Audio-Visual Speech Fusion Using Coupled Hidden Markov Models

Stephen M. Chu
IBM T. J. Watson Research Center
1101 Kitchawan Road, NY 10598
schu@us.ibm.com

Thomas S. Huang
Beckman Institute and Department of ECE
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

1. Introduction

The fusion of audio and visual speech is an instance of the general sensory fusion problem. The sensory fusion problem arises in the situation when multiple channels carry complementary information about different components of a system. In the case of audio-visual speech, the two modalities manifest two aspects of the same underlying speech production process. From an observer's view, the audio channel and the visual channel represent two interacting stochastic processes. We seek a framework that can model the two individual processes as well as their dynamic interactions.

One interesting aspect of audio-visual speech is the inherent asynchrony between the audio and visual channels. Most *early integration* approaches to the fusion problem assume tight synchrony between the two. However, studies have shown that human perception of bimodal speech does not require rigid synchronization of the two modalities. Furthermore, humans appear to use the audio-visual asynchronies as multimodal features. For example, it is well known that the voice onset time is an important cue to the voicing feature in stop consonants. This information can be conveyed bimodally by the interval between seeing the stop release and hearing the vocal cord vibration. Therefore, a successful fusion scheme should not only be tolerant to asynchrony between the audio and visual cues, but also be apt to capture and exploit this bimodal feature.

2. Overview of the work

It is possible to just use conventional hidden Markov model (HMM) to model and fuse multiple information sources. This can be accomplished by attaching multiple observation variables to the state variable, with each observation variable corresponding to one of the information sources. Because both channels share the single state variable, this approach in effect assumes the two information sources always evolves in lockstep. Therefore, it is not able to model asynchronies between the two channels.

An interesting instance of the dynamic Bayesian Networks is the *coupled hidden Markov model* (CHMM). The name CHMM comes from the fact that these networks can be viewed as parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional probabilities. An n-chain CHMM has n hidden nodes in a time slice, each connected to itself and its nearest neighbors in the next time slice. For the purpose of audio-visual speech modeling, we considered the case of n=2, or the 2-chain CHMM. Figure 1 shows the inference graph of such a model.

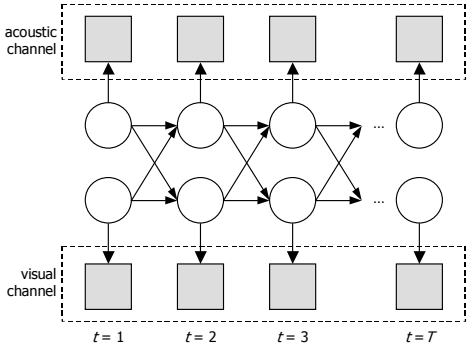

Figure 1. Audio-visual fusion using CHMM

There are two state variables in the graph. The state of the system at certain time slice is jointly determined by the states of these two multinomial variables. More importantly, the state of each state variable is dependent on both of its two parents in the previous time slice. This configuration essentially permits unsynchronized progression of the two chains, while encouraging the two sub-processes to assert temporal influence on each other's states. Note that the Markov property is not jettisoned by introducing the additional state variable and the directed links. Given the current state of the system, the future is conditionally independent of the past. Furthermore, given its two parents, a state variable is also conditionally independent of the other state variable.

In the context of audio-visual speech fusion, the audio and visual channels are associated with the two state variables respectively through the observable nodes. Inter-channel asynchrony is allowed. The overall dynamics of the audio-visual speech is determined by both modalities.

In general, the time complexity of exact inference for DBN is exponential in the number of state variables per time slice. For systems with large number of state variables, exact inference quickly becomes computationally intractable. Consequently, much attention in the literature has been paid to approximation methods that aim to solve the general problem. Existing approaches include the *variational methods* [4] and the *sampling methods* [5]. However, these methods usually exhibit nice computational properties in an asymptotic sense. When the number of states is very small, the computational overhead embedded in the approximation method is often large enough to offset the theoretical reduction in time complexity. In this situation, the approximation becomes superfluous and exact inference becomes more desirable. In this work, we propose a model transformation strategy that facilitates inference and learning in CHMM.

3. Results

Evaluation of the bimodal speech recognition system was performed on an audio-visual speech dataset [1] collected by Chen *et al.* at the Carnegie Mellon University. The visual features were derived from the lip-tracking data provided with the bimodal speech dataset. The results are summarized in Table 1.

Table 1. Summary of recognition results (measured in %word accuracy). 'A' indicates the audio-only system; 'V' indicates the visual-only system; 'A+V' indicates the bimodal system using early integration; and 'CHMM' indicates the CHMM-based system.

SNR	10dB	20dB	30dB
A	4.03	43.61	99.10
V	42.95	42.95	42.95
A+V	10.58	72.79	99.74
CHMM	35.32	86.58	93.32

An important cue the visual modality provides in bimodal speech perception is the information about boundary locations of the speech units within an utterance. We computed forced alignment of a speech segment in the 20 dB test set using both the acoustic only recognizer and the CHMM-based bimodal recognizer. The results are illustrated in Figure 2. The two subplots on the bottom show the word boundaries superimposed with the speech waveform. The upper one is the alignment obtained using audio-visual CHMM; the lower one shows the alignment obtained using acoustic only HMM. The three subplots on the top display the static visual features used in the bimodal system. All five plots are time-aligned so that the correspondence among them can be visualized.

From the plot, we see that the audio-only recognizer almost always give the incorrect end-of-word boundary at

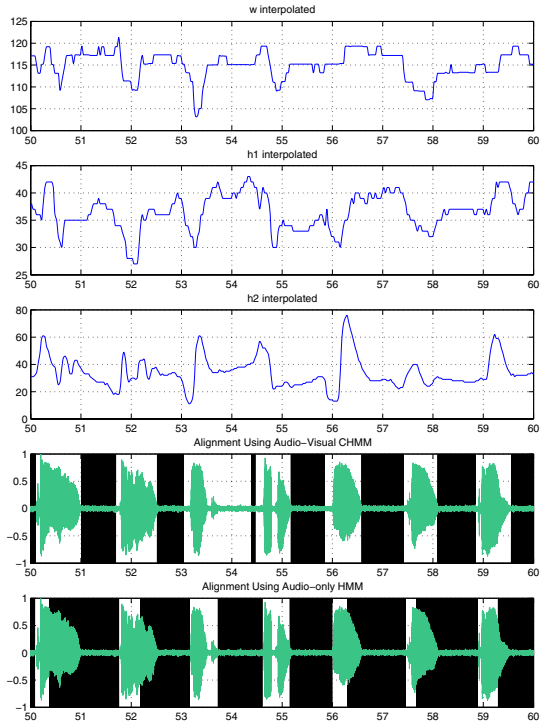

Figure 2. Forced alignment using audio only HMM and audio-visual CHMM

this noise level. In contrast, the bimodal system was able to precisely determine the end boundaries in 6 out of 7 cases. It is interesting to observe that the bimodal recognizer consistently introduced a lead-time before the audible starting point of a word. This observation is consistent with the finding from human speech perception, that the visual speech usually leads the visual speech by a varying time window.

References

[1] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18(1), pp. 9-21, 2001.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2(3), pp. 141-150, 2000.

[3] Z. Grahramani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information* (C. L. Giles and M. Gori, eds.), Lecture notes in artificial intelligence, Springer-Verlag, 1997.

[4] M. Jordan, Z, Grahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.

[5] D. J. C. Mackay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.