CCIT Report #524   March 2005

Pixels that Sound

Einat Kidron, Yoav Y. Schechner
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

Michael Elad
Department of Computer Science
Technion - Israel Institute of Technology
Haifa 32000, Israel

Abstract

People and animals fuse auditory and visual information to obtain robust perception. A particular benefit of such cross-modal analysis is the ability to localize visual events associated with sound sources. We aim to achieve this using computer-vision aided by a single microphone. Past efforts encountered problems stemming from the huge gap between the dimensions involved and the available data. This has led to solutions suffering from low spatio-temporal resolutions. We present a rigorous analysis of the fundamental problems associated with this task. Then, we present a stable and robust algorithm which overcomes past deficiencies. It grasps dynamic audio-visual events with high spatial resolution, and derives a unique solution. The algorithm effectively detects the pixels that are associated with the sound, while filtering out other dynamic pixels. It is based on canonical correlation analysis (CCA), where we remove inherent ill-posedness by exploiting the typical spatial sparsity of audio-visual events. The algorithm is simple and efficient thanks to its reliance on linear programming and is *free of user-defined parameters*. To quantitatively assess the performance, we devise a localization criterion. The algorithm capabilities were demonstrated in experiments, where the algorithm overcame substantial visual distractions and audio noise.

Keywords: CCA (Canonical Correlation Analysis), Sparse Representation, Localization, Multi-Modal Processing.

1 Introduction

There is a growing interest in multi-sensor processing. A particularly interesting sensor combination involves visual motion in conjunction with associated *audio*. Activity in computer vision involving audio analysis has various research aspects [5, 35], including lip reading [3, 34], analysis and synthesis of music from motion [31], audio filtering based on motion [8], and source separation based on vision [20, 24, 28, 32, 36]. We note that physiological evidence and analysis of biological systems show that fusion of audio-visual information is used to enhance perception [14, 18, 23].

In this work, we focus on accurately *pinpointing the visual localization* of image pixels that are associated with audio sources. These pixels should be distinguished from other moving objects. We do *not* limit the problem to talking faces [3, 5, 28, 32] or other specific classes of sources [31], but seek a general and effective algorithm to achieve this goal. Some existing methods use several microphones (emulating binaural hearing), where stereo triangulation indicates the spatial location of the sources [2, 25, 33, 37]. In contrast, we seek a very sharp spatial localization of the sound source, using a single microphone (monaural hearing) and a video stream. Moreover, we wish the localization method to perform well, even if interfering sounds exist, unrelated to the desired object.

As indicated in Fig. 1, audio and visual data are inherently difficult to compare because of the huge dimensionality gap between these modalities. To overcome this, a common practice is to project each modality into a one-dimensional (1D) subspace [28, 34, 36]. Thus, two 1D variables represent the audio and the visual signals. Localization algorithms typically seek 1D representations that best correlate [24, 28, 34]. However, as shown in this paper, this approach has a fundamental flaw. The projection of the visual data is controlled by many degrees of freedom. Hence, a substantial amount of data is necessary to reliably learn the cross-relationships. For this reason, some methods use a very aggressive pre-pruning of visual areas or features [3, 5, 34] to reduce the number of unknowns. Others consider acquisition of very long sequences to ensure sufficient data quantities [8, 28]. Those approaches result in a severe loss of either spatial or temporal resolutions, or both.

Audio-visual association can also be performed by optimizing the mutual information (MI) of modal representations [19], while trading off ℓ^2-based regularization terms. This approach requires multiple tune-up parameters, and suffers from the complexity of estimating MI using

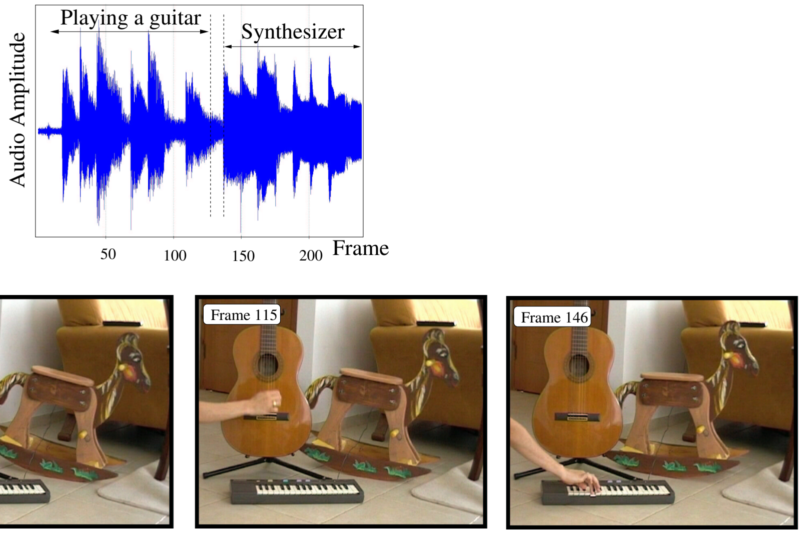


Figure 1: The audio data [Top] is sequential, requiring $\mathcal{O}(10^4)$ samples/sec. Corresponding video frames [Bottom] are highly parallel (multi-pixel), requiring $\mathcal{O}(10^7)$ samples/sec. Pinpointing the sound source in the images by correlation requires dimensionality reduction of the visual signal. This reduction involves of too many degrees of freedom.

Parzen windows. While MI better indicates cross-modal statistical dependency, there is no guarantee for a unique solution, due to the non-convexity of MI.

In this paper we describe an algorithm that overcomes all those difficulties. It results in high spatio-temporal localization, which is unique and stable. We exploit the fact that typically visual cues that correspond to audio sources are *spatially localized*, and thus *sparsity* of the solution is an appropriate prior. This makes the problem well-posed, even-though the analysis is based on very short time intervals. The resulting sparsity does not compromise at all the *full* correlation of audio-visual signals. The algorithm is essentially *free of user-defined parameters*. The numerical scheme is efficient, based on linear programming. To analyze performance, we propose a quantitative criterion for the visual localization of sounds. We then demonstrate the merits of the algorithm in experiments using real data.

This paper is organized as follows. Sec. 2 describes the tool of canonical correlation analysis (CCA), which is a natural choice in multi-modal processing. In Sec. 3 we show an alternative yet equivalent formulation to CCA. This formulation serves our method, as it highlights the ill-posedness of the problem in a clear form, exhibiting a need for regularization. Sec. 4 is dedicated to the exploration of several standard regularizations. We argue that while such regularizations lead to unique solutions, the results are far from satisfactory in general. In Sec. 5

we present the main contribution of this paper, describing how sparsity of the solution can lead to more effective localization and fully correlated results. In Sec. 6 we extend the analysis to cases where no solution is fully correlated. Sec. 7 unveils the fundamental *chorus ambiguity*. Sec. 8 presents a measure of locality that enables us to evaluate localization performance when comparing different algorithms. Sec. 9 presents some experimental demonstrations based on real data. We conclude in Sec. 10 with a brief discussion.

2 Canonical Correlation: Limitations

An important tool for understanding the relationship between sound and video is *canonical correlation analysis* (CCA). In this section we describe CCA, and the reason for its importance and popularity in multi-modal analysis. We then expose a fundamental limitation of it in the context of our problem.

CCA deals with correlation between two random vectors. The vectors can be of different nature, such as audio and visual signals. Let $\mathbf{v}$ represent an instantaneous visual signal corresponding to a single frame, e.g., by pixel values or by wavelet coefficients. Let $\mathbf{a}$ represent a corresponding audio signal, e.g., by the intensity of different audio bands (temporal slices of the periodogram) covering a temporal interval that matches a video frame. Both signals are considered as random vectors, due to their temporal variations.[1] Each of these vectors is projected onto a one dimensional subspace $\mathbf{w}_v$ and $\mathbf{w}_a$, respectively. The result of these projections is a pair of random variables, $\mathbf{v}^T\mathbf{w}_v$ and $\mathbf{a}^T\mathbf{w}_a$, where T denotes transposition. The normalized correlation coefficient of these two variables defines the canonical correlation [26, 27] between $\mathbf{v}$ and $\mathbf{a}$,

$$\rho \equiv \frac{E\left[\mathbf{w}_v^T\mathbf{v}\mathbf{a}^T\mathbf{w}_a\right]}{\sqrt{E\left[\mathbf{w}_v^T\mathbf{v}\mathbf{v}^T\mathbf{w}_v\right]E\left[\mathbf{w}_a^T\mathbf{a}\mathbf{a}^T\mathbf{w}_a\right]}} = \frac{\mathbf{w}_v^T\mathbf{C}_{va}\mathbf{w}_a}{\sqrt{\mathbf{w}_v^T\mathbf{C}_{vv}\mathbf{w}_v\mathbf{w}_a^T\mathbf{C}_{aa}\mathbf{w}_a}}\ , \tag{1}$$

where E denotes expectation. Here $\mathbf{C}_{vv}$ and $\mathbf{C}_{aa}$ are the covariance matrices of $\mathbf{v}$ and $\mathbf{a}$, respectively, while $\mathbf{C}_{va}$ is the cross-covariance matrix of the vectors.

Maximization of the correlation seeks the subspaces $\mathbf{w}_v$ and $\mathbf{w}_a$ that optimize Eq. (1). Note that the solution is scale invariant due to the normalization. This optimization problem has

[1] Each of the vectors $\mathbf{v}$ and $\mathbf{a}$ is assumed to have zero expectation. Numerically, this can be achieved by removal of each vectors' mean.

a closed form solution since it can be posed as an eigenvalues problem [26]:

$$\begin{aligned} \mathbf{C}_{vv}^{-1}\mathbf{C}_{va}\mathbf{C}_{aa}^{-1}\mathbf{C}_{av}\mathbf{w}_v &= \rho^2\mathbf{w}_v \\ \mathbf{C}_{aa}^{-1}\mathbf{C}_{av}\mathbf{C}_{vv}^{-1}\mathbf{C}_{va}\mathbf{w}_a &= \rho^2\mathbf{w}_a \ . \end{aligned} \tag{2}$$

Maximizing the absolute correlation $|\rho|$ is equivalent to finding the largest eigenvalue and its corresponding eigenvectors. Inspecting the optimal $\mathbf{w}_v$, the components which have the largest magnitude indicate the visual components that best correlate with the projection of $\mathbf{a}$, and vice-versa. We should note that a correlation ρ and its opposite $-\rho$ correspond to the same eigenvalue and eigenvectors, and thus to the same solution. Hence, the range $0 \leq \rho \leq 1$ is equivalent to $-1 \leq \rho \leq 0$.

At first sight, CCA may appear as a good tool for correlating audio and visual signals. The projection of feature vectors can bridge the huge dimensionality gap between sound and pictures. Moreover, CCA amounts to an eigensystem solution. Owing to these attractive characteristics, methods based of projections of feature vectors have been the core of several audio-visual algorithms [20, 24, 28, 34]. However, CCA and its related methods [28] have a serious shortcoming. The fundamental problem is the *scarcity of data* available in short time intervals, which is often *insufficient* for reliably estimating the statistics of the signals. To see this, note that $\mathbf{C}_{vv}$, $\mathbf{C}_{aa}$ and $\mathbf{C}_{va}$ should be learned from the data. In practical, $\mathbf{C}_{vv}$ is estimated as the empirical matrix

$$\widehat{\mathbf{C}}_{vv} = (1/N_F)\sum_{t=1}^{N_F}\mathbf{v}(t)\mathbf{v}^T(t) \ , \tag{3}$$

where $\mathbf{v}(t)$ is the vector of visual features at time (frame) t and N_F is the total number of frames used for the estimation. For a reliable representation of typical images, at least thousands of visual features are needed. To reliably learn the statistics of $\mathbf{v}$ and invert $\widehat{\mathbf{C}}_{vv}$ (making Eq. (3) full rank), we must use at least that number of frames. This imposes minutes-long sequences, while assuming stationarity.

To grasp dynamic events, short time intervals should be used (small N_F), but then we run into a problem of data shortage. The matrix $\widehat{\mathbf{C}}_{vv}$ becomes highly rank deficient, hence Eq. (2) cannot be solved, making CCA ill-posed. Technically, the rank deficiency of $\widehat{\mathbf{C}}_{vv}$ can be bypassed by regularization, e.g., by weighted averaging of $\widehat{\mathbf{C}}_{vv}$ with an identity matrix [1, 6, 30]. Such operations do not overcome the fundamental problem of unreliable statistics. They yield an arbitrary solution, which somewhat compromises the correlation ρ. As we show in Sec. 4, such regularization suffers from serious shortcomings, in the context of our problem.

The gap between the amount of data and degrees of freedom is not limited to CCA. It affects methods based on MI as well [19]. Hence, very small images $\mathcal{O}(50 \times 50)$ have been commonly used [3, 28, 32, 34], out of which only a few dozen features were selected by aggressive pruning or face detection algorithms (the latter limiting audio analysis to speech). In contrast, we seek localization of general unknown audio-visual sources, while handling intricate details and motion.

3 An Equivalent Formulation

Before approaching our suggested solution, let us first present an equivalent formulation to CCA that provides insight. The motivation for this alternative formulation will become evident as we turn to the end of Secs. 4 and 5, to handle the ill-posedness of CCA. Let N_v be the number of visual features. Define the matrix $\mathbf{V} \in \mathcal{R}^{N_F \times N_v}$, where row t contains the vector $\mathbf{v}^T(t)$. Similarly, define $\mathbf{A} \in \mathcal{R}^{N_F \times N_a}$, where row t contains the coefficients of the audio signal $\mathbf{a}^T(t)$, and N_a is the number of audio features. Defining the empirical covariances matrices $\hat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, $\hat{\mathbf{C}}_{aa} = \mathbf{A}^T\mathbf{A}$ and $\hat{\mathbf{C}}_{va} = \hat{\mathbf{C}}_{av}^T = \mathbf{V}^T\mathbf{A}$, the empirical canonical correlation[2] (Eq. 1) becomes

$$\hat{\rho} = \frac{\mathbf{w}_v^T(\mathbf{V}^T\mathbf{A})\mathbf{w}_a}{\sqrt{\mathbf{w}_v^T(\mathbf{V}^T\mathbf{V})\mathbf{w}_v\mathbf{w}_a^T(\mathbf{A}^T\mathbf{A})\mathbf{w}_a}} \quad . \tag{4}$$

CCA seeks to maximize $|\hat{\rho}|$. As we show next, maximizing $|\hat{\rho}|$ is equivalent to minimizing the penalty function

$$G(\mathbf{w}_v, \mathbf{w}_a) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2} \tag{5}$$

with respect to $\mathbf{w}_v$ and $\mathbf{w}_a$, where $\|\cdot\|_2$ is the ℓ^2-norm.[3] To prove this, we null the derivatives of $G(\mathbf{w}_v, \mathbf{w}_a)$:

$$\frac{\partial}{\partial \mathbf{w}_v} G(\mathbf{w}_v, \mathbf{w}_a) = 0 \quad , \quad \frac{\partial}{\partial \mathbf{w}_a} G(\mathbf{w}_v, \mathbf{w}_a) = 0 \ \ . \tag{6}$$

Eq. (6) yields,

$$2\mathbf{V}^T(\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a)(\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2) - 2\mathbf{V}^T\mathbf{V}\mathbf{w}_v\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|^2 = 0 \tag{7}$$

$$-2\mathbf{A}^T(\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a)(\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2) - 2\mathbf{A}^T\mathbf{A}\mathbf{w}_a\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|^2 = 0 \tag{8}$$

[2] Strictly speaking, the definition for $\widehat{\mathbf{C}}_{vv}$, $\widehat{\mathbf{C}}_{aa}$ and $\widehat{\mathbf{C}}_{va}$ should be normalized by N_F. However, this constant is factored out in Eq.(4), and is thus discarded throughout the paper.

[3] Note that $0 \le G(\mathbf{w}_v, \mathbf{w}_a) \le 2$. The proof is given in App. A.

leading to

$$\mathbf{V}^T\mathbf{V}\mathbf{w}_v - \mathbf{V}^T\mathbf{A}\mathbf{w}_a = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|^2}{\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2}\mathbf{V}^T\mathbf{V}\mathbf{w}_v \tag{9}$$

$$-\mathbf{A}^T\mathbf{V}\mathbf{w}_v + \mathbf{A}^T\mathbf{A}\mathbf{w}_a = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|^2}{\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2}\mathbf{A}^T\mathbf{A}\mathbf{w}_a \quad . \tag{10}$$

Using the empirical covariances matrices and G,

$$\hat{\mathbf{C}}_{vv}\mathbf{w}_v - \hat{\mathbf{C}}_{va}\mathbf{w}_a = G\hat{\mathbf{C}}_{vv}\mathbf{w}_v \tag{11}$$

implying

$$\mathbf{w}_v = \frac{1}{1-G}\hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\mathbf{w}_a \quad . \tag{12}$$

Analogously,

$$-\hat{\mathbf{C}}_{av}\mathbf{w}_v + \hat{\mathbf{C}}_{aa}\mathbf{w}_a = G\hat{\mathbf{C}}_{aa}\mathbf{w}_a \tag{13}$$

implying

$$\mathbf{w}_a = \frac{1}{1-G}\hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\mathbf{w}_v \quad . \tag{14}$$

Eqs. (12) and (14) yield the following set of equations,

$$\begin{aligned} \hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\mathbf{w}_v &= (1-G)^2\mathbf{w}_v \\ \hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\mathbf{w}_a &= (1-G)^2\mathbf{w}_a \quad . \end{aligned} \tag{15}$$

Note that Eq. (15) is equivalent to the CCA set of equations given in Eq. (2), with $\rho^2 = (1-G)^2$. Thus, an extremum of G is equivalent to an extremum of ρ. Moreover, finding the maximum correlation (e.g., the largest eigenvalue ρ^2) is equivalent to finding the minimal[4] G. It can be shown that the range of $0 \leq G \leq 1$ is equivalent to the range $0 \leq \rho \leq 1$, while $1 \leq G \leq 2$ is equivalent to $-1 \leq \rho \leq 0$. As we discussed in Sec. 2, these two ranges are equivalent. Thus, the solution that maximizes G in the domain $1 \leq G \leq 2$ is equivalent to the one minimizing G when $0 \leq G \leq 1$. Hence, in this paper we can focus on minimizing G towards zero.

One way to obtain intuition into this equivalence is by noting that minimizing Eq. (5) implies that the projected video $\mathbf{V}\mathbf{w}_v$ should be as close as possible to the projected audio $\mathbf{A}\mathbf{w}_a$ in the ℓ^2 sense. It means that we are looking for linear dependency between $\mathbf{V}\mathbf{w}_v$ and $\mathbf{A}\mathbf{w}_a$, which is what we indeed expect in high correlation. The denominator in Eq. (5) serves

[4] Note that G is real and non-negative, by definition.

to avoid trivial solutions, and to properly use the energies of the two projections. This is in analogy to the correlation normalization in Eq. (1).

Before proceeding, we would like to note that there is an alternative formulation to CCA, called *principal angles* [7, 21, 38]. For the principal angles approach an alternative formulation was proposed in [38], which is the constraint optimization

$$\max_{\mathbf{w}_a,\mathbf{w}_v} \{\mathbf{w}_v^T \mathbf{V}^T \mathbf{A} \mathbf{w}_a\} \quad \text{subject to} \;\; \|\mathbf{V}\mathbf{w}_v\|^2 = 1, \; \|\mathbf{A}\mathbf{w}_a\|^2 = 1 \tag{16}$$

The Ill-Posedness of CCA

CCA has limitations, when working with a rank deficient matrix $\mathbf{V}$. This occurs in audio-visual correlation, when short time intervals are used. The number of representation features (at each frame) is expected to be much larger than the number of frames in the time interval ($N_F \ll N_v$). Let us analyze this ill-posedness using a formulation based on the minimization of G. First, we focus on the cases where $N_a = 1$, i.e., the audio is characterized by a single feature. Multiple audio bands cases, i.e., $N_a > 1$, are treated in Sec. 5.2.

When $N_a = 1$ we may set $\mathbf{w}_a = 1$ (where $\mathbf{w}_a$ is a scalar), since the penalty function in (5) is scale invariant (multiplying $\mathbf{w}_v$ and $\mathbf{w}_a$ by the same constant does not change the function's value). Thus, Eq. (5) becomes,

$$G(\mathbf{w}_v) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} \quad . \tag{17}$$

We assume that $\mathbf{A} \neq 0$ (audio modulation exists). Hence the denominator of Eq. (17) cannot be nulled. Thus, we can concentrate on the numerator and minimize

$$g(\mathbf{w}_v) = \|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 \; . \tag{18}$$

As shown later, the denominator is usually unimportant.

Suppose for a moment that a vector $\mathbf{w}_v$ exists such that $g(\mathbf{w}_v) = 0$. This vector yields $G(\mathbf{w}_v, \mathbf{w}_a) = 0$ since the denominator of Eq. (5) is necessarily non-zero.[5] Hence, this solution yields complete coherence, $|\hat{\rho}| = 1$, as desired. Requiring $g(\mathbf{w}_v) = 0$ implies

$$\mathbf{V}\mathbf{w}_v = \mathbf{A} \; . \tag{19}$$

Since $N_a = 1$, $\mathbf{A}$ is a *column* vector of length N_F. Eq. (19) is illustrated in Fig. 2. As discussed in Sec. 2, $N_v \gg N_F$, where N_v is the length of $\mathbf{w}_v$. Therefore, in the set of linear

[5] This is true since $\mathbf{A}$ is a non-zero vector.

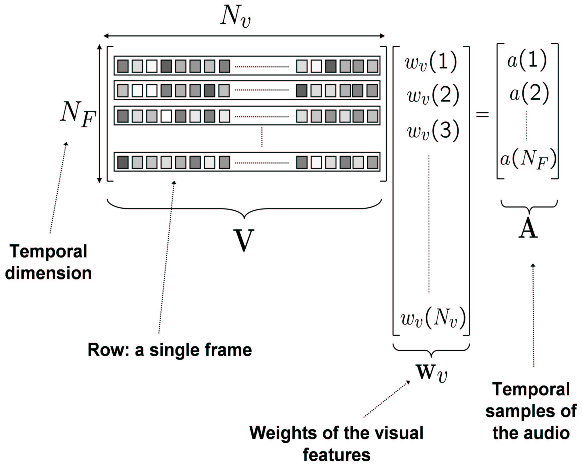


Figure 2: Illustration of Eq. (19). Each row in $\mathbf{V}$ represents visual features of a single frame. There are N_F temporal slices (frames) that are analyzed simultaneously. Since $N_v \gg N_F$ the number of rows (equations) is much smaller than the number of unknowns, yielding an underdetermined linear set of equations. Here $\mathbf{A}$ is the vector of temporal audio samples.

equations (19), the number of equations is much smaller than the number of unknowns, yielding an underdetermined linear set of equations. The number of possible solutions is infinite. To conclude: due to the scarce data, there are infinite number of combinations of visual features that appear to completely correlate with the audio!

How probable is the scenario of having $g(\mathbf{w}_v) = 0$? For $N_v \gg N_F$, most chances are that $\text{rank}(\mathbf{V}) = N_F$, guaranteeing that $\mathbf{A}$ is in the span of the $\mathbf{V}$ column space. Thus, it is highly probable that $g(\mathbf{w}_v)$ has a zero. In fact, noise in the visual data guarantees this outcome, as it causes the rank to become full. However, visual noise implies strong correlation of “junk” features to the audio. A similar situation occurs in the case where $g(\mathbf{w}_v)$ cannot be zero. We prove this is Sec. 6.

4 Attempting Standard Regularization

There are several ways to overcome the ill-posedness of CCA. Since we have infinite number of solutions to CCA of scarce data, some kind of regularization should be imposed. Regularization has the role of choosing the best vector among the infinite space of potential solutions, according to some criterion. Next, several types of standard regularization techniques are discussed, as well as their weaknesses. Our alternative approach, which is stronger in the context of our scenario, is introduced in Sec. 5.

4.1 Regularization Using an ℓ^2 Term

A common regularization of underdetermined problems is to prefer the minimal energy solution [17, 21]. In our case this would be

$$\min \|\mathbf{w}_v\|_2 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ . \tag{20}$$

The constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ nulls the numerator of Eq. (17), thus leading to a solution having full correlation. The ℓ^2 term in Eq. (20) is the imposed regularization. We may find the $\mathbf{w}_v$ that solves Eq. (20) using several techniques, as pseudo inverse, SVD, and QR factorization [21]. Here we shall simply show the pseudo-inverse solution. The rest minimize $\|\mathbf{w}_v\|_2$ as well, and thus suffer from the same major weakness, as we detail in the following.

Pseudo-Inverse

To solve Eq. (20), define the Lagrangian

$$L(\mathbf{w}_v, \lambda) = \frac{1}{2}\|\mathbf{w}_v\|_2^2 + \lambda^T(\mathbf{V}\mathbf{w}_v - \mathbf{A}) \ . \tag{21}$$

Minimization of Eq. (21) implies,

$$0 = \frac{\partial L}{\partial \mathbf{w}_v} = \mathbf{w}_v + \mathbf{V}^T\lambda \quad , \quad 0 = \frac{\partial L}{\partial \lambda} = \mathbf{V}\mathbf{w}_v - \mathbf{A} \ . \tag{22}$$

Combining the above equations yields the $\hat{\mathbf{w}}_v$ having a minimum ℓ^2-norm, i.e., the least square solution

$$\hat{\mathbf{w}}_v^{\mathrm{LS}} = \mathbf{V}^+\mathbf{A} \ , \tag{23}$$

where

$$\mathbf{V}^+ = \mathbf{V}^T(\mathbf{V}\mathbf{V}^T)^{-1} \tag{24}$$

is the pseudo-inverse of matrix $\mathbf{V}$.

In the context of the audio-visual problem, this results in visual *poor localization.* The reason is that the ℓ^2 criterion seeks to spread the energy of $\mathbf{w}_v$ over many small-valued visual components, rather than concentrating energy on a few dominant ones. To obtain some intuition, this phenomenon is depicted in Fig. 3 for $N_v = 2$ and $N_F = 1$. In this figure, a straight line describes the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$. The minimum of the ℓ^2-norm is obtained in point B, which has substantial energy in all components. This nature is contrary to common audio-visual scenarios, where visual events associated with sound are often very *local.* They typically reside in small areas (few components) of the frame. Indeed, the inadequacy of this criterion is demonstrated in the experiments detailed in Sec. 9.

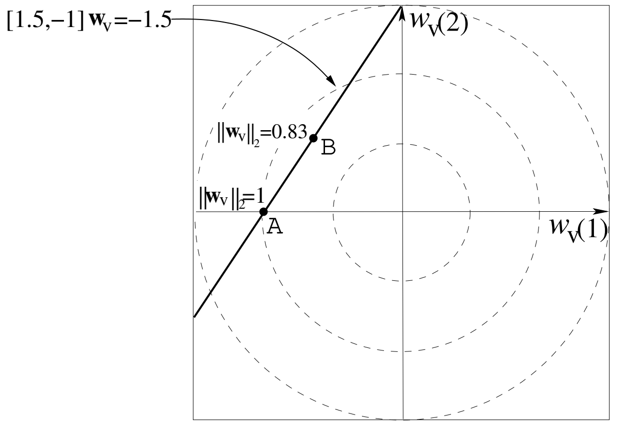


Figure 3: A 2D example of optimization under ℓ^2-norm. The dashed contours represent iso-norm levels. On the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ (solid line), point B minimizes $\|\mathbf{w}_v\|_2$, but it has a substantial energy in all components.

4.2 Regularization Using the Identity Matrix

As described in Eq. (2), CCA may involve inversion of matrix[6] $\hat{\mathbf{C}}_{vv}$. While the audio stream per frame may often be represented by a small number of features, the video representations needs a large number of features, i.e., $N_v \gg N_F$. The covariance matrix $\hat{\mathbf{C}}_{vv}$, defined by $\hat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, is a huge low rank (and thus singular) matrix. Solving CCA using (2) is impossible due to the inability to invert $\hat{\mathbf{C}}_{vv}$. One way to overcome this problem is to regularize matrix $\mathbf{V}$, and hence $\hat{\mathbf{C}}_{vv}$ as follows.

We may use the identity matrix and define an invertible version as

$$\widetilde{\mathbf{C}}_{vv} = \hat{\mathbf{C}}_{vv} + \epsilon\ \mathbf{I}\ , \tag{25}$$

where $\mathbf{I}$ is the identity matrix and ϵ is an arbitrary small number [1, 6, 30]. Such regularization brings the covariance matrix to be a huge full rank matrix, and thus invertible.[7] Inserting the regularized version of the covariance matrix $\widetilde{\mathbf{C}}_{vv}$ into the correlation expression given in (1) we see that it enlarges the denominator of (1). Hence, this regularization reduces the correlation value (destroying the complete coherence). Thus, we expect that such an operation will lead to results that have lower correlation.

Another way to judge this regularization is to examine the penalty function. Recalling that

[6] Inversion of $\mathbf{C}_{aa}$ is not a problem in our audio-visual localization problem. The reason is that the number of audio features is comparable to the number of temporal samples, i.e., $N_a \sim N_F$.

[7] The covariance matrix in its new formulation (25) can be inverted efficiently using the Sherman-Morrison theorem. However, we still have to face a huge eigenproblem.

$\hat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, we can obtain Eq. (25) by defining a matrix $\widetilde{\mathbf{V}}$ of size $N_v \times N_v$, having the form

$$\widetilde{\mathbf{V}} = \left[\begin{array}{c} \mathbf{V}_{N_F \times N_v} \\ \sqrt{\epsilon}\ \mathbf{I}_{(N_v - N_F)\times N_v} \end{array} \right] , \tag{26}$$

and then $\widetilde{\mathbf{C}}_{vv} = \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}}$. Suppose we use $\widetilde{\mathbf{C}}_{vv}$ and $\widetilde{\mathbf{V}}$ defined by Eqs. (25,26) instead of the original matrices $\mathbf{V}$ and $\hat{\mathbf{C}}_{vv}$. Inserting matrix $\widetilde{\mathbf{V}}$ into Eq. (17) yields

$$\widetilde{G}(\mathbf{w}_v) = \frac{\|\widetilde{\mathbf{V}}\mathbf{w}_v - \widetilde{\mathbf{A}}\|_2^2}{\|\widetilde{\mathbf{V}}\mathbf{w}_v\|_2^2 + \|\widetilde{\mathbf{A}}\|_2^2} = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2} \tag{27}$$

where

$$\widetilde{\mathbf{A}} = \left[\begin{array}{c} \mathbf{A} \\ \mathbf{0}_{N_v - N_F} \end{array} \right] . \tag{28}$$

Since ϵ is chosen as a small number, while the audio data given in $\mathbf{A}$ is assumed to contain significant energy, then $\epsilon\|\mathbf{w}_v\|_2^2 \ll \|\mathbf{A}\|_2^2$. Thus, the term $\epsilon\|\mathbf{w}_v\|_2^2$ can be neglected in the denominator. However, it can not be neglected in the numerator since $\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2$ is a small number (as we are close to full correlation). The penalty function becomes

$$\widetilde{G}(\mathbf{w}_v) \approx \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} = G(\mathbf{w}_v) + \tilde{\epsilon}\|\mathbf{w}_v\|_2^2 \ , \tag{29}$$

where $G(\mathbf{w}_v)$ is the penalty function in the non-regularized case and $\tilde{\epsilon}$ is an arbitrary small number

$$\tilde{\epsilon} = \frac{\epsilon}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} \ . \tag{30}$$

Recall that maximizing the correlation is equivalent to minimizing G. Hence minimization of (29) tends to minimize $\|\mathbf{w}_v\|_2^2$ and not only G. Thus, we can see that such a regularization has resemblance to the ℓ^2 regularization given in (20), hence suffering from a similar weakness. This brings us to the next section where we propose a different way to rectify the ill-posedness of our problem.

5 Sparsity as A Key

> *"Out of clutter, find simplicity.*
> *From discord, find harmony."* - Albert Einstein

As we showed in the previous section, solving the audio-video correlation problem using the traditional ℓ^2-norm solution, leads to poorly localized results. We now describe our approach, which leads to a unique solution based on a spatial sparsity criterion. We progress by first looking at cases where $N_a = 1$, i.e., the audio is characterized by a single feature. In Sec. 5.2 we extend the analysis to multiple audio bands.

5.1 A Single Audio Band

With a single audio band our goal is to minimize Eq. (17). We start by first discussing the case where the minimum of the this function is zero. The case of a non-zero cost function value is discussed in Sec. 6.

As discussed in Sec. 4.1, Eq. (17) suffers from poor localization. To overcome this problem, we express locality as a requirement that the sought solution should be *sparse*.[8] Our goal is that the optimal solution will have a minimal number of components. Thus, out of the entire space of possible correlated projections, we aim to solve:

$$\min \|\mathbf{w}_v\|_0 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ \ , \tag{31}$$

where $\|\cdot\|_0$ is the ℓ^0-norm of a vector space (the number of non-zero vector coefficients). In the simple example depicted on the left of Fig. 4, the optimal solution according to this criterion (point A) has a single component. Unfortunately, this criterion is not convex, and the complexity of its optimization is exponential [10, 16, 22] in N_v.

We bypass this difficulty by convexizing the problem and solving

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ \ , \tag{32}$$

where ℓ^1 is used instead of ℓ^0. In the right part of Fig. 4, the solution optimizing this criterion has a single component (point A), just as under the ℓ^0 criterion. All other points in the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ have a larger ℓ^1-norm. Thus, it appears that there is some equivalence between ℓ^0 and ℓ^1 since both lead to the same optimal vector. Moreover, this figure illustrates the convexity of the ℓ^1 criterion.

In general, the equivalence of the ℓ^0 and ℓ^1 problems (31,32) has been studied in depth during the last couple of years from a pure mathematical perspective. Preliminary contributions in this direction considered deterministic sufficient conditions for this equivalence [10, 15, 16, 22]. More recently, a probabilistic approach has been introduced, showing that equivalence holds true far beyond the limits determined by these sufficient conditions [9]. Further details about

[8] Sparsity is enhanced using a wavelet representation of temporal-difference images.

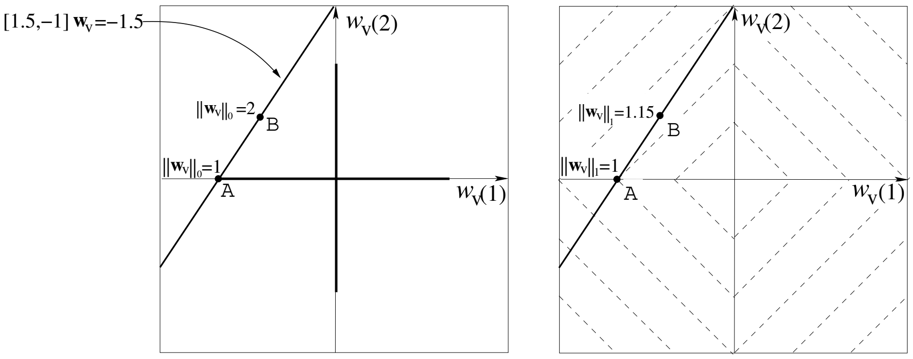


Figure 4: A 2D example of optimization under [Left] ℓ^0-norm [Right] ℓ^1-norm. The dashed contours represent iso-norm levels. On the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ (solid line), point A is the sparsest (minimum $\|\mathbf{w}_v\|_0$), and also minimize $\|\mathbf{w}_v\|_1$. The ℓ^1 criterion is convex, in contrast to ℓ^0.

the equivalence and its conditions are given in App. B. Owing to this theoretical progress, formulating sparsity using the ℓ^1-norm is reliable.

The newly defined formulation (32) can be posed as a *linear programming* problem, and thus can be solved *efficiently* even for $N_v \gg 1$. This formulation influences the solution energy to concentrate on few visual features which strongly correlate with the audio. It penalizes for dispersed components, such as the random "junk" features described at the end of Sec. 3, e.g., image noise. Moreover, the solution is *unique*, because of the convexity of the ℓ^1-norm, except for special cases discussed in Sec. 7.

5.2 Multiple Audio Bands

We now generalize the single-band analysis of Sec. 5.1 to audio signals that are divided into multiple bands. We postpone to Sec. 6 the analysis of scenarios in which the optimal value of the cost function G is non-zero. Here, we analyze cases where the cost function has zeros. This allows us to concentrate on the numerator of Eq. (5). The numerator is zero if and only if

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a \quad . \tag{33}$$

As before, if $\text{rank}(\mathbf{V}) = N_F$, a zero solution of G is guaranteed. As we have claimed in Sec. 5.1, this is a highly probable event, especially for noisy visual data. In the unlikely event that no intersection exists between the subspace spanned by the columns of $\mathbf{V}$ and the subspace spanned by $\mathbf{A}$, the cost function G cannot be nulled (See Sec. 6).

Similarly to Sec. 5.1, Eq. (33) is prone to a scale ambiguity. To overcome this problem and avoid the trivial solution $\mathbf{w}_a = 0$, we use normalization. A way to achieve this is to limit the search to the audio ℓ^1-ball, $\|\mathbf{w}_a\|_1 = 1$. The set $\|\mathbf{w}_a\| = 1$ is not convex. To keep enjoying the benefits of a convex problem formulation, the following process is performed. We break the problem into 2^{N_a} separate ones, where each handles a single face of the audio ℓ^1-ball and is thus convex. As depicted in Fig. 5, the optimization over each face $q \in [1, 2^{N_a}]$ can be posed as

$$s_q = \min \|\mathbf{w}_v\|_1 \quad \text{subject to} \\ \Big\{\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a \ , \ \mathbf{h}_q^T\mathbf{w}_a = 1 \ , \ \mathbf{H}_q\mathbf{w}_a \geq 0\Big\} \tag{34}$$

where $\mathbf{h}_q$ is a vector and $\mathbf{H}_q$ is a diagonal matrix whose diagonal is $\mathbf{h}_q$. The vector set $\{\mathbf{h}_q\}_{q=1}^{2^{N_a}}$ comprises the 2^{N_a} different combinations of the N_a-tuples binary sequences with ±1 as their entries. Since all the constraints are linear, Eq. (34) is solved for each q using linear programming.

Recall that for our audio-visual localization method, we should optimize the visual sparsity over the audio ℓ^1-ball. This is done by running Eq. (34) over all[9] values of q, and then selecting the optimal q by

$$\hat{q} = \arg\min s_q \ \ . \tag{35}$$

The unique vectors $\mathbf{w}_v$ and $\mathbf{w}_a$ which we seek are then derived by using this specific $\hat{q}$ in Eq. (34). We stress that our goal is to localize *visual* events (based on audio cues), while processing of audio is of secondary importance here. This distinction enables us to use a coarse representation of the audio. Hence, only a small number of audio bands N_a is required. For this reason, the computations are tolerable despite the $\mathcal{O}(2^{N_a})$ complexity.

6 A Non-Zero Cost Function Value

So far we considered solutions $\mathbf{w}_v$ that null $g(\mathbf{w}_v)$. We stress that this nulling is very likely due to noise in $\mathbf{V}$, as explained in Sec. 3. In this section we refer to the case where no solution is fully correlated, i.e., $g(\mathbf{w}_v) \neq 0$ for all $\mathbf{w}_v$. Indeed, in our experiments we did not encounter such cases. Still, for the sake of completeness we show that this case can be handled well by our approach.

[9] Actually, there is no need to scan all 2^{N_a} values of q. Due to the scale ambiguity mentioned above, $\mathbf{h}_q$ and $-\mathbf{h}_q$ will yield the same results. Hence it is sufficient to scan 2^{N_a-1} nonequivalent values of q.

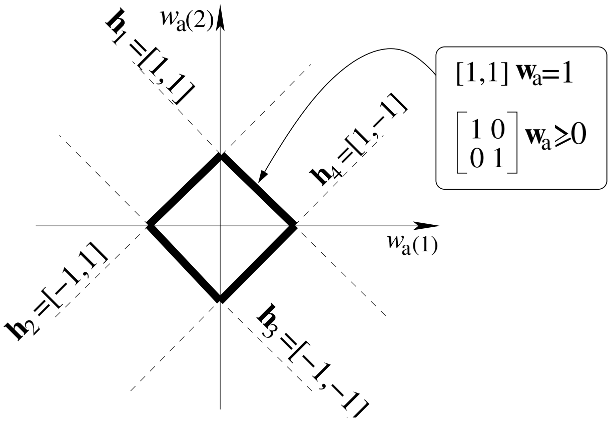


Figure 5: A 2D illustration of the faces of the ℓ^1-ball in the audio space.

6.1 A Single Audio Band

It follows from Sec. 3 that $[\min\, g(\mathbf{w}_v)] \neq 0$ only if $\mathrm{rank}(\mathbf{V}) < N_F$, and if $\mathbf{A}$ is not in the column span of $\mathbf{V}$. In such cases we can decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{A}_{||} + \mathbf{A}_{\perp}$. Here $\mathbf{A}_{||}$ is in the subspace spanned by the columns of $\mathbf{V}$, while $\mathbf{A}_{\perp}$ is orthogonal to $\mathbf{V}$. Thus, the function $g(\mathbf{w}_v)$ becomes

$$g(\mathbf{w}_v) = \|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|^2 = \|\mathbf{V}\mathbf{w}_v - \mathbf{A}_{||}\|^2 + \|\mathbf{A}_{\perp}\|^2, \tag{36}$$

and Eq. (17) becomes

$$G(\mathbf{w}_v) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}_{||}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}_{||}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2}\ . \tag{37}$$

Note that the audio component $\mathbf{A}_{\perp}$ does not correlate with any of the visual features. As such, it can be discarded as irrelevant. The remaining audio signal $\mathbf{A}_{||}$ is a projected version of the original audio for which the solution to $\mathbf{V}\mathbf{w}_v = \mathbf{A}_{||}$ exists. Thus, $\mathbf{A}$ is essentially *projected* to the column space of $\mathbf{V}$, as a “denoising” pre-process. This explanation suggests that we handle the rank-deficient $\mathbf{V}$ matrix case by such a projection, and then proceed as in Eq. (32) when we use $\mathbf{A}_{||}$ instead of $\mathbf{A}$.

As we show now, that line of reasoning is in fact optimal up to a scale. We are interested in characterizing the set of minimizers $\mathbf{w}_v$ of (37). Recall that $\mathbf{A}_{\perp}$ is not spanned by the columns of $\mathbf{V}$, thus no matter what $\mathbf{w}_v$ is, the term $\mathbf{V}\mathbf{w}_v$ is necessarily orthogonal to $\mathbf{A}_{\perp}$. In general, the solutions $\mathbf{w}_v$ satisfies the relation $\mathbf{V}\mathbf{w}_v = \alpha\mathbf{A}_{||} + \mathbf{Z}$, where α is a scalar, and $\mathbf{Z}$ is an arbitrary vector perpendicular to both $\mathbf{A}_{||}$ and $\mathbf{A}_{\perp}$. Thus, $G(\mathbf{w}_v)$ in Eq. (37) becomes the function $G(\alpha, \mathbf{Z})$,

$$G(\alpha, \mathbf{Z}) = \frac{\|\alpha \mathbf{A}_{\|} + \mathbf{Z} - \mathbf{A}_{\|}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2}{\|\alpha \mathbf{A}_{\|} + \mathbf{Z}\|_2^2 + \|\mathbf{A}_{\|}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2} \quad . \tag{38}$$

We need to find α and $\mathbf{Z}$ that minimize this function. We thus derive equations that null the partial derivatives of $G(\alpha, \mathbf{Z})$ with respect to α and $\mathbf{Z}$. Handling $\mathbf{Z}$ first, we rearrange G

$$G(\alpha, \mathbf{Z}) = \frac{(\alpha-1)^2\|\mathbf{A}_{\|}\|_2^2 + \|\mathbf{Z}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2}{(1+\alpha^2)\|\mathbf{A}_{\|}\|_2^2 + \|\mathbf{Z}\|_2^2 + \|\mathbf{A}_{\perp}\|_2^2} = \frac{(\alpha-1)^2 + \frac{\|\mathbf{Z}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2} + \frac{\|\mathbf{A}_{\perp}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2}}{(1+\alpha^2) + \frac{\|\mathbf{Z}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2} + \frac{\|\mathbf{A}_{\perp}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2}} \quad . \tag{39}$$

Here we have exploited the fact that the ℓ^2-norm is separable when dealing with two orthogonal vectors ($\mathbf{A}_{\|}$ and $\mathbf{Z}$ in this case). To simplify this expression, let us define $r \equiv \|\mathbf{A}_{\perp}\|_2^2/\|\mathbf{A}_{\|}\|_2^2$ and $k(\mathbf{Z}) \equiv \|\mathbf{Z}\|_2^2/\|\mathbf{A}_{\|}\|_2^2$,

$$G(\alpha, \mathbf{Z}) = \frac{(\alpha-1)^2 + k(\mathbf{Z}) + r}{(1+\alpha^2) + k(\mathbf{Z}) + r} \quad . \tag{40}$$

Thus,

$$\frac{\partial G(\alpha, \mathbf{Z})}{\partial \mathbf{Z}} \quad = \quad \frac{2\mathbf{Z}}{\|\mathbf{A}_{\|}\|_2^2} \cdot \frac{2\alpha}{[(1+\alpha^2) + k(\mathbf{Z}) + r]^2} \tag{41}$$

where we used the relation $\partial k(\mathbf{Z})/\partial \mathbf{Z} = 2\mathbf{Z}/\|\mathbf{A}_{\|}\|_2^2$. Nulling this derivative leads to $\mathbf{Z} = 0$ as the solution. Handling α leads

$$\frac{\partial G(\alpha)}{\partial \alpha} \quad = \quad \frac{2(\alpha^2 - 1 - k(\mathbf{Z}) - r)}{[(1+\alpha^2) + k(\mathbf{Z}) + r]^2} \quad , \tag{42}$$

hence,

$$\frac{\partial G(\alpha)}{\partial \alpha} = 0 \quad \Rightarrow \quad \alpha_{\text{opt}} = \sqrt{1 + k(\mathbf{Z}) + r} \quad , \tag{43}$$

i.e.,

$$\alpha_{\text{opt}} = \sqrt{1 + \frac{\|\mathbf{Z}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2} + \frac{\|\mathbf{A}_{\perp}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2}} \quad . \tag{44}$$

Since we obtained that $\mathbf{Z} = 0$, then

$$\alpha_{\text{opt}} = \sqrt{1 + \frac{\|\mathbf{A}_{\perp}\|_2^2}{\|\mathbf{A}_{\|}\|_2^2}} \quad . \tag{45}$$

We proved that if $\mathbf{V}\mathbf{w}_v = \alpha \mathbf{A}_{\|} + \mathbf{Z}$, minimization of Eq. (37) as a function of $\mathbf{Z}$ and α yields $\mathbf{Z} = 0$ and $\alpha = \sqrt{1 + \|\mathbf{A}_{\perp}\|_2^2/\|\mathbf{A}_{\|}\|_2^2}$. This result means that the correlated audio-visual features satisfy

$$\mathbf{V}\mathbf{w}_v = \alpha \mathbf{A}_{\|} = \left(\sqrt{1 + \|\mathbf{A}_{\perp}\|_2^2/\|\mathbf{A}_{\|}\|_2^2}\right) \cdot \mathbf{A}_{\|} \quad . \tag{46}$$

Thus, if G cannot be nulled, the set of minimizers $\mathbf{w}_v$ of (37) is given by (46). Since (46) minimizes G, it maximizes the correlation. The scalar α does not influence the localization result, but only the overall scale of $\mathbf{w}_v$. Thus, the results obtained using our algorithm are consistent, up to a scale.

6.2 Multiple Audio Bands

In the multiple audio band problem, the vector $\mathbf{w}_a$ is unknown. However, from the single band analysis discussed in the previous section, we know that whatever the optimal $\mathbf{w}_a$ is, the eventual solution $\mathbf{V}\mathbf{w}_v$ must be parallel to $\mathbf{A}\mathbf{w}_a$. Thus, we should force this parallelism in Eq. (34), and rephrase the problem to the case where the penalty function is non-zero.

Let the space spanned by the columns of $\mathbf{A}$ be $\mathcal{A}$. Decompose this space into two orthogonal subspaces $\mathcal{A}_{\|}$ and $\mathcal{A}_{\perp}$, where $\mathcal{A}_{\|}$ spans the projected audio subspace $\mathbf{A}\mathbf{w}_a$. Define $\mathbf{A}_{\|}$ and $\mathbf{A}_{\perp}$ as matrices whose columns span $\mathcal{A}_{\|}$ and $\mathcal{A}_{\perp}$, respectively. Similarly to Eq. (46), parallelism means that

$$\mathbf{V}\mathbf{w}_v = \beta \mathbf{A}_{\|}\mathbf{w}_v \tag{47}$$

where β is a scalar. Thus, the inner product between $\mathbf{V}\mathbf{w}_v$ and the orthogonal audio space spanned by $\mathbf{A}_{\perp}$ must be zero

$$\mathbf{A}_{\perp} \cdot \mathbf{V}\mathbf{w}_v = 0 \ . \tag{48}$$

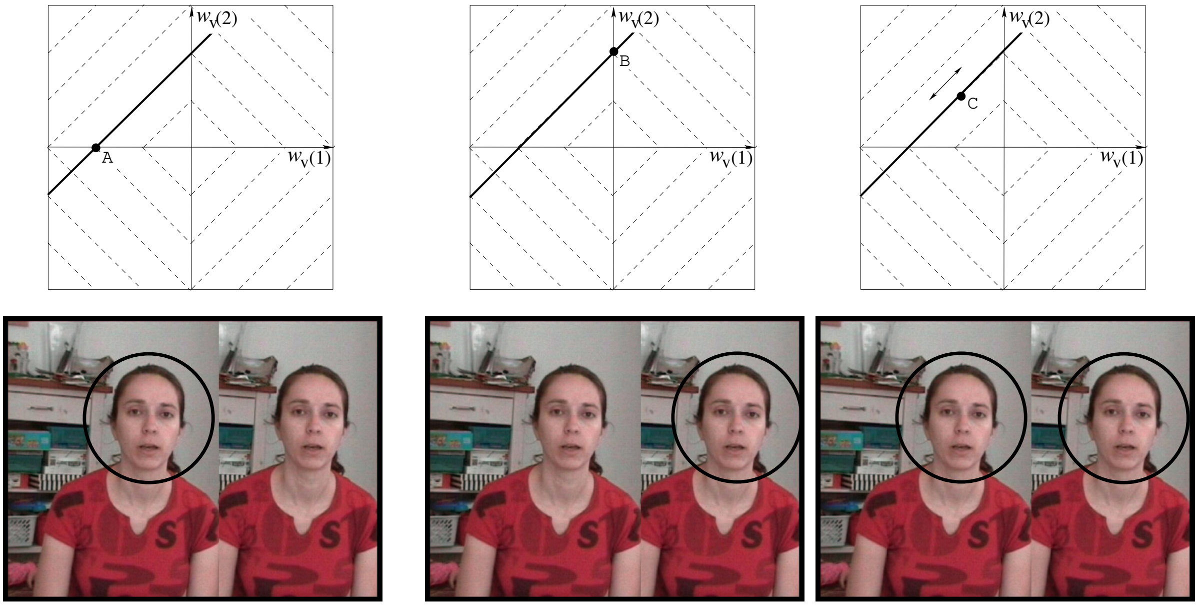
We use (47) and (48) as new constrains. Combining these constraints, Eq. (34) becomes

$$\begin{aligned} s_q = \min \|\mathbf{w}_v\|_1 \quad & \text{subject to} \\ & \left\{ \mathbf{V}\mathbf{w}_v = \mathbf{A}_{\|}\mathbf{w}_a \ , \ \mathbf{A}_{\perp} \cdot \mathbf{V}\mathbf{w}_v = 0 \ , \ \mathbf{h}_q^T\mathbf{w}_a = 1 \ , \ \mathbf{H}_q\mathbf{w}_a \geq 0 \right\} \ . \end{aligned} \tag{49}$$

7 The Chorus Ambiguity

Consider a chorus of identical people singing in synchrony the same song. In this case the audio track corresponds well to several spatially distinct clusters of pixels (faces of the chorus members). Which pixels would you choose as the ones achieving successful localization? We claim that this scenario poses a fundamental ambiguity for any localization algorithm: the result could pinpoint any single person or several of them. In this special scenario all these results are equally acceptable. Thus, we term this phenomenon as the *chorus ambiguity*.

Our algorithm (32,34,35) has this characteristic, just as well. Referring to Fig. 4, this case occurs when the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ aligns with a face of a visual ℓ^1 ball.



(a)

(b)

(c)

Figure 6: The chorus ambiguity under the ℓ^1-norm. The top row is the two pixels scene and the bottom shows a human chorus. (a) Detecting the left person (b) Detecting the right person (c) Detecting both.

Mathematically, this implies that for this special scenario, the problem in (32) does not have a unique solution, but rather a set of them. This case is demonstrated in Fig. 6 for a two-pixels scene (top row) and for a chorus of two people (bottom row). In this illustration, three solution types in the two-pixels scene are represented, denoted by `A`, `B` and `C`. Types `A` and `B` represent exclusive detection of only a single pixel, while type `C` represents all solutions that are a convex superposition of `A` and `B`. Analogously, in the two people chorus, types `A` and `B` represent an exclusive detection of a single person, while type `C` represents detection of the entire chorus (with some weight ratio between members).

We can see that the problem of (31) has only one type of solutions, as demonstrated in Fig. 7 - that of exclusive detection. In the general chorus case, the ℓ^0 criterion can lock into any single person in the chorus, while the ℓ^1 result can spread the detections between several of them. Thus, in this case the equivalence between ℓ^1 and ℓ^0 breaks down. A mathematical insight to this phenomenon can be found in [10, 16, 22]. Still, this effect does not hinder the optimization process (32,34,35): the linear programming converges to one of these solutions, depending on the initialization.

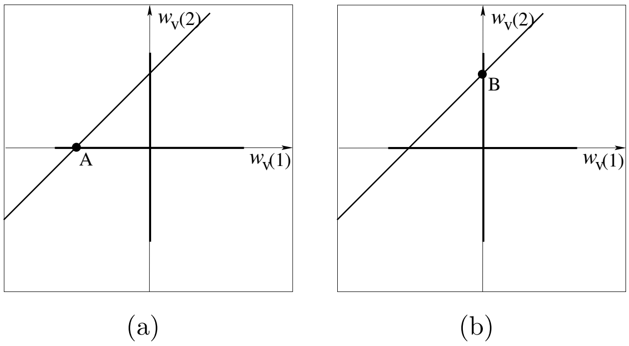


Figure 7: The chorus ambiguity under the ℓ^0-norm. There are only exclusive detections, which correspond to points `A` and `B` in Fig. 7.

8 A Quantitative Localization Criterion

So far we discussed approaches to solve the audio-visual correlation problem, where the signals are represented by some visual and audio features. Once the problem is solved, the results should be transferred from the feature space back into the image domain (pixels domain). In this section we describe this transformation and develop a quantitative criterion to measure performance.

8.1 Back to the Image Domain

The output of the localization algorithm is a weight $w_v(k)$ for each component k of the vector $\mathbf{v}$. The weights are transformed into an image $\mathbf{w}_v^{\text{Image}}$. For example, if wavelets are the domain of $\mathbf{v}$, then an inverse wavelet transform of $\mathbf{w}_v$ brings it to the pixel domain:

$$\mathbf{w}_v^{\text{Image}} = \mathcal{W}^{-1}\mathbf{w}_v \quad . \tag{50}$$

Note that the image $\mathbf{w}_v^{\text{Image}}$ can have positive and negative components. We thus display the energy of the components:

$$e(\vec{x}) = |\mathbf{w}_v^{\text{Image}}(\vec{x})|^2 \quad , \tag{51}$$

where $\vec{x}$ is the pixel coordinate vector.

8.2 Defining A Quantitative Criterion

The energy distribution described in (51) forms the basis for a localization criterion. High localization is obtained if most of the energy of the image $e(\vec{x})$ is concentrated in small areas that are *correct.*

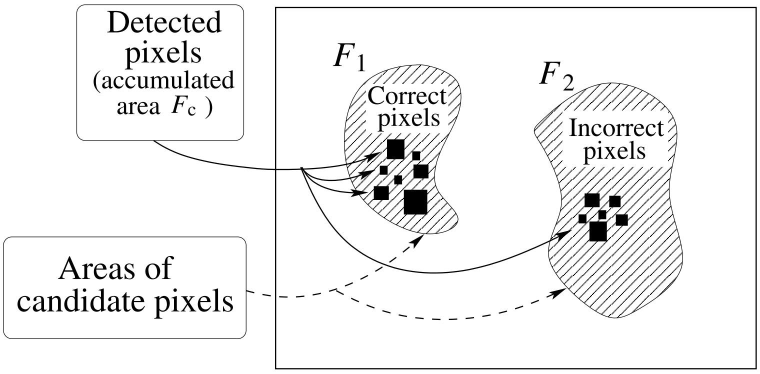

Detected pixels (accumulated area F_c)
F_1
Correct pixels
F_2
Incorrect pixels
Areas of candidate pixels

Figure 8: The candidate dynamic pixels occupy areas F_1 and F_2. Some of them are detected by the audio-visual localization algorithm (marked here in black). If detection is based on a multiresolution representation, then the area of detected pixels typically comprises of blocks of several fixed sizes.

Before audio-visual localization is attempted, all the dynamic pixels[10] are *candidates* for detection. In Fig. 8, they are depicted as residing in regions F_1 and F_2. It must be stressed that all the pixels in those regions are dynamic, since pixels having values with negligible temporal variation are excluded. The pixels detected by the localization algorithm have $e(\vec{x}) > 0$. Some of them are in irrelevant areas. We determine a *correct* detection by manually defining F_1 as the area (of dynamic pixels) corresponding to the sound. For instance, in the sequence appearing in Fig. 1, F_1 includes only pixels in which the hand is moving. The set of correctly detected pixels

$$\mathcal{D}_c \doteq \{\vec{x} \ : \ e(\vec{x}) > 0 \ \text{ and } \ \vec{x} \in F_1\} \tag{52}$$

occupies a cumulative area F_c. The localization criterion is

$$L_c = \frac{\sum_{\vec{x} \in \mathcal{D}_c} e(\vec{x})}{\sum_{\vec{x}} e(\vec{x})} \cdot \frac{F_1 + F_2}{F_c} \quad . \tag{53}$$

It can be easily seen that if there is no preference for localization at the correct region, then $L_c = 1$. The case where $L_c < 1$ indicates failure, as most of the energy is outside the correct region. We seek $L_c \gg 1$, meaning that the energy is concentrated in small areas of correct identity.

[10] The dynamic areas can have a variety of features. These features are not limited for pixels or wavelet components. Features can be corners detected by preprocessing.

9 Experiments

In this section we present results of experiments based on real video sequences. The sequences were sampled at 25 frames/sec at resolution of 576 × 720 pixels.[11] The audio was sampled at 44.1KHz. **Movie #1** features a hand playing a guitar and then a synthesizer. Such an example gives a good demonstration of *dynamics.* The hand playing motion is distracted by a rocking wooden-horse. Some raw data of this sequence appears in Fig. 1. **Movie #2** features a talking face and a distracting rocking wooden-horse as well. The audio plot and a representative frame of this sequence are shown in Fig. 9. Both movies can be linked through *http://www.ee.technion.ac.il/∼ yoav/AudioVisual.html* .

The experiments had the following features, aimed at demonstrating the capability of our approach:

- **Handling dynamics.** Each sequence was ≈ 10 seconds long. However, analysis was performed on intervals of $N_F = 32$ frames (≈ 1 second).

- **Handling false-positives and noise.** The sequences deliberately include strong visual distractions (a rocking wooden-horse), challenging the algorithm. Moreover, in some experiments we added strong audio noises (SNR=1), in the form of unseen talking people (via a recording), broadband noise, or background beats.

- **High spatial resolution (localization).** In some of the prior work, pruning of visual features had been very aggressive, greatly decreasing spatio-temporal resolution. Our algorithm does *not* need this, thanks to the sparsity criterion. Nevertheless, memory limits currently restricted the number of visual features to $N_v = 3000$. The dynamic pixels in our frames were effectively represented by wavelet coefficients of such dimensions, as described below. The dynamic pixels are shown in Fig. 10. It is stressed that pruning was done only for reducing the computational load. However, we observed in experiments that using a larger number of features has a diminishing return. We aim to demonstrate high spatial resolution in the resulting visual localization.

- **No parameters to tweak.** The implementation has essentially no parameters. The selection of $N_F = 32$ represents our desire to localize brief events, but longer time intervals can be used as well. The selection of $N_v = 3000$ stems from hardware limits, but the results

[11] We used only the pixel intensities, and discarded the chromatic channels.

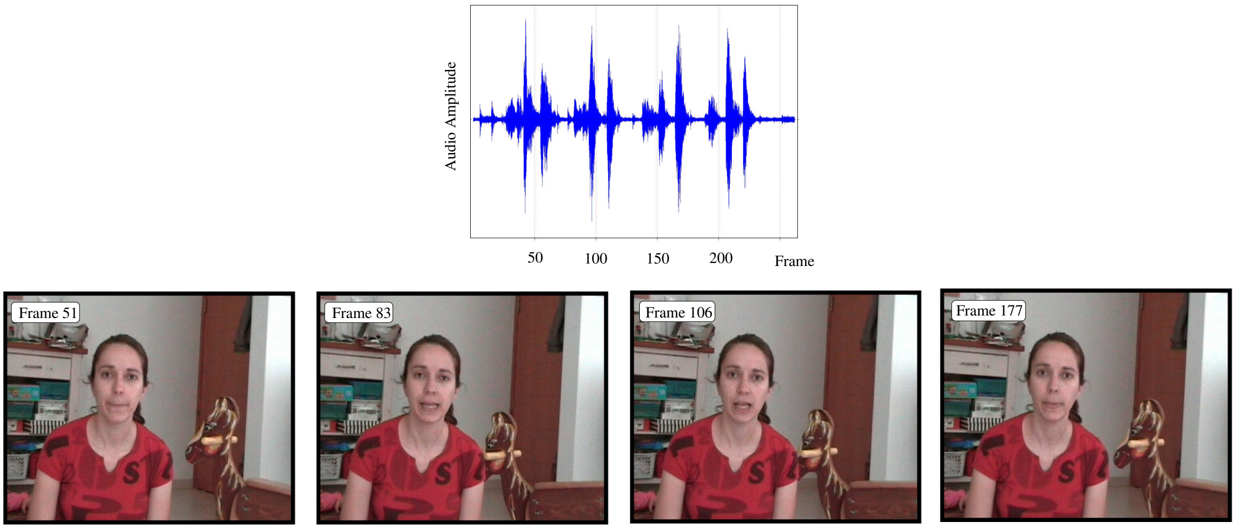


Figure 9: Movie #2 includes a talking face and a moving wooden horse. [Top] The audio signal. [Bottom] Sample frames.

of our experiments were robust to this choice, as verified in experiments.

- **Simple audio representation.** Our experiments *did not attempt to filter sounds*, but rather to filter the visual signals. Hence, only a few audio bands were used. We analyzed the sequences using a single wide band ($N_a = 1$), averaging sound energy at each frame (1/25'th second). We then re-analyzed the data using $N_a = 4$ audio bands, selected as the strongest periodogram coefficients.

Since a sparse representation is desired, we worked on temporal-difference images, applying a wavelet transform to each of these difference-frames [13, 29]. We choose to use wavelet decomposition up to level 3. Coarser levels may inclines the algorithm to choose coarser level coefficients, which reduces the ℓ^1 value but expands the spatial spread in the image domain.

Fig. 11 shows sample frames resulting from the analysis of **Movie #1**. At each frame, we overlaid the energy distribution of the detected pixels $e(\vec{x})$ with the corresponding raw image. The algorithm pinpointed the source of the sound on the motion of the *fingers*, demonstrating both high spatial accuracy and temporal resolution. Compared to the large area occupied by dynamic pixels in Fig. 10, the detected pixels in Fig. 11 are concentrated in much smaller areas. Thus, high localization is achieved. Note that the algorithm handls *dynamics*. First, the guitar is detected, corresponding its audio tones. When the hand played the synthesizer,

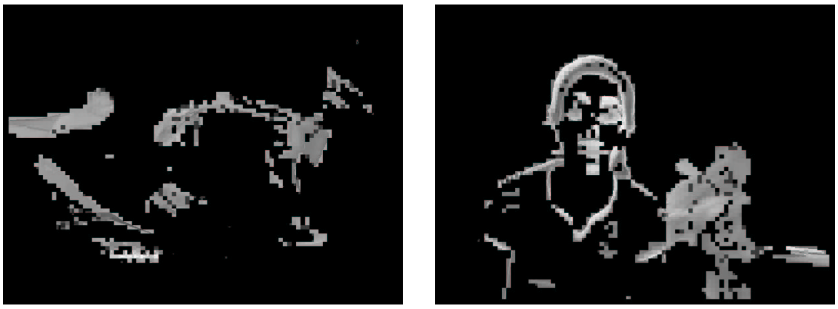


Figure 10: Dynamic pixels expressed by the wavelet components, using level 3, in [Left] Movie #1 [Middle] Movie #2 [Right] Movie #3. Graylevels indicate the temporal average of pixels values. Black regions represent static pixels.

	Using ℓ^1-norm	Using ℓ^2-norm
Movie #1	58 ± 20	4.0 ± 0.8
Movie #2	81 ± 20	2.9 ± 0.6

Table 1: The localization criterion L_c obtained in the experiments. The reported numbers are the mean and standard deviation of the measurements. The use of the ℓ^1-norm leads to sharp localization, much better than that resulting from ℓ^2.

the algorithm managed to shift its focus accordingly. The motion distractions (rocking horse) were successfully filtered out by our audio-visual localization algorithm.

Similarly, Fig. 12 shows sample frames resulting from the analysis of **Movie #2**. Here pixels in the *mouth* were predominantly detected as correlated with the audio. Similarly to the results of Movie #1, the motion distractions are successfully filtered out.

To judge the results, we compare our algorithm to the performance obtained using ℓ^2 regularization, as in (20). Typical sample frames are shown in Fig. 13. They suffer from very poor localization and detection rate: there are many false-positives (especially detection of the moving horse), while the energy spreads over a large area. Table 1 reports the temporal mean and standard deviation of the empirical localization values L_c, resulting from the use of either the ℓ^1 or ℓ^2-based localization algorithms. These quantitative results indicate that using the ℓ^2-based solution achieves poor localization, compared to the ℓ^1-norm counterpart.

As mentioned above, we repeated our experiments by sequentially adding three types of audio disturbances. The results were within the standard deviation of the L_c values reported in Table 1. Moreover, the multiple audio representation using $N_a = 4$ was tested. The performance was very similar to that described in Figs. 11, 12 and Table 1.

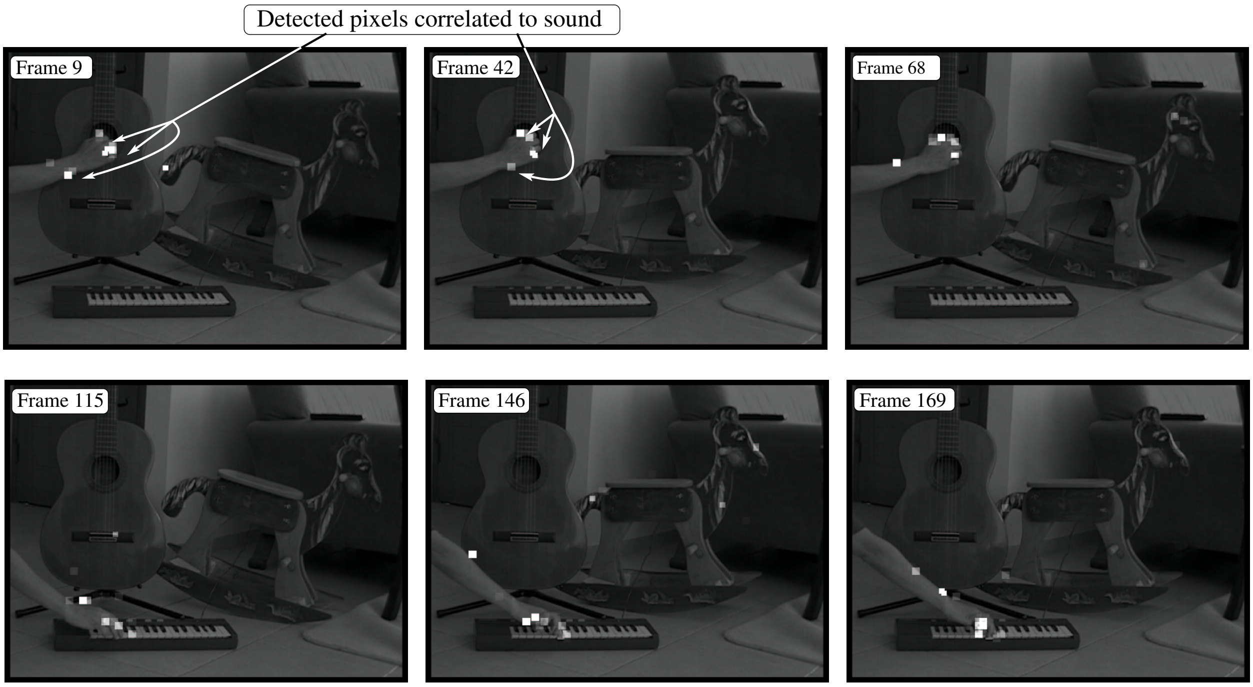


Figure 11: The algorithm results, when run on Movie #1. For visualization, we overlayed the detected energy distribution with the corresponding sample raw frames. Localization concentrates on the playing fingers, which dynamically move from the guitar to the synthesizer. Sporadic detections exist in other areas, usually with much lower energies.

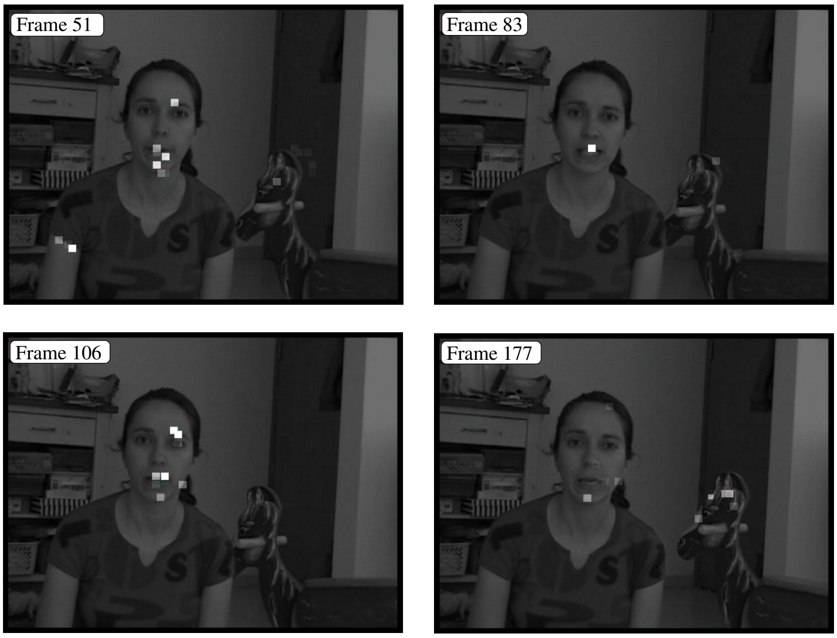


Figure 12: Sample frames resulting from the algorithm, when run on Movie #2. The visualization is as described in Fig. 11. Localization in the mouth area is consistent. Sporadic detections exist in other areas, usually with much lower energies.

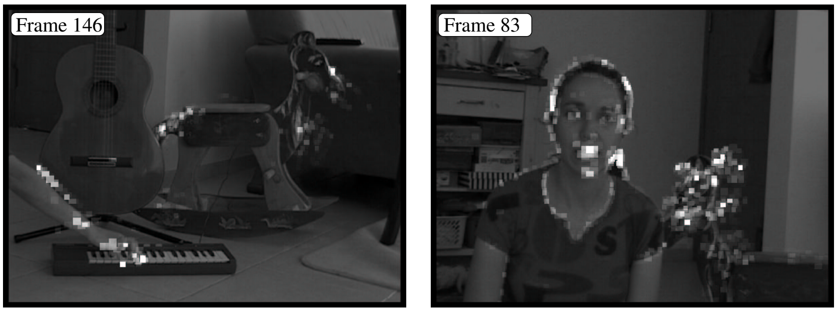


Figure 13: Typical results of using ℓ^2 as a criterion. Compared to the corresponding frames shown in Fig. 11 and 12, the detected energy is much more spread, particularly in non-relevant areas (see the wrong detection of the horse on the right frame).

Post Processing for Visualization

The algorithm described above hardly exploits spatial coherence and temporal consistency, which are typical to audio-visual events. Still, it yields good results. Nevertheless, the performances can be improved by further development of these aspects. This can be done by reformulating the optimization problem using priors expressing spatial coherence and temporal consistency. That option is elegant, but solving it is complex. We opted for an alternative option, in which post-processing is applied to the results of our algorithm, to filter out inconsistent behavior in time and space. This option is simpler and faster, since it involves concatenation of two relatively simple stages. As the post processing stage, we performed temporal median filtering (in windows of 10 frames), followed by spatial convolution with a 5×5 Gaussian kernel. The first step deletes temporal outliers, while the second stabilizes spatial positions and filters out fluctuations. Samples of resulting frames are shown in Fig. 14 and Fig. 15.

10 Discussion

We have presented a robust approach for audio-visual dynamic localization, based on a single microphone. It overcomes the lack of sufficient data (ill-posedness) associated with short time intervals. The algorithm exploits the spatial sparsity of audio-visual events. Furthermore, leaning on recent results that show the relation between sparsity and the ℓ^1-norm, we are able to convexize the problem. Our algorithm is parameter-free, and is thus robust to scenario variability. Nevertheless, the principles posed here can become the base for a more elaborate

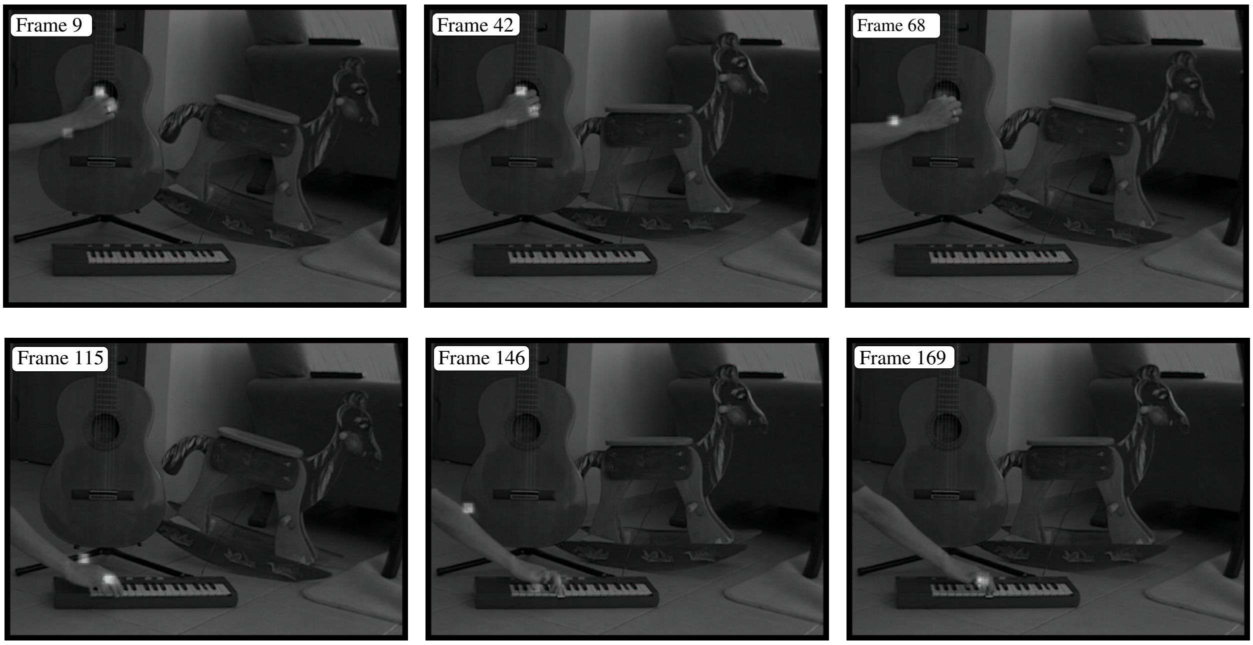


Figure 14: Results of post processing of the algorithm output when the input is Movie #1. Compared to Fig. 11, the detected regions are much more stable and contain much less false-positives. **Movie results are linked via http://www.ee.technion.ac.il/∼yoav/AudioVisual.html** .

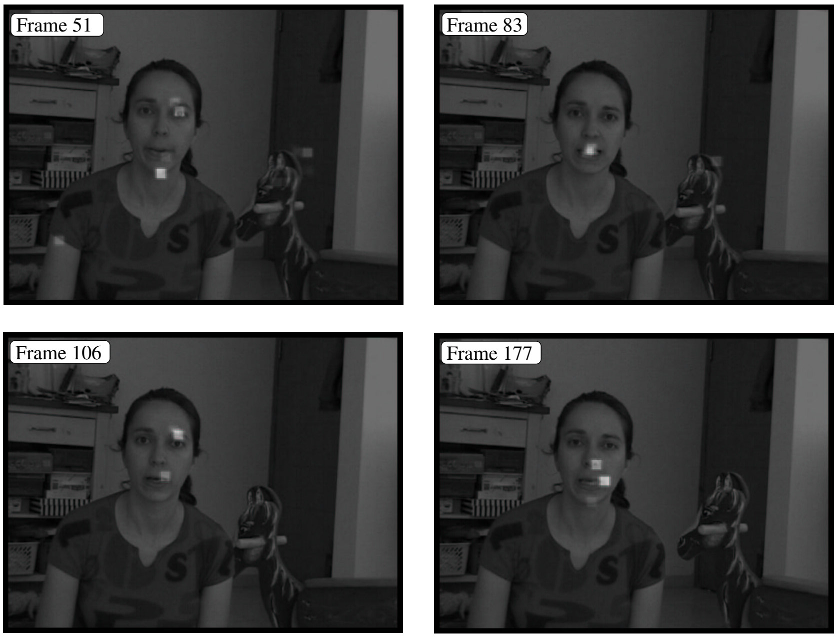


Figure 15: Results of post processing of the algorithm output when the input is Movie #2. Compared to Fig. 11, the detected regions are much more stable and contain much less false-positives.

localization approach, that uses spatial temporal consistency as a prior, as done in tracking methods.

It is possible to extend this approach, e.g., by a kernel version for treating nonlinear relations between the modalities [1, 30, 38]. One may go further and generalize audio-visual localization to multiple simultaneous visual events. In addition, time-lag between the audio and the video data can be introduced as a variable in the optimization. Based on the speed of sound, this would enable estimation of object distances from the camera. Furthermore, our sparsity-based approach may be helpful in other scientific domains that aim to correlate arrays of measurement vectors (unrelated to sound), such as climatology.

A Bounds of G

In this appendix we prove that the penalty function given by (5) is bounded by 0 and 2. Looking at $G(\mathbf{w}_v, \mathbf{w}_a)$ in (5) it is clear that this nonnegative function becomes zero when $\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a$. Hence, the lower bound of G is zero.

For the upper bound, consider the special case $\mathbf{V}\mathbf{w}_v = -\mathbf{A}\mathbf{w}_a$. Using this in (5) yields $G(\mathbf{w}_v, \mathbf{w}_a) = 2$. Next, we prove that $G(\mathbf{w}_v, \mathbf{w}_a)$ cannot be larger then 2.

$$\begin{aligned}\frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|^2}{\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2} &\leq \frac{(\|\mathbf{V}\mathbf{w}_v\| + \|\mathbf{A}\mathbf{w}_a\|)^2}{\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2} = 1 + \frac{2\|\mathbf{V}\mathbf{w}_v\|\|\mathbf{A}\mathbf{w}_a\|}{\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2} \qquad (54)\\ &\leq 1 + \frac{\sqrt{\|\mathbf{V}\mathbf{w}_v\|^2\|\mathbf{A}\mathbf{w}_a\|^2}}{(\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2)/2} \leq 1 + \frac{(\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2)/2}{(\|\mathbf{V}\mathbf{w}_v\|^2 + \|\mathbf{A}\mathbf{w}_a\|^2)/2} = 2\ .\end{aligned}$$

The last inequality is valid since the geometric average is always smaller or equal to the corresponding arithmetic average. As explained in Sec. 3, the ranges $0 \leq G \leq 1$ and $1 \leq G \leq 2$ are equivalent.

B Sparsity using ℓ^1

Suppose we seek to solve

$$\min\ \|\mathbf{w}_v\|_0 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A}\ . \qquad (55)$$

This task is highly complex (known to be NP-hard) [10, 16, 22], being a combinatorial problem whose complexity grows exponentially with the number of columns in $\mathbf{V}$. Fortunately we may use an approximation method that replaces the ℓ^0-norm with an ℓ^1 norm, yielding

$$\min\ \|\mathbf{w}_v\|_1 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A}\ . \tag{56}$$

This approximation is known as the *basis pursuit* algorithm [11, 12]. Replacing the ℓ^0-norm with ℓ^1-norm can be seen as a way of convexizing the target problem (55). The advantage of such a change is that it can be cast as a linear programming problem and be solved by modern interior point methods, even for very large N_v.

Recent studies have established that if the solution of (55) is sparse enough, then (i) no other solution exists with the same or lower cardinality (*uniqueness*); and (ii) solving Eq. (56) yields a solution which is identical to the solution of Eq. (55) (*equivalence*) [22]. Both the uniqueness and the equivalence results are derived from the properties of the matrix $\mathbf{V}$. Defining $\mathbf{v}_n$ as the n-th column in this matrix, the *mutual incoherence* is defined as

$$M = \max_{n\neq j}\ \frac{|\mathbf{v}_n^T\mathbf{v}_j|}{\|\mathbf{v}_n\|_2\|\mathbf{v}_j\|_2} \tag{57}$$

for $n, j = 1, 2, ..., N_v$. The work reported in [10] shows that uniqueness and equivalence of Eqs. (55) and (56) hold true if the solution satisfies[12]

$$\|\mathbf{w}_v^{\text{optimal}}\|_0 < 0.5\left(1+\frac{1}{M}\right) \tag{58}$$

In this case, the solution is considered to be a *highly sparse solution*, and solving Eq. (56) can replace Eq. (55). Thus, if we obtain the solution of Eq. (56) and observe that it happens to be sparse beyond the threshold (Eq. 58), then we know that we have also solved (55).

The bound in Eq. (58) is rather restrictive. It is very conservative since it relates to worst-case scenarios. There are, however, cases where this restriction in meaningless. Consider an extreme case where the matrix $\mathbf{V}$ includes two identical columns. In this case, Eq. (57) yields $M = 1$, implying that uniqueness and equivalence hold true for $\mathbf{w}_v$ vectors having less than a single non-zero component (i.e., the entire vector is zero). Such observation is useless. Apparently, in this special case, the equivalence between (55) and (56) may break down if we rely only on the bound in Eq. (58). However, empirical tests show that the basis pursuit algorithm (56) recovers the solution of (55) for cases far exceeding this bound.

Encouraged by these empirical observations, very recent theoretical analysis [4, 9, 16] addressed the above questions from a probabilistic point of view. This analysis has replaced a

[12] It can be shown [10] that $\sqrt{\frac{N_v-N_F}{N_F(N_v-1)}} \leq M \leq 1$.

deterministic claim of "guaranteed uniqueness and equivalence" with a claim of "guaranteed uniqueness and equivalence with probability one". These studies establish a much higher bound on the cardinality of the solution to guarantee success.[13] These new results stand as support to our experiments (Sec. 9), where basis pursuit succeeded in locking on a very sparse solution.

Acknowledgments

Yoav Schechner is a Landau Fellow - supported by the Taub Foundation, and an Alon Fellow. The work was supported by the US-Israel Binational Science Foundation (BSF), and the Ollendorff Minerva Center in the Elect. Eng. Dept. at the Technion. Minerva is funded through the BMBF.

References

[1] F. Bach and M. Jordan, 2002, "Kernel independent component analysis," J. of Machine Learning Research **3**, pp. 1-48.

[2] M. J. Beal, N. Jojic, and H. Attias, 2003, "A graphical model for audiovisual object tracking," IEEE Tran. on PAMI **25**, pp. 828-836.

[3] C. Bregler, and Y. Konig, 1994, "Eigenlips for robust speech recognition," In Proc. IEEE ICASSP, vol. **2**, pp. 667-672.

[4] E. Candes, J. Romberg, and T. Tao, 2004, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," submitted to IEEE Trans. on Information Theory.

[5] R. Cutler, and L. Davis, 2000, "Look who's talking: speaker detection using video and audio correlation," Proc. IEEE Int. Conf. on Multimedia and Expo (ICME), vol. **3**, pp. 1589-1592.

[6] T. De Bie, and B. De Moor, 2003, "On the regularization of canonical correlation analysis," Int. Sympos. ICA and BSS, pp. 785-790.

[7] K. De Cock, and B. De Moor, 2002, "Subspace angles and distances between ARMA models," in Systems and Control Letters **46**, pp. 265-270.

[8] S. Deligne, G. Potamianos, and C. Neti, 2002, "Audio-visual speech enhancement with AVCDCN (audio-visual codebock dependent cepstral normalization)," IEEE Workshop on Sensor Array and Multichannel Signal Processing, pp. 68-71.

[9] D. L. Donoho, 2004, "For most large underdetermined systems of linear equations, the minimal ℓ^1-norm solution is also the sparsest solution," Tech. Rep., Statistics Dept., Stanford U.

[10] D. L. Donoho, and M. Elad, 2003, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," Proc. Nat. Aca. Sci. **100**, pp. 2197-2202.

[13] Those results assume a special structure of $\mathbf{V}$ and an asymptotic behavior, which do not necessary exist in our case.

[11] D. L. Donoho, and M. Elad, 2005, "On the stability of the basis pursuit in the presence of noise," accepted to the EURASIP Signal Processing Journal.

[12] D. L. Donoho, M. Elad, and V. Temlyakov, 2005, "Stable recovery of sparse overcomplete representations in the presence of noise," accepted to the IEEE Trans. on Information Theory.

[13] D. L. Donoho, and A. G. Flesia, 2001, "Can recent innovations in harmonic analysis explain key findings in natural image statistics?," Network: Comput. Neural. Syst. **12**, pp. 371-393.

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

[15] M. Elad and A.M. Bruckstein, 2002, "A generalized uncertainty principle and sparse representation in pairs of bases," IEEE Trans. on Information Theory **48**, pp. 2558-2567.

[16] M. Elad, and M. Zibulevsky, 2004, "A probabilistic study of the average performance of the basis pursuit", submitted to the IEEE Trans. on Information Theory.

[17] G. Farnebäck, 1999, "A unified framework for bases, frames, subspace bases, and subspace frames", Proc. Scand. Conf. Image Analysis pp. 341-349.

[18] D. E. Feldman, and E. I. Knudsen, 1996, "An anatomical basis for visual calibration of the auditory space map in the barn owl's midbrain," The J. Neuroscience, vol. **17** pp. 6820-6837.

[19] J. W. Fisher III, and T. Darrell, 2004, "Speaker association with signal-level audiovisual fusion," IEEE Trans. Multimedia **6**, pp. 406-413.

[20] J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola, 2001, "Learning joint statistical models for audio-visual fusion and Segregation," Advanced in Neural Inf. Process. Syst. **13**, pp. 772-778.

[21] G. H. Golub, and C. F. Van Loan, *Matrix Computations*, 3th edition, The Johns Hopkins University Press, Baltimore, 1996.

[22] R. Gribonval, and M. Nielsen, 2003, "Sparse representations in unions of bases," IEEE Trans. on Information Theory **49**, pp. 3320-3325.

[23] Y. Gutfreund, W. Zheng, and E. I. Knudsen, 2002, "Gated visual input to the central auditory system," Science **297**, pp. 1556-1559.

[24] J. Hershey, and J. Movellan, 1999, "Audio-vision: using audio-visual synchrony to locate sound," Advances in Neural Inf. Process. Syst. **12**, pp. 813-819.

[25] B. Kapralos, M. R. M. Jenkin, and E. Milios, 2003, "Audiovisual localization of multiple speakers in a video teleconferencing setting," Int. J. Imaging Systems and Technology **13**, pp. 95-105.

[26] H. Knutsson, M. Borga, and T. Landelius, 1995, "Learning canonical correlations," Tech. Rep. LiTH-ISY-R-1761, Computer Vision Laboratory, S-581 83 Linköping Univ., Sweden.

[27] H. Knutsson, M. Borga, and T. Landelius, 1998, "Learning multidimensional signal processing," Proc. ICPR, vol. II, pp. 1416-1420.

[28] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, 2003, "Multimedia content processing through cross-modal association," Proc. ACM Int. Conf. Multimedia, pp. 604-611.

[29] S. G. Mallat, 1989 "A theory multiresolution signal decomposition: The wavelet representation," IEEE Trans. on PAMI, vol. **11**, pp. 674-693.

[30] T. Melzer, M. Reiter, and H. Bischof, 2003, "Appearance models based on kernel canonical correlation analysis," Patt. Rec. **36**, pp. 1961-1971.

[31] D. Murphy, T. H. Andersen, and K. Jensen, 2003, "Conducting audio files via computer vision," Proc. Gesture Workshop, pp. 529-540.

[32] H. J. Nock, G. Iyengar, and C. Neti, 2002, "Assessing face and speech consistency for monologue detection in video," Proc. ACM Int. Conf. Multimedia, pp. 303-306.

[33] C. Schauer, and H. M. Gross, 2003, "A computational model of early auditory-visual integration," Proc. Patt. Rec. Sympos., Lecture Notes in Computer Science **2781** pp. 362-369.

[34] M. Slaney, and M. Covell, 2000, "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," Advanc. in Neural Inf. Process. Syst. **13**, pp. 814-820.

[35] M. Song, J. Bu, C. Chen, and N. Li, 2004, "Audio-visual based emotion recognition-a new approach," Proc. IEEE CVPR, vol. 2, pp. 1020-1025.

[36] P. Smaragdis, and M. Casey, 2003, "Audio/Visual independent components," Int. Sympos. ICA and BSS, pp. 709-714.

[37] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, 2001, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," Proc. IEEE ICCV, vol. 1, pp. 741-746.

[38] L. Wolf, A. Shashua, 2003, "Learning over Sets using Kernel Principal Angles," J. of Machine Learning Research **4**, pp. 913-931.