BIMODAL FUSION IN AUDIO-VISUAL SPEECH RECOGNITION

Xiaozheng Zhang, Russell M. Mersereau, Mark Clements

ABSTRACT

Extending automatic speech recognition (ASR) to the visual modality has been shown to greatly increase recognition accuracy and improve system robustness over purely acoustic systems, especially in acoustically hostile environments. An important aspect of designing such systems is how to incorporate the visual component into the acoustic speech recognizer to achieve optimal performance. In this paper, we investigate methods of integrating the audio and visual modalities within HMM-based classification models. We examine existing integration schemes and propose the use of a coupled hidden Markov model (CHMM) to exploit audio-visual interaction. Our experimental results demonstrate that the CHMM consistently outperforms other integration models for a large range of acoustic noise levels and suggest that it better captures temporal correlations between the two streams of information.

1. INTRODUCTION

Speech is bimodal in nature: there is both an audio and a visual component. While the audio signal is a major source of speech information, the visual component is considered to be a valuable supplementary information source in noisy environments because it remains unaffected by acoustic noise. One major advantage of the visual component is that it carries information that is complementary to the acoustic signal — many phonemes that are acoustic confusable are easily distinguished visually. Perceptual studies [1] have shown that using the visual information leads to more accurate speech perception even in noise-free environments.

Purely acoustic speech recognizers work quite well for many applications, but their performance degrades significantly when the speech is corrupted by acoustic noise. In order to overcome their limitation, much research has been directed toward systems for noisy speech environments that use noise robust methods such as feature-normalization algorithms, microphone arrays, and representations based on human hearing [2, 3, 4].

Another way to increase the robustness against acoustic noise is to incorporate the visual modality. Since the pioneering work by Petajan in 1984 [5], automatic speechreading through its use of visual information to augment acoustic counterpart has drawn much attention [6, 7]. Various automatic speechreading systems developed so far demonstrated that the visual modality yields information that is not always present in the acoustic signal and enables improved recognition accuracy over purely ASR systems, especially in environments corrupted by acoustic noise and multiple talkers.

Automatic speechreading mainly involves two research areas — one is the design of a visual front end where visual speech features are accurately and reliably extracted, the other is the development of an effective strategy to integrate the two separate information sources. In our previous studies [8, 9], we addressed the first issue. In this paper, we focus on combining the audio-visual modalities to improve speech recognition performance. Most current speech recognition systems employ HMMs to model feature sequences. In this paper we examine all audio-visual integration schemes within HMM-based classification models.

2. BIMODAL FUSION

Existing audio-visual recognizers fuse the information from the acoustic and visual channels in different ways. Two main integration models have been reported in the literature: early integration and late integration [6].

In early integration, the fusion process takes place prior to any classification. The integration forms composite audio-visual feature vectors (often by simply concatenating the vectors from each modality) and the recognition is performed in the audio-visual feature space. Late integration uses two parallel unimodal classifiers, one for audio and one for video. The final recognition is based on the combined results from the two modalities. Figs. 1 and 2 show the HMM topologies for early integration and late integration, respectively. Here, S_i represents the hidden state variable, and O_i the sequence of feature vectors. There is no general consensus as to which model is the best in achieving speech recognition, though evidence in human speech perception suggests that the fusion takes place somewhere between the peripheral input level and the categorical level [10].

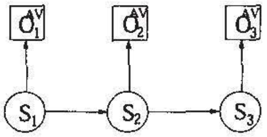

Fig. 1. HMM topology in early integration.

Late integration offers several advantages over early integration because its implementation is simple and it does not require precise synchronization of the acoustic and visual features. In late integration, each independent subsystem can be developed and trained separately. However, the use of separate models assumes conditional independence between the two feature sets and therefore it fails to model the correlations between the visual and acoustic channels. On the other hand, early integration provides a more general model by integrating the two components before recognition. However, the classification is based on a single HMM on the concatenated vectors of audio and visual features. It forces the same state sequences upon the audio and visual components and this does not correspond to the way that people talk. Often the lips start moving before voicing commences. Therefore an early integration model restricts the asynchrony between the two streams of information which occurs in speech production.

Based on the above analysis, early and late integration models are not suitable for the composite modeling of multiple time-series. We therefore propose the use of a more generalized model — the coupled hidden Markov model (CHMM) to model the audio-visual interaction for speech recognition. The coupled hidden Markov model was first introduced by Brand in 1996 and was successfully used for modeling Tai Chi gestures [11]. In a coupled HMM, as shown in Fig. 3, the traditional left-right HMM is expanded to a model containing two Markov chains, representing the audio and visual channels. The coupling between the two subprocesses is introduced by conditional probabilities between the hidden state variables $\Pr(S_t^A|S_{t-1}^A, S_{t-1}^V)$ and $\Pr(S_t^V|S_{t-1}^A, S_{t-1}^V)$. On one hand, this architecture relaxes the restriction of the early integration by allowing asynchrony between the two channels. On the other hand, unlike late integration, it incorporates temporal coupling terms across the two subsystems. Intuitively this model better captures the interprocess influences between multiple processes.

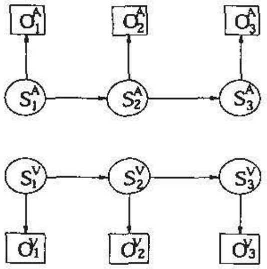

Fig. 2. HMM topology in late integration.

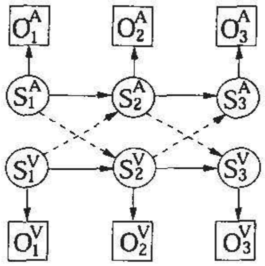

Fig. 3. Coupled hidden Markov model.

Exact inference through naive inference reduces the two-modal coupled HMM to an ordinary HMM by performing a cartesian product of all sub-HMMs' state spaces. This results in an exponentially increased state space dimension. Assuming that each HMM has a state space of dimension k, the resulting HMM would require k^2 distinct states to model this system. This representation is not only computationally inefficient, but it requires tremendously large training data to achieve parameter estimation accuracy.

To solve the inference problem in a coupled HMM, we employ the approximate approach proposed by Boyen and Koller [12, 13]. The key ingredient of the BK algorithm is the propagation of an approximate probability distribution over the entire system using factorized products over independent clusters. The accumulated error arising from the repeated approximation was proved to remain bounded indefinitely over time. The BK algorithm has been shown to be an efficient approach to solving inference problems in general dynamic Bayesian networks.

For learning parameters in the CHMM, forward and backward variables are first approximated. The BK algorithm represents the forward variable $\alpha_t = \Pr(i_t, \mathbf{o}_1, \cdots, \mathbf{o}_t)$ as a product of marginals over two subprocesses $\alpha_t \approx \Pr(i_t^A, \mathbf{o}_1^A, \cdots, \mathbf{o}_t^A)\Pr(i_t^V, \mathbf{o}_1^V, \cdots, \mathbf{o}_t^V)$. The approximated forward variable at time t is then propagated through the transitional model and conditioned on evidence at time $t+1$ using the junction tree algorithm [14]. To allow the algorithm to continue, the forward variable at $t+1$ is approximated using one that admits a compact representation by computing marginals over each cluster. The same procedures can be applied to approximating the backward variable β_t. These two variables are then used in an EM algorithm that learns the model in an iterative manner.

3. EXPERIMENTS

We perform experiments on audio-visual speech recognition using the audio-visual database from Carnegie Mellon University [15]. This database includes ten test subjects (three females, seven males) speaking 78 isolated words repeated ten times. These words include numbers, weekdays, months, and others that are commonly used for scheduling applications.

In the visual subsystem, we use six geometric features as defined in Fig. 4: mouth width (w_2), upper/lower lip width (h_1, h_3), lip opening height/width (h_2, w_1), and the distance between the horizontal lip line and the upper lip (h_4). Besides the geometric dimensions of the lips, we include two other features characterizing the visibility of the tongue and teeth. These two features are measured by the number of pixels of the tongue and tooth colors within the lip inner contour. Delta features are also included in the visual features, forming a 16-dimensional feature vector. They are obtained by using a regression formula drawing over a few frames before and after the current frame. The visual feature vectors were preprocessed by normalizing against the average mouth width w_2 of each speaker to account for the difference in scale between different speakers and different record settings for the same person.

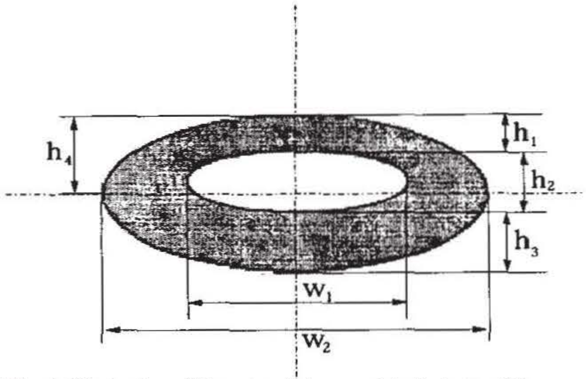

Fig. 4. Illustration of the extracted geometric features of the lips.

In the acoustic subsystem, we use 12 Mel Frequency Cepstral Coefficients (MFCCs) and their corresponding delta parameters as features — a 24-dimensional feature vector. MFCCs are derived from FFT-based log spectra with a frame period of 11 msec and a window size of 25 msec.

We conducted tests for both speaker-dependent and independent tasks. For the speaker-dependent task, the test was set up by using a leave-one-out procedure, i.e., for each person, nine repetitions were used for training and the tenth for testing. This was repeated ten times. The recognition rate was averaged over the ten tests and again over all ten speakers. For the speaker-independent task, we use different speakers for training and testing, i.e., nine subjects for training and the tenth for testing. The whole procedure was repeated ten times, each time leaving a different subject out for testing. The recognition rate was averaged over all ten speakers.

In all cases, the HMMs have ten states, and we model the observation vectors using two Gaussian mixtures for the speaker-independent task. Because of the limited training data available, we use one Gaussian mixture in the speaker-dependent case. In early integration, the classification is based on training a traditional HMM on the concatenated audio-visual observation vectors. The video has a frame rate of 33 ms. To match the audio frame rate of 11ms, linear interpolation was used on the visual features to fit the data values between the existing feature data points. In the late integration fusion, the combined score takes the following form: $\log P_{av} = \lambda \log P_a + (1 - \lambda) \log P_v$, where P_a and P_v are the probability scores of the audio and visual components and the weighting factor λ is set to 0.7 in our experiments. Model training and Viterbi decoding of the HMMs were implemented using the HTK Toolkit [16]. The BK algorithm for the coupled HMM was implemented using the Bayes Net Toolbox [17]. Prior to employing the BK algorithm, the model parameters need to be well initialized, which is essential in achieving good model estimates. For this, we apply the traditional EM algorithm on the two separate HMMs and use the model parameters trained on the separate HMMs as the initial parameters in the coupled HMM.

In the following, we present our experimental results on audio-visual speech recognition over a range of noise levels using these three models. Artificial white Gaussian noise was added to simulate various noise conditions. The experiment was conducted under a mismatched condition — the recognizers were trained at 30dB SNR, and tested under varying noise levels. Tables 1 and 2 summarize the recognition performance using the three integration schemes for the speaker-dependent and independent tasks, respectively. For comparison, the visual-only and audio-only results are also included. As can be seen, all three integration models demonstrate improved recognition accuracy over audio only performance. The coupled HMM consistently outperforms the early and late integration over a wide range of SNRs.

S.D.	v-only	a-only	early int.	late int.	CHMM
0dB	45.59	3.24	31.21	25.76	33.86
10dB	45.59	27.86	71.12	43.47	76.86
30dB	45.59	86.82	89.42	80.26	94.59

Table 1. Audio-visual speech recognition performance in the speaker-dependent mode. The numbers represent the percentage of correct recognition.

S.I.	v-only	a-only	early int.	late int.	CHMM
0dB	21.08	3.69	8.9	3.81	11.91
10dB	21.08	14.58	35.50	20.27	38.43
30dB	21.08	43.77	62.14	56.92	66.17

Table 2. Audio-visual speech recognition performance in the speaker-independent mode. The numbers represent the percentage of correct recognition.

4. SUMMARY

We proposed the use of a coupled hidden Markov model for temporal fusion of the audio and visual modalities in a speech recogntion task. We analyzed the HMM structures in conventional AV integration models — early and late integration, and argued that the coupled HMM better captures temporal correlations between audio and visual sources of information. Our experimental results verified this assumption and suggest that the coupled HMM is a better model for fusing data from multiple channels.

Acknowledgments

We would like to acknowledge the use of audio-visual data [15] from the Advanced Multimedia Processing Lab at the Carnegie Mellon University.

5. REFERENCES

[1] D. Reisberg, J. McLean, and A. Goldfield, "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by Eyes: The Psychology of Lipreading*, Barbara Dodd and Ruth Campbell, Eds., pp. 97–113. Lawrence Erlbaum Associates, London, 1987.

[2] T. Sullivan and R. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proc. IEEE ICASSP*, April 1993, pp. 91–94.

[3] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition. Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Kluwer Academic Publishers, 1996.

[4] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[5] E. D. Petajan, *Automatic lipreading to Enhance Speech Recognition*, Ph.D thesis, Univ. of Illinois, Urbana-Champaign, 1984.

[6] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series F*, Springer Verlag, 1996.

[7] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep., CLSP/Johns Hopkins University, Baltimore, 2000.

[8] X. Zhang and R. M. Mersereau, "Lip feature extraction towards an automatic speechreading system," in *Proc. IEEE Int. Conf. on Image Processing*, 2000.

[9] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer-interfaces," Submitted to EURASIP Journal on Applied Signal Processing, Special issue on Audio-Visual Speech Processing, 2002.

[10] D. Massaro, "Bimodal speech perception: A progress report," *in [6]*, 1996.

[11] M. Brand, "Coupled hidden Markov models for modeling interacting process," Tech. Rep., MIT Media Lab, TR 405, 1996.

[12] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Madison, Wisconsin, July 1998, pp. 33–42.

[13] X. Boyen and D. Koller, "Approximate learning of dynamic models," in *Proceedings of the 11th Annual Conference on Neural Information Processing Systems (NIPS-98)*, Denver, Colorado, December 1998, pp. 396–402.

[14] C. Huang and A. Darwiche, "Inference in belief networks: a procedural guide," *International Journal of Approximate Reasoning*, vol. 11, pp. 1–158, 1994.

[15] "URL: http://amp.ece.cmu.edu/projects/ audiovisual-speechprocessing," .

[16] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Ltd., Cambridge, 1999.

[17] K. Murphy, "The Bayes Net Toolbox for Matlab," in *Computing Science and Statistics: Proceedings of Interface*, 2001, vol. 33.