we propose to use sparse signal representations in order to reduce the dimensionality of the problem and to make it more effective and intuitive to process multimodal entities.

Germinal instances of this approach have been developed in [16], [17], and [24], where audio and video signals are expressed in terms of salient, relevant data structures by decomposing each modality over a redundant dictionary of functions. Important multimodal structures are, thus, intuitively defined as synchronous relevant audio-video features that can be effectively detected and extracted. In this paper, we further develop this concept and we introduce a new model to represent multimodal signals.

Instead of separately decomposing each data modality over a general dictionary of functions, as in [16], [17], and [24], here we propose to represent a multimodal signal as a sparse sum of *multimodal basis functions*. Considering for example the audiovisual signal case, a multimodal basis function could be the signal couple depicted in Fig. 1. The function is composed of an audio and a video component: the audio part expresses a digit in English, while the corresponding video part shows a moving edge that could represent the lower lip during the utterance of the digit. The two components share a common temporal axis, and, thus, they exist in the same temporal interval even though they are sampled with a different time resolution. The challenge here is to generate a collection (or *dictionary*) of such *meaningful* multimodal structures that can be used to efficiently represent multimodal signals or, as we will do in this paper, to analyze multimodal data. The manual design of a *multimodal dictionary* is indeed complex, and, thus, we propose an algorithm that allows to learn dictionaries of such multimodal functions.

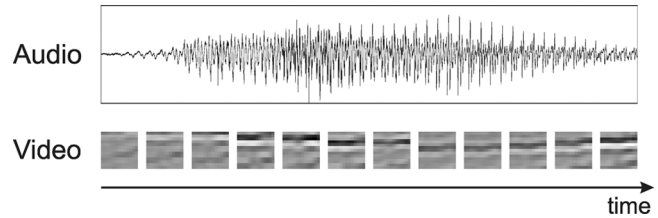


Fig. 1. Multimodal function composed of an audio and a video part sharing a common temporal axis.

This paper features three major contributions.

- We first define a general signal model to represent multimodal data using sparse representations over dictionaries of multimodal functions. We then refine such model adding two properties that are useful in order to represent real-world multimodal data: synchrony between the different components of multimodal functions and shift invariance of the basis waveforms.
- We propose an efficient algorithm to learn dictionaries of basis functions representing recurrent multimodal structures. Such patterns are learned using a recursive algorithm that enforces synchrony between the different modalities and decorrelation between the dictionary elements. The learned multimodal functions are translation invariant, i.e., they are *generating functions* defining a set of structures corresponding to all their translations. The algorithm learns the generating functions by alternatively localizing and learning interesting signal structures on training multimodal signals. Considering the audiovisual signal case, the algorithm iteratively solves the following four steps until convergence.
  1) ***Localize***: For a given audio waveform (word), find the temporal position that maximizes the correlation between such word and the training audio signals.
  2) ***Learn***: Find, at the time instant found in step 1 for the audio, the visual structure that best represents on average the training video signals.
  3) ***Localize***: Find the temporal position that maximizes the correlation between the video structure found at step 2 and the video signals.
  4) ***Learn***: Find, at the time instant found in step 3 for the video, the audio word that best represents on average the audio signals.
- Finally, we apply the proposed signal model and the learning method to audiovisual data. Results show that

the proposed algorithm allows to learn meaningful audio-video signal patterns from a training dataset. The training set being made of audiovisual patches extracted from sequences of talking mouths, the emerging multimodal generating functions represent salient audio patterns (words or phonemes) and associated video components showing synchronous movements of mouth parts during the utterances, like the function shown in Fig. 1. We will see that detecting such structures in audiovisual sequences makes it possible to effectively localize audio-video sources, overcoming severe acoustic and visual noise. Localization results favorably compare with those obtained by state-of-the-art audiovisual localization algorithms.

To summarize, the structure of the paper is the following. Section II describes the proposed model for multimodal signals. Section III constitutes the central part of this work, presenting the learning algorithm for multimodal signals. In Section 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. Modeling and Understanding

### A. Sparse Approximations of Multimodal Signals

Multimodal data are made up of $M$ different modalities and they can be represented as $M$-tuples $s = (s^{(1)}, \ldots, s^{(M)})$ 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 multimodal data such as multispectral images or biomedical sequences could be made of images, time-series and video sequences at various resolutions.

To date, methods dealing with multimodal 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 modeling 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 [16], [17], and [24]. 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 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 multimodal 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 multimodal events by the means of an intrinsically *multimodal dictionary* $\mathcal{D} = \{\phi_k\}$ made of *multimodal atoms* $\phi_k = (\phi_k^{(1)}, \ldots, \phi_k^{(M)})$, yielding a multimodal sparse signal model

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

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

Given the multimodal dictionary $\mathcal{D} = \{\phi_k\}$ and the multimodal 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 [25], [26] can provide generally good sparse approximations. We defer the challenge of multimodal signal approximation using dictionaries until future work and, in the next section, we further detail the proposed multimodal data model.

### B. Synchrony and Shift Invariance in Multimodal Signals

Very often, the various modalities in a multimodal 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 multimodal 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, multimodal signal could share higher dimensions, and the notion of synchrony could refer to spatial co-localization, for example, in multispectral 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

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

be a multimodal function whose modalities $\phi^{(m)}$, $m = 1, \ldots, M$ 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 $\Delta$.

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 $\phi$ be a multimodal function localized in an interval centered on $t = 0$. The operator $T_p$ shifts $\phi$ to time $p \in \mathbb{R}$ in a straightforward way

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

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 multimodal 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 multimodal 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 [27]–[29], natural images [29]–[33] and video sequences [34]. In the next section, we propose a learning strategy adapted to synchronous multimodal signals.

## III. Learning Multimodal Dictionaries

Our goal is to design an algorithm capable of learning sets of multimodal synchronous functions adapted to particular classes of multimodal signals. However, the design of an algorithm for learning dictionaries of multimodal atoms is nontrivial 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 multimodal signals that involve huge arrays of data, the computational complexity of the algorithm becomes a challenging issue.
- Natural multimodal 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/s), while the video clip is a 3-D signal sampled with considerably lower temporal resolution ($\mathcal{O}(10^1)$ frames/s).

We will design a novel learning algorithm that captures the underlying structures of multimodal signals overcoming both of these difficulties. We propose to learn *synchronous multimodal generating functions* as introduced in the previous section using a generalization of the MoTIF algorithm [29]. In [29], the authors propose a method to learn generating functions successively. A constraint that imposes low correlation between the learned functions is also considered, such that no function is picked several times. Each 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 fasten the computation, the MoTIF algorithm learns the generating functions by alternatively localizing and learning interesting signal structures.

In this paper, we extend the MoTIF learning method by generalizing it to multimodal signals. Such generalization is not trivial, since in general different signal modalities have different dimensions, which makes them complicated to handle and compare. As will be shown in the following, the recursive nature of the MoTIF method and its "localize and learn" strategy makes it suitable to handle complex multimodal data. Moreover, the structure of the algorithm allows to enforce synchrony between modal structures in an easy and intuitive fashion.

The goal of the learning algorithm is to build a set $\mathcal{G} = \{g_k\}_{k=1}^K$ of multimodal 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 *resampling 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

$$\mathcal{D} = \{\{\mathcal{T}_p g_k\}_{p\in\mathbb{R}}, k = 1 \ldots K\}. \tag{3}$$

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:

$$g_1 = \underset{\|g^{(a)}\|_2 = \|g^{(v)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} \sum_i \left| \left\langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \right\rangle \right|^2 \tag{4}$$

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 (4).

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

$$\mathrm{UP}' : g_1^{(i)} = \underset{\|g^{(i)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} \left| \left\langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \right\rangle \right|^2 \tag{5}$$

where $i = a$. We can then solve (5) 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

$$\mathrm{CP} : g_k^{(i)} = \underset{\|g^{(i)}\|_2 = 1}{\arg\max} \frac{\sum_{n=1}^{N} \max_{p_n} \left| \left\langle f_n^{(i)}, \mathcal{T}_{p_n}^{(i)} g^{(i)} \right\rangle \right|^2}{\sum_{l=1}^{k-1} \sum_{q\in\mathbb{Z}} \left| \left\langle g_l^{(i)}, T_q g^{(i)} \right\rangle \right|^2} \tag{6}$$

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 (5), with the only difference that a decorrelation 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 (6) 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 $g_l (l = 1, \ldots, k-1)$, which has at the same time to be minimized, being represented by the denominator.

Finding the best solution to the unconstrained problem $(\mathrm{UP}')$ or the constrained problem (CP) is indeed hard. However, the problem can be split into several simpler steps following a *localize and learn* paradigm [29]. 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

$$q_n^{(a)}[j] := \underset{q\in\mathbb{Z}}{\arg\max} \left| \left\langle f_n^{(a)}, T_q g_k^{(a)}[j-1] \right\rangle \right|.$$

2) ***Learn***: Update $g_k^{(v)}[j]$ by solving $\mathrm{UP}'$ (5) or CP (6) only for modality $(v)$, with the translations fixed to the values $p_n = p_n^{(a)}[j]$ found at step 1, i.e., $q_n^{(v)} := \mathrm{nint}(\mathrm{RF} \times q_n^{(a)}[j])$.
3) ***Localize***: Find the best translations $p_n^{(v)}[j] := \nu^{(v)} \cdot q_n^{(v)}[j]$ using the function $g_k^{(v)}[j]$

$$q_n^{(v)}[j] := \underset{q\in\mathbb{Z}}{\arg\max} \left| \left\langle f_n^{(v)}, T_q g_k^{(v)}[j] \right\rangle \right|.$$

4) ***Learn***: Update $g_k^{(a)}[j]$ by solving $\mathrm{UP}'$ (5) or CP (6) only for modality $(a)$, with the translations fixed to the values $p_n = p_n^{(v)}[j]$ found at step 3, i.e., using $q_n^{(a)} = \mathrm{nint}(q_n^{(v)}[j]/\mathrm{RF})$.

The first and third steps consist of 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

$$\mathbf{g}_k^{(i)}[j] = \underset{\|\mathbf{g}^{(i)}\|_2=1}{\arg\max}\, \mathbf{g}^{(i)^T} \mathbf{A}^{(i)}[j] \mathbf{g}^{(i)}. \tag{7}$$

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 (5) can be expressed as the quadratic form (7), 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 decorrelated as possible from all the atoms in $\mathcal{D}_{k-1}$. This corresponds to minimizing

$$\sum_{l=1}^{k-1} \sum_{q \in \mathbb{Z}} \left| \left\langle T_{-q} g_l^{(i)}, g^{(i)} \right\rangle \right|^2 \tag{8}$$

or, denoting

$$\mathbf{B}_k^{(i)} = \sum_{l=1}^{k-1} \sum_{q \in \mathbb{Z}} \mathbf{g}_{l,-q}^{(i)} \mathbf{g}_{l,-q}^{(i)\,T} \tag{9}$$

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

$$\mathbf{g}_k^{(i)}[j] = \underset{\|\mathbf{g}^{(i)}\|_2=1}{\arg\max}\, \frac{\mathbf{g}^{(i)^T} \mathbf{A}^{(i)}[j] \mathbf{g}^{(i)}}{\mathbf{g}^{(i)^T} \mathbf{B}_k^{(i)} \mathbf{g}^{(i)}}. \tag{10}$$

The best generating function $\mathbf{g}_k^{(i)}[j]$ is the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem defined in (10). 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 (6) can be solved as the generalized eigenvector problem (10) because of the particular quadratic form imposed to the objective function to optimize, when the translations $p_n$ are fixed.

The proposed multimodal learning algorithm is summarized in **Algorithm 1**.

**Algorithm 1** Principle of the Multimodal Learning Algorithm

$k = 0$, training set $\{(f_n^{(a)}, f_n^{(v)})\}$
**for** $k = 1$ to $K$ **do**
  $j \leftarrow 0$
  random initialization of $\{(g_k^{(a)}[j], g_k^{(v)}[j])\}$
  compute constraint matrices $\mathbf{B}_k^{(a)}$ and $\mathbf{B}_k^{(v)}$ as in (9)
  **while** no convergence reached **do**
    $j \leftarrow j + 1$
    **localize in modality** $(a)$:
    for each $f_n^{(a)}$, find the translation $p_n^{(a)}[j] \leftarrow \nu^{(a)} \cdot \arg\max_q |\langle f_n^{(a)}, T_q g^{(a)}[j-1] \rangle|$ maximally correlating $f_n^{(a)}$ and $g^{(a)}[j-1]$
    **learn modality** $(v)$:
    set $\mathbf{A}^{(v)}[j] \leftarrow \sum_{n=1}^{N} \mathbf{f}_{n,-p_n^{(a)}[j]}^{(v)} \mathbf{f}_{n,-p_n^{(a)}[j]}^{(v)\,T}$
    find $\mathbf{g}_k^{(v)}[j]$, the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem $\mathbf{A}^{(v)}[j]\mathbf{g} = \lambda \mathbf{B}_k^{(v)} \mathbf{g}$, using (10)
    **localize in modality** $(v)$ :
    for each $f_n^{(v)}$, find the translation $p_n^{(v)}[j] \leftarrow \nu^{(v)} \cdot \arg\max_q |\langle f_n^{(v)}, T_q g^{(v)}[j] \rangle|$ maximally correlating $f_n^{(v)}$ and $g^{(v)}[j]$
    **learn modality** $(a)$:
    set $\mathbf{A}^{(a)}[j] \leftarrow \sum_{n=1}^{N} \mathbf{f}_{n,-p_n^{(v)}[j]}^{(a)} \mathbf{f}_{n,-p_n^{(v)}[j]}^{(a)\,T}$
    find $\mathbf{g}_k^{(a)}[j]$, the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem $\mathbf{A}^{(a)}[j]\mathbf{g} = \lambda \mathbf{B}_k^{(a)} \mathbf{g}$, using (10)
  **end while**
**end for**.

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 [29] and the multimodal constrained algorithm typically converge in few steps to a stable solution independently of the initialization.

## IV. Experiments

The described framework is of a wide scope and both signal model and learning algorithm can be applied to different types of multimodal data. In this section, we demonstrate them for audio-visual analysis. In the first experiment, we want to show that the learning algorithm is capable of discovering salient audiovisual patterns from a set of training patches. Audio-video patches are extracted from sequences of talking mouths; thus, we expect the emerging multimodal generating functions to represent meaningful audio patterns like words or phonemes with corresponding video components showing movements of mouth

parts during the utterances. We will see that these are exactly the kind of patterns that the algorithm recovers. With the second experiment, we want to confirm the intuition that the learned functions effectively capture important signal structures. We will show that detecting the learned multimodal patterns in audiovisual sequences exhibiting severe acoustic and visual distractors, it is possible to localize audiovisual sources. The localization algorithm is effective and it outperforms existing audio-video localization methods.

### A. Audiovisual Dictionaries

This experiment demonstrates the capability of the 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 subsampled to 8 kHz, while the gray-scale video was recorded at 29.97 frames/second (fps) and at a resolution of $70 \times 110$ pixels. The total length of the training sequences is 1060 video frames, i.e., approximately 35 s, for $\mathcal{D}_1$, and 1140 video frames, i.e., approximately 38 s, 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 resampling factor RF equal to the ratio between the two frequencies has to be applied. In this case, the value of the resampling factor is $\mathrm{RF} = 8000/29.97 \approx 267$. Video sequences are filtered following the procedure suggested in [34], 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 [31], [35], 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 $31 \times 31$ 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 $16 \times 16$ pixels in space and 12 frames in time. The 20 elements of $\mathcal{D}_2$ are shown in Fig. 2. 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), and *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 [29], both components of generating function #2 are mainly high frequency due to the decorrelation constraint with the first atom.

The video components have characteristics similar to those found in [34]. Instead, audio waveforms are quite different to those learned in [27], where emerging speech basis functions are similar to sines and cosines functions. We suppose that there are at least three aspects that influence this result. First, the learned functions here are much longer than in [27] (3204 against 64 samples at the same sampling frequency of 8 kHz); thus, we have the opportunity to find complex phonemes and not only sines and cosines-like functions. Second, the proposed learning method searches for signal patches that are as similar as possible between them in the *localization* step. Thus, each generating function is a sort of principal component but computed for patches representing similar structures. It is, thus, not surprising that the "average" dominant components of such structures represent some entities averaging words or phonemes. Finally, the constraint of synchrony between audio and video patterns encourages the algorithm to focus on training patches containing audio structures synchronous with visual movements, naturally highlighting words and phonemes.

To summarize, the learning algorithm captures well high-level signal structures representing the synchronous presence of meaningful acoustic and visual patterns. All the learned multimodal functions consist of 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. This result is indeed interesting, considering that audio-video generating functions are randomly initialized and no constraint on their shape is imposed.

### B. Audiovisual Speaker Localization

In this experiment, we want to test if the learned dictionaries are able to recover meaningful audiovisual patterns in real mul-

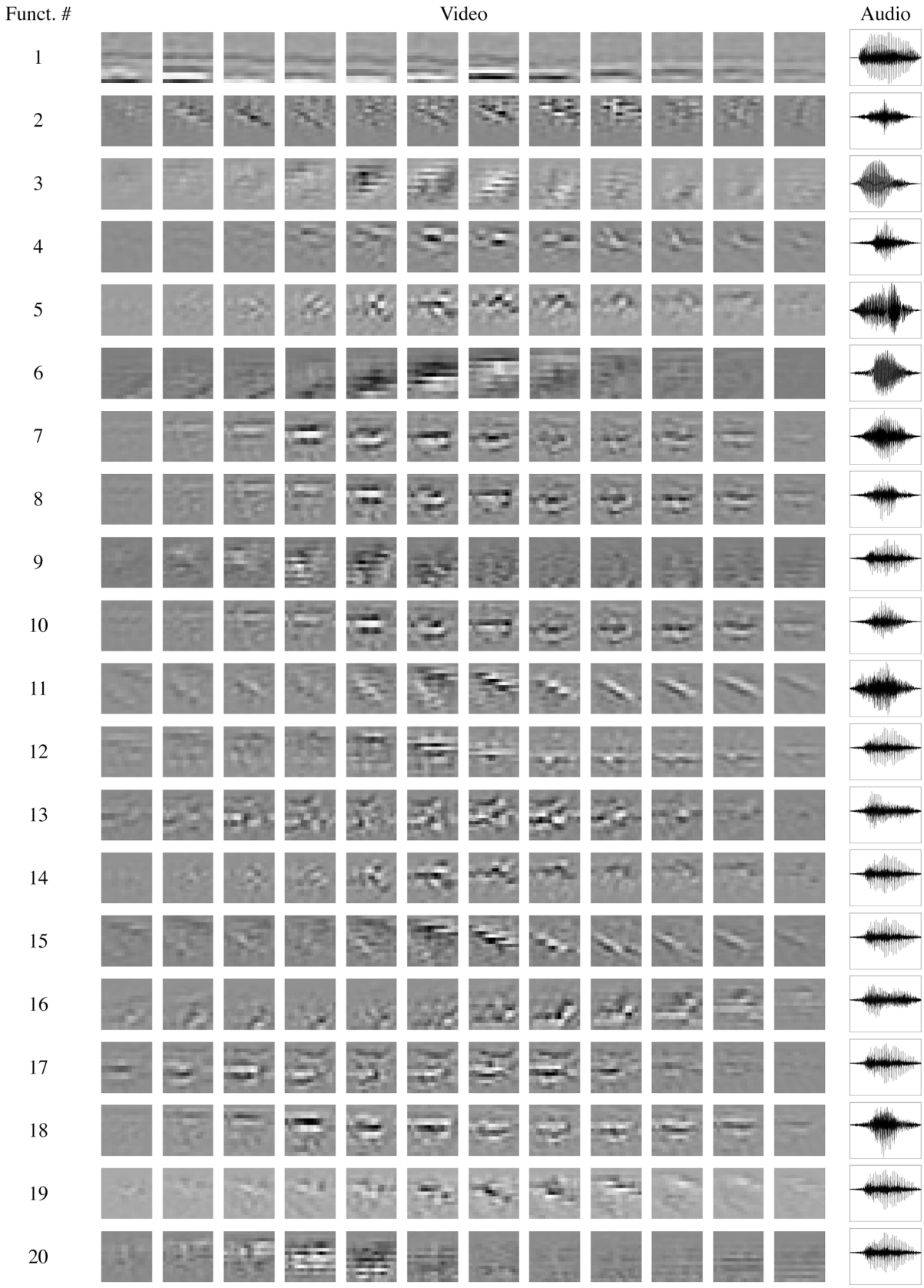


Fig. 2. Audio-video generating functions of *Dictionary 2*. Shown are the 20 learned functions, each consisting on an audio and a video component. Video components are on the left, with time proceeding left to right. Audio components are on the right, with time on the horizontal axis.

timedia 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`

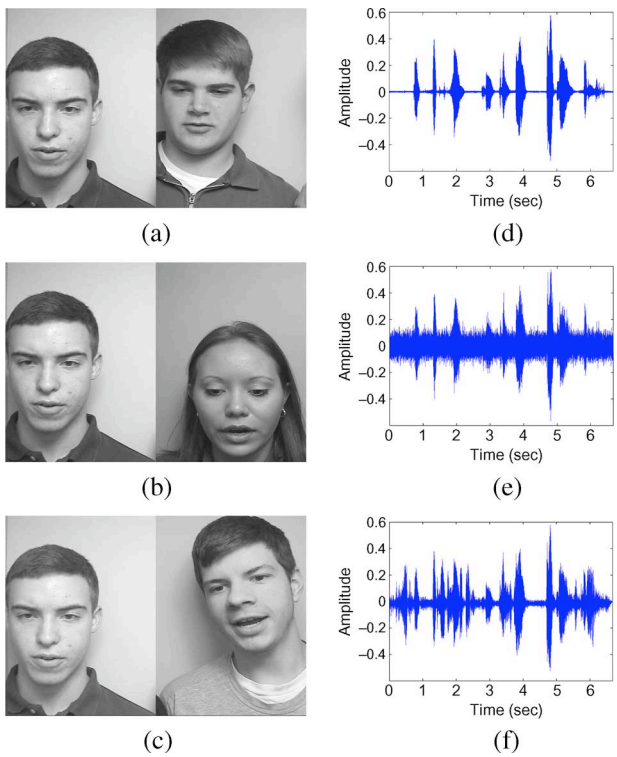


Fig. 3. Test sequences. Sample frames of (a) `Movie 1`, (b) `Movie 2` , and (c) `Movie 3` are shown on the left. (d) The original audio track a, together with (e) its noisy versions with additive gaussian noise a + AWGN and (f) added distracting speech and music a + speech are plotted on the right.

`2` and `Movie 3`, consisting of two people placed in front of the camera arranged as in Fig. 3. One of the subjects is uttering digits in English, while the other one is mouthing *exactly the same words*. Test sequences consist of 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 nontrivial. The 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 the following.

- For both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$, the positions of maximal projection between the dictionary atoms $\phi_k$ and the test sequences are localized on the actual location of the audio-visual 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 $(\mathrm{SNR} = 1\ \mathrm{dB})$ 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 $\mathcal{D}_2$.

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 $(\widehat{\vec{x}}_k, \hat{p}_k^{(v)})$ 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. 4, 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 position of the mouth center of the correct speaker has been manually annotated for each test sequence. The sound source location is considered to be correctly detected if it falls within a circle of radius 100 pixels centered in the labelled mouth. The sound source is correctly localized 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. 3(d)], as well as when considerable acoustic noise $(\mathrm{SNR} = 1\ \mathrm{dB})$ is present [signals a + AWGN and a + speech in Fig. 3(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

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

[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 noncritical.

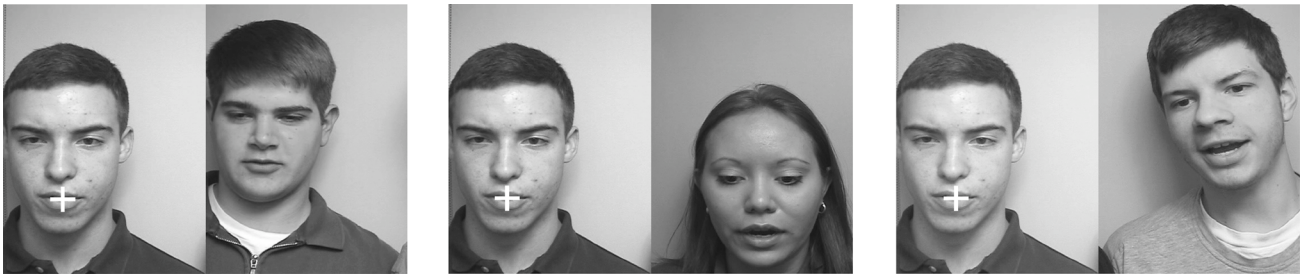

Fig. 4. Sample frames of (left) `Movie 1`, (center) `Movie 2`, and (right) `Movie 3`. The left person is the real speaker, the right subject mouths the same words pronounced by the speaker but his audio track has been removed. The white cross highlights the estimated position of the sound source, which is correctly placed over the speaker's mouth.

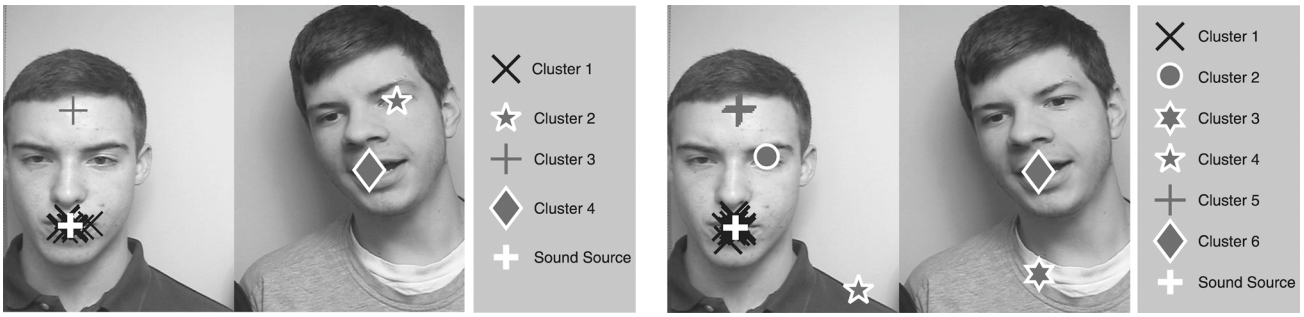


Fig. 5. Sample frames of `Movie 3`. The positions of maximal projection between video functions and test sequence are plotted on the image plane. Points belonging to the same cluster are indicated with the same marker. The biggest cluster is in both cases *Cluster 1*; it contains (left) 17 elements when $\mathcal{D}_1$ is used and (right) 13 when $\mathcal{D}_2$ is used.

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.

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. 5. 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.

At this point, it is interesting to compare the localization performances achieved using the learned dictionaries with those obtained by the *audiovisual Gestalts* detection method presented in [16]. The interest of such a comparison is twofold. First, the cross-modal localization algorithm introduced in [16] relies on signal representation techniques that model *separately* audio and video modalities using sparse decompositions over *general* dictionaries of Gabor and edge-like functions, respectively. This comparison is the occasion to check if a modeling of cross-modal correlations done at a level that is closer to the signals themselves (the model proposed here) than to the features (the model presented in [16]) is advantageous or not. Second, the audiovisual Gestalts localization algorithm is a generalization of our previous work on audiovisual signal representation [17]. Both algorithms exhibit state-of-the-art performances on the CUAVE database [23], outperforming the method presented in the only previously published systematic study on audiovisual speaker localization [15]. The comparison, thus, is significant *per se*.

The Gestalt detection algorithm [16] is tested on the same audiovisual sequences shown in Fig. 3, but resized to a resolution of $120 \times 176$ pixels to be more quickly processed. They are decomposed using 50 video atoms retrieved from a redundant dictionary of edge-like functions using the video approximation algorithm proposed in [36]. Each atom has a feature associated describing its displacement. The audio tracks are represented using 1000 Gabor atoms and a monodimensional feature that estimates the average acoustic energy is extracted. Meaningful audiovisual events are then defined as synchronous activations of audio and video features [16]. The video atoms exhibiting the highest degree of correlation with the audio are detected using a simple relevance criterion and the sound source location over the image sequence is estimated. Mouth positions have been manually labelled in the resized clips and the region of correct source

TABLE I
SUMMARY OF THE SOURCE LOCALIZATION RESULTS FOR ALL THE TESTED SEQUENCES

| Video | Audio | Dict. | Localization *learning* | $r$ | Localization *gestalts* [16] |
|---|---|---|---|---|---|
| Movie 1 | a | $\mathcal{D}_1$ | ✓ | 0.65 | ✓ |
| | | $\mathcal{D}_2$ | ✓ | 0.50 | |
| | a+AWGN | $\mathcal{D}_1$ | ✓ | 0.65 | ○ |
| | | $\mathcal{D}_2$ | ✓ | 0.50 | |
| | a+speech | $\mathcal{D}_1$ | ✓ | 0.65 | ✓ |
| | | $\mathcal{D}_2$ | ✓ | 0.50 | |
| Movie 2 | a | $\mathcal{D}_1$ | ✓ | 0.90 | ✓ |
| | | $\mathcal{D}_2$ | ✓ | 0.60 | |
| | a+AWGN | $\mathcal{D}_1$ | ✓ | 0.90 | ✓ |
| | | $\mathcal{D}_2$ | ✓ | 0.65 | |
| | a+speech | $\mathcal{D}_1$ | ✓ | 0.85 | ✓ |
| | | $\mathcal{D}_2$ | ✓ | 0.60 | |
| Movie 3 | a | $\mathcal{D}_1$ | ✓ | 0.85 | ○ |
| | | $\mathcal{D}_2$ | ✓ | 0.65 | |
| | a+AWGN | $\mathcal{D}_1$ | ✓ | 0.80 | ○ |
| | | $\mathcal{D}_2$ | ✓ | 0.65 | |
| | a+speech | $\mathcal{D}_1$ | ✓ | 0.85 | ○ |
| | | $\mathcal{D}_2$ | ✓ | 0.70 | |

detection is defined as a circle of diameter 25 pixels centered in the "real" mouth. Considering the down-sampling factor of 4 applied to these clips, the areas of correct mouth detection are the same for the two algorithms.

Table I summarizes the experimental results for all tested sequences and both localization methods (denoted as *learning* and *Gestalts*). 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 using the learned dictionaries and the fifth column indicates the reliability $r$ of the localization. In all cases, the audio source position is correctly found on the image plane, as indicated by the ticks ($\surd$). Finally, the sixth column reports the localization results for the audiovisual Gestalt detection method [16]. In this case, the speaker's mouth is erroneously detected on four out of nine clips, as indicated by the circles (○).

These results highlight that detecting the learned multimodal atoms, it is possible to effectively localize audiovisual sources in challenging real-world sequences. The algorithm proposed here outperforms the localization method presented in [16], which is more general (no specific assumption on the type of sequences is made and no training is required) but less robust to audio and video distractors. The audiovisual Gestalt model relies on the assumption that in general audio-video synchronous events occur randomly, except if a meaningful audiovisual source is observed. The test sequences employed here do not satisfy this hypothesis: visual distractors exhibit some strong correlation with the audio signal since the characters on the right in the test clips utter the same words pronounced by the real speaker. The proposed localization method overcomes these difficulties exploiting the temporal proximity between adapted audio and video patterns.

## V. CONCLUSION

In this paper, we present a new method to learn translation invariant multimodal functions adapted to a class of multicomponent 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 decorrelation 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 multimodal 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 multimodal signals, like climatologic or EEG-fMRI data, are foreseen.

## REFERENCES

[1] 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, Feb. 1997.

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

[3] I. R. Farah, M. B. Ahmed, and M. R. Boussema, "Multispectral satellite image analysis based on the method of blind separation and fusion of sources," in *Proc. Int. Geoscience and Remote Sensing Symp.*, 2003, vol. 6, pp. 3638–3640.

[4] K. C. Partington, "A data fusion algorithm for mapping sea-ice concentrations from special sensor microwave/imager data," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 4, pp. 1947–1958, Apr. 2000.

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

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

[7] 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, Sep. 2003.

[8] 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, Sep. 2005.

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

[10] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 813–819, 1999.
[11] M. Slaney and M. Covell, "FaceSync: A linear operaior for measuring synchronization of video facial images and audio tracks," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 814–820, 2000.
[12] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. ICA*, Apr. 2003, pp. 709–714.
[13] J. W. Fisher, III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 406–413, Jun. 2004.
[14] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, Apr. 2007.
[15] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. Int. Conf. Image and Video Retrieval*, 2003, pp. 488–199.
[16] G. Monaci and P. Vandergheynst, "Audiovisual Gestalts," in *Proc. IEEE CVPR Workshop*, 2006, p. 200.
[17] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Process.*, vol. 86, no. 12, pp. 3534–3548, 2006.
[18] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, vol. 381, pp. 66–68, 1996.
[19] 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," *Exp. Brain Res.*, vol. 158, pp. 252–258, 2004.
[20] S. Watkins, L. Shams, S. Tanaka, J.-D. Haynes, and G. Rees, "Sound alters activity in human V1 in association with illusory visual perception," *Neurolmage*, vol. 31, no. 3, pp. 1247–1256, 2006.
[21] A. Violentyev, S. Shimojo, and L. Shams, "Touch-induced visual illusion," *Neuroreport*, vol. 10, no. 16, pp. 1107–1110, 2005.
[22] J.-P. Bresciani, F. Dammeier, and M. Emst, "Vision and touch are automatically integrated for the perception of sequences of events," *J. Vis.*, vol. 6, no. 5, pp. 554–564, 2006.
[23] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1189–1201, 2002.
[24] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proc. IEEE Int. Conf. Image Processing*, 2005, vol. 3, pp. 46–49.
[25] 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. IEEE Int. Conf. Image Processing*, 2002, vol. 3, pp. 3057–3060.
[26] J. Tropp, A. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proc. IEEE ICASSP*, 2005, vol. 5, pp. 721–724.
[27] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.
[28] S. Abdallah and M. Plumbley, "If edges are the independent components of natural images, what are the independent components of natural sounds?," in *Proc. ICA*, 2001, pp. 534–539.
[29] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "MoTIF: An efficient algorithm for learning translation invariant dictionaries," in *Proc. IEEE ICASSP*, 2006, vol. 5, pp. 857–860.
[30] A. Bell and T. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vis. Res.*, vol. 37, no. 23, pp. 3327–3338, 1997.
[31] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, pp. 3311–3327, 1997.
[32] M. Lewicki and B. A. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Amer. A*, vol. 16, no. 7, pp. 1587–1601, 1999.
[33] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, pp. 349–396, 2003.
[34] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. IEEE Int. Conf. Image Processing*, 1, 2003, pp. 41–44.
[35] D. Dong and J. Atick, "Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Network: Comput. Neural Syst.*, vol. 6, pp. 159–178, 1995.
[36] O. Divorra Escoda, "Toward sparse and geometry adapted video approximations" Ph.D. dissertation, Swiss Fed. Inst. Technol., Lausanne, Switzerland, Jun. 2005 [Online]. Available: http://lts2www.epfl.ch/.

**Gianluca Monaci** received the degree in telecommunication engineering from the University of Siena, Siena, Italy, in 2002, and the the Ph.D. degree in signal and image processing from the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, in 2007.

He is currently a Postdoctoral Researcher at EPFL. His research interests include image, video, and multimodal signal representation and processing, learning algorithms design, and source separation problems.

Dr. Monaci received the IBM Student Paper Award at the 2005 ICIP Conference, Genova, Italy, for his work on multimodal signal analysis.

**Philippe Jost** received the M.S. degree in communication systems from the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, in 2002, where he is currently pursuing the Ph.D. degree in the Signal Processing Laboratory and where his research focuses on sparse representations.

**Pierre Vandergheynst** (M'01) received the M.S. degree in physics and the Ph.D. degree in mathematical physics from the Université catholique de Louvain, Louvain-la-Neuve, Belgium, in 1995 and 1998, respectively.

From 1998 to 2001, he was a Postdoctoral Researcher with the Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland. He is now an Assistant Professor at EPFL, where his research focuses on harmonic analysis, information theory, sparse approximations, and mathematical image processing with applications to higher dimensional and complex data processing.

Dr. Vandergheynst was Co-Editor-in-Chief of *Signal Processing* (2002–2006).

**Boris Mailhé** was born in 1983 in Nice, France. He graduated from ENS Cachan and Rennes I University in 2005. He is currently pursuing the Ph.D. degree at IRISA, Rennes, under the direction of F. Bimbot, R . Gribonval, and P. Vandergheynst.

**Sylvain Lesage** received the engineer degree in signal processing from Supélec, Metz, France, in 2003, and the Ph.D. degree in signal processing and telecommunications from Rennes I University, Rennes, France, in 2007.

He is currently an Associate Researcher in IRISA, Rennes. His research focuses on dictionary learning for sparse modeling and separation of multichannel signals.

**Rémi Gribonval** graduated from the École Normale Supérieure, Paris, France, in 1997, and received the Ph.D. degree in applied mathematics from the University of Paris-IX Dauphine in 1999.

From 1999 until 2001, he was a Visiting Scholar in the Department of Mathematics, Industrial Mathematics Institute (IMI), University of South Carolina, Columbia. He is currently a Research Associate with the French National Center for Computer Science and Control (INRIA) at IRISA, Rennes, France. His research interests are in adaptive techniques for the representation and classification of audio signals with redundant systems, with a particular emphasis in blind audio source separation.