BIMODAL SPEECH RECOGNITION USING COUPLED HIDDEN MARKOV MODELS

Stephen M. Chu and Thomas S. Huang

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

ABSTRACT

In this paper we present a bimodal speech recognition system in which the audio and visual modalities are modeled and integrated using coupled hidden Markov models (CHMMs). CHMMs are probabilistic inference graphs that have hidden Markov models as sub-graphs. Chains in the corresponding inference graph are coupled through matrices of conditional probabilities modeling temporal influences between their hidden state variables. The coupling probabilities are both cross chain and cross time. The later is essential for allowing temporal influences between chains, which is important in modeling bimodal speech. Our bimodal speech recognition system employs a two-chain CHMM, with one chain being associated with the acoustic observations, the other with the visual features. A deterministic approximation for maximum a posteriori (MAP) estimation is used to enable fast classification and parameter estimation. We evaluated the system on a speaker independent connected-digit task. Comparing with an acoustic-only ASR system trained using only the audio channel of the same database, the bimodal system consistently demonstrates improved noise robustness at all SNRs. We further compare the CHMM system reported in this paper with our earlier bimodal speech recognition system in which the two modalities are fused by concatenating the audio and visual features. The recognition results clearly show the advantages of the CHMM framework in the context of bimodal speech recognition.

1. INTRODUCTION

Incorporating visual information into automatic speech recognition (ASR) has been demonstrated as an effective approach to improve the performance and robustness over the audio-only systems, and has received much attention in recent years [1-4]. One of the most challenging issues in bimodal ASR is how to fuse the audio (e.g. acoustic speech signal) and the visual (e.g. lip motion) modalities.

The approaches to the fusion problem can be classified into three categories in terms of which stage the fusion takes place.

1. The fusion is carried out before the recognition, usually through the concatenation of audio and visual features.
2. The fusion occurs after the recognition by combining the separate outputs from the audio and visual channels in some manner.
3. The fusion takes place during the recognition, allowing the two modalities influence each other in the recognition process.

The approaches in the first category treat the two modalities as one, while the approaches in the second category essentially consider the two modalities independently. In either case, the sensor fusion problem is not explicitly addressed. A number of approaches of the third category have shown encouraging results [3-5]. In [4], discriminative training is used to find the hidden Markov model (HMM) stream exponents that weight and combine the likelihood of the audio stream and the likelihood of the visual stream at the state level. In [5], two synchronous HMM chains, each modeling one of the modalities, are linked using the Boltzmann zipper.

Coupled hidden Markov models (CHMMs), proposed in [6], provide a promising framework for coupling HMMs to model systems of multiple interacting processes [7]. In particular, the framework can be directly applied to model the class of couplings known as the sensor fusion problem. The audio-visual speech is a perfect example.

The speech sound and the lip movement carry complimentary information about different components of the same underlying speech-production process. The two modalities are tightly coupled, and they influence and interact with each other via temporal dependencies. To be able to capture this inter-modality dynamics is critical to a desired fusion scheme.

CHMMs capture this inter-modality dynamics by temporally coupling parallel HMMs. Moreover, CHMMs provide a consistent probabilistic framework that allows us to address both the modeling and fusion in a unified fashion. CHMMs are in essence probabilistic inference graphs that have hidden Markov models as sub-graphs. Chains in the corresponding inference graph are coupled through matrices of conditional probabilities modeling temporal influences between their hidden state variables. The internal dynamics of each modality is modeled by one of the component hidden Markov models, while the interactions between modalities are reflected in the coupling probabilities. Exact inference in CHMMs is not attractive due to its high computational complexity. We use a deterministic approximation for maximum a posteriori (MAP) estimation that enables fast classification and parameter estimation.

In our experiments, the bimodal speech recognition system employs a two-chain CHMM, with one chain being associated with the acoustic observations, the other with the visual features. To obtain the visual features, we implemented a lip-tracking algorithm based on the Bézier volume deformation model.  

1 Alternatively, one could view [3] and [4] as being in the second category. However, due to the fact that with these methods, the likelihood scores are combined during the decoding process rather than at the end, we feel it is more appropriate to put them in the third category.
This paper is organized as follows. In the next section we first introduce the CHMM topology, which is followed by a brief discussion on inference in the CHMM framework. The visual features and the lip-tracking algorithm are explained in section 3. We present the experimental results in section 4 and conclude the work in section 5.

2. Coupled Hidden Markov Models

The original CHMM framework as developed in [6] consists of two parts. The first is the “shoelace” topology that makes temporal coupling of HMMs possible. The second part is a deterministic approximation method for MAP inference in CHMM. We will give more attention to the former since it is of direct interest with regard to the modeling and fusion of audio-visual speech.

2.1. CHMM Topology

Let’s first consider a conventional one-chain HMM. A typical left-to-right HMM is shown in Figure 1 with the commonly used state machine representation. In this form of representation, each node denotes one of the \( N \) discrete hidden states in \( S = \{ s_1, s_2, s_3, \ldots, s_N \} \). The arrows indicate the transition probability from the state \( i \) to state \( j \), \( P(j|i) \).

![Figure 1. State machine diagram of a three state left-to-right HMM. Output variables are not shown for simplicity.](image)

Alternatively, HMMs can also be represented graphically using probabilistic inference graphs, which are obtained by rolling out the state machine in time. Figure 2 shows an HMM as a probabilistic inference graph.

![Figure 2. Probabilistic inference graph for three time slices of an HMM. Circular nodes denote state variables, and square nodes represent output variables.](image)

Under this representation, a circular node denotes the value of the state variable \( s(t) \in S \) at time \( t \), and the square represents the output variable \( o(t) \). The arrows connecting the circles are the state-to-state transition probability \( P(s(t)|s(t-1)) \). The output density \( p(o(t)|s(t)) \) for state \( s(t) \) is depicted as the arrows going from the circular nodes to the square nodes.

It is convenient to use probabilistic inference graphs to represent the independence structure of a distribution over random variables. In particular, they provide a handy way to illustrate the coupling of multiple HMMs.

There are various ways by which multiple HMMs can be coupled [6]. Two of the possibilities that could be applied to audio-visual speech recognition are shown in Figure 3.

![Figure 3. Different approaches of coupling interacting processes. a. HMMs are coupled through joint probabilities between synchronous states. b. The interaction of the two processes is modeled by cross-time and cross-chain conditional probabilities.](image)

In Figure 3a, the coupling of two HMMs is carried out by connecting synchronous states from the opposite chains with joint probabilities. These coupling probabilities determine how likely a pair of states occur together. In general, inference architectures with synchronous links assume lockstep processes that do not have temporal influences on each other [6]. However, in the case of audio-visual speech, the interaction between the audio and visual channels is evidently temporal. For example, the coarticulation visible from the lip motion does not coincide with the vocalized speech sound [8]. In fact, it is not difficult to see that the temporal relationship between the two modalities can be properly captured by the inference structure shown in Figure 3b. In the two-chain CHMM, the current state of one chain not only depends on the previous state in the same process, but also is conditioned on the previous state in the opposite chain, which by the Markov property encapsulates the context of the parallel process. The cross-chain and cross-time dependencies are expressed as a conditional probability table.
The ability to model temporally interacting processes makes CHMM a suitable framework for sensor fusion in bimodal speech recognition. However, the cross-work connections also lead to an implicit state space that is exponential in the number of state variables. Exact inference in CHMM therefore is impracticable.

2.2. Inference in CHMM

Given a particular combined state sequence $\Sigma = \{s_1; s_2\}$, where $s_1 = \{s_1(1), s_1(2), \ldots, s_1(T)\}$ is the state sequence of the first HMM, and $s_2 = \{s_2(1), s_2(2), \ldots, s_2(T)\}$ is the state sequence of the other HMM, the posterior of $\Sigma$ through the two-chain CHMM is

$$P(\Sigma | O) = \frac{P(O | \Sigma) P(\Sigma)}{P(O)}$$

where

$$P(O | \Sigma) = \prod_{t=1}^{T} p(o_t | s_t) p(s_t | s_{t-1})$$

and

$$P(\Sigma) = P(s_1(1)) P(s_2(1)) \prod_{t=2}^{T} \left( P(s_1(t) | s_1(t-1)) \cdot P(s_2(t) | s_2(t-1)) \cdot P(s_1(t) | s_1(t-1)) \cdot P(s_2(t) | s_2(t-1)) \right)$$

For a two-chain CHMM with $N$ states in each chain, one can show that exact MAP inference is an $O(TN^4)$ computation. Various approximation methods to the inference problem have been proposed, including the Monte Carlo sampling method and the $N$-head dynamic programming method [6]. The later is a deterministic approximation for MAP inference, with a reduced computational complexity of $O(T(2N)^2)$. We use the $N$-head algorithm in this experiment.

3. VISUAL FEATURES

A lip motion tracker is implemented based the piecewise Bézier volume deformation model [9]. In this model, a face is divided into multiple connected deformable surfaces. Each surface is characterized as a deformable mesh embedded in a Bézier volume. The face mesh is deformed by moving the control points of the volumes. Given a video segment, the tracking algorithm estimates the 2-D displacements of the nodal points on the mesh using template matching. And from these displacement vectors, 3-D global motion parameters (rotation and translation) and non-rigid motion parameters (action units) are computed using a least square estimator. One important property of this model is that it can accurately track facial motions without relying on a person-dependent 3-D head model. In fact, it can be used to track any sub-region of a face. A simple initialization procedure suffices the adaptation of the model to a new face. Therefore, this tracking algorithm is well suited for speaker-independent audio-visual speech recognition experiments. An example of the lip-tracking result is shown in Figure 4.

4. EXPERIMENTS

4.1. Audio-Visual Speech Database

The experiments were carried out using an audio-visual speech database created by the authors. We collected and processed bimodal speech data from 100 speakers, including 50 speakers of each gender. All of the subjects are native speakers of American English. The data were first recorded on analog videotapes and later digitized using the AVID system. The video frame rate is 30Hz; the audio is sampled at 44.1KHz.

The database consists of two parts. The first part is made of continuous speech of military command sentences, with a vocabulary size of 101. The second part contains connected digits. Each digit string is 11-digit in length. There are a total of 21,000 sentences in the database, of which 1,000 are devoted to connected digits. The work reported in this paper is based on the second part of the database.
4.2. Experimental Results

The acoustic data are first down-sampled to 16KHz and processed using a 25ms window, with the frame shift set to 11.1ms. This setting results in an equivalent frame rate of 90Hz. For each frame, 13 MFCC coefficients with an analysis order of 24 and the corresponding delta coefficients are computed. The visual features generated by the lip tracker are interpolated from 30Hz to 90Hz to synchronize with the acoustic features. The first-order and second-order delta coefficients for the visual features are also computed. We run the lip-tracking algorithm on an 8-processor SGI ONYX computer with a VTX graphics engine. The acoustic analysis is performed using HTK.

The architecture of the recognizer is based on connected word models. For each digit, a two-chain CHMM with 5 emitting states in each chain is constructed. The intra-chain transition probabilities are initialized such that left-to-right state progression within a given chain is ensured. The audio HMM and visual HMM are trained separately at first. Upon convergence, the inter-chain coupling transition probabilities are enabled and the resulting CHMM is continuously trained using the N-head algorithm.

We evaluate the proposed audio-visual speech recognition system on a speaker-independent connected-digit task. Seven testing conditions of different SNR levels ranging from 0dB to 30dB are simulated by adding white noise to the clean speech data. The system is trained and tested using cross-validation. The recognition results are listed in Table 1.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>A</th>
<th>A+V</th>
<th>CHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12.4</td>
<td>36.6</td>
<td>53.3</td>
</tr>
<tr>
<td>5</td>
<td>24.2</td>
<td>56.0</td>
<td>67.7</td>
</tr>
<tr>
<td>10</td>
<td>52.6</td>
<td>76.1</td>
<td>84.5</td>
</tr>
<tr>
<td>15</td>
<td>77.4</td>
<td>89.3</td>
<td>93.1</td>
</tr>
<tr>
<td>20</td>
<td>90.9</td>
<td>93.8</td>
<td>95.9</td>
</tr>
<tr>
<td>25</td>
<td>94.3</td>
<td>95.1</td>
<td>97.1</td>
</tr>
<tr>
<td>30</td>
<td>95.6</td>
<td>95.1</td>
<td>97.2</td>
</tr>
</tbody>
</table>

Table 1. Word accuracy (%) at different SNR levels. “A” is an acoustic only system, whereas “A+V” is a bimodal system using concatenated audio and visual features. “CHMM” indicates the recognition results of the CHMM system.

Comparing with an acoustic-only ASR system trained using only the audio channel of the same database, the CHMM bimodal system consistently demonstrates improved noise robustness at all SNRs. Furthermore, the CHMM system outperforms our earlier bimodal speech recognition system by a clear margin. In the earlier system, the two modalities are integrated by concatenating the audio and visual features and modeled by a single-chain HMM.

6. CONCLUSIONS

We considered the coupled hidden Markov models in the context of bimodal speech recognition. The framework is applied to integrate the audio and visual modalities in a bimodal speech recognition system. Experimental results confirm the superior sensor fusion capabilities of the CHMM framework.

7. ACKNOWLEDGMENTS

We acknowledge the support from the Army Research Laboratory. This work was also supported in part by National Science Foundation Grant CDA 96-24396.

8. REFERENCES