

# Multichannel and Multimodality Person Identification

*Ming Liu, Yanxiang Chen, Xi Zhou, Xiaodan Zhuang,  
Mark Hasegawa-Johnson, Thomas Huang*

Department of Electrical and Computer Engineering  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801

{mingliu1}@ifp.uiuc.edu

May 3, 2007

## Abstract

Person's identity is a very important high level information for video analysis and retrieval. Along the growth of multimedia data, the recording is not only multimodality and also multichannel(microphone array, camera array). In this paper, we describe a multimodal person identification system of UIUC team for CLEAR 2007 evaluation. The audio only system is based on a new proposed model – Chain of Gaussian Mixtures. The visual only system is a face recognition module based on nearest neighbor classifier at appearance space. Final system fuses 7 channel microphone recordings and 4 camera recordings at decision level. The experimental results indicate the effectiveness of speaker modeling methods and the fusion scheme.

## 1 Introduction

Person identification is a task of identifying a particular person out of a group of people based on physiology cues such as speech, facial images, finger print and iris, etc. Because of great commercial potential, this topic has brought many research and engineering efforts in both academia and industry. Based on speech signal, the identification of person is also called speaker identification[1][2][3]. Based on facial image, the identification of person is also called face identification[4]. Either category has been extensively addressed, and is traditionally formulated as a pattern recognition problem in some feature vector space, tackled by statistical classification and machine learning algorithms.

Fusing audio and visual cues can substantially boost the performance of the person identification system. The concept of multimodal person identification has been brought to the attention of the speech and computer vision communities. In CLEAR evaluation, the multichannel recordings are also available.

To fuse the microphone array recordings and multiple camera recordings is a challenge and interesting research problem. This paper describes a system fuse multimodal cues as well as multichannel recording so that person identification achieve significant better performance. The experimnts are conducted on the CLEAR 2007 Evaluation corpus[?]. The results show that the fusion of multichannel and multimodality do improve the performance significantly. The accuracy of 1sec testing utterance is boosted from 70% to 89%. For longer testing utterance, the fused system can achieve 100% accuracy. These results clearly demonstrate the effectiveness of multichannel and multimodal fusion. The detailed algorithms and implementation of the system are described in the following sections.

## 2 Audio Only Person Identification Subsystem

The Gaussian Mixture Model (GMM) has been considered one of the best modeling method for text-independent speaker identification. In the domain of speaker identification, the Mel Frequency Cepstral Coefficient (MFCC) is widely used as the acoustic feature. Although MFCC is not exclusively designed as a sort of speaker-distinguishing features, it capture the vocal tract structure which is essentially one speaker distinguish feature.

### 2.1 GMM

An  $M$ -mixture GMM is defined as a weighted sum of  $M$  component Gaussian densities

$$p(\bar{x}|\lambda) = \sum_{m=1}^M w_m N(\bar{x}|\bar{\mu}_m, \Sigma_m) \quad (1)$$

where  $\bar{x}$  is a  $D$ -dimensional feature vector,  $w_m$  is the  $m^{th}$  mixture weight, and  $N(\bar{x}|\bar{\mu}_m, \Sigma_m)$  is a multivariate Gaussian density, with mean vector  $\bar{\mu}_m$  and covariance matrix  $\Sigma_m$ . Note that  $\sum_{m=1}^M w_m = 1$ .

A speaker model  $\lambda = \{w_m, \bar{\mu}_m, \Sigma_m\}_{m=1}^M$  is obtained by fitting a GMM to a training utterance  $X = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$  using the expectation-maximization (EM) algorithm. The log likelihood of a testing utterance  $Y = \{\bar{y}_1, \bar{y}_2, \dots, \bar{y}_T\}$  on a given speaker model  $\lambda$  is computed as follows.

$$LL(Y|\lambda) = \frac{1}{T} \sum_{t=1}^T \log p(\bar{y}_t|\lambda) \quad (2)$$

where  $p(\bar{y}_t|\lambda)$  is the likelihood of the  $t^{th}$  frame of the utterance. To identify an utterance as having been spoken by a person out of a group of  $N$  people, we compute its utterance scores against all  $N$  speaker models and pick the maximum

$$\hat{\lambda} = \arg \max_{\lambda_n} LL(Y|\lambda_n) \quad (3)$$

where  $\lambda_n$  is the model of the  $n^{th}$  speaker.

## 2.2 UBM-GMM

The GMM algorithm described in the previous subsection requires that every speaker model be trained independently with the speaker's training data. In the case when the available training data is limited for a speaker, the model is prone to singularity. In the UBM-GMM algorithm, a different scheme is adopted to train the speaker models. A single speaker-independent Universal Background Model (UBM)  $\lambda_0$  is trained with a combination of the training data from all speakers, and a speaker model  $\lambda$  is derived by updating the well-trained UBM with that speaker's training data via Maximum A Posteriori (MAP) adaptation[5]. The final score of the testing utterance is computed by the log likelihood ratio between target model and background model.

$$LLR(Y) = LLR(\bar{y}_1^T) = \frac{1}{T} \sum_{t=1}^T \log \frac{P(\bar{y}_t|\lambda_1)}{P(\bar{y}_t|\lambda_0)} \quad (4)$$

where  $(\bar{y}_1^T)$  are the feature vectors of the observed utterance – test utterance  $Y$ ,  $\lambda_0$  is the parameter of UBM and  $\lambda_1$  is the parameter of the target model. Essentially, the verification task is to construct a generalized likelihood ratio test between hypothesis  $H_1$  (observation drawn from the target) and hypothesis  $H_0$  (observation not drawn the target).

The UBM is trained with a large amount of data of many speakers. It is considered as a speaker-independent model. Ideally, it describes the all possible acoustic features of human speech. A speaker model, obtained by adapting the parameters of the UBM with a small amount of new data, is expected to be focusing on the difference between this specific speaker and the speaker independent model. Hence, the GMM-UBM can localized the most unique feature for each target speaker.

## 2.3 Chain of Gaussian Mixture

Beside the simplicity and effectiveness of GMM-UBM, the most referred limitation is that the independence between frames is implied. The temporal information of speech signal is found useful in text dependent speaker recognition. In this paper, a new modeling technique named a chain of Gaussian Mixture Model is proposed. The motivation for the new model is to encode the temporal correlation in the chain structure. With this chain structure, a special decoding network is established to ensure that the optimal matches will be found for every segments of test utterance. In stead of conventional frame based system, we are trying to match the trial utterance and the training utterance in segment level. Bascially, the goal is to find the longest possible matching segments between training and testing uttearnces. For example, the training utterance is a phone number sequence "two-one-seven-two-three-four-nine" and the testing utterance is "two-one-seven-three-four-seven-six". The chain model is able to find the first three digits are matched between training and testing, and the 4th and 5th digits in testing uttearnce match the 5th and 6th digits in training, and 6th digit of

testing utterance matches the 3rd digit of training utterance. In this sense, the proposed method is kind of speech decoding with variable length units. Compared with speech recognizer based speaker recognition, our method has the potential advantage of incorporate much long range correlation for speaker recognition. If there are matched long segments between training and testing, the method is going to use the match score of the whole segment. If there are no matched long segments, the method becomes conventional frame based method.

The training of this new model have two steps. First of all, the global Gaussian Mixture is generated by MAP adaptation from the UBM with the whole utterance. Then, the local Gaussian Mixtures are generated by MAP adaptation from the UBM with only the current segment. In our experiment, we set the segment length to be 40 frames(400ms) and overlap between segments are 20frames. After buidling up the model, to match a test utterance is simply decoding on the chain model with one addition background state which has the global Gaussian Mixture as the observation density.

In order to combine different microphone channel of the microphone array, a linear fusion is adopted to fuse these channels. The MarkIII microphone array in our task has 64 channels which is linear configured with 2cm distance between conjacent channels. In order to have more variety between two channels, we select one channel out of every 10 channels. The channels used in fusion module are 01, 10, 20, 30, 40, 50, 60. The fusion is conducted directly on the log-likelihood score of each individual channel with equal weight.

### 3 Face Recognition

Our face recognition subsystem is based on the idea of  $K$ -Nearest Neighbor (KNN) algorithm. As a typical face recognition system, our system also has these modules: cropping, alignment, metric measurement, and KNN. The big difference is that, instead of determining the person ID based on a single face image, our system makes the decision by processing all face samples in a clip of a video. In other words, our system has a module of fusing multiple face samples.

#### 3.1 Face Cropping

For both training and testing videos, the faces are cropped according to the bounding boxes and positions of nose bridge provided by the organizers. In spite of the big variation of the view angles, the face images are then scaled to a fixed size ( $20 \times 20$  in our experiment) with the nose bridge fixed to the center of the image. The face images without the positions of nose bridge are omitted in the experiment because most of those face images have bad quality and may induce extra errors to the system. Figure 1 shows some cropped face samples. These images have varying face angles, changing illumination, and varying background which make the face recognition a big challenge.



Figure 1: Examples of cropped face samples.

### 3.2 Face Alignment

In a typical face recognition system, an alignment procedure should be applied to the cropped faces such that the main facial feature points (such as eye corners, nose point, mouth corners) are aligned image by image. However, face alignment is extremely difficult for this CHIL data, because face angles vary a lot and face resolution is too low. Therefore we use shifting procedure to partly substitute for the alignment procedure. In detail, the training samples are repeatedly shifted by one or two pixels in all directions to generate new training samples. We assume, after shifting, any test sample has a counterpart in the training data set (including the shifted samples) that both of them come from the same person while having same alignment.

### 3.3 Affinity Measurement

As we know, all face recognition algorithms depends heavily on the choice of the metric measurement. In our work, we transform the color face images into gray scale, then expand them into vectors, and finally calculate the  $l^p$  distances.

$$d_p(f_1, f_2) = \left[ \sum_{i=1}^D (f_1(i) - f_2(i))^p \right]^{1/p} \quad (5)$$

where  $f_1$  and  $f_2$  are the face samples and the  $D$  is the total dimension.  $f_1(i)$  is the  $i^{th}$  dimension of the face sample. It is worth to mention that  $l^1$  distance generates better performance than  $l^2$  distance in our work.

### 3.4 KNN and Fusion of Multiple Faces

As mentioned above, unlike the typical face recognition systems, our system makes the decision by processing all face samples in a clip of a video. We call the face samples in the same clip as “test subset”. To determine the person ID by considering the entire “test subset”, we first apply the KNN algorithm to

each sample in subset separately, and then fuse the output of KNN algorithm to make the final decision.

We choose the standard KNN algorithm. For each face sample  $f$  in a “test subset”  $S$ ,  $K$  training samples with smallest distance (with  $f$ ) are selected as candidates. These  $K$  samples is called candidate set of sample  $\Omega(f)$ . Therefore, given that  $S$  contains  $N_S$  samples, the subset  $S$  will have  $K \times N_S$  candidates from the training set which forms a candidate set  $\Omega = \bigcup_{f \in S} \Omega(f)$ . Then we use voting to generate the id for one test subset.

## 4 Audio Visual Fusion

In order to fuse the two modalities for better performance, an audio/visual fusion module is applied to combine these two modalities. There are different kinds of fusion strategies proposed in the literature[6][7][8]. There are mainly three level of fusion: feature-level, state-level and decision level. Fusion at feature-level mainly concatenate the features from different modalities as a single big feature vector. Some dimension reduction techniques such as PCA, LDA can be used to reduce the dimensionality of the final feature vector. The modeling is then conducted on the final feature vectors. Usually the feature level fusion is most simple fusion strategies and often result in moderate improvement after fusion. State-level fusion is considered the best strategies from the reports by audio/visual speech recognition literatures. The basic idea is to fuse the observation likelihood of different modalities on the same state. By searching the right confidence measure of two streams, the fusion can achieve best improvement.

However, the text-independent ID task make it difficult to find the same state for audio and visual streams. To circumvent this difficulties, we explore the decision level fusion for this task. The decision output of audio and visual stream are the similarity scores of the testing utterance on 28 target speaker models. By tuning the weighting factor between two streams, we obtain very good improvement after fusion. Intuitively, the weighting factor should not be static between audio/visual streams. In principle, the optimal weighting factor should be estimated based on the SNRs of different modalities. However, the estimation of SNRs usually is also difficult to obtain. However, the duration of the speech utterance is correlate to the performance of audio-only system in consistent way and so is the number of face frames. In this task, we searching the optimal weighting factor for different testing conditions(1sec, 5sec, 10sec, 20sec) indivisually based on the experiments on CHIL development dataset.

## 5 Experiment Results

The CHIL 2007 ID task corpus[?] contain 28 speakers video sequences from 5 cites. The audio recording is far-field microphone array recording. In our experiments, only one microphone array – MarkIII recording is considered. There are 64 channels and linear array configuration with 2cm apart. The video recording

Methods	$test_1$	$test_5$	$test_{10}$	$test_{20}$
GMM-UBM	62.4	85.1	88.9	95.5
Chain of GMM	69.5	89.6	92.5	95.5

Table 1: GMM-UBM vs Chain of GMM on single channel

Methods	$test_1$	$test_5$	$test_{10}$	$test_{20}$
GMM-UBM	71.4	92.1	94.6	98.3
Chain of GMM	76.6	93.6	95.6	98.3

Table 2: GMM-UBM vs Chain of GMM on single channel

includes four cameras located at four corner of ceiling. Both of audio and visual recording are far-field, therefore noisy and low resolution. The performance of each individual modality will not be sufficient. It contains seminar recording as well as interactive dicussion recording.

There are two training conditions varing with respect to the duration of the enrollment. The train set A has 15sec training enrollment while train set B has 30 sec enrollment. The testing conditions varies in term of testing durations. The four testing conditions are corresponds to 1sec, 5sec, 10sec and 20sec.

A 128 component UBM is trained from approximate CHIL development data. To improve the audio only system by multichannel recording, we fuse the channels based on decision level fusion and all 7 channels(01,10,20,30,40,50,60) are treated with equal weighting factor. The experiment results(Table ?? and Table 3) shows the improvement is significance by multichannel fusion, especially for short testing utterarncce conditions(acurracy boost from 69.5% to 79.2%). For visual only part, we have try different distance measure ( $l_1, l_2$  and normalized cross correlation) and different neighborhood size ( $N = 1, 3, 5, 7, 10$ ). It turns out the  $l_1$  norm combined with  $N = 1$  is the optimal based on the CHIL development data, Table ??. The performance of Audio/Visual fusion is listed in Table 5. The improvement due to A/V fusion is as large as 6% in absolute percentage compared to the multichannel fused audio only system and 16% in absolute percentage compared to the single channel audio only system.

## 6 Conclusion and future work

In this paper, we describe a Multimodal person ID system base on multichannel and multimodal fusion. The audio only system is combining 7 channels microphone recording at decision output individual audio-only system. The modeling technique of audio system is UBM-GMM and the visual only system works directly on the appearance space via  $l_1$  norm and nearest neighbor classifier. The linear fusion is then combining the two modalities to improve the ID performance. The experiments indicate the effectiveness of micropohone array fusion and audio/visual fusion. Although the CHIL07 corpus is quite large database(200 giga bytes for all evaluation data), the number of speakers might be few. In the near future, we are going to including more speakers from CHIL07

TrainSet	$test_1$	$test_5$	$test_{10}$	$test_{20}$
A	79.2	93.3	95.5	98.2
B	82.2	97.3	97.3	100

Table 3: Microphone Array Audio-only System Performance

TrainSet	$test_1$	$test_5$	$test_{10}$	$test_{20}$
A	58.9	65.4	67.9	70.2
B	66.2	70.6	72.6	73.7

Table 4: Visaul-only System Performance( $N = 1$ )

TrainSet	$test_1$	$test_5$	$test_{10}$	$test_{20}$
A	85.4	95.3	96.9	99.1
B	88.7	97.5	98.7	100

Table 5: Final Auido Visual Fusion System Performance

development corpus to futher verify our framework. Also, linear fusion is simple yet useful solution for multichannel and multimodal fusion. More sophisticate fusion schemes are under investigation.

## Acknowledgments

This work was supported in part by National Science Foundation Grant CCF 04-26627 and ARDA VACE program.

## References

- [1] Doddington, G.: Speaker recognition - identifying people by their voices. (1985) 1651–1664
- [2] Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* **17** (1995) 91–108
- [3] FURUI, S.: An overview of speaker recognition technology. (1996) 31–56
- [4] Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. *ACM Comput. Surv.* **35**(4) (2003) 399–458
- [5] Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. *Digital Signal Processing* (2000)
- [6] Dupont, S., Luettin, J.: Audio-visual speech modelling for continuous speech recognition. *IEEE Transactions on Multimedia* (2000) to appear.
- [7] Garg, A., Potamianos, G., Neti, C., Huang, T.S.: Frame-dependent multi-stream reliability indicators for audio-visual speech recognition. In: *Proc.*

of international conference on Acoustics, Speech and Signal Processing (ICASSP). (2003)

- [8] Potamianos, G.: Audio-Visual Speech Recognition. Encyclopedia of Language and Linguistics (2005)