Signal Level Fusion for Multimodal Perceptual User Interface

John W. Fisher III
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
fisher@ai.mit.edu

Trevor Darrell
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
trevor@ai.mit.edu

ABSTRACT

Multi-modal fusion is an important, yet challenging task for perceptual user interfaces. Humans routinely perform complex and simple tasks in which ambiguous auditory and visual data are combined in order to support accurate perception. By 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. We present an information theoretic approach for fusion of multiple modalities. Furthermore we discuss a statistical model for which our approach to fusion is justified. We present empirical results demonstrating audio-video localization and consistency measurement. We show examples determining where a speaker is within a scene, and whether they are producing the specified audio stream.

Keywords

multimodal fusion, information theory, nonparametric statistics

1. INTRODUCTION

Multi-modal fusion is an important, yet challenging task for perceptual user interfaces. Humans routinely perform complex and simple 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 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.

PUI 2001 Orlando, FL USA

We discuss a nonparametric statistical approach to fusion which jointly models audio-visual phenomena. Using principles from information theory we show an approach for learning maximally informative joint subspaces for multi-modal fusion. Specifically, we simultaneously learn projections of images in the video sequence *and* projections of sequences of periodograms taken from the audio sequence. The projections are computed adaptively such that the video and audio projections have maximum mutual information (MI). The approach uses the methodology presented in [2, 6, 4] which formulates a learning approach by which the entropy, and by extension the MI, of a differentiable map may be optimized. We also discuss a statistical model for which the approach can be shown to be optimal.

Combining audio and video signals for dialog interface applications is an important goal for perceptual user interfaces. There has been substantial progress on feature-level integration of speech and vision. However, many of these systems assume that no significant motion distractors are present and that the camera was "looking" at the user who was uttering the audio signal.

Indeed, 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. If a second person says "shut down" near a voice-enabled workstation, the primary user may not be pleased with the result! In general, it is clear that multi-modal cues can aid the segmentation of multiple speakers into separate channels (e.g. the "cocktail party" effect)..

In this paper we show how signal level fusion of audio and video data using nonparametric models can capture useful joint structure. Specifically, we show results on two tasks, one localizing a speaker in a video stream, and the second measuring audio/video consistency–whether the audio and video came from the same source.

1.1 Related Work

As mentioned above, there has been much work on feature level audio-visual speech recognition. For example, Meier *et al* [9] and Stork [12] (and others) have built visual speech reading systems that can improve speech recognition results dramatically. It is not clear whether these systems could be used to localize the speaker as they implicitly rely on localization having already been performed. In theory, these systems could be modified to verify if the sequence of observed visemes was consistent with the detected phonemes.

We are not aware of a system which has been reported to do this to date, though it may be a successful approach. The method we will present works at a pre-feature level and does not presume detection of phonemes or visemes, so it may be advantageous in cases where a person-independent viseme model is hard to obtain. Also, since our method is not dependent on speech content, it would have the advantage of working on non-verbal utterances.

Other work which is more 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. An inherent assumption of this method was that the joint statistics were gaussian. Slaney and Covell [11] looked at optimizing temporal alignment between audio and video tracks, but did not address the problem of detecting whether two signals came from the same person or not. Their technique was more general than [7] in that pixels changes were considered jointly, although there is also an implicit Gaussian assumption. Furthermore, this technique makes use of training data.

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 primary 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 in perceptual user interfaces which addresses signal-processing level estimators to do both video localization and classify audio-visual synchrony among individuals.

2. INFORMATIVE SUBSPACES

We now give a brief description of our information theoretic fusion approach. While the algorithm has been described in previous work [3], that discussion focused primarily on the information theoretic intuition which motivated the method. In this section we also present a statistical model from which the method can be derived and the conditions under which our fusion approach is optimal.

2.1 Information Theoretic Fusion

Figure 1 illustrates our audio/visual fusion approach. Each image in the measured video sequence is treated as a single sample of a high-dimensional random variable (i.e. the dimension equals the number of pixels) . We denote ith image as V_i. The audio signal is converted to a sequence of periodograms (i.e. magnitude of windowed FFTs). Peridograms are computed at the video frame rate using a window equal to twice the frame period. Similarly to the video sequence, each periodogram "frame" is also treated as a sample of a high dimensional random variable (whose dimension is equal to the number of frequency bins) and whose ith frame is denoted U_i.

Using the approach described in [3] we learn projections of the audio/video frames, denoted,

$$f_{V\,i} = h_V^T V_i \quad (1)$$

$$f_{U\,i} = h_U^T U_i \quad (2)$$

resulting in samples of low-dimensional features $f_{V\,i}$ and $f_{U\,i}$ (whose dimensionality is determined by the matrices h_V and h_U, respectively). The criterion for learning the projection vectors, h_V and h_U, is to maximize the MI between the resulting audio and video features$f_{V\,i}$ and $f_{U\,i}$.

Mutual information (in the case of continuous features) is defined as [1]

$$\begin{aligned} I\left(f_V, f_U\right) &= h\left(f_U\right) + h\left(f_V\right) - h\left(f_U, f_V\right) \quad (3)\\ &= \int_{R_U} p_{f_U}\left(x\right) \log\left(p_{f_U}\left(x\right)\right) dx + \\ &\quad \int_{R_U} p_{f_V}\left(x\right) \log\left(p_{f_V}\left(x\right)\right) dx - \\ &\quad \int\int_{R_U \times R_V} p_{f_U, f_V}\left(x, y\right) \log\left(p_{f_U, f_V}\left(x, y\right)\right) dxdy \end{aligned}$$

The 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. Consequently, we replace equation 3 with the approximation of [6]

$$\begin{aligned} \hat{I}\left(f_V, f_U\right) &= \hat{H}(f_U) + \hat{H}(f_V) - \hat{H}(f_V, f_U) \quad (4)\\ &= \int_{R_U} \left(\hat{p}_{f_U}(x) - p_u(x)\right)^2 dx \\ &\quad + \int_{R_V} \left(\hat{p}_{f_V}(x) - p_u(x)\right)^2 dx \\ &\quad - \int_{R_U \times R_V} \left(\hat{p}_{f_V, f_U}(x, y) - p_u(x, y)\right)^2 dxdy \end{aligned}$$

where R_U is the support of one feature output, R_V is the support of the other, p_u is the uniform density over that support, and $\hat{p}(x)$ is the Parzen density [10] estimate computed from the projected samples:

$$\hat{p}\left(x\right) = \frac{1}{N} \sum_i \kappa\left(x - x_i, \sigma\right) \quad (5)$$

where $k(\,)$ is a gaussian kernel in our case and σ is the standard deviation.

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). The consequence of using this approximation is that its gradient 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 4 of the ith feature vector at iteration k as a function of the value of the feature vector at iteration $k-1$ is (where f_i denotes a sample of either f_U or f_V or their concatenation depending on which term of 4 is being computed)

$$\begin{aligned} \Delta {f_i}^{(k)} &= b_r({f_i}^{(k-1)}) - \\ &\quad \frac{1}{N} \sum_{j \neq i} \kappa_a\left({f_i}^{(k-1)} - {f_j}^{(k-1)}, \sigma\right) \quad (6)\\ b_r(f_i)_l &\approx \frac{1}{d^M} \prod_{j \neq l} \left(\kappa_1\left(f_{ij} + \frac{d}{2}, h\right) - \right. \\ &\quad \left. \kappa_1\left(f_{ij} - \frac{d}{2}, h\right)\right) \quad (7)\\ \kappa_a\left(f_i, \sigma\right) &= \kappa(f_i, \sigma) * \kappa'(f_i, \sigma) \quad (8)\\ &= -\frac{\exp\left(-\frac{{f_i}^T f_i}{4h^2}\right)}{\left(2^{M+1} \pi^{M/2} h^{M+2}\right)} f_i \quad (9) \end{aligned}$$

where M is the dimensionality of the feature vector f_i. Both $b_r(f_i)$ and $\kappa_a(f_i, \sigma)$ are M-dimensional vector-valued func-

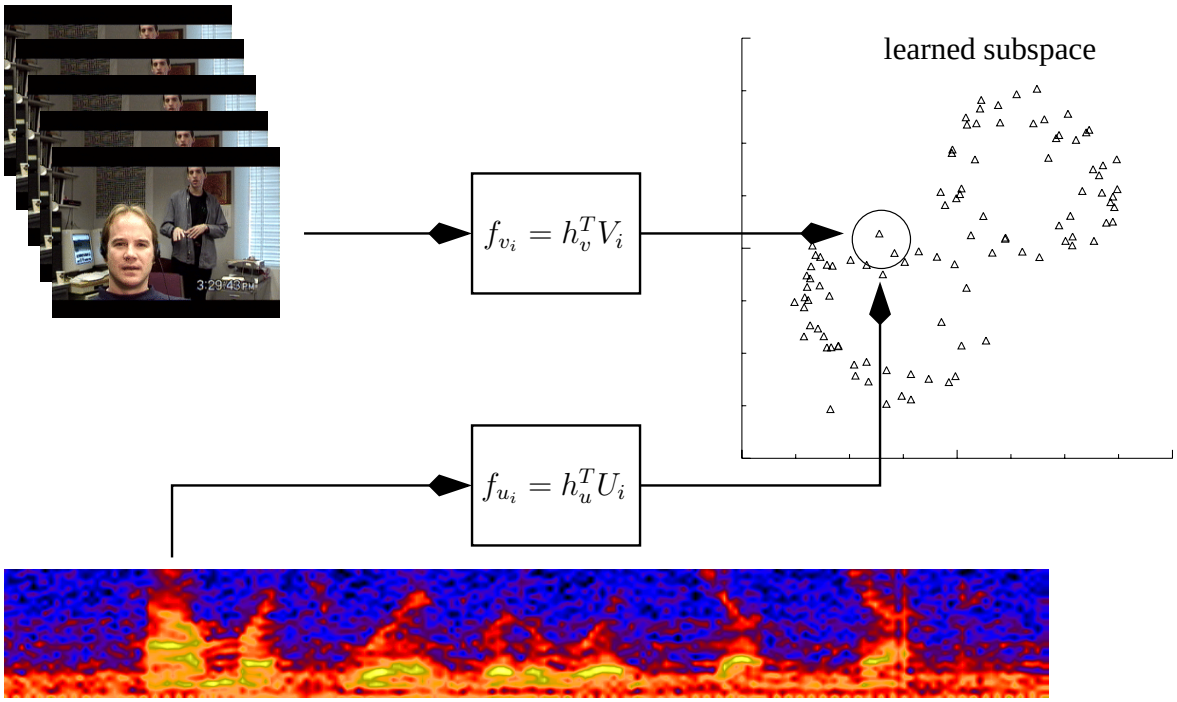

Figure 1: Maximally Informative Joint Subspace

tions and d is the support of the output of the mapping (i.e. a hyper-cube with sides of length d centered at the origin). The notation $b_r(y_i)_l$ indicates the lth element of $b_r(f_i)$ [6].

The process is repeated iteratively until a local maximum is reached using the update rule above. In the experiments that follow the dimensionality of f_U and f_V are set to unity while the number iterations is typically 150 to 300 iterations.

2.1.1 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. In practice we impose an L_2 penalty on the projection coefficients of h_U 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 4) and the regularization terms:

$$J = \hat{I}\left(f_V, f_U\right) - \alpha_v h_V^T h_V - \alpha_u h_U^T h_U - \beta h_V \bar{R}_V^{-1} h_V \quad (10)$$

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, 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_U, are solved in a similar (and simultaneously) without the initial pre-whitening step.

2.2 The Implicit Statistical Model

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

However, we now show that there is a statistical model for which such fusion is optimal. Consider the graphical models shown in figure 2. Figure 2a shows an independent cause model, where $\{A, B, C\}$ are unobserved random variables representing the causes of our (high-dimensional) observations $\{U, V\}$. 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 2a is

$$P(A,B,C,U,V) = P(A)P(B)P(C)P(U|A,B)P(V,B,C) \ .$$

Given the independent cause model a simple application of

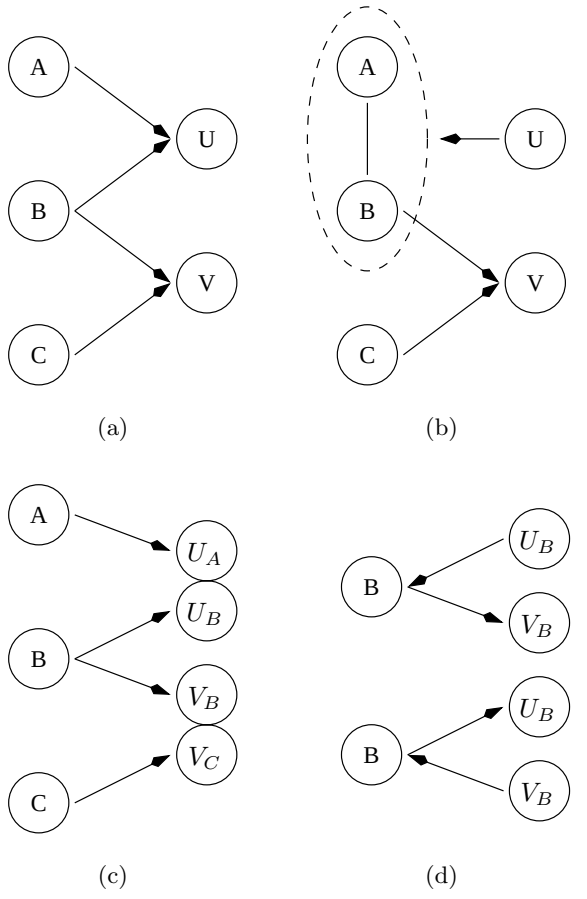

Figure 2: Graphs illustrating the various statistical models exploited by the algorithm: (a) the independent cause model - U and V are independent of each other conditioned on $\{A, B, C\}$, (b) information about U contained in V is conveyed through *joint* statistics of A and B, (c) the graph implied 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.

Bayes' rule (or the equivalent graphical manipulation) yields the graph of figure 2b which is consistent with

$$P(A,B,C,U,V) = P(U)P(C)P(A,B|U)P(V|B,C) \ ,$$

which shows that information about U contained in 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 V. Suppose decompositions of the measurement U and V *exist* such that the following joint densities can be written:

$$\begin{aligned} P(A,B,C,U,V) &= P(A)P(B)P(C)P(U_A|A)P(U_B|B) \\ &\quad P(V_B|B)P(V_C|C) \end{aligned}$$

where $U = [U_A, U_B]$ and $V = [V_B, V_C]$. 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 2c and from that graph we can extract the Markov chain which contains elements related only to B. Figure 2d 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 [1], that the following inequality holds:

$$I(f_U, f_V) \leq I(f_U, B) \quad (11)$$

$$I(f_U, f_V) \leq I(f_V, B) \quad (12)$$

So, by maximizing the mutual information between $I(f_U, f_V)$ we must necessarily increase the mutual information between f_U and B and f_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 $P(f_U, f_V)$ 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.

3. EMPIRICAL RESULTS

We now present experimental results in which the general method described previously is used to first to localize the speaker in the video and second to measure whether the audio signal is consistent with the video signal. 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. 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.

3.1 Video Localization of Speaker

Figure 3a 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 3c 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 3b 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

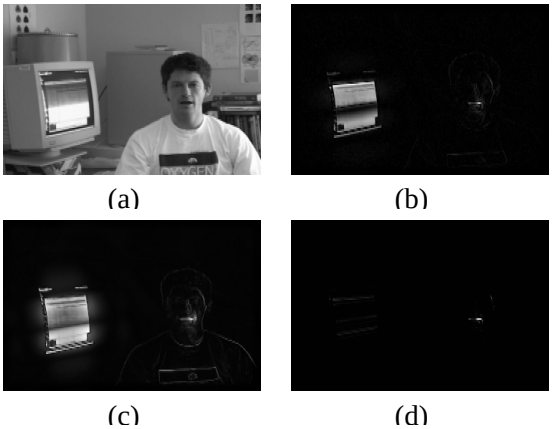

Figure 3: Video sequence contains one speaker and monitor which is flickering: (a) one image from the sequence, (b) magnitude of the image after pre-whitening, (c) pixel-wise image of standard deviations taken over the entire sequence, (d) image of the learned projection, h_V.

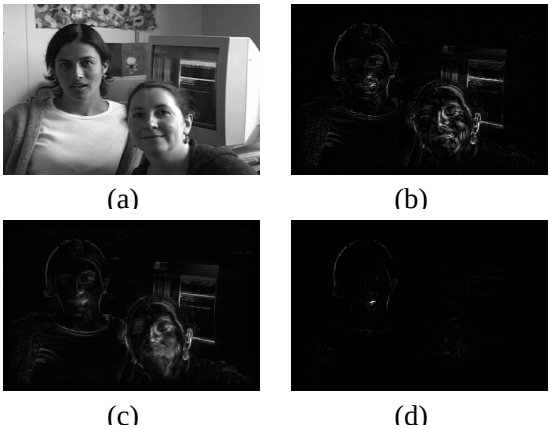

Figure 4: Video sequence containing one speaker (person on left) and one person who is randomly moving their mouth/head (but not speaking): (a) one image from the sequence, (b) magnitude of the image after pre-whitening, (c) pixel-wise image of standard deviations taken over the entire sequence, (d) image of the learned projection, h_V.

use is signed. The absolute value is shown instead because it illustrates the enhancements of edges in the image.

Figure 5a shows the associated periodogram sequence where the horizontal axis is time and the vertical axis is frequency (0-10 Khz). Figure 3d 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 4a 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 4b and 4c show the pre-whitened images as before. There are significant changes about both subjects lips. Figure 4d 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.

3.2 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 sequence of figure 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 5b. Figure 6a (correct audio) and 6b (alternate audio) compares the resulting projections h_v. In the case that the alternate audio was used we see that coefficients related to the video monitor increase significantly.

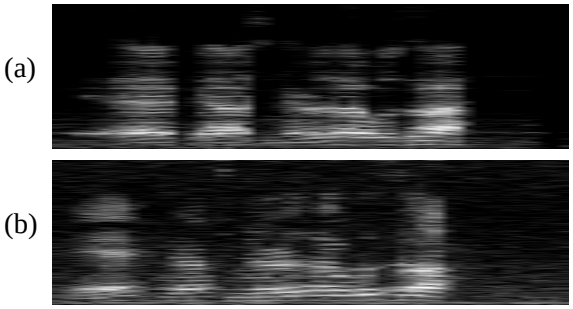

Figure 5: Gray scale magnitude of audio periodagrams. Frequency increases from bottom to top, while time is from left to right. (a) audio signal for image sequence of figure 3. (b) alternate audio signal recorded from different subject.

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 5b , the value drops to 0.08 of maximum. We repeat the same experiment with two speaker video sequence, shown in figure 6c (correct audio) and 6d (alternate audio) and again we see, not surprisingly, the speaker is not localized. The estimate of mutual information for this correct sequence was 0.61 relative to maximum, while it drops to 0.27 when the alternate audio is used.

3.3 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 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.

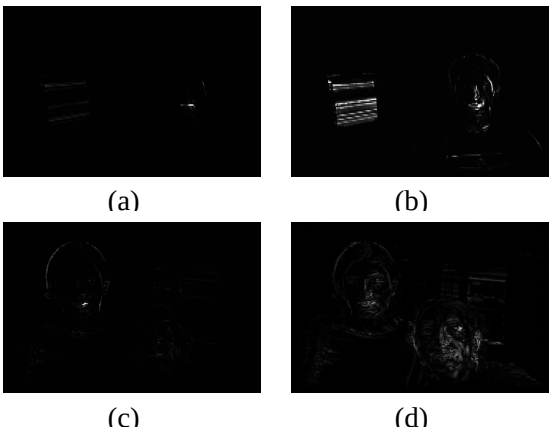

Figure 6: Comparison of learned video projections when correct (left) and incorrect (right)(audio is compared to image sequences of figure 3 (top) and figure 4 (bottom). When correct audio is used energy is concentrated on subject, when incorrect audio is used it is distributed throughout the image.

	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

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.

4. DISCUSSION AND FUTURE WORK

We have presented a method for information theoretic fusion of audio and video data. We have demonstrated over a small set of data, that the method shows promise for detecting audio-video consistency. We are not aware of equivalent results in the literature, although previous multi-modal methods might also work for this application. However, in contrast to previous approaches our method does not make strong assumptions about the underlying joint properties of the modalities being fused (e.g. Gaussian statistics). Consequently, it has the capacity to represent more complex structure which may be present in the data. Furthermore, our method makes no use of training data. While there is an adaptive element to the method, the adaptation occurs in an online fashion over a short sequence (approximately 2-2.5 seconds) of audio-video data. Consequently, the method is applicable when a prior model cannot be trained. As might happen when a multi-modal interface is moved to a new environment. Future work will address the robustness of the method over a larger corpus of data. Another area of interest is to determine the relationship between camera resolution, audio signal-to-noise ratio, and sampling rates for which the method maintains reliability.

5. ACKNOWLEDGEMENTS

This research was supported by MIT Project Oxygen.

6. REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[2] J. Fisher and J. Principe. Unsupervised learning for nonlinear synthetic discriminant functions. In D. Casasent and T. Chao, editors, *Proc. SPIE, Optical Pattern Recognition VII*, volume 2752, pages 2–13, 1996.

[3] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems 13*, 2000.

[4] J. W. Fisher III, A. T. Ihler, and P. A. Viola. Learning informative statistics: A nonparametric approach. In S. A. Solla, T. K. Leen, and K.-R. Mller, editors, *Advances in Neural Information Processing Systems 12*, 1999.

[5] J. W. Fisher III and J. C. Principe. Entropy manipulation of arbitrary nonlinear mappings. In J. Principe, editor, *Proc. IEEE Workshop, Neural Networks for Signal Processing VII*, pages 14–23, 1997.

[6] J. W. Fisher III and J. 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] J. Hershey and J. 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] U. Meier, R. Stiefelhagen, J. Yang, and A. 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. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, 2000.

[12] 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.