Dynamic Dependency Tests: Analysis and Applications to Multi-modal Data Association

Michael R. Siracusa **John W. Fisher III**
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology Cambridge, MA 02139
{siracusa,fisher}@csail.mit.edu

Abstract

The goal of a dynamic dependency test is to correctly label the interaction of multiple observed data streams and to describe how this interaction evolves over time. To this end, we propose the use of a hidden factorization Markov model (HFactMM) in which a hidden state indexes into a finite set of possible dependence structures on observations. We show that a dynamic dependency test using an HFactMM takes advantage of both structural and parametric changes associated with changes in interaction. This is contrasted both theoretically and empirically with standard sliding window based dependence analysis. Using this model we obtain state-of-the-art performance on an audio-visual association task without the benefit of labeled training data.

1 Introduction

Statistical approaches to modeling dynamics and clustering data are well studied research areas. Two classic examples are fitting mixture models and training hidden Markov models (HMMs) [16] using the expectation maximization (EM) algorithm [4]. We consider a complementary problem: given multiple data streams, we are interested in the nature of their interaction as it evolves over time. We refer to all methods that address such problems as dynamic dependency tests. We propose a model in which the interaction is described by graphical models with changing structures, *i.e.*, the presence or absence of edges, but whose parameters (or more generally, parameterization) are not available. In contrast to standard approaches, we explicitly model varying interactions via a dynamically switching graphical structure.

We cast a dynamic dependency test as the problem of inference on a special class of probabilistic models in which a latent state variable indexes a discrete set of possible dependency structures on measurements. We refer to this class of models as dynamic dependence models and introduce a specific implementation via a hidden factorization Markov model (HFactMM). Such models can be described in terms of a Contingent Bayesian Network (CBN) [13] in which the dependency structure of a set a variables is contingent upon the values those variables take on. For general CBNs, exact inference may not be tractable and efficient learning may not be available. We show, however, when we restrict ourselves to the class of HFactMMs, standard methods, specifically EM and Viterbi decoding, can be used with slight modification for efficient learning and exact inference.

We utilize an HFactMM in an audio-visual speaker association task. Consider a scene in which there are several individuals, each of whom may be speaking at any given moment. Given a single audio recording of the scene and a separate video stream for each individual in the scene, we wish to associate the audio data with the proper video stream at any given time. Association is defined by the presence or absence of an edge in a graphical model representing the relationship between the audio and video streams. In this application, each possible dependency structure has a simple semantic interpretation, namely it indicates who is speaking. The HFactMM allows us to exploit the fact that the appearance of all individuals may change depending on which individual is speaking. We learn who, if anyone, is speaking at each point in time and the dynamics of the conversation. In contrast to previous approaches, the method described in this paper does not require prior scene- or user-specific appearance models, which are often not available. We demonstrate, both theoretically and empirically, a clear advantage over standard moving window methods.

The outline of this paper is as follows. An overview of related work is presented in Section 2. In Sec-

tion 3 we present the HFactMM and its use for dynamic dependency tests. We theoretically show how an HFactMM takes advantage of both structural and parametric changes associated with changes in interaction. This is contrasted with standard sliding window based dependence analysis. Illustrative examples are presented in Section 4. Section 5 demonstrates that the proposed method obtains the best performance reported to date on the standard audio-visual CUAVE database [15] for speaker association. In contrast to previous approaches applied to this dataset, superior performance is achieved without benefit of labeled training data, or the use of a specialized silence detector.

2 Related Work

This work fits into the general category of data clustering and dynamic modeling. Typically, models used for such tasks assume a fixed dependency structure for the observed data. The study of models whose graphical structure is contingent upon the values/context of the nodes in the graph can be traced back to Geiger and Heckerman's similarity networks and multinets [6] . This class of models has been further explored and formalized by Boutilier, *et al.*'s Context-Specific Independence [3] and more recently Milch, *et al.*'s CBNs [13]. The HFactMM presented here fits into this class of models and is closely related to Bilmes's Dynamic Bayesian Multinets [2]. The focus of [2] was to show how learning state-indexed structure using labeled training data can yield better models for classification tasks. In contrast, here the dependency structures are defined by the problem and no labeled data is required.

An HFactHMM is also related to switching linear dynamic systems (SLDSs) used by the tracking community [7]. SLDS models are combinations of discrete Markov models and linear state-space dynamical systems. The hidden discrete state chooses between a predefined number of state-space models to describe the data at each point in time. SLDS models are primarily used to help improve tracking and track interpretation by allowing changes to the state-space model parameters. In contrast, an HFactMM explicitly models varying dependence structure over time with no linear Gaussian assumptions.

There are many related techniques for estimating the dependence among a set of random variables. Information theoretic approaches have a long history beginning with Kullback [12]. In the domain of audio-visual association, Hershey and Movellan showed how measured correlation between audio and pixels can help in detecting who is speaking. Nock and Iyengar [14] provided an empirical study of this technique on the CUAVE dataset [15]. Further study of detecting and characterizing the dependency between audio and video was carried out by Slaney and Covell [19] and Fisher,*et al.*[5]. All of these techniques process data using a sliding window over time and assume a single audio source within that window. As such, they do not take advantage of the past or future to learn a audio-visual appearance model of the potential audio sources.

3 Hidden Factorization Markov Model

Let $\mathbf{O}_t = \{\mathbf{o}_t^1, \mathbf{o}_t^2, \ldots, \mathbf{o}_t^N\}$ be an observation of N random variables at time t with $\mathbf{o}_t^i \in \mathbb{R}^{d_i}$. Let $\mathbf{O}_{1:T}$ represent $\mathbf{O}_t$ from time 1 to T. Given $\mathbf{O}_{1:T}$, the goal of a dynamic dependency test is to label the sequence according to the dependency among the N random variables at each time t. To this end, we propose a hidden factorization Markov Model (HFactMM) in which we assume that the observation $\mathbf{O}_t$ is independent of all other observations conditioned on a hidden state S_t, and that the states $S_{1:T}$ are first order Markov. Thus,

$$p(\mathbf{O}_{1:T}, S_{1:T}; \Theta) = p(S_{1:T}; \Theta) \prod_{t=1}^{T} p(\mathbf{O}_t | S_t; \Theta), \qquad (1)$$

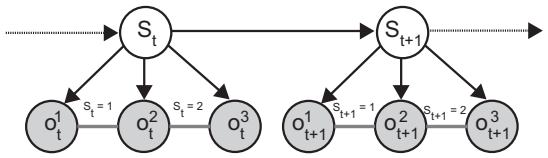
where Θ are the parameters. We adopt the term HFactMM to distinguish such models as a special case of more general HMMs. That is, this model is an HMM with the special property that the value $k \in [1...K]$ of the hidden state variable S_t indicates one of K possible "interactions" between the N random variables at time t. These "interactions" are defined in terms of a particular factorization F^k and parametrization Θ^k:

$$p(\mathbf{O}_t | S_t = k; \Theta) = p_{\Theta^k}(F_t^k) = \prod_{i=1}^{C_k} p(F_{i,t}^k; \Theta^k) \qquad (2)$$

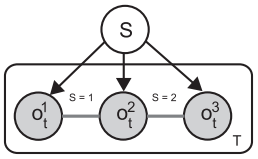
where F^k specifies a partitioning of the full set of N random variables into C_k subsets such that $\bigcup_{i=1}^{C_k} F_i^k = \{\mathbf{o}^1, \ldots, \mathbf{o}^N\}$ and $F_i^k \bigcap F_j^k = \emptyset \ \forall i, j \in [1...C_k]$ when $i \neq j$.

Figure 1(a) shows an HFactMM with two possible factorizations: $F^1 = \{\{\mathbf{o}^1, \mathbf{o}^2\}, \{\mathbf{o}^3\}\}$ and $F^2 = \{\{\mathbf{o}^2, \mathbf{o}^3\}, \{\mathbf{o}^1\}\}$. For this graph, $p_{\Theta^1}(F_t^1) = p(\mathbf{O}_t | S_t = 1; \Theta) = p(\mathbf{o}_t^1, \mathbf{o}_t^2; \Theta^1) p(\mathbf{o}_t^3; \Theta^1)$ and $p_{\Theta^2}(F_t^2) = p(\mathbf{O}_t | S_t = 2; \Theta) = p(\mathbf{o}_t^2, \mathbf{o}_t^3; \Theta^2) p(\mathbf{o}_t^1; \Theta^2)$.

Note that the value of the state S_t determines the probabilistic structure of the observations at time t. For the applications considered here each structure has a simple semantic interpretation. For example, in Figure 1(a), if $\mathbf{o}_t^1$ and $\mathbf{o}_t^3$ are the video observations of two



(a) HFactMM



(b) Static Factorization Model

Figure 1: Example HFactMM and static factorization model. This graph notation follows that of CBNs [13] in which conditionally labeled edges are only present when the condition is true.

individuals at time t and $\mathbf{o}_t^2$ is the corresponding audio observation at time t then when $S_t = 1$ ($S_t = 2$) we assume that the individual corresponding to $\mathbf{o}^1$ ($\mathbf{o}^3$) is speaking.

We consider situations in which the model parameters are not known *a priori*. The Baum-Welch/EM [16, 4] algorithm can be used with a slight modification for learning the parameters of an HFactMM. Subsequently Viterbi decoding can be used for exact inference [16]. We construct and utilize an HFactMM model in the following way:

1. Define the K possible dependency structures and the parameterization of the HFactMM (e.g. Gaussian, discrete/codebook or mixture densities for each factor) for your task.
2. Learning: Calculate $\hat{\Theta} = \arg\max_{\Theta} p(\mathbf{O}_{1:T}; \Theta)$
3. Inference: Find $\hat{s}_{1:T} = \arg\max_{s_{1:T}} p\left(s_{1:T} | \mathbf{O}_{1:T}; \hat{\Theta}\right)$

Note that we assume no training data and perform learning and inference on the data being analyzed. We make the usual assumption that all K states are visited at least once (and typically multiple times) during the observed sequence.

3.1 Learning

Let $\Theta = \{\pi, A, \Theta^1, \ldots, \Theta^K\}$ be the parameter set for the model where $\pi_k = p(S_1 = k)$ are the prior state probabilities, A is a $K \times K$ matrix with transition probabilities $A_{ij} = p(S_{t+1} = i | S_t = j)$, and Θ^k are the parameters for factorization F^k (*i.e.* parameters for $p_{\Theta^k}(F^k) = p(\mathbf{O}_t | S_t = k; \Theta)$). As with typical HMMs [16] the EM algorithm, can be applied to models with this structure in order to find the parameters,$\hat{\Theta}$, that maximize the likelihood of the given data.

While the E-step is unchanged, the HFactMM requires a minor change to the M-step of EM. Since the state conditional model $p_{\Theta^k}(F^k)$ breaks up into the C_k factors of F^k, the structure of the M-step updates simplify accordingly. For example, if each $p_{\Theta^k}(F^k_{f,t})$ is a simple Gaussian with mean $\mu_{k,f}$ and covariance $\Sigma_{k,f}$, the M-step at iteration (i) would yield:

$$\mu_{k,f}^{(i)} = \frac{\sum_{t=1}^{T} \left[F_{f,t}^k\right] \gamma_t(k)}{\sum_{t=1}^{T} \gamma_t(k)}, \tag{3}$$

$$\Sigma_{k,f}^{(i)} = \frac{\sum_{t=1}^{T} (\left[F_{f,t}^k\right] - \mu_{k,f}^{(i)})(\left[F_{f,t}^k\right] - \mu_{k,f}^{(i)})^{\mathsf{T}} \gamma_t(k)}{\sum_{t=1}^{T} \gamma_t(k)} \tag{4}$$

where $\gamma_k(t) = p(S_t = k | \mathbf{O}_{1:T}; \Theta^{(i-1)})$ and the notation $\left[F_{f,t}^k\right]$ is used to denote a stacked vector of the variables in factor F_f^k at time t. Note that, here, a parenthesized superscript indicates the iteration number. This structural break-down by factor holds for all other families of distributions yielding a more structured learning procedure with savings in storage and computation.

3.2 Inference

Having learned the parameters $\hat{\Theta}$, the data sequence is labeled by finding $\{\hat{s}_{1:T}\} = \arg\max_{s_{1:T}} p(s_{1:T} | \mathbf{O}_{1:T}; \hat{\Theta})$. This can be done efficiently with the Viterbi algorithm [16]. Note that choosing the state sequence with the maximum posterior probability via Viterbi decoding is implicitly performing an M-ary hypothesis test over all possible sequences, where $M = K^T$.

The HFactMM can take advantage of both differences in structure and parameters as compared to windowed methods which can be shown to exploit only differences in structure. We illustrate this by considering a binary hypothesis test between two of the K^T different state sequences $S_{1:T}^{H1}$ and $S_{1:T}^{H2}$. Given the learned parameters $\hat{\Theta}$ a hypothesis test deciding between them has the following form:

$$\hat{L}_{1,2} \triangleq \log\left(\frac{p\left(\mathbf{O}_{1:T} | S_{1:T}^{H1}; \hat{\Theta}\right)}{p\left(\mathbf{O}_{1:T} | S_{1:T}^{H2}; \hat{\Theta}\right)}\right) \underset{H2}{\overset{H1}{\gtrless}} \log\left(\frac{p\left(S_{1:T}^{H2}; \hat{\Theta}\right)}{p\left(S_{1:T}^{H1}; \hat{\Theta}\right)}\right) \tag{5}$$

It is easy to show that in expectation, the value of the log likelihood ratio under $H1$ can be expressed as:

$$E_{H1}\left[\hat{L}_{1,2}\right] = \sum_{\substack{\forall t \text{ s.t.} \\ S_t^{H1} \neq S_t^{H2}}} D\left(p_{\hat{\Theta}^{S_t^{H1}}}\left(F^{S_t^{H1}}\right) || p_{\hat{\Theta}^{S_t^{H1}}}\left(F^{\emptyset}\right)\right) + \sum_{\substack{\forall t \text{ s.t.} \\ S_t^{H1} \neq S_t^{H2}}} D\left(p_{\hat{\Theta}^{S_t^{H1}}}\left(F^{\emptyset}\right) || p_{\hat{\Theta}^{S_t^{H2}}}\left(F^{S_t^{H2}}\right)\right) \tag{6}$$

and under $H2$ can be expressed as:

$$E_{H2}\left[\hat{L}_{1,2}\right] = -\sum_{\substack{\forall t \text{ s.t.} \\ S_t^{H1} \neq S_t^{H2}}} D\left(p_{\hat{\Theta}^{S_t^{H2}}}\left(F^{S_t^{H2}}\right) || p_{\hat{\Theta}^{S_t^{H2}}}\left(F^{\emptyset}\right)\right) - \sum_{\substack{\forall t \text{ s.t.} \\ S_t^{H1} \neq S_t^{H2}}} D\left(p_{\hat{\Theta}^{S_t^{H2}}}\left(F^{\emptyset}\right) || p_{\hat{\Theta}^{S_t^{H1}}}\left(F^{S_t^{H1}}\right)\right) \quad (7)$$

where $D(p||q)$ is the Kullback-Leibler divergence between p and q, and $F^{\emptyset}$ is the factorization which contains only edges/factors common to *all* F^k (*i.e.* it represents the common structure among all factorizations). Note that the parameter set of $p_{\Theta^k}(F^{\emptyset})$ can be obtained for any Θ^k by marginalizing over $p_{\Theta^k}(F^k)$ appropriately.

Notice that in both cases the expected log likelihood ratio decomposes into two terms. The first term concerns purely structural differences. The true structure is compared with the common structure, under the true parameters. The second term contains both structural and parameter differences. The common structure is compared with the true parameters to the model for the incorrect hypothesis. It is important to note that all such tests can be decomposed in this way. This decomposition quantifies the contributions of prior (or learned) model differences versus structural differences to separability between hypothesized sequences. We will contrast this with standard windowed methods in the next section.

3.3 Comparison with Windowed Factorization Tests (WFT)

Sliding window methods are an alternative to batch analysis for a dynamic dependency test. These methods hypothesize the dependency structure over a window of time in which the structure is assumed to be held constant. Such tests are referred to as factorization tests in [11]. The model associated with a factorization test is a special case of an HFactMM in which the state is constant over the window analyzed. Figure 1(b) shows an example static factorization model.

Here we summarize some of the key results provided in [11] adapted to the notation used in this paper. The hypothesis test between two factorizations $H1 : S = 1$ and $H2 : S = 2$ of a data sequence of length T with unknown (Θ^1, Θ^2) takes the form of a generalized likelihood ratio test (GLRT) in which:

$$\hat{L}_{1,2} = \frac{1}{T} \log \left(\frac{p\left(\mathbf{O}_{1:T} | S = 1; \hat{\Theta}^1\right)}{p\left(\mathbf{O}_{1:T} | S = 2; \hat{\Theta}^2\right)} \right) \quad (8)$$

$$\text{where} \quad \hat{\Theta}^i = \arg\max_{\Theta^i} p(\mathbf{O}_{1:T} | S = i; \Theta^i). \quad (9)$$

In expectation the (normalized) generalized log likelihood ratio becomes:

$$\begin{aligned} E_{H1}\left[\hat{L}_{1,2}\right] &= D\left(p_{\hat{\Theta}^1}\left(F^1\right) || p_{\hat{\Theta}^1}\left(F^{\emptyset}\right)\right) \\ &\quad + D\left(p_{\hat{\Theta}^1}\left(F^{\emptyset}\right) || p_{\hat{\Theta}^2}\left(F^2\right)\right) \\ &= D\left(p_{\Theta^1}\left(F^1\right) || p_{\Theta^1}\left(F^{\emptyset}\right)\right) + 0 \\ E_{H2}\left[\hat{L}_{1,2}\right] &= -D\left(p_{\hat{\Theta}^2}\left(F^1\right) || p_{\hat{\Theta}^2}\left(F^{\emptyset}\right)\right) \\ &\quad + D\left(p_{\hat{\Theta}^2}\left(F^{\emptyset}\right) || p_{\hat{\Theta}^1}\left(F^1\right)\right) \\ &= -D\left(p_{\Theta^2}\left(F^2\right) || p_{\Theta^2}\left(F^{\emptyset}\right)\right) - 0 \end{aligned} \quad (10)$$

where here $F^{\emptyset}$ represents the common structure between factorizations F^1 and F^2. Note that only the purely structural term remains. This is because one is estimating both $\hat{\Theta}^1$ and $\hat{\Theta}^2$ from the same observation sequence and thus loses the ability the exploit parameter differences. That is, when H1 is true, only the pure structure term remains because the parameter estimates for H2 converge to those consistent with the marginals of H1 and yield a model which factors according to the common structure between the two hypotheses:

$$\text{Under } H1 \begin{cases} p_{\hat{\Theta}^1}\left(F^1\right) \to p_{\Theta^1}\left(F^1\right) \\ p_{\hat{\Theta}^2}\left(F^2\right) \to p_{\Theta^1}\left(F^{\emptyset}\right) \end{cases} \quad (11)$$

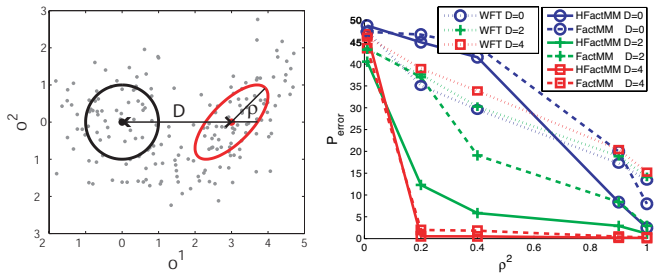
and similarly,

$$\text{Under } H2 \begin{cases} p_{\hat{\Theta}^2}\left(F^2\right) \to p_{\Theta^2}\left(F^2\right) \\ p_{\hat{\Theta}^1}\left(F^1\right) \to p_{\Theta^2}\left(F^{\emptyset}\right) \end{cases} \quad (12)$$

In the special case where $F^1 = \{\{\mathbf{o}^1, \mathbf{o}^2\}\}$ and $F^2 = \{\{\mathbf{o}^1\}, \{\mathbf{o}^2\}\}$, $\hat{L}_{1,2}$ converges to mutual information (MI). All tests that estimate correlation or MI over a sliding window to check for dependence fall into this windowed factorization test (WFT) framework (e.g. [10, 19, 5]). Another issue with windowed factorization tests that is common to GLRTs is how to make a decision when the hypotheses are nested (e.g. F^1 is a fully joint model, F^2 is a fully factored), since the more expressive model (F^1) will always have a higher likelihood. It is common to use permutations to obtain a p-value that can be used to decide between the two hypotheses (see [8, 18]).

4 Illustrative Examples

In this section we present a few simple synthetic examples to illustrate the differences between performing a dynamic dependency test using an HFactMM and a windowed factorization test (WFT). The questions we wish to address are 1) How do changes in both the structural and parametric differences between



(a) (b)

Figure 2: 2D Gaussian results. a) sample draw. Note that is not possible to see the temporal dynamics in this figure b) Average % error over 100 trials for HFactMM, FactMM and WFT as a function of ρ for various D.

the state conditional models affect the performance of these methods. 2) When are state dynamics in the HFactMM important, and 3) What can be done when the correct parameterization is unknown.

Consider a model with two 1-D variables which switch between being dependent and independent: $F^0 = \{\{\mathbf{o}^1\},\{\mathbf{o}^2\}\}$ and $F^1 = \{\{\mathbf{o}^1,\mathbf{o}^2\}\}$. When $S_t = 0$, the observations are i.i.d. Gaussian with zero mean and unit variance. When $S_t = 1$ the observations have a mean of $[0\ \ D]^\mathsf{T}$ and correlation coefficient ρ. We fix the dynamics on the state S_t by setting the parameters $\pi_0 = \pi_1 = .5$, $A_{00} = A_{11} = .95$ and $A_{12} = A_{21} = .05$. This yields a model with a simple state dynamic and a control on structural and parametric differences via ρ and D respectively.

We draw 200 samples for each setting of ρ and D. Figure 2(a) shows one such sampling. Note that these samples are subject to the process dynamics defined by A. Three different techniques are compared: dynamic dependency test using an HFactMM model, a factorization mixture model (FactMM) and a WFT. The FactMM has the same structure as an HFactMM without a dynamic on S_t. The WFT reduces to simply calculating the correlation between the observations in a sliding window and estimating a p-value via permutations. Window sizes of 5, 10, 20, and 40 samples are tested. For each trial, we find the threshold on the p-value that yields the best performance for each window size and then report the best over all window sizes. This represents an unrealistic best-case scenario for the WFT.

Results are shown in Figure 2. Each data point is the average probability of error over 100 trials. Consistent with previous analysis, Figure 2(b) shows the performance of the WFT does not change substantially as non-structural parameters, D, vary. In general all approaches improve in performance with increasing ρ, with more rapid improvements for the HFactMM and FactMM for larger D. Dynamics help most when D is small, *i.e.* when the state conditional distributions overlap. More complex probability models can be used, but as the next example shows, certain ambiguities may arise. Such ambiguities can be overcome when an underlying dynamic is present. Consider the data shown in Figure 3(a). We again assume a two state model (independent shown in thin black vs dependent shown in thick red) using the same dynamics as in the previous example. Note that in this case each state conditional model is a mixture of Gaussians ($p_{\Theta^0}(\mathbf{O}_t)$ is a product of two mixtures of two Gaussians and $p_{\Theta^1}(\mathbf{O}_t)$ is a mixture of four).

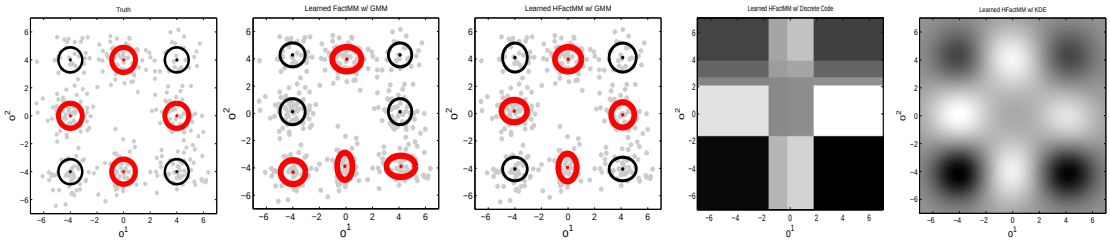
Figures 3(b) and 3(c) show FactMM and HFactMM models learned from 200 samples of this mixture model when they are given the correct parameterization (*i.e.* correct number of mixtures) but unknown parameters. Note that for this particular model there are many possible combinations of independent and dependent mixtures. In fact, the FactMM model picked one of these alternative mixtures in Figure 3(b). This is because by assuming independent samples and ignoring the state dynamic all valid combinations of dependent and independent cluster mixtures are equally likely. By incorporating dynamics the HFactMM finds the correct solution.

Note that in Figure 3(b) and 3(c) we used the correct parameterization. When little is known about the appropriate parameterization for a particular problem one can use other more flexible state conditional distributions. A simple approach is to first create separate codebooks for each observed variable via vector quantization and then use an HFactMM with discrete models for each state conditional distribution. This can be done by fitting a Gaussian mixture model (GMM) or via K-means. Creating an 8 code codebook for each variable ($\mathbf{o}^1$ and $\mathbf{o}^2$) using GMMs and then using an HFactMM with discrete state conditional distributions we obtain the result shown in Figure 3(d).

Alternatively one can utilize sample kernel density estimates. A non-parametric sample based kernel density estimate (KDE) can be used for the each factor in the state condition models. Each factor f's distribution is of the form:

$$p_{\Theta^i}(F^i_{f,t}) = \frac{1}{T}\sum_{j=1}^{T}\alpha^i_j K(F^i_{f,t} - F^i_{f,j};\sigma^i) \qquad (13)$$

where $K()$ is a valid kernel function with kernel size σ^i and $\alpha^i_j = p(S_j = i|\mathbf{O}_{1:T};\Theta)/\sum_t p(S_t = i|\mathbf{O}_{1:T};\Theta)$. Figure 3(e) shows the learned HFactMM model using a KDE with a Gaussian kernel. Leave-one-out likelihood was used to adjust kernel size. It is important to note that one must be careful when using more powerful state-conditional distributions. If a single state conditional distribution is flexible enough to describe all of the data and the state transition probabilities are



(a) (b) (c) (d) (e)

Figure 3: A more complex 2D example. a) True distribution, F^0=thin black, F^1=thick red b) Learned FactMM w/ GMM c) learned HFactMM w/ GMM, d) learned HFactMM with Discrete Code. e) Learned HFactMM w/ KDE. Figures d and e show $p(\mathbf{O}|S=1;\hat{\Theta}^1) - p(\mathbf{O}|S=0;\hat{\Theta}^0)$ Mean accuracy over 50 trials : FactMM w/ GMM=64%, HFactMM w/ GMM=99%, HFactMM w/ Discrete Code=98%, HFactMM w/ KDE=98%

learned, the most likely answer will be that all of the data came from a single state with a complex distribution. One way to deal with this is to set reasonable priors on the state transition matrix A or initial distribution parameters $\pi_i = p(S_t = i)$.

5 Audio-Visual Experiments

In this section demonstrate state-of-the-art results on an audio-visual association task using an HFactMM. Given a single audio stream and separate video streams for each speaker in a scene, our task is to determine who, if anyone, is speaking at each point in time. When person i is speaking it is assumed that the audio stream will be dependent on video stream i, otherwise the streams are independent.

Two different datasets are used. The first is the CUAVE corpus [15], a multiple speaker audio-visual corpus of spoken connected digits. We use the 22 clips from the *groups* set in which two speakers take turns reading digit strings and then proceed to speak simultaneously. In order to compare to [14] and [9] we only consider the section of alternating speech. In each clip both individuals face the camera at all times. We use ground truth from [1]. The second dataset is a single clip recorded in the same style as the CUAVE database in which two individuals take turns speaking digits. However, while the speaker looks into the camera the other subject turns to look at the speaker. This gives yields a dataset in which there is a strong appearance change depending on who is speaking, as may be the case in a meeting where participants look toward the current speaker. Each dataset contains video sampled at 29.97 fps. The audio is resampled at 16kHz. For each of these datasets the video streams are extracted faces normalized to 100×100 pixels. In the CUAVE dataset a face detector and correlation tracking of the nose region is used to get a stabilized face. For the second dataset a fixed region of the video around each person's face is simply extracted. The extracted faces of both datasets are made publicly available [17].

Simple frame-based features are used as observations. The audio is broken into segments corresponding to each video frame. For each stream, at each frame, both static and dynamic features are calculated. At each frame t, Mel-frequency cepstral coefficients are computed from the corresponding audio segment and used as the static audio features. The static video features are PCA coefficients (using 40 principle components) for the images of the segmented faces. The dynamic features for all streams at frame t are the differences between the static features at $t+1$ and $t-1$.

For each of these feature streams a 20-symbol codebook is learned via fitting a 20-component GMM. All methods use a common set of observations,$\mathbf{o}_t^{A_s}, \mathbf{o}_t^{A_d}, \mathbf{o}_t^{V1_s}, \mathbf{o}_t^{V1_d}, \mathbf{o}_t^{V2_s}, \mathbf{o}_t^{V2_d}$, which are the feature streams encoded with their corresponding codebook for the static and dynamic audio and both video streams respectively. This results in a 1D discrete code representation for each static and dynamic feature stream. Note that the dimensionality reduction and codebook learning is done separately for each stream and for each data sequence analyzed (i.e. there is no user or corpus/dataset specific training).

Three possible states are considered with the following factorizations: $F^0 = \{\{\mathbf{o}^{A_d}\}, \{\mathbf{o}^{V1_d}\}, \{\mathbf{o}^{V2_d}\}, F^s\}$, $F^1 = \{\{\mathbf{o}^{A_d}, \mathbf{o}^{V1_d}\}, \{\mathbf{o}^{V2_d}\}, F^s\}$, and $F^2 = \{\{\mathbf{o}^{A_d}, \mathbf{o}^{V2_d}\}, \{\mathbf{o}^{V1_d}\}, F^s\}$ where $F^s = \{\{\mathbf{o}^{A_s}\}, \{\mathbf{o}^{V1_s}\}, \{\mathbf{o}^{V2_s}\}\}$. F^0 is fully independent corresponding to neither person speaking. F^1 and F^2 correspond to persons 1 and person 2 speaking respectively. Note that the structural differences between these 3 states are only in the dynamic features. The assumption is that the dependence information is mainly in the dynamics of the audio-visual speech process and static features mainly change in their appearance / parameters not in their dependence structure.

For all 22 sequences in the CUAVE groups set a dynamic dependency test is performed with an

(a) (b) (c) (d)

Figure 4: Example frames and extracted faces from the audio-visual datasets. (a) and (b) are from CUAVE, (c) and (d) show the second dataset.

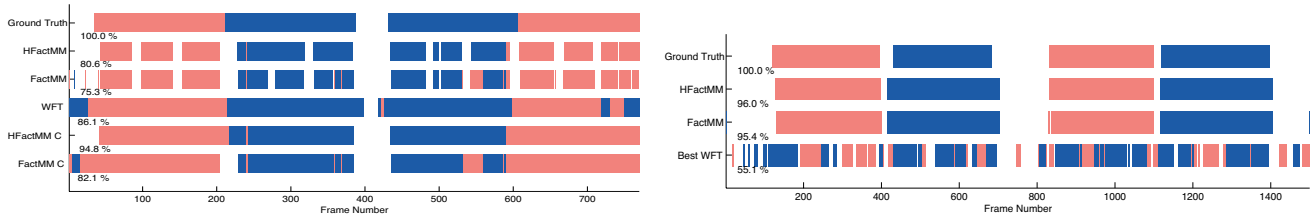
HFactMM, FactMM, and a WFT with window lengths of 8,15,30,60,90, and 120 frames. For the WFT, at each frame, the likelihood ratios $\hat{L}_{1,2}$, $\hat{L}_{1,0}$ and $\hat{L}_{2,0}$ (see Equation 8) are calculated using a window of samples centered around that frame. Additionally p-values for $\hat{L}_{1,0}$ and $\hat{L}_{2,0}$ are calculated via permutations [8, 18]. If $\hat{L}_{1,2}$ is positive (negative) then we eliminate $S_t = 2$ ($S_t = 1$) as a possible hypothesis and use the p-value for $\hat{L}_{1,0}$ ($\hat{L}_{2,0}$) to choose between $S_t = 1$ ($S_t = 2$) and $S_t = 0$. The parameter learning for the HFactMM and FactMM was set to try 100 different random starting points and run until convergence or a maximum of 80 EM iterations in order to combat the potential local maxima problems with EM. In most cases EM converged before 40 iterations.

The first row of Table 1 shows that all techniques yield around 80% accuracy. The maximum average performance of the WFT was obtained with a window length of 30 frames. This shows with some training data to set window length and thresholds the WFT method would do well with these features. However, these results are somewhat misleading as we explain. Figure 5(a) shows the estimated labels for a typical sequence in the corpus (g09). The top line shows the ground truth labeling. The next two are the outputs of the HFactMM and FactMM. Notice that these methods disagree with the ground truth by consistently putting non-speaking (fully independent) blocks between speaker transitions and within speaking blocks. Examination of these sections in the original video reveals that they are actually short periods of silence. In actuality the HFactMM and FactMM *correctly* labeled these sections. The WFT does not exhibit this behavior and smooths over the short silence regions.

The disagreement stems from an artifact of the procedure used to ground-truth the data where periods of silence that are less than 25 frames within a speech block are considered to be part of speech [1]. This constraint is not part of the HFactMM model and thus it produced a more accurate and fine scaled labeling of the periods of "silience" ($S_t = 0$). Nevertheless, to be consistent with the publicly available ground truth we can easily impose this silence constraint by post processing the outputs to remove any periods of labeled silence ($S_t = 0$) less than 25 consecutive frames. The constrained outputs are shown in the last two lines of Figure 5(a). With this constraint the HFactMM and FactMM outperform all other techniques improving to 88% and 86% respectively as shown in Table 1. Note that applying this constraint to the outputs of the WFTs does not affect performance. This is because the WFT smoothes over short silence regions as an artifact of having a sliding window.

To the best of our knowledge these results are equivalent to or better than all other reported results for speaker labeling on the CUAVE group set. Nock and Iyengar [14] obtain 75% accuracy with a windowed Gaussian MI measure and Gurban and Thiran [9] get 87.4% with a trained audio-visual speech detector. However, it is important to note that [9] utilizes a training corpus while the method described here does not. Additionally, both methods use a silence/speech detector and only perform a dependence test when speech is detected. A dynamic dependency test with an HFactMM obtains better performance without the benefit of separate training data or a silence detector.

In the CUAVE database most of the information about who is speaking comes from the changes in dependency structure between the audio and the video. (WFT gives similar performance to the HFactHMM as in the D=0 case in the synthetic example). In the second dataset there is a significant appearance change. When one person is speaking the other subject changes their gaze. The results for this sequence are shown in Figure 5(b). Both the HFactMM and FactMM greatly outperformed the WFT. The poor results of the WFT show that there is not sufficient dependency information in the features at all times. However the HFactMM and FactMM take advantage of the static appearance differences (in this case head pose) to help group/cluster the data and correctly label the video.



(a) CUAVE g09 (b) Dataset 2

Figure 5: AV Results. White = neither person is speaking, light red = person 1, dark blue = person 2. C=silence constraint imposed

	HFactMM	FactMM	Best WFT
Mean Accuracy (%)	80.24	78.51	83.86
Mean Accuracy C(%)	**88.11**	86.38	83.42

Table 1: Results Summary for CUAVE. The Best WFT accuracy corresponds to the WFT with settings that maximized the average performance for the entire dataset. C= silence constraint imposed. All results are based on the ground truth provided in [1].

6 Conclusion

In this paper we have introduced the use of an HFactMM for dynamic dependency tests. We have shown both theoretically and empirically that an HFactMM can exploit both structural and parameter differences to distinguish between hypothesized states of interaction. This is in contrast to sliding window methods which can only discriminate based on structural differences. We have shown state-of-the art performance on a standard dataset for audio-visual association. Significantly this was achieved without benefit of training data.

References

[1] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt. Experimental framework for speaker detection on the CUAVE database. In *Tech. Rep. 2006-003, EPFL, Lausanne, Switzerland*, 2006.

[2] J. A. Bilmes. Dynamic bayesian multinets. In *In Proc. of the 16th conf. on UAI*, pages 38–45, 2000.

[3] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *UAI*, pages 115–123, 1996.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society, Series B*, volume 39, pages 1–38, 1977.

[5] J. Fisher, III, T.Darrell, W. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NIPS*, pages 772–778, 2000.

[6] D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and bayesian multinets. In *Artificial Integlligence*, volume 82, pages 45–74, 1996.

[7] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. In *Neural Computation*, volume 12, pages 963–996, 1998.

[8] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, 1994.

[9] M. Gurban and J. Thiran. Multimodal speaker localization in a probabilistic framework. In *Proc. of EUSIPCO*, 2006.

[10] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 1999.

[11] A. Ihler, J. Fisher, and A. Willsky. Nonparameteric hypothesis tests for statistical dependency. In *Trans. on signal processing, special issue on machine learning*, 2004.

[12] S. Kullback. *Information Theory and Statistics*. Dover, 1968.

[13] B. Milch, B. Marthi, D. Sontag, S. Russell, D. Ong, and A. Kolobov. Approximate inference for infinite contingent bayesian networks. In *AIStats*, 2005.

[14] H. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *Proc. Intl. Conf. on Image and Video Retrieval*, 2003.

[15] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. Technical report, Department of ECE, Clemson University, 2001.

[16] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. IEEE*, volume 77 No. 2, pages 257–286, 1989.

[17] M. R. Siracusa and J. W. Fisher. Extracted faces for CUAVE and other audio-visual datasets. In *http://people.csail.mit.edu/siracusa/avdata/*, 2006.

[18] M. R. Siracusa, K. Tieu, A. T. Ihler, J. W. Fisher, and A. S. Willsky. Estimating dependency and significance for high-dimensional data. In *ICASSP*, volume 5, pages V/1085– V/1088, 2005.

[19] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *NIPS*, 2000.

This material is based upon work supported by the AFOSR under Award No. FA9550-06-1-0324. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Air Force.