Speaker Association With Signal-Level Audiovisual Fusion

John W. Fisher, III, *Member, IEEE,* and Trevor Darrell, *Member, IEEE*

***Abstract*—Audio and visual signals arriving from a common source are detected using a signal-level fusion technique. A probabilistic multimodal generation model is introduced and used to derive an information theoretic measure of cross-modal correspondence. Nonparametric statistical density modeling techniques can characterize the mutual information between signals from different domains. By comparing the mutual information between different pairs of signals, it is possible to identify which person is speaking a given utterance and discount errant motion or audio from other utterances or nonspeech events.**

***Index Terms*—Audiovisual correspondence, multimodal data association, mutual information.**

I. Introduction

CONVERSATIONAL dialog systems have become practically useful in many application domains, including travel reservations, traffic information, and database access. However most existing conversational speech systems require *tethered* interaction, and work primarily for a single user. Users must wear an attached microphone or speak into a telephone handset, and do so one at a time. This limits the range of use of dialog systems, since in many applications users might expect to freely approach and interact with a device. Worse, they may wish to arrive as a group, and talk among themselves while interacting with the system. To date it has been difficult for speech recognition systems to handle such conditions, and correctly recognize the utterances intended for the device. We are interested facilitating *untethered* and casual conversational interaction, and address the problem of how to temporally segregate the speech of multiple users interacting with a system.

With a single modality, properly associating speech from multiple unknown speakers is quite difficult. However, if other modalities are available they can often provide disambiguating information. In particular, visual information can be valuable for deciding whether an individual user is speaking a particular utterance. We wish to solve a conversational audiovisual correspondence problem: given sets of audio visual signals, decide which audiovisual pairs are consistent and could have come from a single speaker. We approach this problem from a signal-processing perspective, and develop a statistical measure of whether two signals come from a common source. We make no assumptions about the content of the audio signal or the visual appearance, and use a general information-theoretic approach. Our method works without learning a specific lip or language model, and is therefore robust to a range of appearances and acoustic environments.

The core of our approach is a technique for jointly modeling audio and video variation to identify cross-modal correspondences. It is driven by the simple hypothesis of whether a region of interest in an image sequence (perhaps the entire image) is associated with a separately measured audio signal. We formulate the problem within a nonparametric hypothesis testing framework, from which information theoretic quantities naturally arise as the measure of association between two high-dimensional signals. We show how this approach can detect which user is speaking when several are facing a device and distracting motion is present. This allows the segregation of users' utterances from each other's speech, and from background noise events.

II. 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 statistical processing of multimodal 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 multimodal fusion at a signal processing level often 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.

An information theoretic approach motivates fusion at the measurement level without regard to specific parametric densities. The idea of using information-theoretic principles in an adaptive framework is not new (e.g., see [1] for an overview) with many approaches suggested over the last 30 years. Critical distinctions in most information theoretic approaches lie 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 versus nonlinear). Early approaches used a Gaussian assumption, e.g., Plumbley [2], [3] and Becker [4].

There has been substantial progress on feature-level integration of speech and vision. For example, Meier *et al.* [5], Stork [6], and others have built visual speech reading systems that can improve speech recognition results dramatically. Our goal is not recognition, but to be able to detect and disambiguate cases where audio and video signals are coming from different

Manuscript received December 1, 2002; revised November 15, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jun Ohya.

The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: trevor@ai.mit.edu).

Digital Object Identifier 10.1109/TMM.2004.827503

sources. Hershey and Movellan [7] addressed this problem using the per-pixel correlation relative to the energy of an audio track as a measure of their dependence. An inherent assumption of this method was that the joint statistics were Gaussian. As this is a per-pixel measure there is no straightforward way to integrate the measure over an image region for purposes of *association* without making simplifying assumptions which will not hold in practice (e.g., pixels are independent of each other conditioned on the speech signal). We should note that the objective of their work was to *locate* the source of an audio signal in an image sequence, association is an implicit step. A more general approach was taken by Slaney and Covell [8] which looked specifically at optimizing temporal alignment between audio and video tracks using canonical correlations which is equivalent to the maximum mutual information projection in the jointly Gaussian case. They did not address the problem of detecting whether two signals came from the same person, although their method could be adapted to do so. Nock *et al.* [9] consider two mutual information approaches and one HMM based approach for assessing face and speech consistency. The mutual information approaches compare a histogram based estimate over vector quantized codebooks to a Gaussian estimate over feature vectors. They report that the Gaussian method gave superior results when using a cepstral representation of the audio and a discrete cosine transform representation of the video. All three methods utilize a training corpus in order estimate a prior model, thereafter associations and/or likelihoods are computed under the trained model. A time-delay neural network approach was suggested in [10] demonstrating location detection for a single visual appearance on a small test set. Each of [8]–[10] require training data in order to estimate model parameters. Here, and in contrast to the previous methods, we develop a methodology for testing audio–video association in the *absence* of either a prior model and without the requirement of training data with which to construct one.

III. Signal-Level Audiovisual Association

We propose an independent cause model to capture the relationship between generated signals in each individual modality. 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 first show how audiovisual association problem can be formulated as a hypothesis test and giving a relationship to mutual information based association methods (see [11] for an extensive treatment). Following that we present an information theoretic analysis of a graphical model of multimodal signal generation which gives some incite on the relationship between data association and learning a generative audiovisual model.

Given an audio–video sequence, let us denote the sequence of N images (or a region within each image) as x_t^v where t indicates (discrete) time. Similarly denote audio measurements as x_t^a. For our purposes, x_t^a will be vectors of spectral measurments. Treating the audio and video measurements as *i.i.d.* samples from the random variables X^a and X^v, respectively, allows us to cast the audiovisual association problem as a simple hypothesis test:

$$\begin{aligned} H_0&: x_t^v, x_t^a \sim p(x^v)\,p(x^a) \\ H_1&: x_t^v, x_t^a \sim p(x^v, x^a) \end{aligned} \tag{1}$$

where H_0 states that the measurements are statistically *independent* (i.e., their joint density is expressed as a product of marginal densities) and H_1 states that the measurements are statistically *dependent* (or equivalently associated). Perceptual grouping problems, in which there are multiple sources of both video and audio can be stated in a similar, albeit more complicated, fashion [12], [13]. Plugging the measurements into a (normalized) log-likelihood ratio statistic, using a consistent probability density estimator for $p(x^v)$, $p(x^a)$, $p(x^v, x^a)$, and taking the expectation with respect to the joint probability density of X^a and X^v yields

$$\begin{aligned} &E\left\{\frac{1}{N}\sum_{t=0}^{N-1}\log\left(\frac{p(x_t^v, x_t^a|H_1)}{p(x_t^v, x_t^a|H_0)}\right)\right\} \\ &\quad = E\left\{\frac{1}{N}\sum_{t=0}^{N-1}\log\left(\frac{p(x_t^v, x_t^a)}{p(x_t^v)\,p(x_t^a)}\right)\right\} && (2) \\ &\quad = I(X^a; X^v) && (3) \\ &\quad = h(X^a) + h(X^v) - h(X^a, X^v) && (4) \end{aligned}$$

where $I(X^a; X^v)$ is the mutual information between the random variables X^a and X^v. Mutual information can be expressed as a combination of the differential entropy terms $h(X^a)$, $h(X^v)$, $h(X^a, X^v)$ [14]. Consequently, estimating the mutual information between signals is, in this sense, equivalent to computing log-likelihood ratio statistic for the hypothesis test of (1). For more complex perceptual grouping hypotheses consisting of only pairwise relationships it has been shown [12], [13] that the sufficient statistics involve pairwise mutual information estimates. We elaborate on this in the empirical section. A significant issue, and what distinguishes our approach from others, is how one models the probability density terms of (2). Another important issue, which we address later, arises when direct density estimation is infeasible as is the case when measurements are of high dimension (e.g., audio video measurements).

Nonparametric density estimators, such as the Parzen kernel density estimator [15], are useful for capturing complex statistical dependencies between random variables. The resulting models can then be used to measure the degree of mutual information in complex phenomena [16] 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 sequence. The projections are computed adaptively such that the video and audio projections have maximum mutual information (MI).

We now review our basic method for audiovisual fusion and information theoretic adaptive methods. We then present a probabilistic model for cross-modal signal generation, and show how audiovisual correspondences can be found by identifying components with maximal mutual information. In an experiment comparing the audio and video of every

combination of a group of eight users, our technique was able to perfectly match the corresponding audio and video for each 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.

IV. Probabilistic Models of Audiovisual Fusion

We consider multimodal scenes which can be modeled probabilistically with one joint audiovisual source and distinct background interference sources for each modality. 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 (Fig. 1) to represent this relationship. In the diagrams, B represents the joint source, while A and C represent single modality background interference. Recall that the test of association is formulated as a measure of dependence between the measurements X^a and X^v. By conjecturing a latent variable structure via $\{A, B, C\}$ measurement dependence is explained solely through the hidden cause B. Our purpose here is to analyze under which conditions and in what sense our methodology uncovers the underlying cause of our observation without explicitly defining B or its exact relationship to X^a and X^v.

Fig. 1(a) 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 *common* dependence on a single cause. The joint statistical model consistent with the graph of Fig. 1(a) 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 the equivalent graphical manipulation) yields the graph of Fig. 1(b) 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, however, that decompositions of the measurement X^a and X^v *exist* such that the following joint densities can be written:

$$P(A,B,C,X^a,X^v) = P(A)P(B)P(C)P(X_A^a|A)P(X_B^a|B)P(X_B^v|B)P(X_C^v|C)$$

where $X^a = [X_A^a, X_B^a]$ and $X^v = [X_B^v, X_C^v]$. 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 Fig. 1(c) and from that graph we can extract the Markov chain which contains elements related only to B. Fig. 1(d) shows equivalent graphs of the extracted Markov chain. As a consequence, there is no influence due to A or C.

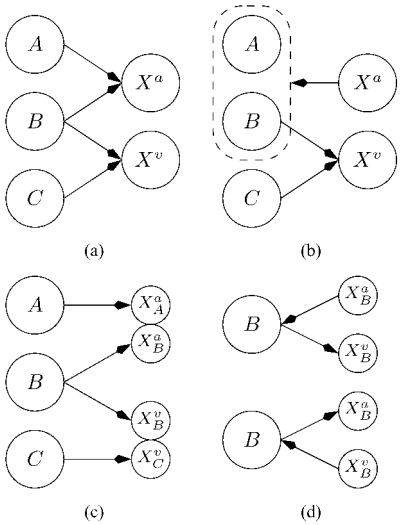

Fig. 1. Graphs illustrating the various statistical models exploited by the algorithm: (a) the independent cause model—X^a and X^v are independent of each other conditioned on $\{A, B, C\}$, (b) information about X^a contained in X^v is conveyed through *joint* statistics of A and B, (c) the graph implied 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.

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 [14], that the following inequality holds:

$$I(X_B^a; X_B^v) \le I(X_B^a; B)$$
$$I(X_B^a; X_B^v) \le I(X_B^v; B).$$

More importantly, these inequalities hold for *any* functions of X_B^a and X_B^v (e.g. $Y^a = f(X^a; h_a)$ and $Y^v = f(X^v; h_v)$). That is

$$I(Y^a; Y^v) \le I(Y^a; B) \quad (5)$$
$$I(Y^a; Y^v) \le I(Y^v; B) \quad (6)$$

and finally one can show (see [12]) that

$$I(Y^a; Y_B^v) \le I(X_B^a; X_B^v) = I(X^a : X^v). \quad (7)$$

The inequalities of (5) and (6) show that by maximizing the mutual information between $I(Y^a; Y^v)$ we 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 and in that sense captures elements of the generative model of audio and video. Note that this is the case without ever specifying the exact form of B or its relationship to the measurements. Additionally, the inequality of (7) shows that by maximizing $I(Y^a; Y^v)$ we are also maximizing a lower bound on the likelihood statistic, (3), of the association hypothesis test. Finally, with an approximation we describe shortly, 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. From

the perspective of information theory, estimating separate projections of the audio–video measurements which have high mutual information has intuitive appeal as such features will be predictive of each other. An additional 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.

V. Maximally Informative Projections

We now describe a method for learning maximally informative projections. The method uses a technique that maximizes the mutual information between the projections of the audio-visual measurements. Following [17], we use a nonparametric model of joint density for which an analytic gradient of the mutual information with respect to projection parameters is available. 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 [17]). Here, we restrict ourselves to linear functions of the measurements resulting in a significant computational savings at a minimal cost to the representational power. Note that while the projections are linear, the joint density is estimated nonparametrically allowing for more complex joint dependencies than can be captured by Gaussian assumptions. We parameterize the projections as

$$y_t^v = h_v^T x_t^v \tag{8}$$

$$y_t^a = h_a^T x_t^a \tag{9}$$

where $x_t^v \in \Re^{N_v}$ and $x_t^a \in \Re^{N_a}$ 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 $y_t^a \in \Re^{M_a}$. Treating x_t's and y_t'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 for continuous random variables can be expressed in several ways as a combination of differential entropy terms [14]

$$\begin{aligned} I(Y^v; Y^a) &= h(Y^a) + h(Y^v) - h(Y^a, Y^v) \\ &= \int_{R_{Y^a}} p_{Y^a}(y) \log\left(p_{Y^a}(y)\right) dy \\ &\quad + \int_{R_{Y^v}} p_{Y^v}(y) \log\left(p_{Y^v}(y)\right) dy \\ &\quad - \int\int_{R_{Y^a} \times R_{Y^v}} p_{Y^a, Y^v}(x, y) \\ &\quad \times \log\left(p_{Y^a, Y^v}(x, y)\right) dx dy. \end{aligned} \tag{10}$$

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 (10) with a second-order Taylor-series approximation as in [16], [18]

$$\hat{I}(Y^v, Y^a) = \hat{H}(Y^a) + \hat{H}(Y^v) - \hat{H}(Y^v, Y^a) \tag{11}$$

$$\begin{aligned} &= \int_{R_{Y^a}} \left(\hat{p}_{Y^a}(y) - p_u(y)\right)^2 dy \\ &\quad + \int_{R_{Y^v}} \left(\hat{p}_{Y^v}(y) - p_u(y)\right)^2 dy \\ &\quad - \int_{R_{Y^a} \times R_{Y^v}} \left(\hat{p}_{Y^v, Y^a}(x, y) - p_u(x, y)\right)^2 \\ &\quad \times dx dy \end{aligned} \tag{12}$$

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 $\hat{p}(y)$ is a Parzen density [15] estimated over the projected samples. The Parzen density estimate is defined as

$$\hat{p}(y) = \frac{1}{N} \sum_t \kappa(y - y_t, \sigma) \tag{13}$$

where $k(\,)$ 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 (12)) 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 [16], [18]. The update term for the individual entropy terms in (12) (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_t denotes a sample of either Y^a or Y^v or their concatenation depending on which term of (12) is being computed)

$$\begin{aligned} \Delta {y_t}^{(k)} &= b_r({y_t}^{(k-1)}) \\ &\quad - \frac{1}{N} \sum_{j \neq t} \kappa_a\left({y_t}^{(k-1)} - {y_j}^{(k-1)}, \Sigma\right) \end{aligned} \tag{14}$$

$$b_r(y_t)_j \approx \frac{1}{d}\left(\kappa\left(y_t + \frac{d}{2}, \Sigma\right)_j - \kappa\left(y_t - \frac{d}{2}, \Sigma\right)_j\right) \tag{15}$$

$$\begin{aligned} \kappa_a(y, \Sigma) &= \kappa(y, \Sigma) * \kappa'(y, \Sigma) \\ &= -\left(2^{M+1} \pi^{M/2} \sigma^{M+2}\right)^{-1} \exp\left(-\frac{y^T y}{4\sigma^2}\right) y \end{aligned} \tag{16}$$

where $M = M_a$, M_v, or $M_a + M_v$ depending on the entropy term. Both $b_r(y_t)$ and $\kappa_a(y_t, \sigma)$ 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 $b_r(y_t)_j$ indicates the jth element of $b_r(y_t)$. 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.

A. Capacity Control

In [17] early results were demonstrated using this method for the video-based localization of a speaking user. However, the technique lacked robustness as 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. The method of [16] requires that the projection be differentiable, which it is in this case. The specific means of capacity control that we utilize is to 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.* [19] 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 (11) and the regularization terms:

$$J = \hat{I}\left(Y^v, X^a\right) - \alpha_v {h_v}^T h_v - \alpha_a {h_a}^T h_a - \beta {h_v}^T \bar{R}_V^{-1} h_v \quad (17)$$

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 [19]) and is equivalent to prewhitening 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. Note that there is a straightforward probabilistic interperetation of the each of terms where $\hat{I}\left(Y^v, X^a\right)$ relates to the hypothesis test and the remaining terms represent Gaussian priors on the *coefficients* of the projections (but **not** on the resulting the projections of the measurements).

Computing h_v can be decomposed into three stages:

1) Prewhiten the images **once** (using the average spectrum of the images) followed by iterations of
2) Updating the feature values (y_t^v's) using (14), and
3) Solving for the projection coefficients using least squares and the L_2 penalty.

The prewhitening interpretation has intuitive appeal for the images 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. Furthermore, by including a prewhitening filter as a preproecessing step one can exclude the final term of (17) which is what we do in practice.

The projection coefficients related to the audio signal, h_a, are solved in a similar fashion (simultaneously) without the initial prewhitening step.

VI. Experiments

Our motivating scenario for this application is a group of users interacting with an anonymous handheld device or kiosk using spoken commands. Given a received 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. 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 interaction may be anonymous, we presume no prior model of the voice or appearance of users are available to perform the verification and localization.

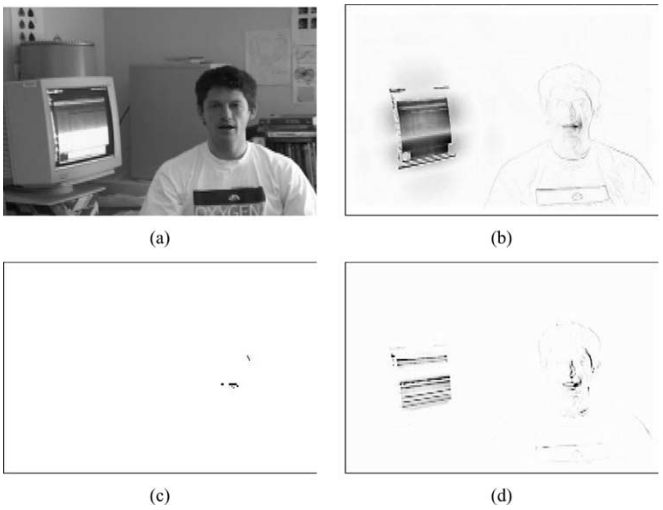

Fig. 2. Video sequence contains one speaker and monitor which is flickering: (a) one image from the sequence, (b) pixel-wise image of standard deviations taken over the entire sequence, (c) image of the learned projection, h_v, and (d) image of h_v for incorrect audio.

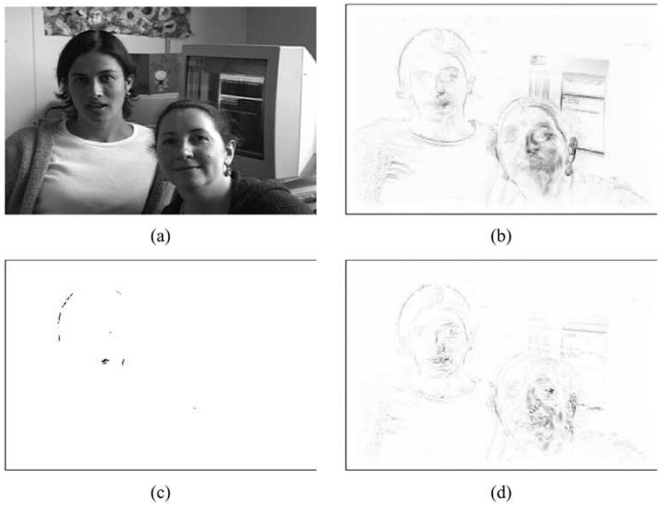

Fig. 3. Video sequence containing one speaker (person on left) and one person who is randomly moving their mouth/head (but not speaking): (a) one image from the sequence, (b) pixel-wise image of standard deviations taken over the entire sequence, (c) image of the learned projection, h_v, and (d) image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

In the first experiment[1] 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 360×240. The audio signal was collected at 48 000 KHz, but only 10 KHz of frequency content was used. All subjects were asked to utter a specific phrase. This typically yielded 2–2.5 s of data. Video frames were processed as is, while the audio signal was transformed to a series

[1]A portion of these results appear in [20]

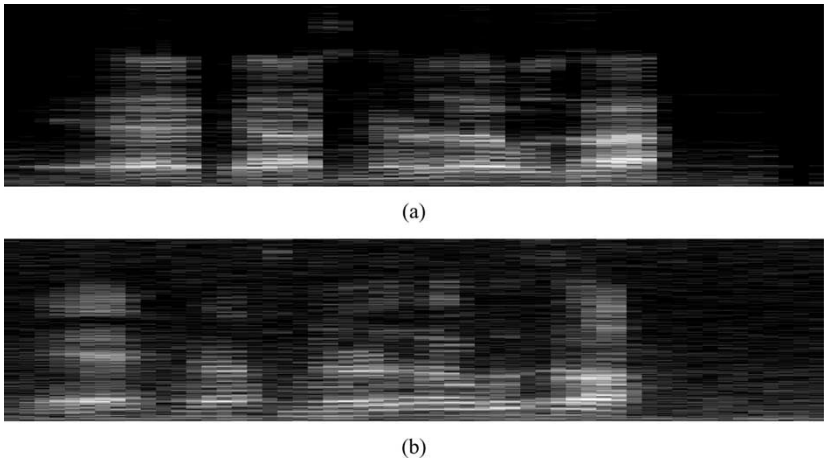

Fig. 4. Grayscale magnitude of audio periodograms. Frequency increases from bottom to top, while time is from left to right. (a) Audio signal for image sequence of Fig. 2. (b) Alternate audio signal recorded from different subject.

of periodograms. The window length of the periodogram was 2/29.97 s (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 marginal densities of the two variables are uniform while the conditional density is that of the kernel. In all cases we assume that there is no 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.

Fig. 2(a) shows a single video frame from one sequence of data. In the figure there is a single speaker and a video monitor. Throughout the sequence the video monitor exhibits significant flicker. Fig. 2(b) 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. Fig. 4(a) shows the associated periodogram sequence where the horizontal axis is time and the vertical axis is frequency (0–10 KHz). Fig. 2(c) shows an image of 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. Fig. 3(a) 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. Fig. 3(b) shows the pixel variance as before. There are significant changes about both subjects lips. Fig. 3(c) shows an image of the learned projection coefficients when the video is fused with the audio. Again the region about the correct speaker's lips is highlighted as well as some associated head motion.

While we can localize the audio source in the image sequence by inspecting the amplitudes of the projection coefficients, the resulting statistic also quantifies the likelihood of the association. Consequently, 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 using it to quantify the audio/video consistency.

TABLE I
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**

Using the sequences of Figs. 2 and 3, we compared the fusion result when using a separately recorded audio sequence from another speaker. The periodogram of the alternate audio sequence is shown in Fig. 4(b). Figs. 2(d) and 3(d) 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 Fig. 4(d) while energy is distributed throughout the image of Fig. 3(d). For Fig. 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 Fig. 4(b), the value drops to 0.08 of maximum. For the sequence of Fig. 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. The drop in the mutual information statistic tells us directly that the mismatched audio and video are less likely to be associated. This is further illustrated by inspecting the projection coefficients and noting that the projections of Figs. 2(d) and 3(d) extract information from the entire image rather than from a localized region.

Data was collected from six additional subjects for this experiment, and each video sequence was compared to each

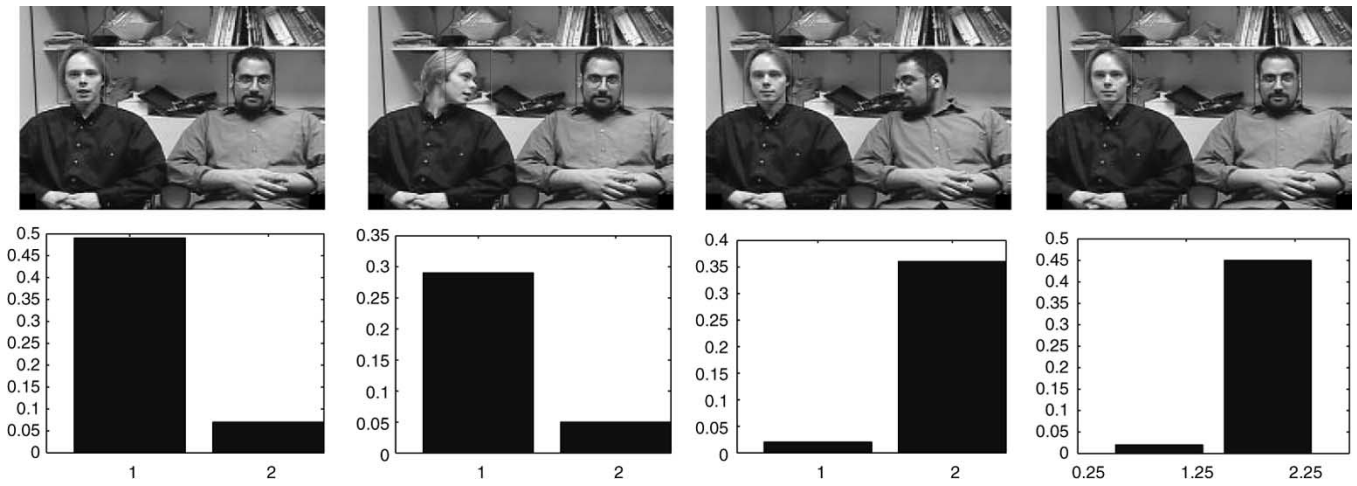

Fig. 5. Top row presents four frames from a video sequence with two speakers in front of a single camera and microphone. Audiovisual consistency is measured using a mutual information criteria. In the first two frames the left person is speaking, while in the last two the right person is speaking. The consistency measure shown in the bottom row for each frame correctly detects who is speaking.

audio sequence. (No attempt was made to temporally align the mismatched audio sequences at a fine scale, but they were coarsely aligned). Table I 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 using the full 2–2.5 s utterance.

In our second experiment, we test how this method can segregate speech of multiple users in a single field of view in concert with a face detection module. We show it is possible to detect which user is speaking by comparing the association statistics for each region. This problem has a slightly different formulation than the hypothesis test of (1). Let x_{lt}^v and x_{rt}^v denote the image sequence subregion on the left and right, respectively. These are the sequence of pixels contained in the moving boxes in Fig. 5. The single audio signal is denoted x^a. The hypothesis test is then stated as

$$\begin{aligned} H_0&: x_{lt}^v, x_{rt}^v, x_t^a \sim p(x_l^v, x^a)\, p(x_r^v) \\ H_1&: x_{lt}^v, x_{rt}^v, x_t^a \sim p(x_l^v)\, p(x_r^v, x^a)\,. \end{aligned} \tag{18}$$

H_0 states that the audio signal is associated with the region on the left, while H_1 states that the audio signal is associated with the signal on the right. Following similar analysis as (4) yields

$$\begin{aligned} &E\left\{\frac{1}{N}\sum_{t=0}^{N-1}\log\left(\frac{p(x_{rt}^v, x_{lt}^v, x^a|H_1)}{p(x_{rt}^v, x_{lt}^v, x^a|H_0)}\right)\right\} \\ &\quad = E\left\{\frac{1}{N}\sum_{t=0}^{N-1}\log\left(\frac{p(x_{lt}^v)\,p(x_{rt}^v, x_t^a)}{p(x_{lt}^v, x_t^a)\,p(x_{rt}^v)}\right)\right\} \qquad (19) \\ &\quad = E\left\{\frac{1}{N}\sum_{t=0}^{N-1}\log\left(\frac{p(x_{lt}^v)\,p(x_{rt}^v, x_t^a)\,p(x_t^a)}{p(x_{lt}^v, x_t^a)\,p(x_{rt}^v)\,p(x_t^a)}\right)\right\} \\ &\quad = I(X^a; X_r^v) - I(X^a; X_l^v). \qquad (20) \end{aligned}$$

Consquently, the difference between the estimated mutual information quantities (or their bounds) is equivalent to the log-likelihood ratio of the hypothesis test. The audiovisual mutual information method is able to match the visual speech motion with the acoustic signal, and ignore confounding motions of the other user's head or other motions in the scene. Fig. 5 shows the result tracking two users speaking in turns in front of a single camera and microphone, and detecting which is most likely to be speaking based on the measured audiovisual consistency. In the figure the bar graph under each image shows the individual mutual information estimates. In each case the difference favors the correct speaker.

VII. Conclusions and Future Work

We have developed a technique for measuring the signal-level correspondence between audio and video observations. The method is based on estimating a bound on mutual information. Addtionally we were able to put approaches to audio-visual association based on mutual information into the context of nonparametric hypthesis testing. Measuring the degree of mutual information between signals yields a useful cue indicating whether they come from a common source. This cue is useful to know whether a microphone and camera are receiving information from a single user, or whether signals from multiple co-located users are being confused. Nonparametric statistical density models can be used to represent complex joint densities of projected signals, and to successfully estimate mutual information.

Experiments using our approach demonstrated that the method is able to correctly determine which audio and video fragments come from the same speaker. Our technique is a step toward natural and untethered interfaces, where multiple users can interact with conversational systems without attachments or explicit segmentation cues. In this work we specifically explored signal-level fusion, and did not make any assumptions about acoustic or visual models. In domains where such assumptions are viable, or when prior models of individual users are available, such information could be profitably used in concert with our approach.

REFERENCES

[1] G. Deco and D. Obradovic, *An Information Theoretic Approach to Neural Computing*. New York: Springer-Verlag, 1996.

[2] M. Plumbley and S. Fallside, "An information-theoretic approach to unsupervised connectionist models," in *Proceedings of the 1988 Connectionists Models Summer School*, D. Touretzky, G. Hinton, and T. Sejnowski, Eds., San Mateo, CA, 1988, pp. 239–245.

[3] M. Plumbley, "On Information Theory and Unsupervised Neural Networks," Cambridge Univ., Eng. Dept., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR. 78, 1991.

[4] S. Becker, "An Information-Theoretic Unsupervised Learning Algorithm for Neural Networks," Ph.D., Univ. Toronto, Toronto, ON, Canada, 1992.

[5] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Toward unrestricted lipreading," in *Second Int. Conf. Multimodal Interfaces (ICMI99)*, 1999.

[6] G. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke, "Lipreading by neural networks: Visual preprocessing, learning and sensory integration," in *Proc. Neural Information Syst. NIPS-6*, 1994, pp. 1027–1034.

[7] J. Hershey and J. Movellan, "Using audio-visual synchrony to locate sounds," in *Adv. Neural Inform. Process. Syst.*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds., 1999, vol. 12, pp. 813–819.

[8] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Adv. Neural Inform. Process. Syst.*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., 2000, vol. 13.

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

[10] R. Cutler and L. S. Davis. Look who's talking: Speaker detection using video and audio correlation. presented at IEEE Int. Conf. Multimedia and Expo (III). [Online]. Available: citeseer.nj.nec.com/cutler00look.html

[11] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[12] A. T. Ihler, J. W. Fisher, and A. S. Willsky, "Hypothesis testing of factorizations for data association," in *Information Processing in Sensor Networks*, F. Zhao and L. Guibas, Eds. Berlin, Germany: Springer-Verlag, 2003, Lecture Notes in Computer Science, pp. 239–253.

[13] ——, "Nonparametric hypothesis testing for statistical dependency," *IEEE Trans. Signal Processing*, to be published.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[15] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math Statist.*, vol. 33, pp. 1065–1076, 1962.

[16] J. W. Fisher III and J. C. Principe, "A methodology for information theoretic feature extraction," in *Proc. IEEE Int. Joint Conf. Neural Networks*, vol. 3, A. Stuberud, Ed., 1998, pp. 1712–1716.

[17] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Adv. Neural Inform. Process. Syst.*, 2000, vol. 13.

[18] J. W. Fisher III and J. C. Principe, "Entropy manipulation of arbitrary nonlinear mappings," in *Proc. IEEE Workshop, Neural Networks for Signal Processing VII*, J. Principe, Ed., 1997, pp. 14–23.

[19] A. Mahalanobis, B. Kumar, and D. Casasent, "Minimum average correlation energy filters," *Appl. Opt.*, vol. 26, no. 17, pp. 3633–3640, 1987.

[20] J. W. Fisher III and T. Darrell, "Probabalistic models and informative subspaces for audiovisual correspondence," in *7th Eur. Conf. Computer Vision*, vol. 2352, Lecture Notes in Computer Science, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds., Copenhagen, Denmark, May 28–31, 2002, pp. 592–603.

John W. Fisher, III (M'91) received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 1997.

He is currently a Principal Research Scientist in the Computer Science and Artificial Intelligence Laboratory and affiliated with the Laboratory for Information and Decision Systems, both at the Massachusetts Institute of Technology (MIT), Cambridge. Prior to joining MIT, he was with the University of Florida as both a faculty member and graduate student since 1987, during which time he conducted research in the areas of ultra-wideband radar for ground penetration and foliage penetration applications, radar signal processing, and automatic target recognition algorithms. His current area of research focus includes information theoretic approaches to signal processing, multimodal data fusion, machine learning and computer vision.

Trevor Darrell (M'96) received the B.S. degree in 1988 from the University of Pennsylvania, Philadelphia, and the M.S. and Ph.D. degrees in 1991 and 1996, respectively, from the Massachusetts Institute of Technology (MIT), Cambridge.

He heads the Vision Interface Project at the MIT Artificial Intelligence Laboratory, which investigates the use of computer vision for better interfaces between people and computers. Previously, he was a Member of the Research Staff at Interval Research Corportation, where he developed perceptual interfaces based on the integration of multiple visual processing modalities. His interests include the detection, tracking, and recognition of people using visual sensors, image motion and correspondence estimation, automatic morphing and image-based rendering, and perceptually-enabled intelligent environments.