VIDEO-BASED PERSON AUTHETICATION WITH RANDOM PASSWORDS

Chia-Wei Liao[1], *Wei-Yang Lin*[1], *and Chia-Wen Lin*[2]

[1]Department of Computer Science & Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan
[2]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan

ABSTRACT

The proposed system aims at providing a novel framework of speaker authentication using lip-motion data. T The system is divided into three steps: feature extraction and modeling, model synthesis, and probabilistic model matching. First, the visual features are obtained by locating a region-of-interest and extracting geometric and textural features from the region. The extracted feature vectors are then classified by *K*-means clustering to obtain a reduced number of observation vectors. These observation vectors are fed into a set of Hidden Markov Models (HMMs) classifiers to capture the temporal characteristics of the features. The main contribution of this paper lies in the introduction of using random passwords in performing speaker verification. Our results have demonstrated that random passwords provide useful information for speaker verification. Also, by using the proposed method, we observe a significant improvement in verification rate.

Index Terms— Face authentication, speaker verification, biometrics, security, face recognition

1. INTRODUCTION

Because of increasing concerns over security, the biometric authentication technologies have received unprecedented interests in the recent years. Many human biometric characteristics, including face, speech, iris, fingerprints, and palm prints, etc., have been utilized for verifying the identity of an individual. In the development of a biometric authentication system, an inevitable issue is to choose an appropriate biometric based on all the critical constraints. In particular, for a designer or a commercial product provider, it is vital to have a thorough understanding of design tradeoffs pertaining to a biometric authentication system. For example, the intrusiveness is an important factor. If a system makes users feel uncomfortable, then the system is intrusive and therefore becomes less desirable. For example, for public areas requiring middle or low security levels (e.g., hotels, hospitals), an iris recognition system which scans eyeballs will annoy people and thus practical deployments will be deterred. Similarly, when we use fingerprints and palm prints to perform identity verification, the direct contact with data acquisition devices is neither user-friendly nor sanitary.

Compared to iris, fingerprint, and palm print, the modality of human speech generally provides less intrusiveness, i.e., better convenience. However, the study has shown that the average error rate of speech recognition is 10 to 100 times higher than that of fingerprint [1]. Therefore, a designer of security system also needs to consider the issue of accuracy, i.e., different biometrics yield different error rates. The accuracy is arguably the most important factor for choosing a biometric characteristic. Ideally, a good verification system should have both low intrusiveness and low error rate. Practically, a tradeoff has to be made between these two competing factors. In [2], Kung et al. illustrate the tradeoffs among various approaches for biometric authentication. Among the popular biometrics (face, palm, iris and fingerprint), face is the most non-intrusive modality. Thus, it has long been receiving broad attentions in the field [4].

In this paper, the biometric modality called *visual speech* is used for performing person authentication. The terms visual speech refers to the features extracted from a sequence of mouth images when a person is speaking. Compared with audio speech (i.e., acoustic waveform) and face modalities, the visual speech based authentication methods can potentially deliver a higher accuracy and reliability. Note that visual speech is bimodal in nature; while visual refers to the texture and shape of a mouth region, speech refers to the lips movements produced by a speaker. Moreover, the visual speech enjoys the same level of convenience or non-intrusiveness as audio speech and face do. In other words, the modality of visual speech provides the best convenience with the least compromising on system accuracy. Hence, the visual speech has been subject to extensive research in the recent years [4]-[9]. Both speech content and the identity of a speaker can be revealed from visual speech behaviors. Therefore, it can be applied to speech and speaker recognition.

Here, we provide a brief review on the previous works which utilize visual speech data in speaker verification. In [4], the DCT coefficients of gray-level lip images are utilized as lip features. This feature is relatively easy to obtain but it suffers from illumination variations. Lip geometry has been employed in [5], where lip is segmented by computing the accumulated difference images. In [6], the authors find lip contour by active shape model and extract features from the model parameters. In the speaker verification system presented in [7], the lip contour is first extracted and then contour pixels are associated with color information.

Besides lip geometry, lip motion has proven to be useful in speaker verification. In [8], the emphasis has been placed on the discrimination analysis of lip motion features. The authors reported that the performance of a speaker verification system can be improved by selecting discriminative lip motion features. Our previous method [9] uses HMMs to characterize the lip motions of visual speech with fixed passwords, where the lip shape is extracted using the Active Appearance Models (AAMs). The models need to be rebuilt when the password is changed.

Although several approaches have been proposed to utilize visual speech information in speaker verification, there is no framework for integrating visual speech and random passwords in

the literature. This paper aims at answering the following questions:

1. Do random passwords, instead of fixed passwords for each user, carry useful information for verification purpose?
2. If so, how do we process the visual speech data obtained from random passwords?

In order to answer these questions, a novel framework for speaker verification is presented in this paper. Furthermore, experiments are conducted on the dataset of 10 subjects and we observe some promising results. Thus, the main contribution of this paper lies in the introduction of using random passwords in performing speaker verification.

The rest of this paper is organized as follows. In Sec. 2, we give an overview of the proposed speaker verification system. Sec. 3 describes the detailed steps of the proposed systems, including feature extraction, feature modeling, and model synthesis. In Sec 4, we describe experimental procedures and demonstrate promising results. Finally, we provide concluding remarks in Sec 5.

2. SYSTEM OVERVIEW

The proposed system provides a novel framework of speaker verification and authentication using lip motion features extracted from the face video of a person speaking a password. During the authentication process, the person is asked to show his/her front face to the camera and pronounce a randomly generated password (e.g., a sequence of several symbols) shown on the display panel. The system identifies the speaker by comparing the visual features extracted from the captured visual speech (pronouncing the random password) with the ones synthesized from a pre-trained model database. Since the password is randomly generated, the user does not know the password prior to the authentication, thus he/she neither needs to remember the password nor needs to keep it secret. The risk of disclosing or forgetting a secret password or of being faked with a pre-captured/forged face video is thus drastically reduced, making the system well suited for practical applications.

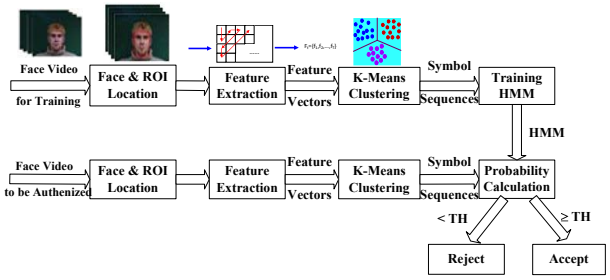


Fig. 1. Block diagram of the proposed speaker verification system.

Fig. 1 shows the block diagram of our proposed system. A visual speech corresponding to a random password is captured for authentication. The system is divided into four steps: feature extraction, feature modeling, model synthesis of the password, and probabilistic model matching. First, the visual features are obtained by locating a region-of-interest (ROI) in the input visual speech (e.g., the mouth region in our work) and then by extracting geometric and textural features of the ROI. The extracted feature vectors are subsequently classified by *K*-means clustering to obtain a reduced number of observation vectors. These observation vectors are fed into a set of HMM classifiers to capture the temporal characteristics of the features.

While training the models of authorized persons, each person is asked to pronounce a complete set of individual symbols (e.g., 10 digits {0,…,9} in our case) one by one and a few patterns of continuously spoken symbols. The training patterns of continuously spoken symbols are used to capture a person's unique styles/features of transitions when pronouncing two symbols continuously rather than pronouncing them separately. Such features are used in synthesizing/interpolating the model of continuous visual speech for a random password from the visual speech models of individual symbols as will be explained in Sec. 3.3. Typically, the model database does not need to be changed once it is established. The successfully authenticated visual speech of a person can be used to update the database to further improve the authentication accuracy.

3. FEATURE EXTRACTION AND MODELING

3.1. Visual Feature Extraction

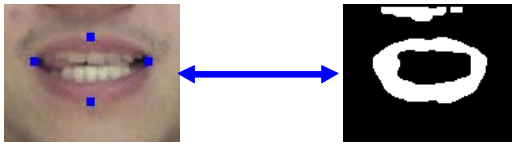
3.1.1. Preprocessing

In the preprocessing step, our system detects the face region as well as locates the lips region and the four corner points of lips using our previous method [9].

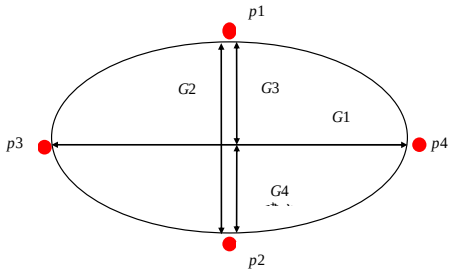
After extracting the corners of lips, the mouth region is aligned using the 4-parameter affine transform as follows:

$$x' = a_0 x + a_1 y + a_2$$
$$y' = a_1 x + a_0 y + a_3 \quad (1)$$

where (x,y) and (x',y') represent the current and target coordinates, respectively. $\mathbf{a} = [a_0,\ldots, a_3]$ denotes the parameter set of affine transform.



(a)



(b)

Fig. 2. Illustration of geometric feature extraction: (a) extracted lip region, and (b) feature points and geometric features used to characterize the lip motion behaviors.

3.1.2. Extraction of geometric features of lips

As shown in Fig. 2(b), the four lip centers, $p1$~$p4$, are extracted as feature points. For feature values G_1~G_4 computed from the coordinates of $p1$~$p4$ are concatenated to form the lip geometric feature vector $\boldsymbol{F_g} = \{G_1,G_2,G_3\}$.

3.1.3. Extraction of ROI textural features

We also propose to extract the textural features from the ROI region shown in the left-hand image of Fig. 3. To mitigate the

effect of illumination variation, after aligning the lip region, histogram equalization is performed to enhance the contrast of ROI prior to extracting the features. As shown in Fig. 3, the DCT coefficients of the $M \times N$ ROI region are computed as follows [4]:

$$F(u,v) = C(u)C(v)\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\cos\left[\frac{(2m+1)u\pi}{2M}\right]\cos\left[\frac{(2n+1)v\pi}{2N}\right],$$

$$u,v = 0,1,\cdots,N-1,$$

$$\text{where } C(u) = \sqrt{1/M}, C(v) = \sqrt{1/N}, \text{ for } u,v = 0;$$

$$C(u) = \sqrt{2/M}, C(v) = \sqrt{2/N}, \text{ otherwise.}$$

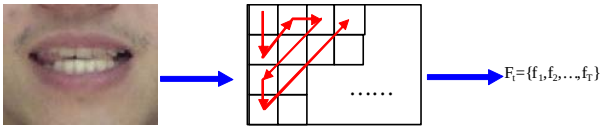


Fig. 3. Illustration of textural feature extraction.

Consequently, after zig-zag scanning, the first T low-frequency coefficients $f_1 \sim f_T$ out of the $M \times N$ coefficients are extracted to form the textural feature vector $\boldsymbol{F_t} = \{f_1, f_2, \ldots, f_T\}$.

3.2. Visual Feature Modeling

After extracting visual feature vector $\boldsymbol{F} = \{\boldsymbol{F_t}, \boldsymbol{F_g}\}$, the next step is to characterize the temporal behavior of the feature vector. Our method characterizes the temporal behavior of lip motion using HMMs. First, the visual feature vectors are classified by K-means clustering to obtain a reduced number of observation vectors before they are fed into the HMM classifiers. The K-means clustering classifies data such that the data within the same class are as close as possible in the Euclidean distance.

An HMM can be viewed as an unobservable Markov chain with a finite number of states. An HMM can be described by a transition probability matrix **A**, an initial state probability distribution $\boldsymbol{\pi}$, and a set of probability density functions for observations **B** (or emission matrix). The whole HMM modeling process is divided into the following steps, the first three steps and for training, and the final step is for testing (decoding) [10]. (a) Apply K-means clustering to classify the feature vectors into M classes of observation vectors, each with a mean vector. (b) Initialize **A**, **B**, and $\boldsymbol{\pi}$. (c) Use the Baum-Welch (Expectation Maximization, EM) algorithm to find the optimal HMM models for the training data set. (d) In the testing stage, evaluate the probability values of the input feature vector with the H HMMs and use the Viterbi algorithm to find the best-match HMM by

$$P(\mathbf{O} \mid \boldsymbol{\lambda}_h) = \max_h P(\mathbf{O} \mid \boldsymbol{\lambda}_h), \ 1 \le h \le H \qquad (2)$$

where $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, and $\mathbf{O} = \{o_1, o_2, \ldots, o_H\}$ denoting the sequence of observation symbols.

3.3. Model Synthesis of a Random Password

Basically, a random password is synthesized by combining discrete symbols. Before combining the feature sequences of discrete symbols, each feature sequence has to be normalized. The details of normalization are discussed in section 3.3.1. After the normalization, we will compute the merging points, namely startPoint and endPoint, from a normalized feature sequence. The definitions of these two points are given in section 3.3.2. According to merging points, two discrete symbols can be combined by interpolation. In section 3.3.3, we present the method for interpolating the transition between two discrete symbols.

3.3.1. Normalization of a spoken symbol

Given the feature sequence of a spoken symbol $\boldsymbol{\Phi} = \{\boldsymbol{F_1}, \boldsymbol{F_2}, \ldots, \boldsymbol{F_K}\}$, where $\boldsymbol{F_i}$ is the feature vector extracted from the i-th frame and K is the number of frames in the spoken symbol, the distance d_i between two consecutive frames is computed by

$$d_i = \left\|\mathbf{F_{i+1}} - \mathbf{F_i}\right\|, \ i = 1, \ldots, K-1 \qquad (5)$$

where $\|\cdot\|$ denotes L_2 norm. Then, we can divide the original sequence $\boldsymbol{\Phi}$ into L partitions according to the following rule.

$$(j-1)\bullet u < \sum_{i=1}^{p-1} d_i \le j \bullet u \text{ , then } \mathbf{F_p} \text{ belong to the } j-\text{th partition}$$

where j = 1,2,…,L, and the average distance u is given by

$$u = \frac{1}{L}\sum d_i \qquad (6)$$

Then we can compute the averaged feature vector $\boldsymbol{G_j}$ within each partition and obtain the normalized feature sequence $\boldsymbol{\Phi_n} = \{\boldsymbol{G_1}, \boldsymbol{G_2}, \ldots, \boldsymbol{G_L}\}$.

3.3.2. Merging Points

The merging points refer to startPoint and endPoint. These two pints are computed from a discrete symbol and the same symbol segmented from a continuous speech. In Fig. 4, $\boldsymbol{\Phi_n}$ denotes the normalized feature sequence of a symbol and $\boldsymbol{\Phi_c}$ denotes the feature sequence of the same symbol segmented from a password. In the training stage, a subject has to say each symbol for 5 times and 5 random passwords. Hence, for a symbol contained in a password, we can always find the same symbol which is said individually. For the first feature vector of $\boldsymbol{S_i}$, we find the index of the closest feature vector in the first half of $\boldsymbol{\Phi_n}$. Likewise, for the last feature vector in $\boldsymbol{\Phi_c}$, we find the index of the closest feature vector in the second half of $\boldsymbol{\Phi_n}$. Note that the similarity between two feature vectors is measured by using L_2 distance.

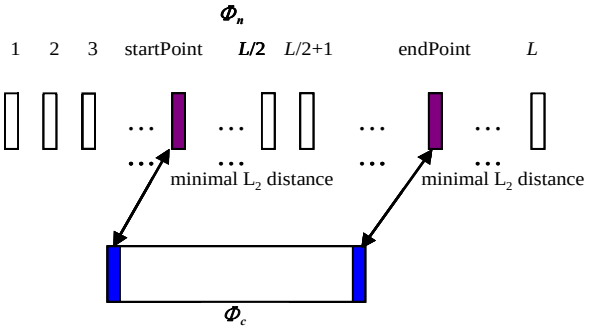


Fig. 4. Illustration of compute change ratios

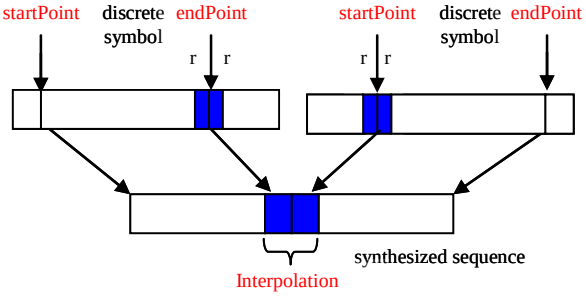


Fig. 5. The transition between two discrete symbols is synthesized by Bezier interpolation.

3.3.3. Interpolation

The transition between discrete symbols is obtained by performing interpolation (shown in Fig. 5). Given the feature sequences of two symbols, we take r features before endPoint and r features after endPoint from the first symbol. Similarly, we take r features before startPoint and r features after startPoint from the second symbol. Hence, we have two set of features and each set contains $2r+1$ features. We can then perform Bezier interpolation [11] between two sets of features features.

4. EXPERIMENTAL RESULTS

Note that a password is a permutation of symbols. It is customary to take any symbol set of the form $\{s_0, \ldots, s_{q-1}\}$ for q different symbols. In the following experiments, we use the set of digits {0, …, 9} pronounced in English as our symbol set. The length of a password is chosen to be four digits. In the training stage, we need to collect data and estimate the parameters of the subject-specific HMMs from the training data. More specifically, the parameter values of a HMM is estimated from 10 synthesized passwords. Note that a synthesized password is obtained by combing discrete symbols as described in section 3.3. For each subject, his or her training data contain 10 symbols (each one is spoken 5 times) and 5 random passwords.

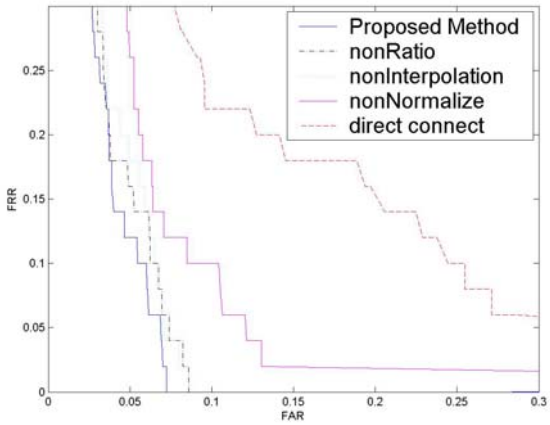


Fig. 6. DET performance obtained by using random passwords.

In the testing stage, we have 50 testing videos recorded from 10 subjects. In particular, each subject says five random passwords. A random password is generated by our system and a subject is assumed to say exactly the same password. Also, a subject has to claim his or her identity. Since we know the content of a password and the identity of a subject, the corresponding HMM can readily be obtained (as described in the previous paragraph). We then synthesize the HMM for each testing sequence. Therefore, we have 50 testing sequence and 50 corresponding synthesized HMMs. Every testing sequence is compared against all the synthesized HMMs. The results are stored in a 50 by 50 matrix, called the similarity matrix. The (i,j)-th entry of this similarity matrix represents the likelihood that the j-th recorded sequence is generated by the i-th synthesized HMM. Performance is evaluated by using a Detection Error Tradeoff (DET) curve that shows the tradeoff between False Acceptance Rate (FAR) and False Rejection Rate (FRR), as shown in Fig. 6. The "direct connect" denotes the results of concatenating individual directly. Similarly, the "nonRatio," "nonInterpolation," and "nonNormalize" denote the result without using change ratio, interpolation, and normalization, respectively. It is obvious that each step in the proposed method has its own contribution to the result. The use of change ratio, interpolation, and normalization improve the performance to 6% Equal Error Rate (EER). Here, we demonstrate that random passwords provide useful information for speaker verification. Also, by using the proposed method, we observe a significant improvement in the verification rate.

5. CONCLUSION

In this paper, we proposed a novel lip-motion based speaker verification scheme using random passwords. After extracting model parameters of lip region in a visual speech sequence, a reduced set of observations of the sequence are obtained by K-means clustering. HMMs are then used to characterize the temporal dynamics of lip motions for speaker verification. Our results have demonstrated that random passwords provide useful information for speaker verification. Also, by using the proposed method, we observe a significant improvement in verification rate. Besides, thanks to the use of random passwords, the risk of disclosing or forgetting a secret password or of being faked with a pre-captured/forged user video is drastically reduced compared to existing approaches, making the proposed system well suited for practical applications.

REFERENCES

[1] B. Miller, "Vital signs of identity," *IEEE Spectrum*, pp. 22-30, 1994.

[2] S. Y. Kung, M. W. Mak, and S. H. Lin, *Biometric Authentication: A Machine Learning Approach*, *Prentice Hall*, 2005.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705-741, May 1995.

[4] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 840-852, Oct. 2005.

[5] C. C. Broun, X. Zhang, R. M. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 685-688, May 2002.

[6] L. L. Mok, W. H. Lau, S. H. Leung, S. L. Wang, and H. Yan, "Lip features selection with application to person authentication," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*,, vol. 3, pp. 397-400, May 2004.

[7] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, pp. 169-186, 2001.

[8] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans Image Process*, vol. 15, no. 10 pp. 2879-91, 2006.

[9] K.-Z. Chen, Y.-J. Chang, and C.-W. Lin, "Video-based authentication using appearance models and HMMs," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2006, Island of Kos, Greece.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[11] M. Mignotte and D. Stefanescu, *Polynomials: An Algorithmic Approach*: Springer, 1999.