Probabalistic Models and Informative Subspaces for Audiovisual Correspondence

John W. Fisher III and Trevor Darrell

Massachusett Institute of Technology
Artficial Intelligence Laboratory
Cambridge, Massachusetts, USA

Abstract. We propose a probabalistic model of single source multi-modal generation and show how algorithms for maximizing mutual information can find the correspondences between components of each signal. We show how non-parametric techniques for finding informative subspaces can capture the complex statistical relationship between signals in different modalities. We extend a previous technique for finding informative subspaces to include new priors on the projection weights, yielding more robust results. Applied to human speakers, our model can find the relationship between audio speech and video of facial motion, and partially segment out background events in both channels. We present new results on the problem of audio-visual verification, and show how the audio and video of a speaker can be matched even when no prior model of the speaker's voice or appearance is available.

1 Introduction

Relating multi-modal signals is an important and challenging task for machine perception systems. For example, given audio and video signals, one would like to find the audio-visual correspondences: that portion of the signals which can be inferred to have come from the same underlying source. Applied to observation of human speakers, this ability can be useful for several different tasks. It can help localize the position of the speaker in the video frame (video localization), enhance the audio quality of the speaker by segmenting it from other sources (audio localization), and finally verify whether the person being observed is the person speaking (audio-visual verficiation).

We propose an independent cause model to capture the relationship between generated signals in each individual modalitiy. Using principles from information theory and nonparametric statistics we show how an approach for learning maximally informative joint subspaces can find cross-modal correspondences. We analyze the graphical model of multi-modal generation and show under what conditions related subcomponents of each signal have high mutual information.

Non-parametric statistical density models can be used to measure the degree of mutual information in complex phenomena [6] which we apply to audio/visual data. This technique simultaneously learns projections of images in the video sequence *and* projections of sequences of periodograms taken from the audio

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 592–603, 2002.

sequence. The projections are computed adaptively such that the video and audio projections have maximum mutual information (MI). Applied to audiovisual data, early results with this technique on video and audio localization have been reported [4], but without any derivation from a probabalistic framework. In practice, these techniques may fail to converge to a useful solution, since the projection parameters were underconstrained.

In this paper we ground the mutual information algorithm in probabalistic model, and extend the informative subspace algorithm to include a prior bias towards small projection coefficients. We also present new results on the problem of audio-visual verification without prior models of user speech or appearance, an application not previously addressed in the literature.

In the next section we review related work on audio-visual fusion and information theoretic adaptive methods. We present our probabalistic model for cross-modal signal generation, and show how audio-visual correspondences can be found by identifying components with maximal mutual information. We then review techniques for efficient estimation of mutual information using non-parametric entropy models. Finally, we show a new application to a verification task where we detect whether audio and video come from a single speaker. In an experiment comparing the audio and video of every combination of a group of eight users, our technique was able to perfecly match the corresponding audio and video from a single user. These results are based purely on the instantaneous cross-modal mutual information between the *projections* of the two signals, and do not rely on any prior experience or model of user's speech or appearance.

2 Related Work

Humans routinely perform tasks in which ambiguous auditory and visual data are combined in order to support accurate perception. In contrast, automated approaches for processing multi-modal data sources lag far behind. This is primarily due to the fact that few methods adequately model the complexity of the audio/visual relationship. Classical approaches to multi-modal fusion usually either assume a statistical relationship which is too simple (e.g. jointly Gaussian) or defer fusion to the decision level when many of the joint (and useful) properties have been lost. While such pragmatic choices may lead to simple statistical measures, they do so at the cost of modeling capacity.

Information theory motivates fusion at the measurement level without reagard to specific parametric densities. The idea of using information-theoretic principles in an adaptive framework is not new (e.g. see [3] for an overview) with many approaches suggested over the last 30 years. A critical distinction in most information theoretic approaches lies in how densities are modeled (either explicitly or implicitly), how entropy (and by extension mutual information) is approximated or estimated, and the types of mappings which are used (e.g. linear vs. nonlinear). Approaches which use a Gaussian assumption include Plumbley [12, 11] and Becker[1]. Additionally, [1] applies the method to fusion of artificially generated random dot sterograms.

With regards to audio/visual processing, there has been substantial progress on feature-level integration of speech and vision. For example, Meier *et al* [9], Stork [14] and others have built visual speech reading systems that can improve speech recognition results dramatically. However, many of these systems assume that no significant motion distractors are present and that the camera was "looking" at the user uttering the audio signal. Speech systems (both those that integrate viseme features and those that do not) are easily confused if there are nearby speakers also making utterances, either directed at the speech recognition system or not. Our system, described below, is designed to be able to detect and disambiguate cases where audio and video signals are coming from different sources. Since our method is not dependent on speech content, it would have the advantage of working on non-verbal utterances.

Other audio/visual work which is closely related to ours is that of Hershey and Movellan [7] which examined the per-pixel correlation relative to an audio track, detecting which pixels have related variation. Again, an inherent assumption of this method was that the joint statistics were Gaussian. Slaney and Covell [13] looked at optimizing temporal alignment between audio and video tracks using canonical correlations (equivalent to mutual information in the jointly Gaussian case), but did not address the problem of detecting whether two signals came from the same person or not.

The idea of simply gating audio input with a face detector is related to ours, but would not solve our target scenerio above where the user is facing the screen and a nearby person makes an utterance that can be mistakenly interpreted as a system command. We are not aware of any prior work which can perform audio-visual verification at a signal-level, without any prior speech or appearance model.

3 Probabalistic Models of Audio-Visual Fusion

We consider multimodal scenes which can be modeled probabalistically with one joint audio-visual source and distinct background interference sources for each modality. We note that the proposed method can be extended to multiple multimodal sources. Each observation is a combination of information from the joint source, and information from the background interferer for that channel. We use a graphical model, figure 1 to represent this relationship. In the diagrams, B represents the joint source, while A and C represent single modality backgroud interference. Our purpose here is to analyze under which conditions our methodology should uncover the underlying cause of our observations.

Figure 1a shows an independent cause model for our typical case, where $\{A, B, C\}$ are unobserved random variables representing the causes of our (high-dimensional) observations in each modality $\{X^a, X^v\}$. In general there may be more causes and more measurements, but this simple case can be used to illustrate our algorithm. An important aspect is that the measurements have dependence on only one common cause. The joint statistical model consistent

with the graph of figure 1a is

$$P(A, B, C, X^a, X^v) = P(A)P(B)P(C)P(X^a|A, B)P(X^v|B, C) \ .$$

Given the independent cause model a simple application of Bayes' rule (or

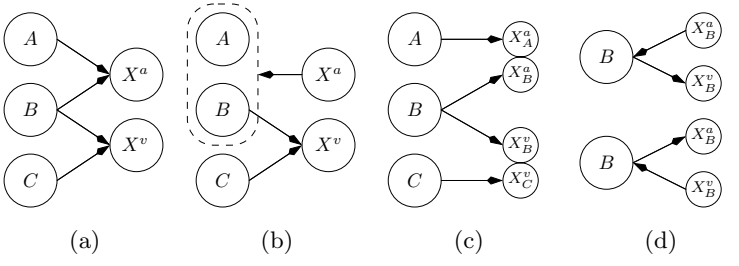

Fig. 1. Graphs illustrating the various statistical models exploited by the algorithm: (a) the independent cause model - X^a and X^v are independent of each other conditioned on $\{A, B, C\}$, (b) information about X^a contained in X^v is conveyed through *joint* statistics of A and B, (c) the graph implied by the existence of a separating function, and (d) two equivalent Markov chains which can be extracted from the graphs *if* the separating functions can be found.

the equivalent graphical manipulation) yields the graph of figure 1b which is consistent with

$$P(A, B, C, X^a, X^v) = P(X^a)P(C)P(A, B|X^a)P(X^v|B, C) \ ,$$

which shows that information about X^a contained in X^v is conveyed through the *joint* statistics of A and B. The consequence being that, in general, we cannot disambiguate the influences that A and B have on the measurements. A similar graph is obtained by conditioning on X^v. Suppose decompositions of the measurement X^a and X^v *exist* such that the following joint densities can be written:

$$P(A, B, C, X^a, X^v) = P(A)P(B)P(C)P(X^a_A|A)P(X^a_B|B)P(X^v_B|B)P(X^v_C|C)$$

where $X^a = [X^a_A, X^a_B]$ and $X^v = [X^v_B, X^v_C]$. An example for our specific application would be segmenting the video image (or filtering the audio signal). In this case we get the graph of figure 1c and from that graph we can extract the Markov chain which contains elements related only to B. Figure 1d shows equivalent graphs of the extracted Markov chain. As a consequence, there is no influence due to A or C.

Of course, we are still left with the formidable task of finding a decomposition, but given the decomposition it can be shown, using the data processing inequality [2], that the following inequality holds:

$$I(X^a_B, X^v_B) \leq I(X^a_B, B) \tag{1}$$
$$I(X^a_B, X^v_B) \leq I(X^v_B, B) \tag{2}$$

More importantly, these inequalities hold for functions of X_B^a and X_B^v (e.q. $Y^a = f(X^a; h_a)$ and $Y^v = f(X^v; h_v)$). Consequently, by maximizing the mutual information between $I(Y^a; Y^v)$ we must necessarily increase the mutual information between Y^a and B and Y^v and B. The implication is that fusion in such a manner discovers the underlying cause of the observations, that is, the joint density of $p(Y^a, Y^v)$ is strongly related to B. Furthermore, with an approximation, we can optimize this criterion without estimating the separating function directly. In the event that a perfect decomposition does not exist, it can be shown that the method will approach a "good" solution in the Kullback-Leibler sense.

4 Informative Subspaces from Linear Projections

From the perspective of information theory, estimating separate projections of the audio video measurements which have high mutual information makes intuitive sense as such features will be predictive of each other. The advantage is that the form of those statistics are not subject to the strong parametric assumptions (e.g. joint Gaussianity) which we wish to avoid.

We can find these projections using a technique that maximizes the mutual information between the projections of the two spaces. Following [4], we use a nonparametric model of joint density for which an analytic gradient of the mutual information with respect to projection parameters is available. First we review that technique, and then present new extensions to add a prior bias term on the projection coefficients. We have found that without this extension the technique may not converge to a useful subspace.

In principle the method may be applied to any function of the measurements, $Y = f(X; h)$, which is differentiable in the parameters h (e.g. as shown in [4]). Here we consider a linear fusion model which results in a significant computational savings at a minimal cost to the representational power (largely due the nonparametric density modeling of the output):

$$\begin{bmatrix} y_1^v \cdots y_N^v \\ y_1^a \cdots y_N^a \end{bmatrix} = \begin{bmatrix} h_v^T & 0^T \\ 0^T & h_a^T \end{bmatrix} \begin{bmatrix} x_1^v \cdots x_N^v \\ x_1^a \cdots x_N^a \end{bmatrix} \tag{3}$$

where $x_i^v \in \Re^{N_v}$ and $x_i^a \in \Re^{N_a}$ are lexicographic samples of images and periodograms, respectively, from an A/V sequence. The linear projection defined by $h_v^T \in \Re^{M_v \times N_v}$ and $h_a^T \in \Re^{M_a \times N_a}$ maps A/V samples to low dimensional features $y_i^v \in \Re^{M_v}$ and $y_i^a \in \Re^{M_a}$. Treating x_i's and y_i's as samples from a random variable our goal is to choose h_v and h_a to maximize the mutual information, $I(Y^a; Y^v))$, of the derived measurements.

Mutual information, which is a combination of entropy terms, is defined for continuous random variables as [2]

$$\begin{aligned} I\left(Y^v;Y^a\right) &= h\left(Y^a\right) + h\left(Y^v\right) - h\left(Y^a,Y^v\right) \\ &= \int_{R_{Y^a}} p_{Y^a}\left(y\right)\log\left(p_{Y^a}\left(y\right)\right)dy + \int_{R_{Y^v}} p_{Y^v}\left(y\right)\log\left(p_{Y^v}\left(y\right)\right)dy \\ &\quad - \int\int_{R_{Y^a}\times R_{Y^v}} p_{Y^a,Y^v}\left(x,y\right)\log\left(p_{Y^a,Y^v}\left(x,y\right)\right)dxdy \quad . \end{aligned} \tag{4}$$

Mutual information indicates the amount of information that one random variable conveys on average about another. The usual difficulty of MI as a criterion for adaptation is that it is an integral function of probability densities. Furthermore, in general we are not given the densities themselves, but samples from which they must be inferred. To overcome this problem, we replace *each* entropy term in equation 4 with a second-order taylor-series approximation as in [6]

$$\hat{I}\left(Y^v,Y^a\right) = \hat{H}(Y^a) + \hat{H}(Y^v) - \hat{H}(Y^v,Y^a) \tag{5}$$

$$\begin{aligned} &= \int_{R_{Y^a}} \left(\hat{p}_{Y^a}(y) - p_u(y)\right)^2 dy + \int_{R_{Y^v}} \left(\hat{p}_{Y^v}(y) - p_u(y)\right)^2 dy \\ &\quad - \int_{R_{Y^a}\times R_{Y^v}} \left(\hat{p}_{Y^v,Y^a}(x,y) - p_u(x,y)\right)^2 dxdy \end{aligned} \tag{6}$$

where R_{Y^a} is the support of one feature output, R_{Y^v} is the support of the other, p_u is the uniform density over that support, and $\hat{p}(x)$ is a Parzen density [10] estimate computed from the projected samples. The Parzen density estimate is defined as

$$\hat{p}\left(y\right) = \frac{1}{N}\sum_i \kappa\left(y - y_i, \sigma\right) \tag{7}$$

where $k\left(\ \right)$ is a gaussian kernel (in our case) and σ is the standard deviation. The Parzen density estimate has the capacity to capture relationships with more complex structure than typical parametric families of densities.

Note that this is essentially an integrated squared error comparison between the density of the projections to the uniform density (which has maximum entropy over a finite region). An advantage of this particular combination of second-order entropy approximation and nonparametric density estimator is that the gradient terms (appropriately combined to approximate mutual information as in 6) with respect to the projection coefficients can be computed *exactly* by evaluating a finite number of functions at a finite number of sample locations in the output space as shown in [5, 6]. The update term for the individual entropy terms in 6 (note the negative sign on the third term) of the ith feature vector at iteration k as a function of the value of the feature vector at iteration $k-1$ is (where y_i denotes a sample of either Y^a or Y^v or their concatenation depending on which term of 6 is being computed)

$$\Delta {y_i}^{(k)} = b_r({y_i}^{(k-1)}) - \frac{1}{N}\sum_{j\neq i} \kappa_a\left({y_i}^{(k-1)} - {y_j}^{(k-1)}, \Sigma\right) \tag{8}$$

$$b_r(y_i)_j \approx \frac{1}{d}\left(\kappa\left(y_i + \frac{d}{2}, \Sigma\right)_j - \kappa\left(y_i - \frac{d}{2}, \Sigma\right)_j\right) \tag{9}$$

$$\begin{aligned}\kappa_a(y, \Sigma) &= \kappa(y, \Sigma) * \kappa'(y, \Sigma) \\ &= -\left(2^{M+1}\pi^{M/2}\sigma^{M+2}\right)^{-1} \exp\left(-\frac{y^T y}{4\sigma^2}\right) y\end{aligned} \tag{10}$$

where $M = M_a$, M_v, or $M_a + M_v$ depending on the entropy term. Both $b_r(y_i)$ and $\kappa_a(y_i, \sigma)$ are vector-valued functions (M-dimensional) and d is the support of the output (i.e. a hyper-cube with volume d^M). The notation $b_r(y_i)_j$ indicates the jth element of $b_r(y_i)$. Adaptation consists of the update rule above followed by a modified least squares solution for h_v and h_a until a local maximum is reached. In the experiments that follow $M_v = M_a = 1$ with 150 to 300 iterations.

In [4] early results were demonstrated using this method for the video-based localization of a speaking user. However, the technique often failed to converge, since the projection coefficients were under-determined. To improve on the method, we thus introduce a capacity control mechanism in the form of a prior bias to small weights.

Capacity Control. The method of [6] requires that the projection be differentiable, which it is in this case. Additionally some form of capacity control is necessary as the method results in a system of underdetermined equations. To address this problem we impose an L_2 penalty on the projection coefficients of h_a and h_v. Furthermore, we impose the criterion that if we consider the projection h_v as a filter, it has low output energy when convolved with images in the sequence (on average). This constraint is the same as that proposed by Mahalanobis *et al* [8] for designing optimized correlators the difference being that in their case the projection output was designed explicitly while in our case it is derived from the MI optimization in the output space.

The adaptation criterion, which we maximize in practice, is then a combination of the approximation to MI (equation 5) and the regularization terms:

$$J = \hat{I}\left(Y^v, X^a\right) - \alpha_v {h_v}^T h_v - \alpha_a {h_a}^T h_a - \beta {h_v}^T \bar{R}_V^{-1} h_v \tag{11}$$

where the last term derives from the output energy constraint and $\bar{R}_V^{-1}$ is average autocorrelation function (taken over all images in the sequence). This term is more easily computed in the frequency domain (see [8]) and is equivalent to pre-whitening the images using the inverse of the average power spectrum. The scalar weighting terms α_v, α_u, β, were set using a data dependent heuristic for all experiments.

The interesting thing to note is that computing h_v can be decomposed into three stages:

1. Pre-whiten the images **once** (using the average spectrum of the images) followed by iterations of

2. Updating the feature values (y_i^v's), and
3. Solving for the projection coefficients using least squares and the L_2 penalty.

The pre-whitening interpretation makes intuitive sense in our case as it accentuates edges in the input image. It is the moving edges (lips, chin, etc.) which we expect to convey the most information about the audio. The projection coefficients related to the audio signal, h_a, are solved in a similar (and simultaneously) without the initial pre-whitening step.

5 Empirical Results

We now present new experimental results in which the general method described previously is used to first to localize the speaker in the video (i.e., audio-based video localization) and second to measure whether the audio signal is consistent with the video signal (i.e., audio-visual verification).

Our motivating scenerio for this application is a user interacting with an anonymous handheld device or kiosk using spoken commands. Given a recieved audio signal, we would like to verify whether the person speaking the command is in the field of view of the camera on the device, and if so to localize which person is speaking. If there are numerous handheld devices in an area, one would like them not all to respond to a command given to one of them. Simple techniques which check only for the presence of a face (or moving face) would fail when two people were looking at their individual devices and one spoke a command. Since future devices may be anonymous and interchangeable, we are interested in the case where no prior model of the voice or appearance of users are available to perform the verification and localization. The technique described in the previous sections is thus an appropriate approach.

We collected audio-video data from eight subjects. In all cases the video data was collected at 29.97 frames per second at a resolution of 360x240. The audio signal was collected at 48000 KHz, but only 10Khz of frequency content was used. All subjects were asked to utter the phrase "How's the weather in Taipei?". This typically yielded 2-2.5 seconds of data. Video frames were processed as is, while the audio signal was transformed to a series of periodograms. The window length of the periodogram was 2/29.97 seconds (i.e. spanning the width of two video frames). Upon estimating projections the mutual information between the projected audio and video data samples is used as the measure of consistency. All values for mutual information are in terms of the maximum possible value, which is the value obtained (in the limit) if the two variables are uniformly distributed and perfectly predict one another. In all cases we assume that there is not significant head movement on the part of the speaker during the utterance of the sentence. While this assumption might be violated in practice one might account for head movement using a tracking algorithm, in which case the algorithm as described would process the images after tracking.

5.1 Video Localization of Speaker

Figure 2a shows a single video frame from one sequence of data. In the figure there is a single speaker and a video monitor. Thoughout the sequence the video monitor exhibits significant flicker. Figure 2c shows an image of the pixel-wise standard deviations of the image sequence. As can be seen, the energy associated with changes due to monitor flicker is greater than that due to the speaker. Figure 2b shows the absolute value of the output of the pre-whitening stage for the video frame in the same figure. Note that the output we use is signed. The absolute value is shown instead because it illustrates the enhancements of edges in the image.

Figure 4a shows the associated periodogram sequence where the horizontal axis is time and the vertical axis is frequency (0-10 Khz). Figure 2d shows the coefficients of the learned projection when fused with the audio signal. As can be seen the projection highlights the region about the speaker's lips. Figure 3a

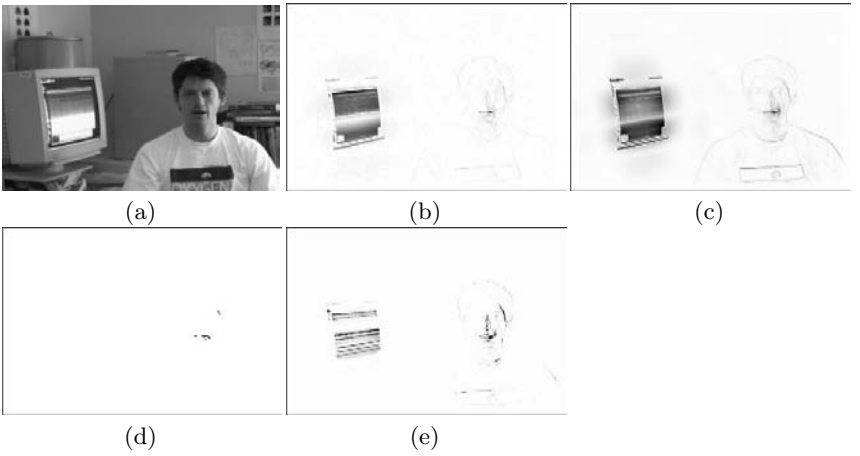

Fig. 2. Video sequence contains one speaker and monitor which is flickering: (a) one image from the sequence, (b) magnitude of the image after pre-whitening, (c) pixel-wise image of standard deviations taken over the entire sequence, (d) image of the learned projection, h_v, (e) image of h_v for incorrect audio

shows results from another sequence in which there are two people. The person on the left was asked to utter the test phrase, while the person on the right moved their lips, but did not speak. This sequence is interesting in that a simple face detector would not be sufficient to disambiguate the audio and video stream. Furthermore, viseme based approaches might be confused by the presence of two faces. Figures 3b and 3c show the pre-whitened images as before. There are significant changes about both subjects lips. Figure 3d shows the coefficients of the learned projection when the video is fused with the audio and again the region about the correct speaker's lips is highlighted.

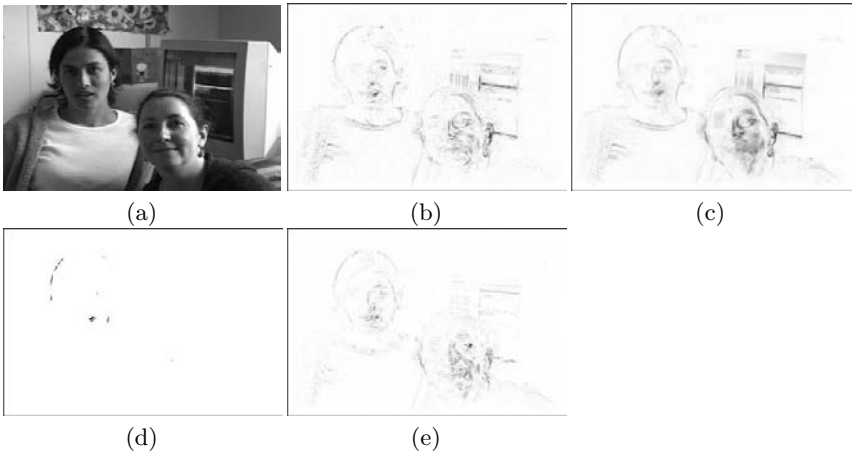

Fig. 3. Video sequence containing one speaker (person on left) and one person who is randomly moving their mouth/head (but not speaking): (a) one image from the sequence, (b) magnitude of the image after pre-whitening, (c) pixel-wise image of standard deviations taken over the entire sequence, (d) image of the learned projection, h_v, (e) image of h_v for incorrect audio.

5.2 Quantifying Consistency between the Audio and Video

In addition to localizing the audio source in the image sequence we can also check for consistency between the audio and video. Such a test is useful in the case that the person to which a system is visually attending is not the person who actually spoke. Having learned a projection which optimizes MI in the output feature space, we can then estimate the resulting MI and use that estimate to quantify the audio/video consistency.

Using the sequences of figure 2 and 3 we compared the fusion result when using separately recorded audio sequence from another speaker. The periodogram of the alternate audio sequence is shown in figure 4b. Figures 2e and 3e show the resulting h_v when the alternate audio sequence is used. In the case that the alternate audio was used we see that coefficients related to the video monitor increase significantly in 4e while energy is distributed throughout the image of 3e. For figure 2 the estimate of mutual information was 0.68 relative to the maximum possible value for the correct audio sequence. In contrast when compared to the periodogram of 4b, the value drops to 0.08 of maximum. For the sequence of figure 3, the estimate of mutual information for the correct sequence was 0.61 relative to maximum, while it drops to 0.27 when the alternate audio is used.

5.3 Eight-Way Test

Finally, data was collected from six additional subjects. These data were used to perform an eight-way test. Each video sequence was compared to each audio sequence. No attempt was made to optimally align the mismatched audio

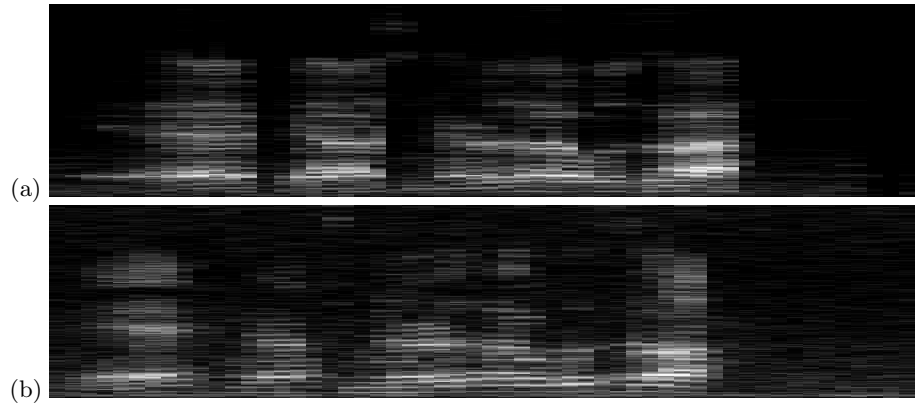

Fig. 4. Gray scale magnitude of audio periodagrams. Frequency increases from bottom to top, while time is from left to right. (a) audio signal for image sequence of figure 2. (b) alternate audio signal recorded from different subject.

sequences. Table 1 summarizes the results. The previous sequences correspond to subjects 1 and 2 in the table. In every case the matching audio/video pairs exhibited the highest mutual information after estimating the projections.

Table 1. Summary of results over eight video sequences. The columns indicate which audio sequence was used while the rows indicate which video sequence was used. In all cases the correct audio/video pair have the highest relative MI score.

	a1	a2	a3	a4	a5	a6	a7	a8
v1	**0.68**	0.19	0.12	0.05	0.19	0.11	0.12	0.05
v2	0.20	**0.61**	0.10	0.11	0.05	0.05	0.18	0.32
v3	0.05	0.27	**0.55**	0.05	0.05	0.05	0.05	0.05
v4	0.12	0.24	0.32	**0.55**	0.22	0.05	0.05	0.10
v5	0.17	0.05	0.05	0.05	**0.55**	0.05	0.20	0.09
v6	0.20	0.05	0.05	0.13	0.14	**0.58**	0.05	0.07
v7	0.18	0.15	0.07	0.05	0.05	0.05	**0.64**	0.26
v8	0.13	0.05	0.10	0.05	0.31	0.16	0.12	**0.69**

6 Conclusions and Future Work

We have presented an information theoretic approach to the problem of finding cross-modal correspondence. A new probabalistic formulation of joint signal generation was proposed, and we showed how maximizing mutual information could find desired correspondences. Informative subspaces can be found with non-parametric density models and second order approximations to mutual information. To find these robustly, we proposed new prior terms on projection

coeficients. Our approach was applied to the problem of audio-visual localization and verification, detecting the correspondence between the speech and appearance of a human speaker without having any prior model of that user in either domain. In all the cases that we tried, our technique was able to correctly pair the video with the corresponding audio from a particular individual, and localize where in a video frame the user's face was present.

References

1. Suzanna Becker. *An Information-theoretic Unsupervised Learning Algorithm for Neural Networks*. PhD thesis, University of Toronto, 1992.
2. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.
3. G. Deco and D. Obradovic. *An Information Theoretic Approach to Neural Computing*. Springer-Verlag, New York, 1996.
4. John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems 13*, 2000.
5. John W. Fisher III and Jose C. Principe. Entropy manipulation of arbitrary nonlinear mappings. In J.C. Principe, editor, *Proc. IEEE Workshop, Neural Networks for Signal Processing VII*, pages 14–23, 1997.
6. John W. Fisher III and Jose C. Principe. A methodology for information theoretic feature extraction. In A. Stuberud, editor, *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1998.
7. John Hershey and Javier Movellan. Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, and K-R. Mller, editors, *Advances in Neural Information Processing Systems 12*, pages 813–819, 1999.
8. A. Mahalanobis, B. Kumar, and D. Casasent. Minimum average correlation energy filters. *Applied Optics*, 26(17):3633–3640, 1987.
9. Uwe Meier, Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Towards unrestricted lipreading. In *Second International Conference on Multimodal Interfaces (ICMI99)*, 1999.
10. E. Parzen. On estimation of a probability density function and mode. *Ann. of Math Stats.*, 33:1065–1076, 1962.
11. M. Plumbley. On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR. 78, Cambridge University Engineering Department, UK, 1991.
12. M. Plumbley and S Fallside. An information-theoretic approach to unsupervised connectionist models. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of the 1988 Connectionists Models Summer School*, pages 239–245. Morgan Kaufman, San Mateo, CA, 1988.
13. Malcolm Slaney and Michele Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In T. K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, 2000.
14. G. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In *Proc. of Neural Information Proc. Sys. NIPS-6*, pages 1027–1034, 1994.