NONNEGATIVE CCA FOR AUDIOVISUAL SOURCE SEPARATION

Christian Sigg, Bernd Fischer, Björn Ommer, Volker Roth and Joachim Buhmann

Institute of Computational Science, ETH Zurich, 8092 Zurich, Switzerland

chrsigg@inf.ethz.ch

ABSTRACT

We present a method for finding correlated components in audio and video signals. The new technique is applied to the task of identifying sources in video and separating them in audio. The concept of canonical correlation analysis is reformulated such that it incorporates nonnegativity and sparsity constraints on the coefficients of projection directions. Nonnegativity ensures that projections are compatible with an interpretation as energy signals. Sparsity ensures that coefficient weight concentrates on individual sources. By finding multiple conjugate directions we finally obtain a component based decomposition of both data modalities. Experiments effectively demonstrate the potential and benefits of this approach.

1. INTRODUCTION

In difficult auditory environments, where many talkers interfere and their spectral characteristics overlap, humans often make use of visual cues to facilitate understanding. In contrast to the auditory signal, the visual input is instantaneous, most often free of reflections, and regions of the visual field can be uniquely assigned to a source. We therefore expect a benefit from incorporating video into a source separation or signal enhancement algorithm. In [1] Yehia et al. have shown that facial behavior and speech acoustics are highly correlated, using infrared markers and a high-speed tracker to obtain precise 3D coordinates of points on the face. Slaney and Covell have shown in [2] that similar results are possible using pixel intensities from a video image of the lower facial region. In this work, we show that trajectories of interest points are another sufficient representation of movement, without the need for specialized equipment or explicit face detection and alignment. Furthermore, we do not restrict our analysis to speech only. For example, the relationship between movement and sound is straightforward with certain musical instruments such as a drum or a strummed guitar. In principle, any auditory source that has corresponding movement (or intensity change) observable in video can be incorporated in the analysis.

In recent years, there have been several proposals to exploit the statistical dependence of synchronous audio and video signals. Methods of this kind typically find projections of both data modalities that either maximize (approximate) mutual information [3, 4] or correlation [2, 5]. These methods, however, have limitations in several aspects, e.g. the restriction to differentiable L_2 penalties [4], or the asymmetric treatment of audio and video [5]. We propose a method that identifies and separates *several* concurrently active sources, by restricting projection directions to the nonnegative orthant and efficiently incorporating sparsity constraints in both modalities. Nonnegativity assures that projected signals define energies, which successively decompose the total audio and video information. E.g. when using pixel intensities as the visual input, projections are valid images themselves. On the audio side, nonnegativity constraints allow the interpretation of the projection direction as time-varying filter weights, which amplify frequency bands that correlate well to their visual counterpart and attenuate others.

The key idea is to include these constraints in a generalized version of *canonical correlation analysis* (CCA) that is based on iterated regression. The method is highly flexible in that it allows the choice of individual regularization strategies for the different data modalities such as sparseness constraints for video and smooth L_2 penalties for audio. Furthermore, it allows us to diminish the influence of outliers by substituting least-squares with robust regression procedures.

2. METHOD OVERVIEW

We perform canonical correlation analysis to locate sources in video and separate their corresponding audio signals by filtering. Modeling audio and video as random vectors, we seek linear projection vectors that maximize the correlation between the two projected signals. To locate a source in the video signal, we identify those components (pixels or interest points) that contribute most to the projection. On the audio side, a properly defined projection onto frequency bands may be interpreted as a frequency-domain filter, amplifying frequencies contained in the source and attenuating others. Other representations of the audio signal, such as mel-frequency cepstral coefficients (MFCCs) or line spec-

1-4244-1566-7/07/$25.00 ©2007 IEEE.

tral pairs (LSPs), are of course possible. For the purpose of source localization and separation, a spatial representation of the auditory scene is of particular importance. Such representations can e.g. be obtained from using a microphone array together with a *filter-and-sum beamformer*, that combines FIR filtered microphone signals such that a certain direction is preferred. In this setting, projections of the input stream associate groups of audio sub-band energies (or any other set of grouped coefficients) with spatial coordinates. If, in addition, these projections are *sparse* in the number of groups, they might be used to localize individual sources in the scene.

CCA maximizes the correlation of random vectors projected to one-dimensional subspaces. In practice, we assume that we are given two samples $\mathbf{A}$ and $\mathbf{V}$ of observations from the audio and video random vectors. In the context of joint audio-video analysis, these samples might be mimicked by considering n subsequent frames of the video/audio streams recorded at time points $t_1, \ldots, t_n$. In order to capture relevant aspects of the randomness in the data, it is clear that one has to find a compromise between the number of sample points (which would favor high sample rates and long time windows) and the absence of correlations and scene changes (which means low sample rates and short windows).

Even for low resolution video (e.g. 160x120 pixels) at 25Hz frame rate, the CCA problem will be severely underdetermined when considering reasonably short time windows. Representing the video as pixel intensity vectors, we end up with a few hundred "samples" of a 19200 dimensional random variable, which means that we will always find trivial projections that correlate perfectly. Using interest points instead of pixel intensities, the deficit in number of samples versus dimensions is reduced, but the fundamental problem of an under-determined setting remains.

In order to find meaningful projections, we need to include a regularization term. Concerning the video signal, it is desirable to have sparse projection vectors $\boldsymbol{\beta}$ so that only those components have nonzero weights, that are associated with the source in question. Penalizing the L_1-norm of $\boldsymbol{\beta}$ as achieved in LASSO regression [6] is a promising candidate for such a learning method to generate sparse representations. The L_1 penalty naturally generalizes to a group sparsity constraint [7] if there are several features per component, e.g. the cartesian and polar coordinates of an interest point. In that case, feature weights within the same group are L_2 penalized, whereas the between-group penalty is the L_1 norm.

On the audio side, the desired regularization properties crucially depend on the data representation and on the application context. If we are interested in *reconstructions* of the audio signals and if we choose e.g. a frequency band representation, sparsity is probably not desirable, because omitting frequency bands will lead to undesired audible artifacts in the reconstructed audio streams. L_2 regularization or a smoothness penalty on coefficients α_i, α_{i+1} of adjacent frequency bands will be more adequate in such a situation. If, on the other hand, we represent the audio signal by directionally grouped components computed from a *beamforming* device, sparsity in the number of groups is again a desired property since it makes it possible to locate individual sources in the scene.

3. NONNEGATIVE CANONICAL CORRELATION ANALYSIS

The classical CCA method finds directions $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, such that the linear projections of two multidimensional random vectors have maximum correlation

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathrm{corr}(\mathbf{A}\boldsymbol{\alpha}, \mathbf{V}\boldsymbol{\beta}). \tag{1}$$

$\mathbf{A}$ and $\mathbf{V}$ are matrices of size $n \times d_a$ and $n \times d_v$, where each row corresponds to one realization of the random variable. In practice, these different realizations are mimicked by using successive frames in the audio and video signal.

From the definition of the sample correlation between two zero mean vectors $\mathbf{x}$ and $\mathbf{y}$

$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, \tag{2}$$

it follows immediately that (assuming centralized matrices, see eq. (7) below) maximizing the correlation in (1) is equivalent to finding $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ as follows:

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{A}\boldsymbol{\alpha} - \mathbf{V}\boldsymbol{\beta}\|^2, \tag{3}$$

$$\text{s.t.} \quad \|\mathbf{A}\boldsymbol{\alpha}\|^2 = 1 \ \wedge \ \|\mathbf{V}\boldsymbol{\beta}\|^2 = 1, \tag{4}$$

where $\|\mathbf{x}\|^2 = \sum_{i=1}^{d} x_i^2$ denotes the squared L_2 norm of $\mathbf{x}$. The solution is readily obtained using the eigenvalue decomposition of the (sample) covariance matrix

$$\mathbf{C} = n^{-1} \begin{bmatrix} \mathbf{A}^\top\mathbf{A} & \mathbf{A}^\top\mathbf{V} \\ \mathbf{V}^\top\mathbf{A} & \mathbf{V}^\top\mathbf{V} \end{bmatrix} \tag{5}$$

$$= \begin{bmatrix} \mathbf{C}_{aa} & \mathbf{C}_{av} \\ \mathbf{C}_{va} & \mathbf{C}_{vv} \end{bmatrix}, \tag{6}$$

where we have assumed that the columns of both matrices are centered, i.e.

$$\sum_{i=1}^{n} A_{ij} = \sum_{k=1}^{n} V_{kl} = 0, \ \forall j = 1, \ldots, d_a \text{ and } l = 1, \ldots, d_v. \tag{7}$$

A full derivation of the procedure can be found e.g. in [8].

From the eigenvalue decomposition of $\mathbf{C}$ we find all direction pairs[1]

$$(\widehat{\boldsymbol{\alpha}}_{(k)}, \widehat{\boldsymbol{\beta}}_{(k)}),\ k = 1, \ldots, \min\big(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{V})\big) \quad (8)$$

that satisfy (1) under the condition that projection directions are mutually conjugate, i.e. these vectors obey the following relations $\forall k \neq l$:

$$\widehat{\boldsymbol{\alpha}}_{(k)}^{\top} \mathbf{C}_{aa} \widehat{\boldsymbol{\alpha}}_{(l)} = \widehat{\boldsymbol{\beta}}_{(k)}^{\top} \mathbf{C}_{vv} \widehat{\boldsymbol{\beta}}_{(l)} = 0. \quad (9)$$

One can show that if these conditions are satisfied, it also holds that $\forall k \neq l$:

$$\widehat{\boldsymbol{\alpha}}_{(k)}^{\top} \mathbf{C}_{av} \widehat{\boldsymbol{\beta}}_{(l)} = 0. \quad (10)$$

In order to simplify the notation, we will neglect the subscripts $_{(k)}$ indicating the k-th projection direction in the following. Unless explicitly stated, the equations are meant to hold for all possible k.

3.1. Iterative regression solver

For a given $\widehat{\boldsymbol{\alpha}}$, the optimization criterion (3) is just the minimum mean-squared error criterion for regression coefficients $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{A}\widehat{\boldsymbol{\alpha}} - \mathbf{V}\boldsymbol{\beta}\|^2, \quad (11)$$

followed by a rescaling step

$$\widehat{\boldsymbol{\beta}} \leftarrow \widehat{\boldsymbol{\beta}} / \|\mathbf{V}\widehat{\boldsymbol{\beta}}\|. \quad (12)$$

This formulation suggest an iterative solution approach to the CCA problem: for iteration $(t+1)$, one set of parameters (for instance $\widehat{\boldsymbol{\alpha}}^{(t)}$) is held constant while a regression step is performed to find the corresponding set of coefficients $\widehat{\boldsymbol{\beta}}^{(t+1)}$. Then $\widehat{\boldsymbol{\beta}}^{(t+1)}$ is held fixed and we determine the corresponding optimal $\widehat{\boldsymbol{\alpha}}^{(t+1)}$. This way of solving the CCA problem has been proposed several times in the literature, see for example [9].

If the dimensionality of the data is larger than the number of samples, it is mandatory to regularize the fits. Therefore an appropriate penalty term is added to constrain the norm of the regression coefficients. If we, for instance, choose group sparsity penalties [7] on the video and audio side (with appropriate group size for each modality), the following update scheme results:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{A}\widehat{\boldsymbol{\alpha}}^{(t)} - \mathbf{V}\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{g=1}^{G} \|\boldsymbol{\beta}_{[g]}\| \right) \\
\widehat{\boldsymbol{\beta}}^{(t+1)} &\leftarrow \widehat{\boldsymbol{\beta}}^{(t+1)} / \|\mathbf{V}\widehat{\boldsymbol{\beta}}^{(t+1)}\| \\
\widehat{\boldsymbol{\alpha}}^{(t+1)} &= \arg\min_{\boldsymbol{\alpha}} \left(\|\mathbf{V}\widehat{\boldsymbol{\beta}}^{(t+1)} - \mathbf{A}\boldsymbol{\alpha}\|^2 + \lambda_2 \sum_{h=1}^{H} \|\boldsymbol{\alpha}_{[h]}\| \right) \\
\widehat{\boldsymbol{\alpha}}^{(t+1)} &\leftarrow \widehat{\boldsymbol{\alpha}}^{(t+1)} / \|\mathbf{V}\widehat{\boldsymbol{\alpha}}^{(t+1)}\|,
\end{aligned}$$

where $\boldsymbol{\beta}_{[g]} = (\beta_{g1}, \ldots, \beta_{gk})^{\top}$ are the coefficients of the g-th group of $\boldsymbol{\beta}$. The group size $c_v = d_v/G$ is chosen to match the input representation, e.g. $c_v = 4$ for the coordinate quadruple of an interest point. This procedure is iterated until convergence of the projection directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

The following three benefits can be realized in this update scheme:

(i) Flexibility in choosing appropriate regularization models. We can perform ridge regression (L_2 penalties), the LASSO (L_1 constrained regression) or combine both in a group penalty, and separately choose an appropriate regularization model for each data modality and each application domain.

(ii) Handling of outliers via robust regression. Techniques for robust regression can be readily incorporated: the quadratic error term $\|\cdot\|^2$ in (11) can be replaced with more robust measures such as the Huber loss function $L_c(\cdot)$ in order to diminish the effect of outliers in the data. In our context, such outliers might occur as audible pops and crackles or temporary mismatches of the interest point tracker.

(iii) Nonnegativity constraints. Finally, it is straightforward to include nonnegativity constraints on the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, i.e. to optimize equation (3) under the additional constraints $\alpha_i \geq 0,\ \beta_j \geq 0,\ \forall i, j$. This problem can be solved via quadratic programming algorithms. These additional sign constraints lead us to the desired model of nonnegative CCA which ensures that $\mathbf{A}\widehat{\boldsymbol{\alpha}}$ and $\mathbf{V}\widehat{\boldsymbol{\beta}}$ are themselves valid audio and video energy signals. It follows that successively found correlation directions decompose the two data modalities into additive energy components.

3.2. Sparse nonnegative regression techniques

Fast approximative techniques for nonnegative regression have been proposed recently. One such method is the *monotone incremental forward stagewise regression* (MIFSR) approach of Hastie et al. [10], which computes the *monotone* LASSO solution. In this variant of standard LASSO, the coefficient weights β_j monotonically increase when relaxing the L_1-constraint. The method inherently finds nonnegative

[1] For the k-th projection vector we typeset the vector in boldface with bracketed subscripts, i.e. $\boldsymbol{\alpha}_{(k)}$, in order to distinguish it from the k-th component γ_k of vector $\boldsymbol{\gamma}$.

weights, which for standard applications (where this non-negativity property is undesirable) is compensated for by replicating the input data with negative sign. For our purposes, we simply omit this replication step. The algorithm is readily extended to the group sparsity case by replacing the L_1 termination criterion $\sum_j \beta_j \leq s$ by $\sum_g \|\boldsymbol{\beta}_{[g]}\| \leq t$. This algorithm is very efficient for sparse solutions, even when the data dimensionality is high.

MIFSR for fitting the video data $\mathbf{V}$ against the target projection $\mathbf{A}\widehat{\boldsymbol{\alpha}}$ proceeds as follows (ϵ is a predefined step-width parameter):

Monotone incremental forward stagewise regression:

1. Start with $\mathbf{r} := \mathbf{A}\widehat{\boldsymbol{\alpha}},\ \boldsymbol{\beta} = \mathbf{0}$.
2. Find the j-th column vector $\mathbf{v}_j$ of matrix $\mathbf{V}$ most positively correlated with $\mathbf{r}$.
3. Update the j-th component of $\boldsymbol{\beta}$ as $\beta_j \leftarrow \beta_j + \epsilon$
4. Update $\mathbf{r} = \mathbf{r} - \epsilon \mathbf{v}_j$. Repeat steps 2 and 3, until either a predefined constraint on the norm of $\boldsymbol{\beta}$ is violated, or the number of nonzero components of $\boldsymbol{\beta}$ equals $\min(n, d_v)$, or there are no columns with positive correlation left.

Note that for column-standardized $\mathbf{V}$ (i.e. every column has zero mean and unit variance), finding the column vector $\mathbf{v}_j$ most positively correlated with $\mathbf{r}$

$$j = \underset{i}{\operatorname{argmax}} \frac{\mathbf{v}_i^\top \mathbf{r}}{\|\mathbf{v}_i\|\|\mathbf{r}\|} = \underset{i}{\operatorname{argmax}} \left(\mathbf{v}_i^\top \mathbf{A}\widehat{\boldsymbol{\alpha}} - \mathbf{v}_i^\top \mathbf{V}\boldsymbol{\beta}\right) \tag{13}$$

is equal to finding the maximum element of the negative gradient of the least-squares problem. In other words, MIFSR is a gradient descent method where at every iteration, $\boldsymbol{\beta}$ is moved an ϵ-step along the coordinate axis where the least-squares error declines most.

3.3. Finding all CCA projections

For the source separation task, we are naturally interested in more than one projection, expecting that distinct sources are retrieved in different projections. We incorporate the constraints (9) on subsequent projection directions by means of an additional quadratic term in the regression function

$$\widehat{\boldsymbol{\beta}}_{(k)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{A}\widehat{\boldsymbol{\alpha}} - \mathbf{V}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \mathbf{O}\boldsymbol{\beta} \tag{14}$$

$$\text{s.t.} \quad \sum_{g=1}^{G} \|\boldsymbol{\beta}_{[g]}\| \leq t \wedge \beta_j \geq 0\ \forall j, \tag{15}$$

where

$$\mathbf{O} = \sum_{l<k} \mathbf{C}_{vv} \widehat{\boldsymbol{\beta}}_{(l)} \widehat{\boldsymbol{\beta}}_{(l)}^\top \mathbf{C}_{vv}. \tag{16}$$

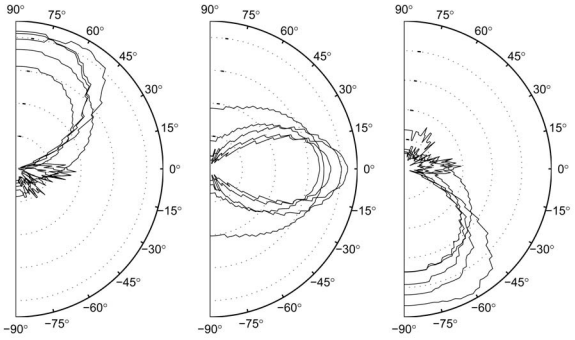

Fig. 1. Simulated directivity response of the beamforming architecture for target directions $[+60°, 0°, -60°]$. Plotted is the directional RMS energy for five bands of band-pass filtered white noise centered at 500Hz, 1000Hz, 1500Hz, 2000Hz and 2500Hz.

This is still a convex problem that can be solved efficiently by the kind of gradient descent we described in the last section.

4. EXPERIMENTS

4.1. Signal representation

We have worked with pixel based and interest point based representations of visual movement:

(i) intensity images were computed from the video frames by converting them to gray-scale, smoothing with a Gaussian mask (for noise reduction) and downsampling to the appropriate size. The images were then linearized to form the rows of $\mathbf{V}$, where each row corresponds to one time point in a sliding window. A pixel based representation is only sensible if the scenes are fairly static or the window size is small, so that a pixel corresponds to the same location on a source (e.g. the chin) over the whole time window. Intensity changes are due to local movement of a source only under such restrictive conditions. To overcome this constraint,

(ii) Shi-Tomasi interest points [11] were tracked using a sparse pyramidal Lucas-Kanade tracker [12]. We used both cartesian and polar coordinates as features. High-pass filtering of the trajectories rejects distracting global movement (e.g. a translation of the whole body) while retaining relevant local movement of lips, chin or hands. A median filter of neighborhood size three was used to obtain smoother trajectories. Here, the rows of $\mathbf{V}$ contain the coordinates of every interest point at each timestep.

On the audio side, we also examined two different representations:

(i) a frequency-domain representation is built from 50 frequency bands equally spaced in mel scale in the range 100 Hz - 8kHz. The signals in each of the bands were rec-

Fig. 2. Three frames from a movie showing one speaker and one moving person. The identified pixels are overlayed as white points in the frames, where intensity is proportional to the corresponding weight β_j. Nonnegative CCA clearly identified the audio source in the presence of uncorrelated movements.

tified, filtered and subsampled to the video frame rate of 25 Hz. Low-pass filtering before the subsampling process avoids aliasing effects in the sampled energy signals.

(ii) A spatial representation of the audio data was derived from a microphone array with a filter-and-sum beamformer. We used a linear array of 4 omni-directional microphones with 6cm spacing between the microphones. We trained a multi-channel Wiener filter with simulated pilot signals from fixed target and interferer directions. The directivity response for this architecture is plotted in figure 1. This very limited architecture provided sufficient separation for three equiangular target directions of the frontal hemisphere in the range of 500Hz to 3000Hz, and serves as proof of concept. Any state-of-the-art fixed beamformer architecture could be used instead for better directivity and more bandwidth. For every target direction, we again extracted sub-band energies equally spaced in mel scale to allow both spatial and spectral filtering. Alternatively, we also worked with a set of LSP coefficients and RMS energy (as motivated in [1]) for every target direction. In that case, the projection coefficients $\boldsymbol{\alpha}$ can be used in a voting scheme to identify the most likely direction of a source.

4.2. Identification in frequency domain

In a first experiment, we used the frequency domain representation and tested our method on a sequence where one person sits and speaks and another person moves around. Nonnegative CCA was performed on sliding windows of size 50 frames. We used L_1 regularization for video in order to identify single pixels, and L_2 regularization on the audio side. Three frames from this sequence are depicted in figure 2. Positive projection coefficients are highlighted as white circles centered at the corresponding pixel (the magnitude of β_j determines the brightness of the spot). One can clearly see that the method was able to discriminate between the sound source and uncorrelated movements.

Fig. 3. A frame from a movie showing the left speaker counting and the right reciting a poem. The identified interest points are marked with circles, squares and triangles, corresponding to the first, second and third projection. Filled markers with black outline are the weighted centroids of the projections.

Using the same audiovisual representation, we have also conducted experiments with two speakers (details omitted due to space restrictions). The first two projections successfully identified the speakers, i.e. coefficient weight for each projection concentrated on a single source. While these results show that nonnegative CCA performs well in finding distinct areas in the image which correspond to different audio sources, the reconstructed audio signals did not provide a good source separation. Such a separation, however, could probably not be expected by solely using frequency bands to represent the audio signal. On a time scale of several seconds (necessary at 25Hz frame rate), the frequency representation alone is no longer discriminative for separating concurrent speakers in audio.

4.3. Identification and separation in spatial domain

In a third experiment, we addressed this problem by using a spatial audio representation derived from a microphone array with beamforming. For each preferred direction of the beamformer, we extracted RMS energy and 10th order LSP coefficients. On the video side, we tracked 560 interest points. Sliding window size was ten seconds. Figure 3 shows the result for two speakers, the left counting numbers and the right reciting a poem. The positive coefficients of the first three projections are depicted as circles, squares and triangles, where the marker size is proportional to coefficient weight. Filled markers with black outline depict

Fig. 4. One person strumming the guitar and one person counting. The first projection correctly identifies the strumming hand. The second and third projections concentrate on the speaker.

the weighted centroids of the projections. The results show that the first two centroids identify the left talker, while the third concentrates on the right talker. There is also one nonzero coefficient from the second projection on the right person. This might be due to the fact that in this case, the second projection corresponds to the center direction of the beamformer, where there is no clear separation between the sources. There are also two more coefficients with small positive weights located in the background.

In a fourth experiment we show that our method makes no strong assumptions about the nature of the source, and can also correctly identify a strummed guitar (see figure 4). The input representations were identical to the third experiment, except for a shorter sliding window of six seconds length. Note that our method does not rely on any geometric information or camera calibration, and can cope with a significant amount of distortion. Furthermore, camera and beamformer position don't have to coincide.

Videos of all experiments can be found at `www.inf.ethz.ch/personal/chrsigg/mlsp2007`.

5. CONCLUSION

We have presented the nonnegative CCA method for jointly analyzing audio and video streams. This technique finds a series of orthogonal projections which successively decompose the signal into a series of additive components. We demonstrate in several experiments that concurrent sources (speech and non-speech) are correctly identified in video and separated in audio.

Although maximizing linear correlation suffices to separate two concurrent speakers, the necessary window size is on the order of several seconds. Clearly, humans not only make use of movement, but integrate other cues such as lip shape. Nonlinear regression could capture higher order dependencies between the two modalities, but reliable parameter estimation would be even more difficult in this setting where dimensionality is far larger than sample size. As another direction of further research, we might seek projections that not only maximize correlation, but also maximize a measure of non-Gaussianity, thus integrating concepts from independent component analysis into CCA.

6. REFERENCES

[1] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[2] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems 13*. 2001, MIT Press.

[3] J. Hershey and J. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*. 2000, pp. 813–819, MIT Press.

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

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

[6] R.J. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, vol. 58, pp. 267–288, 1996.

[7] Min Yuan and Yi Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc. B*, vol. 68, pp. 49–67, 2006.

[8] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, vol. 16, pp. 2639–2664, 2004.

[9] J. Vıa, I. Santamarıa, and J. Perez, "A robust RLS algorithm for adaptive Canonical Correlation Analysis," *Proceedings of ICASSP*, 2005.

[10] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther, "Forward Stagewise Regression and the Monotone Lasso," Available at www-stat.stanford.edu/˜hastie/Papers/.

[11] J. Shi and C. Tomasi, "Good features to track," *Proceedings of CVPR*, pp. 593–600, 1994.

[12] J.Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.