Multi-level Fusion of Audio and Visual Features for Speaker Identification

Zhiyong Wu[1,2], Lianhong Cai[1], and Helen Meng[2]

[1] Department of Computer Science and Technology,
Tsinghua University, Beijing, China, 100084
john.zy.wu@gmail.com, clh-dcs@tsinghua.edu.cn
[2] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
hmmeng@se.cuhk.edu.hk

Abstract. This paper explores the fusion of audio and visual evidences through a multi-level hybrid fusion architecture based on dynamic Bayesian network (DBN), which combines model level and decision level fusion to achieve higher performance. In model level fusion, a new audio-visual correlative model (AVCM) based on DBN is proposed, which describes both the inter-correlations and loose timing synchronicity between the audio and video streams. The experiments on the CMU database and our own homegrown database both demonstrate that the methods can improve the accuracies of audio-visual bimodal speaker identification at all levels of acoustic signal-to-noise-ratios (SNR) from 0dB to 30dB with varying acoustic conditions.

1 Introduction

Human speech is produced by the movement of the articulatory organs. Since some of these articulators are visible, there are inherent correlations between audio and visual speech. There is also loose timing synchronicity between them, for instance, the mouth is opened before producing speech and closed after speech is produced.

While the audio 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. Many studies have shown that the integration of audio and visual features leads to more accurate speaker identification even in noisy environments [1-3].

Audio-visual integration can be divided into three categories: feature fusion, decision fusion and model fusion [3-5]. In feature fusion, multiple features are concatenated into a large feature vector and a single model is trained [4]. However this type of fusion cannot easily represent the loose timing synchronicity between audio visual features. In decision fusion, audio and visual features are processed separately to build two independent models [5], which completely ignore the audio visual correlations. In model fusion, several models have been proposed, such as multi-stream hidden Markov model (HMM) [6], factorial HMM [6], coupled HMM [2], mixed DBN [7], etc. Multi-stream HMM and factorial HMM assume independence between audio

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 493 – 499, 2005.

and visual features. Coupled HMM and mixed DBN force audio visual streams to be in strict synchrony at model boundaries by introducing "anchor-points".

This work attempts to capture the inter-correlations between audio and visual cues as well as the loose synchronicity between them for speaker identification. We propose a new audio-visual correlative model (AVCM) to describe the above relations, which is realized using the DBN. We also explore the fusion of audio and visual evidences through a multi-level hybrid fusion architecture based on DBN, which combines model level and decision level fusion to achieve higher performance.

The outline of this paper is as follows: Section 2 gives the details of the proposed audio-visual correlative model (AVCM). Then the multi-level audio visual fusion architecture is described in section 3. Section 4 presents the experimental results and analysis showing how the proposed approaches improve the speaker identification performance. Finally, section 5 concludes the paper.

2 DBN Based Audio-Visual Correlative Model (AVCM)

Dynamic Bayesian networks are a class of Bayesian networks designed to model temporal processes as stochastic evolution of a set of random variables over time [8]. A DBN is a directed acyclic graph whose topology structure can be easily configured to describe various relations among variables. DBN offers a flexible and extensible means of modeling the feature-based and temporal correlations between audio and visual cues for speaker identification.

We propose the AVCM model as depicted in figure 1. It illustrates a whole sentence model that consists of several words. The square nodes represent discrete variables. The round nodes represent continuous variables. The hollow nodes represent hidden variables and the shaded nodes are observed. The upper part of the model describes the audio stream (audio sub-model) and the lower part describes the video stream (video sub-model). The labeled nodes include:

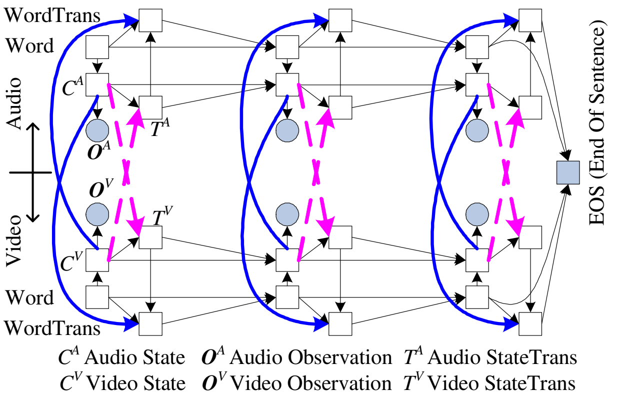
WordTrans
Word
Audio
C^A
T^A
O^A
O^V
T^V
Video
C^V
Word
WordTrans
EOS (End Of Sentence)
C^A Audio State O^A Audio Observation T^A Audio StateTrans
C^V Video State O^V Video Observation T^V Video StateTrans

Fig. 1. Audio-visual correlative model (AVCM) based on DBN

~ the "Word" node stands for the current word which is determined by the sentence;
~ the "State" node (C^A, C^V) indicates the current state and is determined by "Word";
~ the "Observation" node ($\boldsymbol{O}^A$, $\boldsymbol{O}^V$) represents the audio or visual observations;
~ the "State Trans" node (i.e. state transition, T^A, T^V) indicates when the current state ends and switches to the next state;
~ the "Word Trans" node may take the values *true* or *false* to respectively denote whether there is a word transition and is dependent on the "State Trans" node;
~ the "EOS" node (End Of Sentence) represents the end of the whole sentence.

Inter-node dependencies modeled by the proposed AVCM include:

~ the "State Trans" nodes of the audio and video streams are dependent on their "State" nodes from both two models, which describe the inter-correlations between the two streams. This is shown by the thick dashed arrows in figure 1;
~ the "Word Trans" nodes are also dependent on the "State" nodes from both audio and video streams, which capture the loose timing synchronicity in between the two streams. This is shown by the thick solid arrows in figure 1.

The proposed AVCM model differs from previous approaches such as [2] and [7], where the loose timing synchronicity between audio and video streams is restricted by "anchor-points" at word boundaries. In AVCM, audio and video streams have their own independent "Word" and "Word Trans" nodes. Furthermore, the "Word Trans" node of the audio stream is dependent on the "State" node of the video stream, and vice versa. This models the loose synchronicity in between two streams and brings about performance improvements, as will be discussed later.

3 Multi-level Fusion of Audio and Visual Evidences

Incorporation of bimodal, correlated audio-visual features should achieve speaker identification performance that is superior to mono-modal systems. This is because the two modalities, if modeled properly, can complement and reinforce each other. However, under some conditions, the performance of AVCM is not as good as that of decision fusion (see point 4 in section 4.2). Similar observations are reported in [2].

In view of the advantages of model fusion and decision fusion, we proposed a multi-level fusion strategy via DBN, as illustrated in figure 2. There are three models altogether: the audio-only model, the video-only model and the AVCM model that performs model-based audio-video fusion. These three models are further combined by means of decision-level fusion to deliver the final speaker identification result.

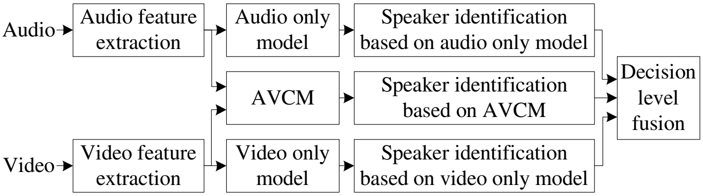
Audio
Audio feature extraction
Audio only model
Speaker identification based on audio only model
AVCM
Speaker identification based on AVCM
Decision level fusion
Video
Video feature extraction
Video only model
Speaker identification based on video only model

Fig. 2. Strategy for audio-visual multi-level fusion

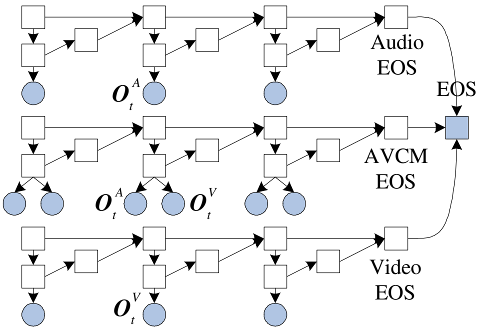

Fig. 3. DBN for audio-visual multi-level fusion

Decision-level fusion for the three models is also achieved through the use of DBN, by virtue of its extensibility. This is shown in figure 3, in which the EOS node of AVCM model (AVCM EOS), audio-only model (Audio EOS) and video-only model (Video EOS) are connected to the global "EOS" node. The EOS nodes of the three models are hidden and the global "EOS" node is observable.

Equation 1 shows the mathematical formula we use for multi-level fusion,

$$P(\boldsymbol{O}^A, \boldsymbol{O}^V \mid M^A, M^V, M^{AV}) = [P(\boldsymbol{O}^A \mid M^A)]^{\lambda_A} \, [P(\boldsymbol{O}^V \mid M^V)]^{\lambda_V} \, [P(\boldsymbol{O}^A, \boldsymbol{O}^A \mid M^{AV})]^{\lambda_{AV}}. \tag{1}$$

where $P(\boldsymbol{O}^A \mid M^A)$ is the recognition formula for audio-only model M^A of audio observation $\boldsymbol{O}^A$, and $P(\boldsymbol{O}^V \mid M^V)$ is the formula for video-only model M^V of video observation $\boldsymbol{O}^V$, and $P(\boldsymbol{O}^A, \boldsymbol{O}^V \mid M^{AV})$ is the formula for AVCM model M^{AV}. λ_A, λ_V and λ_{AV} are stream exponents (fusion weights) for audio-only, video-only and AVCM model, which encode the relative reliability of the models and can be varied according to ambient noise conditions (i.e. SNR). When the SNR is high, AVCM should be more reliable than mono-modal models and should carry higher weights. When SNR is low, the reliability of the audio-only and AVCM models degrade, hence the video-only model should carry the highest weight. The proposed multi-level fusion strategy combines the advantages of model fusion and decision fusion and has the potential of achieving performance improvement.

The estimation of the fusion weights is a key issue. We enforce the parameter constraints of $\lambda_A+\lambda_V+\lambda_{AV}=1$, $\lambda_A=\lambda_{AV}$ and $\lambda_A,\lambda_V,\lambda_{AV}\geq 0$. In addition we impose $\lambda_A=\lambda_{AV}$ by assuming that the performance of both audio-only and AVCM models are equally dependent on the quality of the acoustic speech. We then use a novel methodology known as support vector regression (SVR) to estimate the fusion weights directly from the original audio features [9].

4 Experiments

We perform the text-prompted speaker identification experiments to evaluate the performance of various models including audio-only, video-only, decision fusion, feature fusion, coupled HMM (CHMM), AVCM and multi-level fusion.

The experiments are conducted on the audio-visual bimodal database from Carnegie Mellon University (CMU database) [10] as well as our own homegrown database. The CMU database includes 10 subjects (7 males and 3 females) speaking 78 isolated words repeated 10 times. These words include numbers, weekdays, months, and others that are commonly used for scheduling applications. Our homegrown database includes 60 subjects (38 males and 22 females, aged from 20 to 65) with each subject speaks 30 connect-digit words (the digit length differs from 2 to 6), and each utterance is repeated three times at intervals of 1 month.

The acoustic front-end includes 13 Mel frequency cepstral coefficients (MFCCs) and 1 energy parameter (with frame window size of 25ms and frame shift of 11ms) together with their delta parameters. Hence the audio feature vector has 28 dimensions. The visual front-end includes mouth width, upper lip height, lower lip height [10] and their delta parameters. Thus the visual feature vector has 6 dimensions. The video frame rate is 30 frames per second (fps), which is up-sampled to 90fps (11ms) by copying and inserting two frames between each two original video frames.

Artificial white Gaussian noise was added to the original audio data (SNR=30dB) to simulate various SNR levels. The models were trained at 30dB SNR and tested under SNR levels ranging from 0dB to 30dB at 10dB intervals. We applied cross-validation for every subject's data, i.e. 90% of all the data are used as training set, and the remaining 10% as testing set. This partitioning was repeated until all the data had been covered in the testing set.

Table 1. Accuracies (%) of speaker identification under different SNR on CMU database

audio signal-to-noise-ratio (SNR)	30dB	20dB	10dB	0dB
video-only	77	77	77	77
audio-only	100	64	22	17
feature fusion	99	85	30	20
decision fusion	100	86	78	78
CHMM	100	88	79	60
AVCM	100	92	79	65
multi-level fusion	100	93	81	79
fusion weight for multi-level fusion ($\lambda_A=\lambda_{AV}$)	0.4	0.32	0.1	0.01

Table 2. Accuracies (%) of speaker identification under different SNR on our own database

audio signal-to-noise-ratio (SNR)	30dB	20dB	10dB	0dB
video-only	74	74	74	74
audio-only	99	59	20	15
feature fusion	99	81	26	18
decision fusion	100	83	76	75
CHMM	100	85	77	57
AVCM	100	89	78	61
multi-level fusion	100	91	80	77

All the tested models are implemented as DBNs. A DBN is developed for each word, with a left-to-right no skipping topological structure. The number of the transited state is always equal to or 1 greater than the original state. The audio sub-model has 5 states, and video sub-model has 3 states, each state is modeled using the Gaussian mixture model (GMM) with 3 mixtures. During speaker identification, the words' DBNs are connected to form a whole sentence model, which is then used to identify the speakers. The DBNs are implemented using the GMTK toolkit [11].

The identification accuracies from all the testing data are averaged and reported as the final result. The results on CMU database are summarized in table 1. The experiments are also conducted on our own homegrown database with a larger number of speakers and results are summarized in table 2. Main observations include:

(1) Feature fusion performs worse than other fusion methods, even achieving accuracies lower than video-only model when SNR≤10dB. The main reason is due to the misalignment between audio and video streams;

(2) Decision fusion achieves lower accuracy than CHMM and AVCM when SNR≥10dB. However, when SNR<10dB, the accuracy is higher than CHMM and AVCM, as shown by the shaded table cells;

(3) The AVCM model proposed in this paper has higher accuracy than CHMM, because it describes both inter-correlations and loose timing synchronicity between audio and video features;

(4) Because of the misalignment during model training, when SNR<10dB, the performance of the AVCM and CHMM degrades and the accuracy is even lower than the video-only model;

(5) The audio-visual multi-level fusion strategy has solved the problem in (4) well. Best identification performance is obtained even when SNR=0dB. It can also be seen from both tables that better results are also obtained when SNR is 10dB and 20dB than the AVCM model. This is because that the multi-level fusion strategy combines the results from audio-only, video-only and AVCM model, and the results of audio and video model are supplementary to that of AVCM;

(6) From the results of table 1 and table 2, we can see that the performance of the speaker identification degrades a little with larger speaker numbers, but the conclusions from (1) to (5) can still be drawn. This indicates that the proposed model based on DBN has good extensibility for different databases.

5 Conclusions

This paper investigates the correlations between audio and visual features. A new audio-visual correlative model (AVCM) based on dynamic Bayesian network (DBN) is proposed, which describes both the inter-correlations and the loose synchronicity between audio and visual streams. The experiments on the audio-visual bimodal speaker identification demonstrate that the AVCM model improves the identification accuracies compared to the previous methods.

We also propose a DBN based audio-visual multi-level fusion strategy, which combines the results of audio-only, video-only and AVCM models through decision-level fusion. Experiments on both CMU database and our own homegrown database show that the proposed strategy integrates the advantages of both model level and decision level fusion and achieves the best accuracies of speaker identification at all levels of acoustic signal-to-noise ratio (SNR), ranging from 0dB to 30dB.

Acknowledgments

This work is supported by the research fund from the National Science Foundation of China (NSFC) under grant No. 60433030 and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012 and N-CUHK417 /04.

References

1. Senior, A., Neti, C., Maison, B.: On the use of visual information for improving audio-based speaker recognition. In: Proc. Audio-visual Speech Processing Conf. (1999) 108–111
2. Nefian, A.V., Liang, L.H., Fu, T.Y., Liu, X.X.: A Bayesian approach to audio-visual speaker identification. In: Proc. 4th International Conf. Audio- and Video-based Biometric Person Authentication, Vol. 2688 (2003) 761–769
3. Chibelushi, C.C., Deravi, F., Mason, J.S.D.: A review of speech-based bimodal recognition. IEEE Trans. Multimedia 4 (2002) 23–37
4. Chibelushi, C.C., Mason, J.S.D., Deravi, F.: Feature-level data fusion for bimodal person recognition. In: Proc. 6th IEEE International Conf. Image Processing and its Applications. IEE, Stevenage (1997) 399–403
5. Chatzis, V., Bors, A.G., Pitas, I.: Multimodal decision-level fusion for person authentication. IEEE Trans. Syst. Man Cybern. A 29 (1999) 674–680
6. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia 2 (2000) 141–151
7. Gowdy, J.N., Subramanya, A., Bartels, C., Bilmes, J.: DBN based multi-stream models for audio-visual speech recognition. In: Billene, M. (ed.): Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing, Vol. 1. IEEE, Canada (2004) 993–996
8. Dean, T., Kanazawa, J.: Probabilistic temporal reasoning. In: Proc. 7th National Conf. Artificial Intelligence (1988) 524–528
9. WU, Z.Y.: Audio-visual bimodal modeling for speaker identification and visual-speech synthesis. Ph.D. Dissertation. Department of Computer Science and Technology, Tsinghua University, Beijing, China (2005)
10. Chen, T.: Audiovisual speech processing. IEEE Trans. Signal Processing 18 (2001) 9–21
11. Bilmes, J., Zweig, G.: The graphical models toolkit: An open source software system for speech and time-series processing. In: Proc. IEEE International Conf. Acoustics, Speech and Signal Processing, Vol. 4. IEEE, Florida (2002) 3916–3919