Dynamic Bayesian Networks for Audio-Visual Speaker Recognition

Dongdong Li, Yingchun Yang, and Zhaohui Wu

Department of Computer Science and Technology,
Zhejiang University, Hangzhou 310027, P.R. China
{lidd, yyc, wzh}@cs.zju.edu.cn

Abstract. Audio-Visual speaker recognition promises higher performance than any single modal biometric systems. This paper further improves the novel approach based on Dynamic Bayesian Networks (DBNs) to bimodal speaker recognition. In the present paper, we investigate five different topologies of feature-level fusion framework using DBNs. We demonstrate that the performance of multimodal systems can be further improved by modeling the correlation of between the speech features and the face features appropriately. The experiment conducted on a multi-modal database of 54 users indicates promising results, with an absolute improvement of about 7.44% in the best case and 3.13% in the worst case compared with single modal speaker recognition system.

1 Introduction

Dynamic Bayesian Networks (DBNs) [1] are knowledge representation schemes that can characterize probability relationships among temporal data and make exact or approximate inferences. Some prior knowledge (e.g. gender, noise) can be described by DBNs in a convenient way [2]. A sea of previous research revealed the power of DBNs in fusing visual and audio sensors cues with contextual information and expert knowledge both for speaker detection and other similar applications [3, 4]. It is assumed that DBN is an instrumental tool for information fusion.

D. Li et al. [5] introduced the DBNs to the audio- visual speaker recognition with a specific topology of feature-level fusion. In this paper, we further discuss the topologies by investigate the correlation between multi- features derived from different modalities. Five topologies are explored to model the speech data and face data. This paper is organized as follows: we give a brief description of the architecture of the audio-visual speaker system in Section 2. In Section 3, we give a detailed illustration of the five types of topologies for the feature-level fusion using Dynamic Bayesian Network. The data set is presented in Section 4 with the experimental setup. And the comparisons between the audio-visual speaker based on the proposed topologies and single modal biometrics system with speech features or face features will be presented in this section as well. Section 5 serves as a conclusion.

2 Biometrics Fusion Architecture

The task of identification is to determine if the speaker is a specific one in the group of enrolled users given his utterance. The speech sequences and face image sequences

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 539 – 545, 2005.

are processed by feature extractors. The speech features and the face features are integrated to make a DBN model in this framework. Input speech and face data are matched with DBN models. The final decision is made as the highest scoring decision procedure applied to the DBN matcher module outputs.

In the voiceprint feature extraction, the hamming window is 32 mm and the frame shift is 16mm. The silence and unvoiced segments are discarded based on an energy threshold. The feature vectors are composed by 16 MFCC and their delta coefficients. The face feature extraction method is based on standard Principal Component Analysis (PCA) [6]. The 32 largest eigenvectors are taken from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns to form the feature vector.

3 DBN Based Feature-Level Fusion

Dynamic Bayesian Networks are a special case of singly connected Bayesian networks specifically aimed at time series modeling. Consider in a DBN, every observation V representing the speech features is conditionally dependent on a state variable X. If an N slice data, $V = \{v_1, v_2, ..., v_N\}$, corresponds to a series of states, $X = \{x_1, x_2, ..., x_N\}$, then the conditional probability can be represented as $P(v_n \mid x_n)$. We then attempted to incorporate additional face features, $F = \{f_1, f_2, ..., f_N\}$, into the speaker recognition process.

3.1 Five Topologies

We hypothesize that the face features have a certain relationship with the speech features and the hidden state variables. There are three situations for hidden variables: 1) only have relationship with the speech variables; 2) only have relationship with the face nodes; 3) have correlation with both of them. Similarly, there are two relations between speech variables and face ones, interrelated and irrespective. As a consequence, six (3*2) cases are generated according to the relationship among face features, speech features and hidden state variables. Here we propose five types of topologies that can be used for fusion within the framework of DBN. The case of different state affects speech feature and face information respectively which separates these two kinds of data thoroughly is out of consideration.

In each topology, we conform to standardized measurements: shading nodes are observed; clear nodes are hidden. $X_t^i, t = 1,2,...,T, i = 1,2,..N$ are the hidden nodes with discrete values, N is the number of hidden nodes in one time slice. The observed nodes V_t , $t = 1,2,...,T$, represent the speech features. The observed nodes $F_t, t = 1,2,...,T$, represent the face features. Here T is the length of time slices. V_t and F_t satisfy Gaussian distributions.

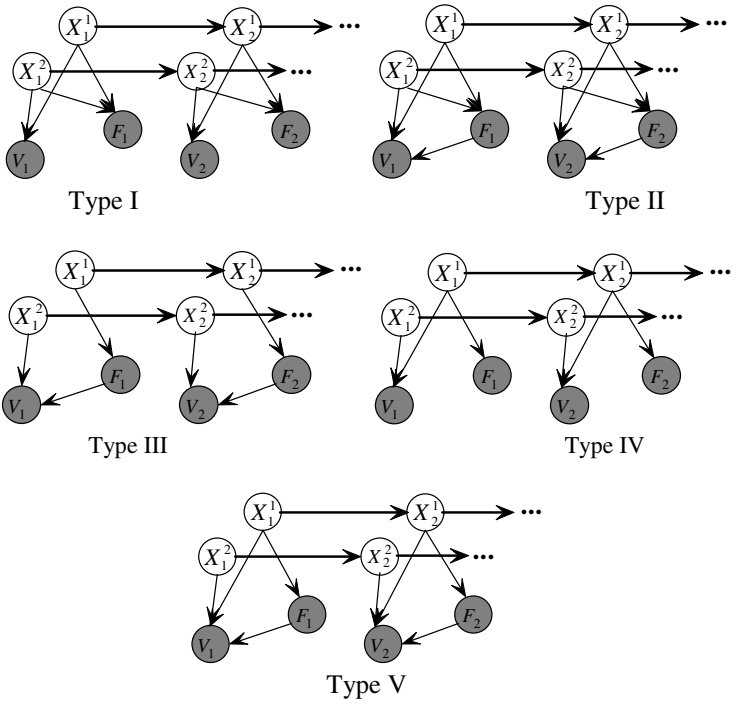

Fig. 1. Five topologies explored for audio-visual speaker recognition. Type I): Identical States and Independent; Type II): Identical States and Dependent; Type III): Different States and Dependent; Type IV): Mixed States and Independent; Type V): Mixed States and Dependent.

The five types of fusion topologies are depicted as follows:

I. Identical States and Independent: The face features are only connected to hidden state variables. Put it another way, both the face features and the speech features are activated by the change of the same hidden state variables. But they themselves have no relation, see Figure 1. Then the joint probability is $\prod_{i=1}^{2} P(v_n \mid x_n^i) \prod_{i=1}^{2} P(f_n \mid x_n^i)$. The topology of this type is the same as in [5].

II. Identical States and Dependent: The face features are connected to speech features and hidden state variables. In other words, observations of the speech features are affected by not only the hidden state variables but also the face features. Then the joint probability is $P(v_n \mid f_n) \prod_{i=1}^{2} P(v_n \mid x_n^i) \prod_{i=1}^{2} P(f_n \mid x_n^i)$.

III. Different States and Dependent: The face features are only connected to speech features. That is to say, the speech features and the face features are controlled by different hidden state variables respectively. However, the face features still has some effect on the speech features. Then the joint probability is $P(v_n \mid f_n) P(v_n \mid x_n^1) P(f_n \mid x_n^2)$.

IV. Mixed States and Independent: The face features are only connected to part of hidden state variables. In this case, different hidden state variables have different relationships with the speech features and the face features. The face features are not affected by all the hidden state variables. Then the joint probability is $P(f_n \mid x_n^1)\prod_{i=1}^{2} P(v_n \mid x_n^i)$.

V. Mixed States and Dependent: The face features are connected to speech features and part of hidden state variables. That is, observations of the speech features are controlled by both the hidden state variables and the face features. But the face features are only affected by part of the hidden state variables. Then the joint probability is $P(v_n \mid f_n)P(f_n \mid x_n^1)\prod_{i=1}^{2} P(v_n \mid x_n^i)$.

3.2 Training and Testing

As in the case of Dynamic Bayesian Networks applied to speaker identification, one may be interested in the following tasks:

Training: Each speaker is modelized by a DBN, and all the models are trained independently in our speaker identification task. We assume the structure of the DBNs be known for simplicity so only the parameters of a DBN are need to be estimated given a sequence of observations so as to model the data in the most appropriate way. The log-likelihood of the training set $C = \{V_1, ...V_M, F_1, ..., F_M\}$ can be calculated as:

$$\begin{aligned} L &= \log \prod_{m=1}^{M} \Pr(Y_m \mid G) \\ &= \sum_{i=1}^{N}\sum_{m=1}^{M} \log P(X_i \mid parent(X_i), V_m, F_m) \end{aligned} \tag{1}$$

Here G is a DBN model with N variables. These marginal posterior probability terms are computed in the inference engine. The computed marginal posterior probability can be used for the expected counts in expectation-maximization (EM) training for learning the mean μ, and the covariance Σ in the case of conditional linear Gaussian distributions, tailored to our needs of speaker identification.

Testing: The testing procedure of speaker identification is concerned with determining the right person whose features best match the features of the person to identify from a closed-set, given a set of observation. The speaker $\hat{i}$ whose DBN model M_i maximizes the posterior probability $p(M_i \mid C)$ is the identified one.

According to the Bayesian rule,

$$p(M_i \mid C) = p(C \mid M_i) * p(M_i) / p(C) \tag{2}$$

The values of $p(M_i) / p(C)$ for all speaker models M_i are treated to be equal as no prior knowledge about the probability could be retrieved. For simplicity, the decision rule can be formulized as:

$$\hat{i} = \arg_i \max\ p(C \mid M_i),\ \ i = 1,2,...N, \tag{3}$$

The posterior probabilities $p(C \mid M_i)$ can be achieved by computing the joint probability using (1).

4 Experiment and Discussion

4.1 Database

We have built a collected multi-modal corpus of 54 people (17 females and 37 males) to evaluate our system [5]. Each visitor is subjected to two recordings: a speech shot and face recording.

The speech content is varied and enormous, including Mandarin, Dialect and English in terms of language types, and prompted texts and free talks in terms of speech types. The corpus contains 54 subjects with 216 images and 2916 sentences.

The utterance set is divided into seven sessions such as personal information, mandarin digits, dialect digits, English digits, province phrase, paragraph and free talk. The image set consist of frontal images and side profile images with 4 shots for every visitor, 2 frontal ones, 2 side ones. Recording is made in an office with a low level of acoustic noise and sufficient lighting. In this way, we obtain a corpus of 54 subjects, with 4 face images and 54 utterances from per subject.

4.2 Setup

We also compare the feature-level DBN fusion method with the feature concatenation method and the single modal recognition system using acoustic features or facial features.

In the baseline strategy, the speaker recognition expert and face recognition expert are evaluated. The approach to recognize the speaker identity is based on the use of the MFCC as the parameters and the Dynamic Bayesian Networks (DBNs) for the classification task. All the settings are the same as described in [7]. BNT toolkit [8] is used as the interface in our source code. The face identification system uses the Eigenface method as the face matcher. Images are compared by means of their corresponding feature vectors extracted as described in Chapter 2.

In the feature concatenation method, the speech features and the face features are extracted using the same approaches as mentioned in baseline strategy. The 32-dimensional face features are then appended to the 32-dimensional speech features directly resulting in a 64-dimensional feature vector. The restructured features are modeled using DBNs, with the same parameters in speaker recognition expert.

The proposed topologies are evaluated as the third setup. The speech features and the face features are the same as the above two setups. The speaker models are trained and tested as presented in Chapter 3.2.

4.3 Results and Discussion

In order to ascertain whether or not the method is robust with different speech contents and different speech types, we make experiments on some subsets of our multi-modal data corpus: Mandarin, Dialect, English, Phrase, and Free talk.

Generally speaking, the speech features and the face features have no direct causal relationship, but they bear some inherent relation, for these two features are produced by one and the same person. We foresee that the case of "Mixed States and Independent" would outperform other topologies with qualitative analyses. The results are listed in table 1.

Table 1. Experimental results with different speech types and contents of test sets. I stands for the identification rate. Man stands for the Mandarin type. Eng stands for the English type. Dia stands for the dialect type. Phr stands for the phrasal type. FTlk stands for free talk.

Fusion Method	I for each speech content and type (%)					
	Man	Dia	Eng	Phr	FTlk	Average
Voice Only	84.63	85.55	91.11	87.78	87.78	87.37
Face Only	85.18					
concatenation	87.59	88.15	91.85	89.07	89.81	89.29
Type I	90.21	91.11	94.81	92.03	92.96	92.22
Type II	89.63	92.03	93.15	91.85	91.66	91.66
Type III	89.68	89.15	92.33	90.21	91.11	90.50
Type IV	93.33	93.70	96.67	95.18	95.18	94.81
Type V	91.11	92.03	94.07	93.33	92.96	92.70

Conclusions can be drawn from our experiments as follows:

- The bimodal speaker authentication system has improved the identification rate by 7.44% and 9.63% compared with the speaker recognition and face recognition system respectively in the best cases. And the simple concatenation method which only enhances the performance by little degree is still far from satisfactory. That's why researchers are still pursuing superior features fusion methods.
- The bimodal speaker identification system based on the feature-level fusion using DBN outperforms the simple concatenation method by 5.52% in the best situation and 1.21% in the worst one. Indications are that it is a promising way of using feature-level DBN fusion in multi-modal problems.
- Among these five types of fusion, "Mixed States and Independent" works the best, which corresponds with our hypothesis.

5 Conclusions

This paper presents a feature-level fusion approach using Dynamic Bayesian Network for audio-visual speaker recognition. Five types of topologies are explored in the framework of bimodal speaker identification based on the correlations between the speech features and the face features. Encouraging experiment findings from multimodal corpus including different speech types and contents reveal that the multibiometric system can be further refined by the DBN-based feature-fusion approach. Further experiments will focus on the automate topology learning of DBN for multimodals fusion.

Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (60273059), Zhejiang Provincial Natural Science Foundation (M603229) and National Doctoral Subject Foundation (20020335025).

References

1. Kevin Murphy.: Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. thesis, U.C. Berkeley (2002)
2. Dongdong Li, Yingchun Yang, Zhaohui Wu, Wenyao Liu,: Add prior knowledge to speaker recognition. Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2005, part of the SPIE Defense and Security Symposium 2005. Vol. 5813 (2005) 192-200
3. Ara V Nefian, Lu Hong Liang, Xiao Xing Liu, Xiaobo Pi and Kevin Murphy.: Dynamic Bayesian networks for audio-visual speech recognition. EURASIP, Journal of Applied Signal Processing, vol. 2002, no 11 (2002) 1274-1288
4. V. Pavlovic, A. Garg, J. Rehg, and T. S. Huang,: Multimodal speaker detection using error feedback dynamic Bayesian networks. Computer Vision and Pattern Recognition vol.2. (2000) 34-41.
5. Dongdong Li, LiFeng Sang, Yingchun Yang and Zhaohui Wu,: Bimodal Speaker Identification Using Dynamic Bayesian Network, 5th Chinese Conf. on Biometric Recognition, Lecture Notes in Computer Science, Vol. 3338. (2004) 577-585
6. Y. Wang, T. Tan and A. K. Jain,: Combining Face and Iris Biometrics for Identity Verification. Proc. of 4th Int'l Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA), Guildford, UK (2003) 805-813
7. Lifeng Sang, Zhaohui Wu, Yingchun Yang, Wanfeng Zhang,: Automatic Speaker Recognition Using Dynamic Bayesian Network. IEEE ICASSP. Vol.1 (2003) 188-191
8. Kevin Murphy. The Bayes Net Toolbox for Matlab. Computing Science and Statistics, vol 33 (2001)