LEARNING SPARSE GENERATIVE MODELS OF AUDIOVISUAL SIGNALS

Gianluca Monaci[*]*, Friedrich T. Sommer*[*] *and Pierre Vandergheynst*[†]

[*] Redwood Center for Theoretical Neuroscience
University of California Berkeley
Berkeley, CA 94720-3190, USA
{gmonaci,fsommer}@berkeley.edu
http://redwood.berkeley.edu

[†] Signal Processing Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
pierre.vandergheynst@epfl.ch
http://lts2www.epfl.ch

ABSTRACT

This paper presents a novel framework to learn sparse representations for audiovisual signals. An audiovisual signal is modeled as a sparse sum of audiovisual kernels. The kernels are bimodal functions made of synchronous audio and video components that can be positioned independently and arbitrarily in space and time. We design an algorithm capable of learning sets of such audiovisual, synchronous, shift-invariant functions by alternatingly solving a coding and a learning procedure. The proposed methodology is used to learn audiovisual features from a set of bimodal sequences. The basis functions that emerge are audio-video pairs that capture salient data structures.

1. BACKGROUND AND SIGNIFICANCE

Everyday tasks involve complex interactions between different sensory modalities. Indeed, a variety of cross-modal integration phenomena occur at various processing levels in our brain [1, 2]. Recently, cross-modal integrating strategies inspired by experimental results in human subjects begin to be successfully used in many signal processing and computer vision problems involving mutually related signals. Examples include speech-speaker recognition [3] and detection [4] aided by video, audio filtering and enhancement based on video [5], or audiovisual sound source localization [6–13].

Typically, audiovisual fusion algorithms exploit the correlation across modalities by looking for structures showing a certain degree of synchrony. In their pioneering work, Hershey and Movellan [6] assessed the interdependency between audio and video simply measuring the correlation coefficient between acoustic energy and the evolution of single pixel values. Since then, more sophisticated audio-video features and audiovisual fusion models have been developed. Audio features are based on audio energy [8, 9, 11] or cepstral representations [4, 7, 10], while video features are pixel intensity values [8, 11] or descriptors of visual changes [4, 7, 9, 10]. Audiovisual interplay is modeled using techniques based on Canonical Correlation Analysis (CCA) [7, 9] or on the estimation of the joint distributions of audiovisual features [4, 8, 10, 11]. In our previous work [5, 12] we propose an audiovisual analysis technique based on sparse features that allows to intuitively define and detect synchronous audiovisual patterns.

All these models make use of hand-designed audio and video features that are correlated using some statistical measure of interdependency. In contrast to previous studies, here we propose a model that *learns* sparse signal representations. The goal is to build codes adapted to the audiovisual signal and that allow to represent relevant data structures in an intuitive and natural way.

This idea has been first explored in our earlier work [13], where audiovisual signals are modeled as sparse sums of *audiovisual basis functions*. An example of audiovisual basis is depicted in Fig. 1. It 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 at different temporal resolution. In [13], a method to learn collections of such multimodal kernels is proposed as well.

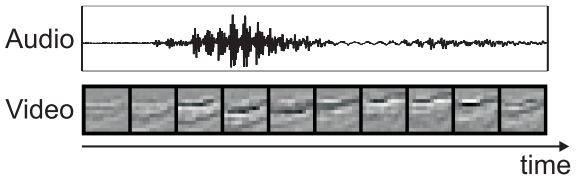


Figure 1: An audiovisual function composed of an audio [Top] and a video part [Bottom] sharing a common temporal axis. Video frames are represented as a succession of images.

In this work we propose a new model where bimodal signal structure is captured by a sparse generative model [14]. The *bimodal signal structure* is the audiovisual signal component that is informative for sensor fusion. Conversely, signal structure that exclusively resides in single modalities is incompletely encoded. An audiovisual signal is thus represented as a sparse sum of audiovisual kernels. Such kernels are bimodal functions like the one shown in Fig. 1 that can be positioned independently and arbitrarily in space and time. We design an algorithm capable of learning sets of such audiovisual, synchronous, shift-invariant features by alternatingly solving a *coding* and a *learning* procedure. This work improves our previous audiovisual learning algorithm in two important aspects:

1. We extend the model in [13] in order to represent audiovisual signals in terms of kernels that are invariant not only to temporal but also to *spatial* translations.
2. In [13], the learned multimodal dictionaries are collections of frequently re-occurring patterns, but the learning does not take into account the sparse coding problems. Here the coding and the learning procedures are alternatingly solved to form sparse audiovisual signal representations.

The paper is structured as follows: Sec. 2 describes the proposed audiovisual signal model. Sec. 3 presents the coding and learning algorithms for audio-video signals. In Sec. 4 experimental results based on synthetic and natural audiovisual data are shown. Sec. 5 concludes the paper with a summary of the achieved results and of the possible developments of this research.

2. CONVOLUTIONAL MODEL FOR AUDIOVISUAL SIGNALS

Audiovisual data structures are made up of two different modalities (audio and video) and they can be represented as couple $s = (a, v)$. The two components a and v are not homogenous in dimensionality: the audio signal is a 1-D stream $a(t)$ and the video sequence is a 3-D signal $v(x, y, t)$ with (x, y) the pixel position. An audiovisual signal can be represented as a sum of audiovisual *atoms* $\phi_k = (\phi_k^{(a)}(t), \phi_k^{(v)}(x, y, t))$ like the one shown in Fig. 1, taken from a multimodal dictionary $\mathscr{D} = \{\phi_k\}$ [13]. Each atom consists of an audio and a video component with unitary ℓ_2 norm.

Audiovisual signals share a common temporal dimension, and

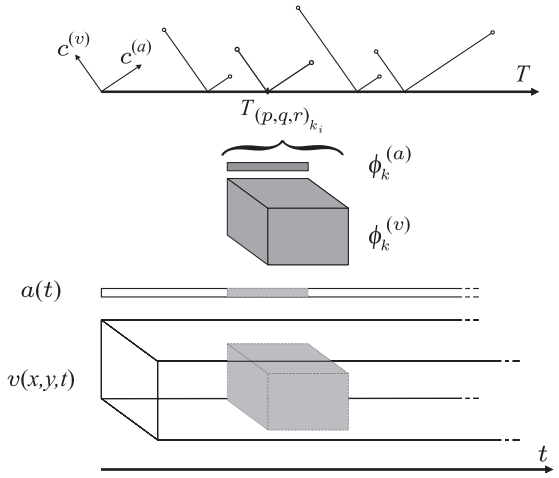


Figure 2: Schematic representation of the audiovisual code. The signal $s = (a(t), v(x,y,t))$ is modeled as a linear sum of kernels $\phi_k = (\phi_k^{(a)}, \phi_k^{(v)})$, $\phi_k^{(a)}$ being a 1-D audio function as in Fig. 1 [Top] and $\phi_k^{(v)}$ a video function as in Fig. 1 [Bottom]. Each kernel is localized in space and time and may be applied at any spatio-temporal position T within the signal.

temporal synchrony between audio and video stimuli is a very important feature, tightly linked to the physics of the problem. Sound in the audio time series is in fact 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. Let $\phi = (\phi^{(a)}(t), \phi^{(v)}(x,y,t))$, be an audiovisual function whose modalities $\phi^{(a)}$ and $\phi^{(v)}$ share a common temporal dimension $t \in \mathbb{R}$. A modality is temporally localized in the interval $\Delta \subset \mathbb{R}$ if $\phi^{(a)}(t) = 0$ and $\phi^{(v)}(x,y,t) = 0$, $\forall t \notin \Delta$. We will say that the modalities are synchronous when $\phi^{(a)}$ and $\phi^{(v)}$ are localized in the same time interval Δ.

Most natural signals exhibit characteristics that are shift-invariant, meaning that they can occur at any instant in time and space. 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 an audiovisual function localized in an interval centered on $t = 0$. The operator $T_{(p,q,r)}$ acts on ϕ in a straightforward way:

$$T_{(p,q,r)}\phi = \left(\phi^{(a)}(t-r), \phi^{(v)}(x-p, y-q, t-r)\right). \qquad (1)$$

This translation is homogeneous in time across channels and thus preserves synchrony. With these definitions, it becomes easy to express a signal as a superposition of synchronous multimodal patterns ϕ_k, occurring at various time instants and in different spatial positions:

$$s \approx \sum_{k=0}^{K-1} \sum_{i=1}^{n_k} c_{k_i} T_{(p,q,r)_{k_i}} \phi_k, \qquad (2)$$

where the pair $c_{k_i} = (c_{k_i}^{(a)}, c_{k_i}^{(v)})$ specify the coefficients of the i-th instance of kernel ϕ_k. The index n_k indicates the number of instances of ϕ_k, which need not be the same across kernels. In general, the audio and video modalities are weighted by different coefficients, $c_{k_i}^{(a)}$ and $c_{k_i}^{(v)}$, since the same audio-video pattern may occur along a sequence with different relative intensities: for example the same mouth movement may produce the same sound but with different acoustic intensity. The model is schematically illustrated in Fig. 2.

3. LEARNING SPARSE AUDIOVISUAL CODES

Our goal is to design a model for preferentially learning bimodal signal structures that are informative in sensor fusion. This can be done alternatingly solving two tasks:

- ***Sparse Coding***, i.e. finding the optimal translations $T_{(p,q,r)_k}$ and coefficients $c_k = (c_k^{(a)}, c_k^{(v)})$;
- ***Learning***, i.e. finding the optimal basis functions $\phi_k = (\phi_k^{(a)}(t), \phi_k^{(v)}(x,y,t))$.

Given a certain dictionary, there are many different methods to find an encoding of a signal, while in general, finding the optimal sparse representation of arbitrary signals using a generic dictionary is a very hard problem, which turns out to be NP-hard. For this reason approximate feasible solutions have to be considered. Here we use an extension of the Matching Pursuit (MP) algorithm [15] to find the values of $T_{(p,q,r)_k}$ and c_k. MP has been used in some recent work to compute sparse codes of audio signals [16] and images [17], showing to yield efficient and robust representations. In the next section we recall the basic concepts of MP and we introduce an extension of the method to audiovisual signals.

3.1 Sparse Coding: the Audiovisual Matching Pursuit

The goal of the sparse coding step is to find the values of $T_{(p,q,r)_k}$ and c_k to represent bimodally informative audiovisual structures according to the sparse model (2). We extend MP to multimodal data as inspired by the Simultaneous Orthogonal MP (S-OMP) algorithm of Tropp and colleagues [18]. Our aim is to represent the audiovisual signal s in terms of functions ϕ_k taken from a redundant dictionary $\mathscr{D}$, as expressed by (2). Since audio and video signals have in general different dimensionality and different temporal sampling rates, S-OMP has to be extended to account for those differences. To this end we introduce the Audiovisual Matching Pursuit algorithm (AV-MP) described in this section.

Since we will deal with digital audio and video signals, in order to proceed we first need to define the time-discrete version $\mathscr{T}_{(p,q,r)}$, $p,q,r \in \mathbb{R}$ of the synchronous translation operator (1). Different modalities are in general sampled at different rates over time. In order to preserve their temporal proximity, the operator $\mathscr{T}$ must shift in time the signals on the two modalities by a different integer number of samples. We define $\mathscr{T}$ as

$$\mathscr{T}_{(p,q,r)} = \left(\mathscr{T}_r^{(a)}, \mathscr{T}_{(p,q,r)}^{(v)}\right) := \left(T_{\rho^{(a)}}, T_{p,q,\rho^{(v)}}\right),$$

where $T_{\rho^{(a)}}$ translates an audio signal by $\rho^{(a)} \in \mathbb{Z}$ samples and $T_{p,q,\rho^{(v)}}$ translates a video signal by $\rho^{(v)}$ time samples and p and q pixels. Therefore $\mathscr{T}_{(p,q,r)}$, $p,q,r \in \mathbb{R}$, is defined with discrete time translations $\rho^{(a)} := \mathrm{nint}(r/\nu^{(a)}) \in \mathbb{Z}$ and $\rho^{(v)} := \mathrm{nint}(r/\nu^{(v)}) \in \mathbb{Z}$, where $\mathrm{nint}(\cdot)$ is the nearest integer function. In the experiments that we will conduct at the end of this paper, $\nu^{(a)} = 1/8000$ for audio signals sampled at 8 kHz and $\nu^{(v)} = 1/29.97$ for videos at 29.97 frames per second (fps).

The Audiovisual Matching Pursuit algorithm iteratively approximates the multimodal signal $s = (a, v)$ with successive projections onto the audiovisual dictionary made of the functions $\phi_k = (\phi_k^{(a)}(t), \phi_k^{(v)}(x,y,t))$. The first step of the AV-MP algorithm decomposes s as

$$\begin{aligned} s &= R^0 s = \left(R^0 a, R^0 v\right) \\ &= \left(\langle a, \phi_0^{(a)}(t-\rho_0^{(a)})\rangle \phi_0^{(a)}, \langle v, \phi_0^{(v)}(x-p_0, y-q_0, t-\rho_0^{(v)})\rangle \phi_0^{(v)}\right) + R^1 s, \end{aligned}$$

where $R^1 s$ is the residual component after projecting s in the subspace described by ϕ_0. The selection of the function ϕ_0 to use and its position $(p,q,r)_0$ are chosen such that the sum of projections $|\langle a, \mathscr{T}_r^{(a)} \phi^{(a)}\rangle| + |\langle v, \mathscr{T}_{(p,q,r)}^{(v)} \phi^{(v)}\rangle|$ is maximal [18]. The projections $(|\langle a, \mathscr{T}_r^{(a)} \phi_0^{(a)}\rangle|, |\langle v, \mathscr{T}_{(p,q,r)}^{(v)} \phi_0^{(v)}\rangle|)$ represent the pair of coefficients $\hat{c}_0 = (\hat{c}_0^{(a)}, \hat{c}_0^{(v)})$. Recursively applying this procedure, after N iterations we can approximate s with $\hat{s}$ as

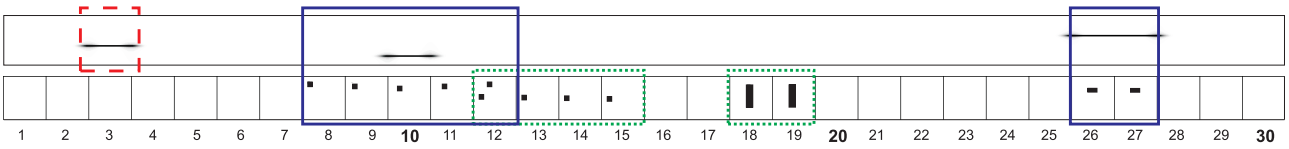


Figure 3: Synthetic example. The top plot is the spectrogram of the audio part, consisting of three sine pulses at different frequencies. The bottom plot shows the video part consisting of 30 video frames. The clip shows four black geometric shapes on a white background. Five events are present in this sequence, one audio-only structure (Red dashed box), two visual-only structures (Green dotted) and two audio-visual structures (Blue).

$$\hat{s} = \sum_{n=0}^{N-1} \left(\hat{c}_n^{(a)} \phi_n^{(a)}(t-\rho_n^{(a)}), \hat{c}_n^{(v)} \phi_n^{(v)}(x-p_n, y-q_n, t-\rho_n^{(v)}) \right), \quad (3)$$

where $\hat{c}_n^{(m)} = \langle R^n m, \mathscr{T}^{(m)} \phi_k^{(m)} \rangle$, $m = a, v$. The algorithm can be stopped either after a fixed number N of iterations either when the value of the projection $\hat{c}$ drops below a certain threshold.

Please note that by separating the sum over n into two sums, one over k (the kernel functions) and one over i (the number of instances of each kernel), we find again the sparse signal model (2):

$$\hat{s} = \sum_k \sum_i \hat{c}_{k_i} \mathscr{T}_{(p,q,r)_{k_i}} \phi_k .$$

3.2 Learning

The AV-MP algorithm provides a way to encode signals given a set of audiovisual kernel functions, but the efficiency of this code depends on how well the kernel functions capture the structure of a given class of signals. To optimize the kernel functions we use unsupervised learning based on gradient descent [14]. Gradient descent algorithms have been successfully employed in recent years for learning sparse signal representations, showing to be able to find biologically plausible codes for acoustic [16] and visual data [14, 17].

We start from the observation [14] that one can rewrite (2) in probabilistic form as $p(s|\mathscr{D}) = \int p(s|\mathscr{D}, c) p(c) dc$, with $p(c)$ a sparse prior on the usage of dictionary elements. It is common to approximate the integral by the maximum of the integrant (its mode), i.e.

$$p(s|\mathscr{D}) = \int p(s|\mathscr{D}, c) p(c) dc \approx p(s|\mathscr{D}, c^\star) p(c^\star) . \quad (4)$$

Here the optimal code $c^\star$ is approximated by the AV-MP decomposition of the signal, $\hat{c}$. Assuming the noise in the likelihood term, $p(s|\mathscr{D}, \hat{c})$, to be Gaussian with variance σ_N^2, the kernel functions can be iteratively updated taking the gradient ascent of the approximate log probability [14]:

$$\begin{aligned} \frac{\partial}{\partial \phi_k} \log(p(s|\mathscr{D})) &\approx \frac{\partial}{\partial \phi_k} \{\log(p(s|\mathscr{D}, \hat{c})) + \log(p(\hat{c}))\} \\ &\approx -\frac{1}{2\sigma_N^2} \frac{\partial}{\partial \phi_k} \left\| s - \sum_{k=0}^{K-1} \sum_{i=1}^{n_k} \hat{c}_{k_i} \mathscr{T}_{(p,q,r)_{k_i}} \phi_k \right\|^2 \\ &= \frac{1}{\sigma_N^2} \sum_{i=1}^{n_k} \hat{c}_{k_i} \{s - \hat{s}\}_{\mathscr{T}_{k_i}} , \end{aligned} \quad (5)$$

where $\{s-\hat{s}\}_{\mathscr{T}_{k_i}}$ indicates the residual error over the extent of kernel ϕ_k at position $\mathscr{T}_{(p,q,r)_{k_i}}$. Thus the functions ϕ_k are updated in Hebbian fashion, simply as a product of activity and residual [14, 16].

To summarize, we randomly initialize the basis functions and we iteratively update them using the rule

$$\phi_k = \phi_k + \eta \Delta \phi_k ,$$

where η is a constant learning rate and $\Delta\phi_k$ is the update step:

$$\begin{aligned} \Delta \phi_k &= \left(\Delta \phi_k^{(a)}, \Delta \phi_k^{(v)} \right) \\ &= \left(\sum_{i=1}^{n_k} \hat{c}_{k_i}^{(a)} \{a - \hat{a}\}_{\mathscr{T}_{r_{k_i}}^{(a)}}, \sum_{i=1}^{n_k} \hat{c}_{k_i}^{(v)} \{v - \hat{v}\}_{\mathscr{T}_{(p,q,r)_{k_i}}^{(v)}} \right) . \end{aligned} \quad (6)$$

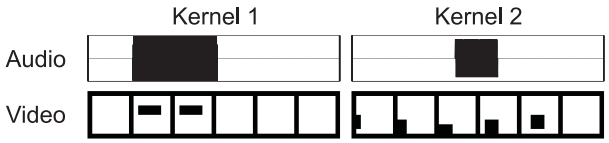


Figure 4: The two audiovisual kernels learned for the synthetic sequence in Fig. 3. Audio components are on the top, with time on the horizontal axis. Video components on the bottom, with time proceeding left to right (each image is a video frame).

After each update step the components of the audiovisual kernels are normalized to 1. The learning is halted after a given number of iterations M or when the change in the ℓ_2 norm of the basis functions is smaller than 1% (whichever comes first).

4. EXPERIMENTS

4.1 A Synthetic Example

We first consider a simple synthetic example to illustrate how the audiovisual sparse coding model works. The soundtrack consists of three sine waves at different frequencies (Fig. 3 [Top]), and the video shows four simple black shapes, static or moving on a white background (Fig. 3 [Bottom]). The sequence represents three possible audiovisual patterns: audio-only structure (Red dashed box), visual-only structures (Green dotted) and audiovisual structures (Blue). We use our algorithm to learn an audiovisual dictionary of 10 functions for this scene. The kernels have an audio component lasting 1602 samples and a video component of size 8×8 pixels and 6 frames in time. The algorithm learns two audiovisual functions (the remaining 8 were not used) that are shown in Fig. 4.

It is clear, observing the results, that the emerging audiovisual bases represent the two cross-modal structures highlighted in blue in Fig. 3. Function 1 shows the audiovisual pattern on frames 26–27, with the static rectangle and the synchronous sine wave, while function 2 depicts the moving square with the short sinusoidal pulse associated appearing on frames 8–12. This simple experiment suggests that our learning algorithm allows extracting meaningful cross-modal structures from data. When learning audiovisual kernels, the algorithm focuses on cross-modal structures, discarding audio-only and video-only components. Since in natural audiovisual streams, visual and auditory parts are often co-occurring, we can learn audiovisual patterns using the proposed method.

4.2 Learning Codes for Audiovisual Speech

The aim of this experiment is to demonstrate the capability of the learning algorithm to recover audiovisual patterns from natural signals. The training database consists of five audiovisual sequences representing the mouth of one speaker uttering the digits from zero to nine in English. Thus we expect the emerging audiovisual kernels to represent audio structures like words or phonemes with corresponding video components showing movements of mouth parts during the utterances.

Training audio tracks were at 44 kHz and down-sampled to 8 kHz and the gray-scale videos were recorded at 29.97 fps and at a resolution of 35×55 pixels. The total length of the training se-

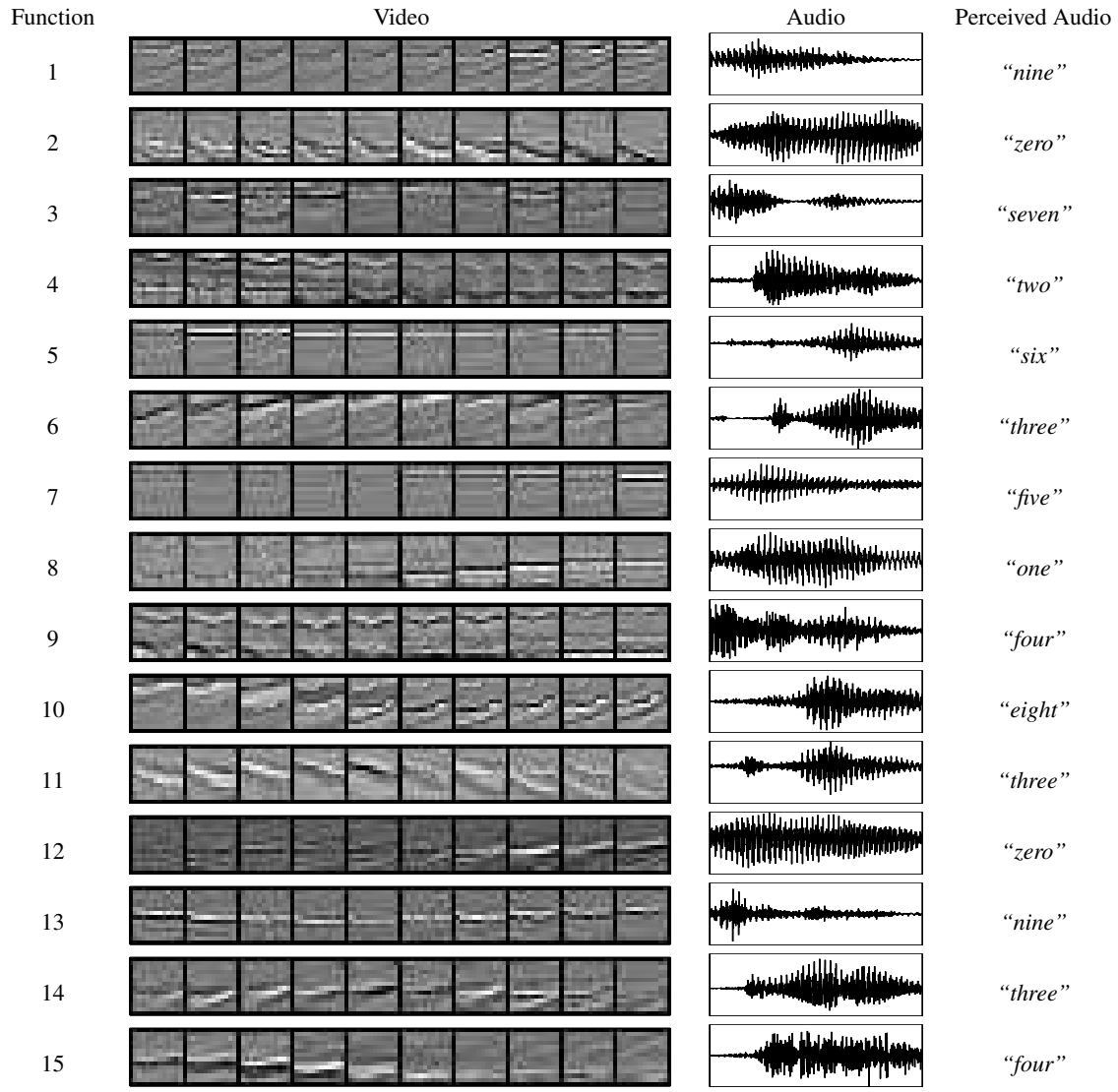


Figure 5: Fifteen learned audiovisual kernels. Video components are on the second column and are represented as a succession of video frames. Audio components are on the third column. Perceived sounds are marked in the forth column.

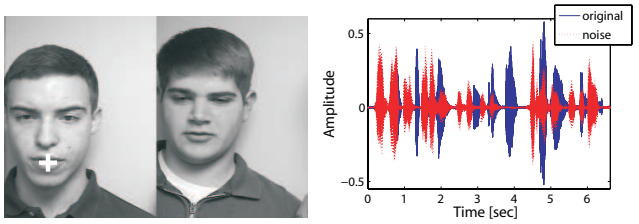
quences is 1310 video frames, i.e. approximately 44 seconds. The audio signal is considered as is while the video is whitened using the procedure described in [14] to speed up the training. We learn 30 audiovisual kernels with an audio component of 2670 samples and a video component of size $12 \times 12 \times 10$. The learned dictionary is shown in Fig. 5. Each function is represented as a video component (on the left), with time proceeding left to right, and an audio part (on the right), with time on the horizontal axis. Video components are spatially localized and oriented edge detector filters shifting from frame to frame. They clearly represent parts of the mouths making distinctive movements during the speech. The audio components feature the numbers present in the training set. Listening to the waveforms, one can hear the digits *zero* (functions 2, 12), *one* (8), *two* (4), *three* (6, 11, 14), *four* (9, 15), *five* (7), *seven* (3), *eight* (10), *nine* (1, 13). Function 5 seems to be a mixture of numbers *six* and *eight*. The digit *six* is difficult to learn because its audiovisual representation has both low acoustic energy and small corresponding lip motion. Different instances of the same digit have either different audio characteristics, like length or frequency content (e.g. functions 6, 11 and 14 all feature a *three*), or different associated video components (e.g. functions 2 and 12).

The set functions shown in Fig. 5 is qualitatively different from the dictionary, learned on the same dataset, reported in our recent paper [13]. The audiovisual kernels that emerge in this study are more heterogeneous, with a great variety of visual motion patterns and sounds. The visual functions exhibit very clear edge-like moving structures describing different visual patterns. Here the learned audio components represent all the digits present in the training set, which was not the case in [13]. Furthermore, the algorithm in [13], due to de-correlation constraints between atoms, learns few spurious audiovisual kernels that do not represent any real data structure. It should be also emphasized that the kernels learned here are invariant to temporal and spatial shifts, while those learned in [13] are only time-invariant. This is probably another reason for the richness of structures learned with our new method.

Overall, the functions learned here seem to depict more clearly underlying data patterns. One reason for this behavior of the algorithm is that the model proposed here integrates learning and coding in a way that is statistically and biologically more consistent [14, 16, 17].

4.3 Audiovisual Source Localization

Here we show that detecting the learned kernels in an audiovisual sequence exhibiting severe acoustic and visual distracters, it is pos-



(a) (b)

Figure 6: (a) Sample video frame. The white cross correctly pinpoints the position of the estimated audiovisual source. (b) Audio signal with the speech of the real speaker (blue line) and added noise signal with SNR = 0 dB (dashed red line). The test audio track is the sum of the two waveforms.

sible to localize the audiovisual source. We consider a clip consisting of two persons in front of the camera arranged as in Fig. 6 (a). One person (on the left) is uttering digits in English, while the other one is mouthing *exactly the same words*. Strong noise (SNR = 0dB) is mixed with the audio track by adding the signal of a male voice pronouncing numbers in English (Fig. 6 (b)). The speaker is the same subject whose mouth was used to train the audiovisual dictionary in Fig. 5; however, the training sequences are different from the test sequence. Such a sequence is difficult to analyze, since both persons mouth the same words at the same time (strong visual distracter) and the audio track is a mixture of two male voices, both uttering digits in English (strong acoustic distracter).

The audio track of the test clip is filtered with the audio component of each learned function. For each audio function we keep the temporal position of the maximum projection and we consider a window of 21 frames around this time position in the video. This restricted video patch is filtered with the corresponding video component and the spatio-temporal position of the maximum projection between the video signal and the video kernel is kept. Thus, for each learned audiovisual function we obtain the location of the maximum projection over the image plane. The locations of the maximal projections on the image plane are grouped into clusters using a hierarchical clustering algorithm, as described in [13]. The centroid of the cluster containing the largest number of points is kept as the estimated location of the sound source. We expect the position of the estimated sound source to be close to the speaker's mouth. Fig. 6 (a) shows a sample frame of the test sequence. The white cross indicates the estimated position of the sound source over the image plane, which coincides with the mouth of the speaker. Thus, the learned code can detect synchronous audiovisual patterns, allowing to localize the sound source on complex multimodal sequences.

5. SUMMARY

In this paper we have presented a new model to represent audiovisual signals as sparse sums of coupled audiovisual functions that are learned from real-world multimodal sequences. The emerging representation includes elements describing typical audiovisual features in the training signals. The proposed framework has been demonstrated on synthetic and natural data, showing that co-occurring audio-video events can be effectively learned, extracted and localized. Applications of the proposed model can range from robust cross-modal source localization [13], to blind audiovisual source separation [5], or to joint encoding of multimedia streams.

Interestingly, the framework developed here relies upon techniques that have been employed for the modeling of perceptual mechanisms [14, 16, 17]. We think that our model might relate to what human perception does. Since an intriguing and still unresolved question arising in the neuroscience community concerns the nature of cross-modal integration mechanisms in human the brain [1, 2], we believe that our audiovisual learning model can provide an interesting starting point for the theoretical analysis of cross-modal interactions in human perception.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through the prospective researcher grant n° PBEL2-118742 to Gianluca Monaci.

REFERENCES

[1] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 505–509, 2001.

[2] D. A. Bulkin and J. M. Groh, "Seeing sounds: visual and auditory interactions in the brain," *Current Opinion in Neurobiology*, vol. 16, no. 4, pp. 415–419, 2006.

[3] 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.

[4] P. Besson, V. Popovici, J.-M. Vesin, J.-Ph. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 63–73, 2008.

[5] A. Llagostera, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation using overcomplete dictionaries," in *Proc. IEEE ICASSP*, 2008.

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

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

[8] 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.

[9] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Proc.*, vol. 55, no. 4, pp. 1390–1404, 2007.

[10] M. Gurban and J.-Ph. Thiran, "Multimodal speaker localization in a probabilistic framework," in *Proc. EUSIPCO*, 2006.

[11] M. R. Siracusa and J. W. Fisher, "Dynamic dependency tests: Analysis and applications to multi-modal data association," in *Proc. AIStats*, 2007.

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

[13] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval, "Learning Multi-Modal Dictionaries," *IEEE Trans. Image Proc.*, vol. 16, no. 9, pp. 2272–2283, 2007.

[14] B. A. Olshausen, "Sparse codes and spikes," in *Probabilistic Models of the Brain: Perception and Neural Function*, R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, Eds. MIT Press, 2002.

[15] S. Mallat and Z. Zhang, "Matching Pursuits with time-frequency dictionaries," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[16] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[17] M. Rehn and F. T. Sommer, "A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields," *Journal of Computational Neuroscience*, vol. 22, no. 2, pp. 135–146, 2007.

[18] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.