DYNAMIC DEPENDENCY TESTS FOR AUDIO-VISUAL SPEAKER ASSOCIATION

Michael R. Siracusa and John W. Fisher III

Computer Science and Artificial Intelligence Laboratory at MIT, MA 02139

ABSTRACT

We formulate the problem of audio-visual speaker association as a dynamic dependency test. That is, given an audio stream and multiple video streams, we wish to determine their dependancy structure as it evolves over time. To this end, we propose the use of a hidden factorization Markov model in which the hidden state encodes a finite number of possible dependency structures. Each dependency structure has an explicit semantic meaning, namely "who is speaking." This model takes advantage of both structural and parametric changes associated with changes in speaker. This is contrasted with standard sliding window based dependence analysis. Using this model we obtain state-of-the-art performance on an audio-visual association task without benefit of training data.

Index Terms— Pattern clustering methods

1. INTRODUCTION

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 (or more commonly a region of a single video) for each individual in the scene, we wish to determine who, if anyone, is speaking at each point in time. The solution to this problem has wide applicability to tasks such as automatic meeting transcription, social interaction analysis, and control of human-computer dialog systems.

We view audio-visual speaker association as a particular example from a general class of problems we call dynamic dependency tests. A dynamic dependency test answers the following question: given multiple data streams, how does their interaction evolve over time? Here, interaction is defined in terms of changing graphical structures, *i.e.*, the presence or absence of edges in a graphical model. In the audio-visual speaker association problem each possible dependency structure has a simple semantic interpretation. Specifically, when video stream i and the audio stream are dependent we say "person i is speaking" and when all streams are independent we assume "the speaker is off camera or no one is speaking."

We cast a dynamic dependency test as a 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). This model allows us to take advantage of both structural and parametric changes associated with changes in speaker. This is contrasted with standard sliding window based dependence analysis [1, 2, 3, 4].

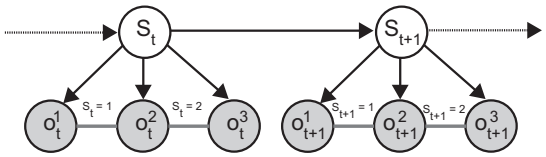
The approach presented in this paper fits into the general category of data clustering and dynamic modeling. Two classic examples in this category are fitting mixture models and training hidden Markov models (HMMs) using the EM algorithm [5]. Typically these models assume fixed dependency structure for 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 the Heckerman and Geiger's similarity networks and multinets [6] . This class of models has been further explored and formalized by Boutilier, *et al.*'s Context-Specific Independence (CSI) [7] and more recently Milch et. al's Contingent Bayesian Networks (CBN) [8]. An HFactMM fits into this class of models and is closely related to Bilmes's Dynamic Bayesian Multinets [9]. The focus of [9] 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.

There are many related techniques for estimating the dependence among a set of random variables. Specific to audio-visual association, Hershey and Movellan showed how measuring correlation between audio and pixels can help in detecting who is speaking [1]. Nock and Iyengar [4] provided an empirical study of this technique on the CUAVE dataset [10]. Further study of detecting and characterizing the dependency between audio and video was carried out by Slaney and Covell [2] and Fisher, *et al.* [3]. All of these techniques process data using a sliding window over time assuming 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. The method presented in this paper can take advantage of the voice and potential user pose and appearance changes associated with changes in speaker. It achieves the best performance we know of on the standard audio-visual CUAVE dataset. It achieves this without any window parameters to set, without a silence detector or lip tracker and without any labeled training data.

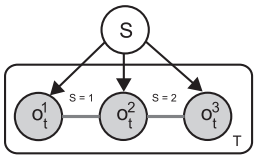
2. HIDDEN FACTORIZATION MARKOV MODEL

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 $\mathbf{O}_{1:T}$, our goal 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 the states $S_{1:T}$ are first order Markov. Thus $p(\mathbf{O}_{1:T}, S_{1:T}; \Theta) = p(S_{1:T}; \Theta) \prod_{t=1}^{T} p(\mathbf{O}_t | S_t; \Theta)$ where Θ are the parameters. 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 factorizations F^k and parameterizations Θ^k:

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



(a) HFactMM



(b) Static Factorization Model

Fig. 1. Example HFactMM and static factorization model. Following the notation of CBNs [8], conditionally labeled edges are only present when the condition is true.

where F^k specifies a partitioning of a 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 $F^2 = \{\{\mathbf{o}^2, \mathbf{o}^3\}, \{\mathbf{o}^1\}\}$. Note that the value of the state S_t determines the probabilistic structure of the observations at time t. Additionally, here we assign a semantic meaning to each structure. For example, in Figure 1(a), if $\mathbf{o}_t^1$ and $\mathbf{o}_t^3$ are the video observations of two 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 infer 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. This necessitates both a learning and inference step. The Baum-Welch/EM algorithm can be used with a slight modification for learning the parameters of a HFactMM, subsequently Viterbi decoding can be used for exact inference [5]. We construct and utilize a HFactMM model in the following way:

1. Define the K possible dependency structures and parameterization of the HFactMM for your task.
2. Learning: Estimate $\hat{\Theta} = \arg\max_{\Theta} p(\mathbf{O}_{1:T}; \Theta)$
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.

2.1. 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 $A_{ij} = p(S_{t+1} = i | S_t = j)$, and Θ^k is the set of 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 [5] 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 $p_{\Theta^k}(F_{f,t}^k)$ is a simple Gaussian with mean $\mu_{k,f}$ and covariance $\Sigma_{k,f}$, the M-step at iteration (i) would yield:

$$\mu_{k,f}^{(i)} = \frac{\sum_{t=1}^{T} \left[F_{f,t}^k\right] \gamma_t(k)}{\sum_{t=1}^{T} \gamma_t(k)}, \tag{2}$$

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

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.

2.2. 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 [5]. First learning the parameters $\hat{\Theta}$ and then inferring the hidden state from the entire data sequence allows us to take advantage of both differences in structure and parameters as compared to windowed methods which only exploit differences in structure. It can be shown that the K^T-ary hypothesis test implicitly being performed when doing viterbi decoding relies on both structural and general statistical model differences between the learned state conditional distributions $p(\mathbf{O}_t | S_t; \hat{\Theta})$ [11].

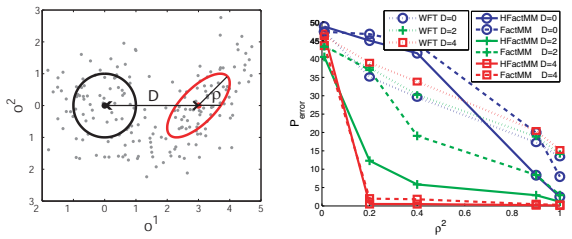
2.3. Comparison with Windowed Factorization Tests (WFT)

Sliding window methods can also be used for dynamic dependency tests. 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 [12]. Dynamics are captured by sliding this windowed test in time. The model associated with a factorization test is a special case of a HFactMM in which the state is constant over the window analyzed. Figure 1(b) shows an example of this type of model.

It has been shown in [12] that an online factorization test can only exploit structural differences between hypothesized factorizations. This is because one is estimating $\hat{\Theta}^k$ for all $k = 1...K$ from the same windowed observation sequence (assumed to have a fixed structure). All tests which estimate correlation or MI over a sliding window to check for dependence fall into this windowed factorization test framework (e.g. [1, 3]). Another issue with windowed factorization tests that is common to generalized likelihood ratio tests 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 make decisions based on estimated p-values in such cases.

3. ILLUSTRATIVE EXAMPLE

In this section we present a simple synthetic example to answer the following questions: 1) How is the performance using an HfactMM and WFT affected as both the structural and parametric differences between the state conditional models are changed, 2) When are the state dynamics used in an HFactMM important. 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) (b)

Fig. 2. 2D Gaussian Experimental Results. a) A sample draw. Note that is not possible to see the temporal dynamics in this figure b) Performance (Avg.% error over 100 trials) of HFactMM,FactMM and WFT as a function of ρ for various D.

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 $A_{12} = A_{21} = .05$.

We draw 200 samples for each setting of ρ and D. Figure 2(a) shows one such sampling. We compare 3 different techniques: 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. Various window sizes and p-value thresholds were tested and for each trial the best result among all settings was recorded. This represents an unrealistic best-case scenario for the WFT.

Results are shown in Figure 2. The performance of the WFT does not change substantially as non-structural parameters, D, vary. This is consistent with the theory in [12]. 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.

In the previous example we assumed simple Gaussian state conditional model for each factorization. When little is known about the appropriate parameterization for a particular problem one can use other more flexible state conditional distributions (e.g. mixture models). The approach taken in this paper is to first create separate codebooks for each observed variable via vector quantization (using cluster labels of a fit GMM or K-means) and then use a HFactMM with discrete models for each state conditional distribution.

4. AUDIO VISUAL EXPERIMENTS

In this section we show how we obtain 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, we determine who, if anyone, is speaking at each point in time. When person i is speaking we assume that the audio stream will be dependent on video stream i, otherwise the streams are independent.

We use two different datasets. The first is the CUAVE corpus [10], 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 [4] and [13] we only consider the section of alternating speech. In each clip both individuals face the camera at all times. We use ground truth from [14]. The second dataset is a single clip recorded in the same style as the CUAVE database in which two individuals take turns

	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.

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 [11].

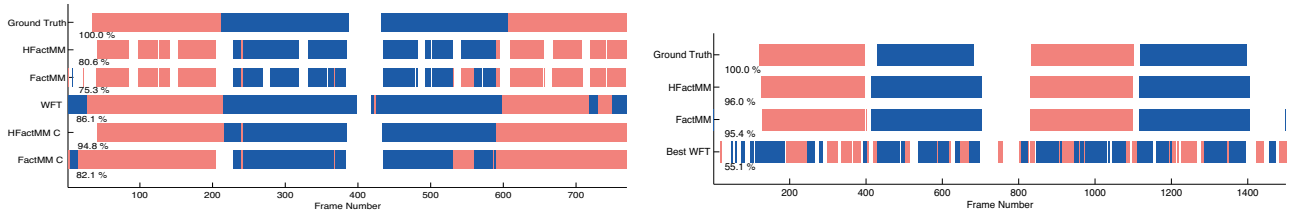
Simple frame-based features are used as observations. The audio is broken into segments corresponding to each video frame. For each stream, at each frame, both static and dynamic features are calculated. At each frame t, Mel-frequency cepstral coefficients are computed from the corresponding audio segment 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 time $t+1$ 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 were 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 was performed with an HFactMM, FactMM, and a windowed factorization with window lengths of 8,15,30,60,90, and 120 frames. For specific details on how the WFT was carried out, the settings for EM training of the HFactMM and FactMM, and a full table of results for all 22 sequences see [11]. Table 1 shows a summary of the average performance for each method. The accuracy percentage is the percentage of frames correctly classified according to the ground truth provided by [14].

The first row of Table 1 shows that all the techniques give around 80% accuracy. The maximum average performance of



(a) CUAVE g09 (b) Dataset 2

Fig. 3. AV Results. White = neither person is speaking, light red = person 1, dark blue = person 2. C=silence constraint imposed

the WFT was obtained with a window length of 30 frames (see [11]). 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 3(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 smoothes over the short silence regions. This disagreement is an artifact of the procedure used for generating the ground truth, which states periods of silence less than 25 frames within speech are considered part of speech [14]. This constraint is easily imposed by post processing the outputs to remove any periods of labeled silence ($S_t = 0$) less than 25 frames. The constrained outputs are shown in the last two lines of Figure 3(a). With this constraint, the HFactMM and FactMM outperform all other methods, improving to 88% and 86% respectively as shown in Table 1. 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 [4] obtain 75% accuracy with a windowed Gaussian MI measure and Gurban and Thiran [13] get 87.4% with a trained audio-visual speech detector. Both methods use a silence/speech detector and only perform a dependence test when there is speech. Using the method described in this paper yields 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 3(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.

5. CONCLUSION

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

6. REFERENCES

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

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

[3] J.W. Fisher, III, T.Darrell, W.T. Freeman, and P.A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *NIPS*, 2000, pp. 772–778.

[4] H.J. 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.

[5] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. IEEE*, 1989, vol. 77 No. 2, pp. 257–286.

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

[7] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller, "Context-specific independence in Bayesian networks," in *UAI*, 1996, pp. 115–123.

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

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

[10] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," Tech. Rep., Department of ECE, Clemson University, 2001.

[11] M.R. Siracusa and J.W. Fisher III, "Dynamic dependency tests," Tech. Rep. LIDS-TR-2706 http://people.csail.mit.edu/siracusa/ddt/, MIT, 2006.

[12] A.T. Ihler, J.W. Fisher, and A.S. Willsky, "Nonparameteric hypothesis tests for statistical dependency," in *Trans. on signal processing, special issue on machine learning*, 2004.

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

[14] 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.