# Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification

Xin Liu, *Member, IEEE*, and Yiu-ming Cheung, *Senior Member, IEEE*

***Abstract*— This paper proposes a concept of lip motion password (simply called lip-password hereinafter), which is composed of a password embedded in the lip movement and the underlying characteristic of lip motion. It provides a double security to a visual speaker verification system, where the speaker is verified by both of the private password information and the underlying behavioral biometrics of lip motions simultaneously. Accordingly, the target speaker saying the wrong password or an impostor who knows the correct password will be detected and rejected. To this end, we shall present a multi-boosted Hidden Markov model (HMM) learning approach to such a system. Initially, we extract a group of representative visual features to characterize each lip frame. Then, an effective lip motion segmentation algorithm is addressed to segment the lip-password sequence into a small set of distinguishable subunits. Subsequently, we integrate HMMs with boosting learning framework associated with a random subspace method and data sharing scheme to formulate a precise decision boundary for these subunits verification, featuring on high discrimination power. Finally, the lip-password, whether spoken by the target speaker with the pre-registered password or not, is identified based on all the subunit verification results learned from multi-boosted HMMs. The experimental results show that the proposed approach performs favorably compared with the state-of-the-art methods.**

***Index Terms*— Lip-password, lip motion segmentation, multi-boosted HMMs, random subspace method, data sharing scheme.**

## I. Introduction

SPEAKER verification has received considerable attention in the community because of its attractable applications such as financial transaction authentication, secure access control, security protection, human-computer interfaces, and so forth [1], [2]. It aims at verifying a claimed speaker using pre-stored information, whereby the speaker will be either accepted as a target speaker or rejected as an impostor under a certain matching criterion.

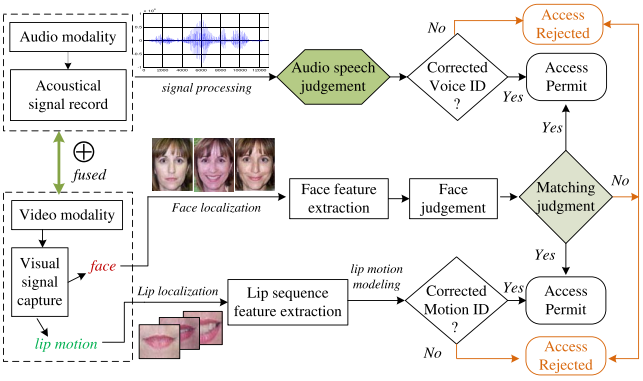


Fig. 1. A speaker verification system based on different modalities, i.e. audio modality (e.g. acoustical signal) and video modality (e.g. face, lip motion).

In general, speech not only conveys the linguistic information but also characterizes the speaker identity, which can thus be utilized for speaker verification [3]. Traditionally, the acoustic speech signals may probably be the most natural modality to achieve speaker verification. Although a purely acoustic-based speaker verification system has shown the effectiveness in its application domain, its performance would be degraded dramatically in the environment corrupted by the background noise or multiple talkers. Under the circumstances, as shown in Fig. 1, speaker verification by taking into account some video information, e.g., the still frames of face, has shown an improved performance over acoustic-only systems [4], [5]. Nevertheless, the access-controlled systems utilizing the still face images are very susceptible to the poor quality of pictures, variations in pose or facial expressions [6]. Further, such a system may be easily deceived by a face photograph placed in front of the camera as well. In recent years, speaker verification utilizing or fused with lip motions has attracted much attention [7]–[9]. As a kind of behavioral biometric trait, the lip motions accompanying with the lip shape variations, tongue and teeth visibility, have been demonstrated to encode rich information to characterize the speaker. For instance, Luettin et al. [10] first extracted a group of visual lip region features and then utilized the Hidden Markov Model (HMM) with the mixtures of Gaussians to build the spatio-temporal models for speaker identification, while Wark et al. [11] utilized the Gaussian Mixture Model (GMM) to build the statistical speaker models for identity identification. Later, Shafait et al. [12] extracted a group of suitable visual features from the sequential mouth regions and utilized

Manuscript received August 3, 2012; revised August 10, 2013 and November 11, 2013; accepted November 11, 2013. Date of publication November 26, 2013; date of current version January 13, 2014. This work was supported in part by the Research Grant Council of Hong Kong SAR under Project HKBU 210309, in part by the Faculty Research Grant of Hong Kong Baptist University under Projects FRG2/11-12/067 and FRG2/12-13/082, and in part by the National Science Foundation of China under Grant 61272366 and Grant 61300138. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo. *(Corresponding author: Y.-M. Cheung.)*

X. Liu is with the Department of Computer Science and Technology, Huaqiao University, Fujian 361021, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: xliu@comp.hkbu.edu.hk).

Y.-M. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, and also with the United International College, Beijing Normal University - Hong Kong Baptist University, Zhuhai 200086, China (e-mail: ymc@comp.hkbu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2013.2293025

the GMM classifier for person authentication. Faraj et al. [13] obtained a group of lip motion features associated with GMM for person verification. Specifically, by considering a unified feature selection and discrimination analysis framework, Ertan et al. [14] have utilized the explicit lip motion information associated with HMM for speaker identification and speech-reading.

Nevertheless, to the best of our knowledge, the performance of the existing lip motion based speaker verification systems is far behind our expectation. The main reasons are two-fold: (1) The principal feature components representing each lip frame are not always sufficient to distinguish the biometric properties between different speakers; (2) The traditional lip motion modeling approaches, e.g. single GMM [11], [12], single HMM [10], [15], often fail to learn the model discriminatively and are thus incompetent to verify some hard-to-classify examples. For instance, some diverse lip motions are so similar that the corresponding models learned from these conventional approaches are not discriminative enough to verify their corresponding speakers. Recently, some researchers have attempted to adopt multi-modal expert fusion systems by combining audio, lip motion and face information to enhance the security and improve the overall verification performance [4], [16]. Nevertheless, the appropriate fusion between different modalities is still a non-trivial task nowadays.

In this paper, we shall concentrate on the single modality only, i.e. lip motion, although the underlying technique can be fused with the other modalities as well. We first propose a concept of lip motion password (simply called *lip-password*[1] hereinafter), which is composed of a password embedded in the lip movement and the underlying characteristic of lip motion. Subsequently, a lip-password protected speaker verification system aiming at holding a double security, is established. That is, the claimed speaker will be verified by both of the password information and the underlying behavioral biometrics of lip motions simultaneously. Accordingly, the target speaker saying the wrong password or an impostor who knows the correct password will be detected and rejected. Further, such a system has at least four merits: (1) The modality of lip motion is completely insensitive to the background noise; (2) The acquisition of lip motions is somewhat insusceptible to the distance; (3) Lip-password protected speaker verification system can be performed silently in a hidden way; (4) It is simply applicable to a speech impaired person.

As for the single modality of lip motions, it should point out that almost all the related speaker verification systems in the literature generally take the whole utterance as the basic processing unit [11], [15]. Note that the design of a lip-password protected system should be able to simultaneously detect both of the following two cases: (1) the target speaker saying the wrong password, and (2) an imposter saying the correct password. Unfortunately, these traditional methods are incompetent for such task. In general, the lip-password always comprises of multiple subunits, i.e. the visibly distinguishable unit of visual speech elements. These subunits indicate a short period of lip motions and always have diverse moving styles between different elements, which should be considered individually, but not as a whole, to describe the underlying lip-password information. To this end, we shall present a multi-boosted HMM learning approach to such a lip-password based speaker verification system. In this paper, we mainly focus on digital lip-password only, although the underlying techniques are extensible for non-digits as well. First, we extract a group of representative visual features to characterize each lip frame, and then propose an effective algorithm to segment the lip-password sequence into a small set of distinguishable subunits. Subsequently, we integrate HMMs with boosting learning framework associated with random subspace method (RSM) and data sharing scheme (DSS) to formulate a precise decision boundary discriminatively for these subunits verification. Finally, the lip-password whether spoken by the target speaker with the pre-registered password or not is identified based on all subunit verification results learned from multi-boosted HMMs. The experimental results have demonstrated the efficiency of the proposed approach. The preliminary version of this paper was reported in [17].

The remaining part of this paper is organized as follows: Section II will overview the related works, i.e. the discrimination analysis, HMM-based speaker verification framework, and the Adaboost learning. Section III presents the proposed multi-boosted HMMs learning framework, in which the visual feature extraction and lip motion segmentation are also introduced. The experimental results are conducted in Section IV. Finally, the concluding remarks are given in Section V.

## II. Overview of Related Works

During the past decade, several techniques, e.g. Neural Networks (NN) [18], GMM [11], [12], and HMM [14], [15], have been developed for lip motion based applications. In general, the successful achievement of lip motion based speaker verification lies in a closer investigation of the physical process and behavioral biometrics within the corresponding lip motion activities, which always incorporate strong temporal correlations between the observed frames. Hence, HMM has been the most popular technique because its underlying state structure can successfully model these temporal correlations. Nevertheless, the performance of the existing lip-motion and HMM-based speaker verification systems is still far behind our expectations. The main reasons are two-fold: (1) The extracted visual features are not so discriminative enough for lip motion investigation and subsequent similarity measurement; (2) The learned models are not sufficient to discriminatively characterize the different lip motion activities. Therefore, the discriminative learning is still desirable. In this paper, we shall integrate HMMs with the boosting learning framework to achieve robust lip-password based speaker verification. Accordingly, the following sub-sections will first survey the discrimination analysis in HMM-based approaches, and then briefly introduce

[1]The concept and the characteristics of lip-password were firstly initiated by the second author of this manuscript.

its framework for speaker verification. Finally, we shall give an overview of a typical boosting learning framework, namely Adaboost algorithm [19].

### A. Discrimination Analysis

In the literature, the discriminative learning of HMM-based speaker verification systems can be roughly summarized along two lines: discriminative feature selection and discriminative model learning. The former methods aiming at minimizing the classification loss will not only emphasize the informative features, but also filter out the irrelevant ones. Ertan et al. [14] have found that the joint discrimination measure of any two features is less than the sum of their individual discrimination power. Accordingly, they utilized Bayesian theory to select the representative features discriminatively provided that the feature components were statistically independent. However, it is very difficult to determine which feature component has more discrimination power. Often, the feature components are not statistically independent of each other.

The latter approaches featuring on parameter optimizations always achieve a better performance than non-discriminative approaches. In HMM, its parameters are normally estimated by Maximum Likelihood Estimation (MLE). Recently, some researches have found that the decision boundary obtained via the discriminative parameters learning algorithms is usually superior than the decision boundary obtained simply from MLE. Typical examples include maximum mutual information (MMI) [20], conditional maximum likelihood (CML) [21] and minimum classification error (MCE) [21]. These methods that maximize the conditional likelihood or minimize the classification error rate always outperform the MLE approach. Nevertheless, their computations are generally laborious and may not be implemented straightforwardly [20].

In the literature, the majority of the existing HMM-based speaker verification systems just employ a single HMM for lip motion analysis and similarity measurement, which may not lead to good performance due to its limited discrimination power. Until most recently, classifier ensemble based systems trained on different data subsets or feature subsets have always generated more discrimination power for better performance [22]–[24]. Differing from the sum rule and majority vote, Adaboost [19] aims at building a strong classifier by sequentially training and combining a group of weak classifiers in such a way that the classifiers can gradually focus more and more on hard-to-classify examples. Accordingly, the mistakes made by such a strong classifier will be decreased. Recently, GMM and HMM have been successfully integrated with boosting framework to form a discriminative sequence learning approaches [25]–[27]. For instance, Siu et al. [26] utilized the boosting method to discriminatively train GMMs for language classification. Foo et al. [27] employed adaptively boosted HMMs to achieve visual speech elements recognition. From their experimental results, it can be found that the traditional single modeling and classification methods cannot identify some samples because of less discrimination capability while the boosted modeling and classification approaches often provide the promising results by successfully identifying these hard-to-classify examples.

### B. Overview of HMM-based Speaker Verification

Let the video databases comprise a group of lip motions and each lip motion contains a series of lip frame sequences. For the HMM of the $e^{th}$ lip motion, its model $\lambda_e = (\pi_e, A_e, B_e)$, is built with $N$ hidden states $\mathcal{S}^e = \{S_1^e, S_2^e, \ldots, S_N^e\}$. Suppose $\lambda_e$ is trained from the observed lip sequence $\mathcal{O}_e = \{o_1^e, o_2^e, \ldots, o_{l_e}^e\}$ and emitted from a sequence of hidden states $s^e = \{s_1^e, s_2^e, \ldots, s_{l_e}^e\}$, $s_i^e \in \mathcal{S}^e$, where $l_e$ is the total number of frames. Let the output of an HMM take $M$ discrete values from a finite symbol set $V^e = \{v_1^e, v_2^e, \ldots, v_M^e\}$. For an N-state-M-symbol HMM, the parameters in the model $\lambda_e$ are summarized as follows:

1) The initial distribution of the hidden states $\pi_e = [\pi_i]_{1\times N} = [P(s_1^e = S_i^e)]_{1\times N}$ $(1 \le i \le N)$ , where $s_1^e$ is the first observed state in the state chain.
2) The state transition matrix $A_e = [a_{i,j}]_{N\times N} = \left[P(s_{t+1}^e = S_j^e \,|s_t^e = S_i^e)\right]_{N\times N}$ $(1 \le i, j \le N,\ 1 \le t \le l_e)$, where $s_{t+1}^e$ and $s_t^e$ represent the states at the $(t+1)^{th}$ and $t^{th}$ frame, respectively.
3) The symbol emission matrix $B_e = [b_j(k)]_{N\times M} = \left[P(v_k^e \ at\ t \,\middle|\, s_t^e = S_j^e)\right]_{N\times M}$ $(1 \le j \le N, 1 \le k \le M)$. It indicates the probability distribution of a symbol output $v_k^e$ conditioned on the state $S_j^e$ at the $t^{th}$ frame.

In general, a typical estimate of $\lambda_e$ can be iteratively computed using Baum-Welch algorithm [28]. Such a method has the advantages of easy implementation and fast convergence. Given the test observation sequence $\mathcal{O}_s = \{o_1^s, o_2^s, \ldots, o_{l_s}^s\}$, the goal of the speaker verification task is to find a decision by computing the likelihood between $\mathcal{O}_s$ with the target speaker model $\lambda(T)$ and imposter model $\lambda(I)$. Suppose the observed variables are conditionally independent of each other, the likelihood can be computed as follows:

$$P(\mathcal{O}_s\,|\lambda_i) = \prod_{t=1}^{l_s} P(o_t^s\,|\lambda_i), \quad \lambda_i \in \{\lambda(T), \lambda(I)\}, \tag{1}$$

where the likelihood score $P(o_t^s\,|\lambda_i)$ is generally measured by means of the forward-backward process while its most probable path is obtained via Viterbi decoding algorithm [28].

In general, the modality for HMM-based speaker verification can be regarded as a binary classification problem, which can be extensionally grouped into closed-set and open-set learning problems. In the closed-set case, the tested speakers are recorded to be known, and the models of both the target-speaker and imposter can be learned during the training phase. Given an observed sequence: $\mathcal{O}_s = \{o_1^s, o_2^s, \ldots, o_{l_s}^s\}$, the classification for this type of speaker verification problem is performed based on the log likelihood ratio (LLR):

$$LLR(\mathcal{O}_s) = \sum_{t=1}^{l_s}\left[\log\frac{P(o_t^s\,|\lambda(T))}{P(o_t^s\,|\lambda(I))}\right] \quad if\ \ LLR(\mathcal{O}_s) \ge \tau : accepted; \quad Otherwise : reject. \tag{2}$$

In the open-set case, the imposters are recorded to be unknown and its model may not be well determined. Given a test observed sequence recorded from unknown utterance, the verification task is to find whether it belongs to the target speaker registered in the database or not. As for the lip-password protected speaker verification system, the password utterances differing from the registered one are also considered to be imposters even they come from the same speaker. Note that, the frame length of the utterance may be slightly different even for the same phrase uttered by the same speaker. Thereupon, this kind of verification problem can be conducted based on normalized log likelihood (NLL):

$$NLL(\mathcal{O}_s) = \frac{1}{l_s}\sum_{t=1}^{l_s}\log P(o_t^s|\lambda(T)) \quad if \;\; NLL(\mathcal{O}_s) \geq \tau : accepted; \quad Otherwise : reject. \tag{3}$$

### C. Overview of Adaboost Learning

Let us consider a binary classification problem. There is a set of $N_t$ labeled training samples $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_{N_t}, y_{N_t})$, where $y_i \in \{1, -1\}$ denotes the class label for the sample $x_i \in \mathbb{R}^n$. The weight $w_i$ is assigned to get the uniform value initially. Let $h(x)$ denote a decision stump (i.e. weaker classifier), the procedure of AdaBoost involves a series of boosting rounds $R$ of weaker classifier learning and weight adjusting under a loss minimization framework, which aims to produce a decision rule as follows:

$$H_R(x) = \sum_{m=1}^{R}\alpha_m h_m(x), \tag{4}$$

where $\alpha_m$ represents the confidence of decision stump $h_m$. The optimal value of $\alpha_m$ can be generally accomplished via minimizing an exponential loss function [19]:

$$Loss(H_R(x)) = \sum_{i=1}^{N_t}\exp(-y_i H_R(x_i)). \tag{5}$$

Given the current ensemble classifier $H_{r-1}(x)$ and newly learned weak classifier $h_r(x)$ at boosting round $r$, the optimal coefficient $\alpha_r$ for the ensemble classifier $H_r(x) = H_{r-1}(x) + \alpha_r \cdot h_r(x)$ is the one which can lead to the minimum cost:

$$\alpha_r = \arg\min_{\alpha}(Loss(H_{r-1}(x) + \alpha h_r(x))). \tag{6}$$

According to the optimization algorithm [29], let $\varepsilon^r$ be the weighted training classification error:

$$\varepsilon^r = \sum_{i=1}^{N_t} w_i^r \cdot [h_r(x_i) \neq y_i], \tag{7}$$

the resultant $\alpha_r$ and updated $w_i$ are computed as:

$$\alpha_r = \frac{1}{2}\log(\frac{1-\varepsilon^r}{\varepsilon^r}) \tag{8}$$

$$w_i^{r+1} = w_i^r \cdot \exp(-y_i \alpha_r h_r(x_i)). \tag{9}$$

In this framework, the updated weights for hard-to-classify examples are increased, meanwhile these weights will also determine the probability of the examples being selected for subsequent component classifier. For instance, if a training sample is classified correctly, the chance of its being selected again for the subsequent component classifier is reduced. Otherwise, its chance will be increased. Accordingly, the training error of the ensemble classifier shall decrease as long as the training error of the component classifier is less than 0.5. In this framework, the individual classifiers are built in parallel and independent of each other. As formulated in Eq. (4), it will generate a strong classifier by linearly combining these component classifiers weighted by their confidence through a sequence of optimization iterations [19].

## III. The Proposed Approach to Lip-Password Based Speaker Verification

This section will present a multi-boosted HMMs learning approach to solving lip-password based speaker verification problem. Before describing the proposed approach in Sub-section III-C, we need to deal with two issues first: (1) Visual Feature Extraction, and (2) Lip Motion Segmentation. The former presents the extracted representative visual features, while the latter aims to separate the visibly distinguishable unit of each password element. We shall address these two issues in Sub-section III-A and III-B, respectively.

### A. Visual Feature Extraction

It is well known that the visual cues of lip movement not only reveal important speech relevant information, but also characterize the significantly behavioral biometrics of the speaker, which can be utilized for speaker verification. Hence, the suitable visual features extracted from the recorded lip image sequences should contain crucial information for password content and behavioral biometric analysis, whereby the different lip-passwords can be well identified. In the last decade, various visual features have been proposed [9], which can be roughly categorized into three branches: contour-based, appearance-based and motion-based features. For the contour-based features, the geometric shape parameters such as mouth area, perimeter, height and width derived form the binary mouth image, can be chosen as the visual features [14]. Their temporal variations can be further utilized to model the lip motion activities. Kaynak et al. [9] have conducted a comprehensive investigation about such features for lip motion analysis. For the appearance-based features, as the teeth and tongue are always appearing during the speaking process, the transforming coefficients such as Principal Component Analysis (PCA), Independent Components Analysis (ICA) and two dimensional Discrete Cosine Transform (2D-DCT) have shown their effectiveness [14], [30], [31]. Differing from the above-mentioned features that are extracted from a single frame level, the motion based features are able to reveal the temporal characteristics of lip movements [7], [13]. For example, Faraj et al. [13] modeled the sequential lip images by moving line patterns and calculated 2D velocity vectors of normal optical flows for person verification.

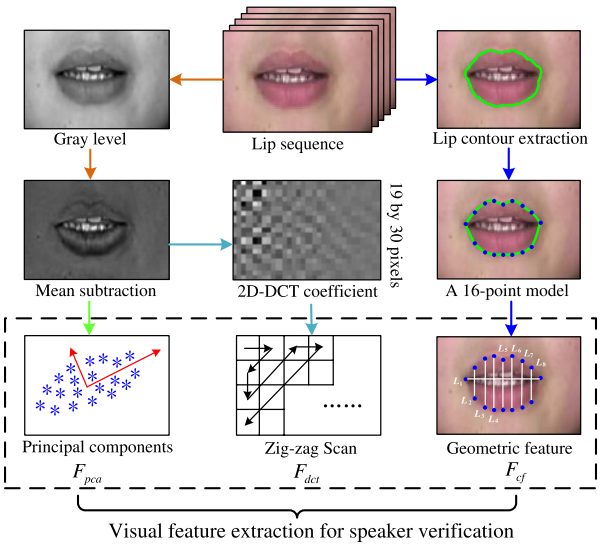


Fig. 2. Visual feature extraction for lip-password based speaker verification.

Nevertheless, it is quite difficult to determine which kind of feature vector has more discrimination power than the others. In general, the motion-based features are quite sensitive to the illumination changes and head movement. As reported in [14], a combination of contour-based and appearance-based features generally yields the acceptable performance for visual speaker identification. Hence, the integration of multiple kinds of features is desired. As shown in Fig. 2, we initially crop the mouth regions of interest (ROI) from the recorded lip sequences and extract the lip contours [32] frame by frame. Then, we employ a 16-point lip model proposed by Wang et al. [33] to compute nine geometric shape parameters, i.e. maximum horizontal distance, seven vertical distances and mouth area, denoted as $L_1$, $L_2$, $L_3$, $L_4$, $L_5$, $L_6$, $L_7$, $L_8$, $\mathcal{A}_c$, respectively, to model the contour-based feature vector $F_{cf}$. These geometric shape parameters are generally normalized with respect to the corresponding values of the first lip frame. Next, the previous cropped raw ROIs are convert to gray level case and illumination equalization method proposed by Liew et al. [34] is adopted to reduce the effects of uneven illuminations. Meanwhile, all the pixel values of incoming lip ROIs are normalized to have a similar distribution and mean subtraction is performed for each pixels to remove the basis effect of unwanted constant variations [35]. Accordingly, the principal components of top $N_{pca}$ numbers are chosen as PCA features $F_{pca}$, while the first $M$ coefficients along the Zig-zag scan order are selected as the 2D-DCT features, denoted as $F_{dct}$ [23]. Often, a size of $N_m \times N_m$ triangular mask is utilized to extract such 2D-DCT coefficients of $M = \frac{N_m(N_m+1)}{2}$ length for each lip frame. Consequently, the feature vector $F = \{F_{cf}, F_{pca}, F_{dct}\}$ is obtained to characterize the contour-based and appearance-based features jointly.

### B. Lip Motion Segmentation

Lip motion segmentation aiming to detect the starting and ending frames of the subunit utterance plays an important role for the lip-password based speaker verification. In the past, a few techniques have been developed to achieve speech segmentation using lip motion solely. For instance, Mak et al. [36] attempted to locate the boundaries of syllable by utilizing the lip velocity that is estimated by a combination of morphological image processing and block matching techniques. Yau et al. [37] computed the motion history images (MHIs) and utilized the Zernike moment features to segment the isolated utterances, in which the magnitude of Zernike moments corresponding to the frames that contain utterances is much greater than the one of the frames within the period of "pause" or "silence". Talea et al. [38] first obtained the mouth areas of the consecutive frames and then made a series of mouth area subtractions associated with a smoothing filtering for syllable separation. Recently, Shaikh et al. [39] have utilized an ad hoc method for temporal viseme segmentation (i.e. 14 different mouth activities) based on the pair-wise pixel comparison of consecutive images. In general, the MHIs, lip velocity and pair-wise pixel comparison are required to compute the pixel intensity variations frame by frame, whose computations are somewhat laborious. By a rule of thumb, the mouth areas of digital lip-password utterance always change significantly over time, where the position with minimum mouth area point always represents the status of mouth closing or intersection point between subunit utterances. Inspired by this finding, we employ the previously extracted mouth area $\mathcal{A}_c$ to achieve subunit segmentation as follows:

First, we obtain the signal $\mathcal{A}_c$ in terms of the mouth area variations via lip tracking [32]. Next, we utilize the forward-backward filtering [40] to process the input area signal $\mathcal{A}_c$ in both the forward and backward directions. Specifically, the resultant signal $\mathcal{A}_c^f$ has precisely zero phase distortion and magnitude while the other filters such as Gaussian filter may change the position of peak or valley point slightly. Accordingly, we can easily obtain the positions of peak points and valley points, where the peak points always represent the mouth opening widely while valley points often denote the mouth closing status. We take such valley points into consideration because these points always represent the connection position between the neighboring subunits. In general, speakers usually keep almost the same speaking pace during the entire utterance such that the frame length of each subunit differs not quite much from the others. Often, the frame length of the whole password sequence and the number of password elements are recorded to be known. Thereupon, the position of the starting frame corresponding to the current subunit, i.e. the ending frame of the previous subunit, can be computed by setting a pre-defined threshold $\Delta T$ as follows:

$$\begin{cases} T_{left} \le P_e^1 \le T_{right} \\ P_e^{i-1} + T_{left} \le P_e^i \le P_e^{i-1} + T_{right} \end{cases} \tag{10}$$

where $T_{left} = \frac{N_{frame}}{N_{element}} - \Delta T$ and $T_{right} = \frac{N_{frame}}{N_{element}} + \Delta T$ denote the left and right range, respectively. For consecutive digit motion separation, $\Delta T$ is often assigned to the values within the interval $[1, \frac{1}{3} \cdot \frac{N_{frame}}{N_{element}}]$.

We have carefully adjusted the parameters of methods [37]–[39] and selected the best results from three runs to make a comparison. Various digital lip-password sequences

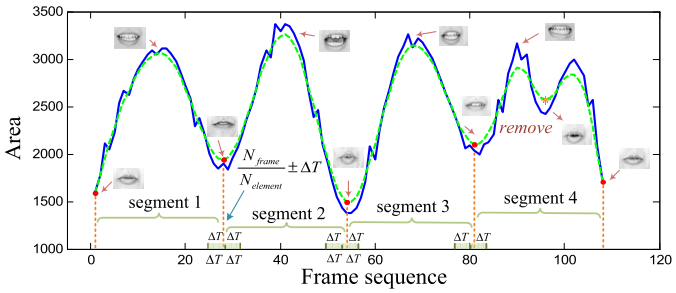


Fig. 3. Lip motion segmentation of the lip-password "6-5-8-7", in which the solid line denotes the original signal of mouth area variations, while the dotted line represents the filtered signal via forward-backward filtering.

TABLE I

THE SUBUNIT SEGMENTATION RESULTS, WHERE EACH CELL SHOWS THE SUBUNIT INTERVAL, THE TOTAL FRAMES AND SEGMENTATION ERROR

| Methods | Lip-password "6-5-8-7" [108 frames] | | | |
|---|---|---|---|---|
| | Subunit "6" | Subunit "5" | Subunit "7" | Subunit "8" |
| Method [37] | 1-27/27/0.04 | 28-54/27/0.07 | 55-81/27/0.04 | 82-108/27/0.00 |
| Method [38] | 1-24/24/0.07 | 25-52/28/0.11 | 53-79/27/0.11 | 80-108/29/0.07 |
| Method [39] | 1-26/26/0.00 | 27-54/28/0.04 | 55-80/26/0.07 | 81-108/28/0.04 |
| Our method | 1-26/26/0.00 | 27-54/28/0.04 | 55-82/28/0.07 | 83-108/26/0.04 |
| Ground truth | 1-26/26 | 27-53/27 | 54-81/28 | 82-108/27 |

have been tested and a typical example is shown in Fig. 3. It can be observed that the solid curve representing the area variations of the lip-password " 6-5-8-7" has many peak or valley points, while the dotted curve describing the processed signal only has some major peak or valley points. Accordingly, the proposed valley point searching scheme can successfully find intersection points between the neighboring subunits and simultaneously remove the irrelevant one. As a result, the lip motion corresponding to each subunit can be successfully separated. The segmentation results and manual annotation (i.e. ground truth) of this example are shown in Table I, where the segmentation error is defined as the sum of error or missing frames between the segmentation result and ground truth to the frame number of ground truth. It can be found that the result obtained by the proposed approach is close to the ground truth and the segmentation errors are small. Apparently, the proposed approach outperforms the method [38] and is even promisingly comparable to the Method [37] and [39]. Importantly, the proposed approach just utilizes the extracted mouth area and does not compute the intensity change of every pixel frame by frame. Therefore, the extra computation cost is not needed any more.

To achieve a more robust and realistic solution, the facial expressions generally tend to appear smoothly during the natural speaking process. Under such circumstance, a bit minor segmentation error (i.e. differ within 3 frames) will not degrade the lip motion analysis dramatically. The reason lies that the frames locating around the intersection points always represent the mouth closing status or transition frames, which are of less importance to the motion investigation. Moreover, the subunit motion differing a bit from the motion of the single digit utterance, would not impact the verification result significantly because the primary motions are preserved.

**Algorithm 1** Random Subspace Ensemble Method.

**Input:** Feature data set, $F = \{F_j, t_j\}, 1 \le j \le n, F_j \in \mathbb{R}^d, t_j \in C$, subspace dimension $k < d$, weak learning classifier $\mathcal{L}$, integer $R$ (i.e. the number of the ensemble learning rounds), and the tested feature sample $x$.

1: **for** $m = 1, \cdots, R$ **do**
2: Generate the random permutation index vector: $p^m = perm(\{1, \cdots, d\})$.
3: Select the index vector: $v^m = \{p_1^m, \cdots, p_k^m\}$.
4: Extract the features indicated by $v^m$: $F_{v^m}^* \subset F$.
5: $h_m = \mathcal{L}(F_{v^m}^*)$. // Step 2-4 denote RSM Projection.
6: **end for**
7: $\hat{h}(x) = \arg\max_{t \in C} \sum_{m=1}^{R} [h_m(x, F_{v^m}^*) = t]$.

**Output:** Final hypothesis $\hat{h}$.

### C. The Proposed Multi-Boosted HMMs Learning Approach

Let the positive value denote the target example and the negative value represent an imposter. According to Eq. (2) and Eq. (3), the decision stump for each weak learner in boosted HMMs can be formulated as:

$$h(\mathcal{O}_s) = \begin{cases} +1, & if\ LLR(\mathcal{O}_s)\ or\ NLL(\mathcal{O}_s) \ge \tau \\ -1, & otherwise. \end{cases} \quad (11)$$

As introduced in Section III-B, the lip motions within the lip-password utterances are usually comprised of several distinguishable units, which can be well separated using the proposed lip motion segmentation algorithm. Specifically, the frame length of each subunit motion can be easily aligned to be the same by using cubic interpolation method. Hence, by integrating the superiority of segmental scheme and boosting learning ability, the whole lip-password sequence can be verified via multi-boosted HMMs, whereby its discrimination power is stronger than a single HMM classifier acted on the whole sequence significantly.

Note that the design for the lip-password protected verification system should be able to reveal the password information and identity characteristics simultaneously. Nevertheless, the motion models learned from the fixed feature vector are incompetent for both of the verification tasks. In addition, the simple utilization of the whole feature vector may not achieve a satisfactory classification performance due to the feature redundant or overfitting problem [22]. As investigated in [41], the random subspace method (RSM) has been successfully utilized in ensemble approaches and demonstrated to perform well when there is a certain redundancy in the collection of feature vectors. The basic random subspace ensemble method is given in Algorithm 1. This ensemble operates by taking the majority vote of a predefined number of classifiers, each of which is built based on a different feature subset sampled randomly and uniformly from the original feature set. This type of approach will enhance the diversity of the base classifiers and often improve the overall classification accuracy within the ensemble approaches. Inspired by these findings, as shown in Fig. 4, we employ RSM to select different feature subsets so that various kinds of subunit motion models can be

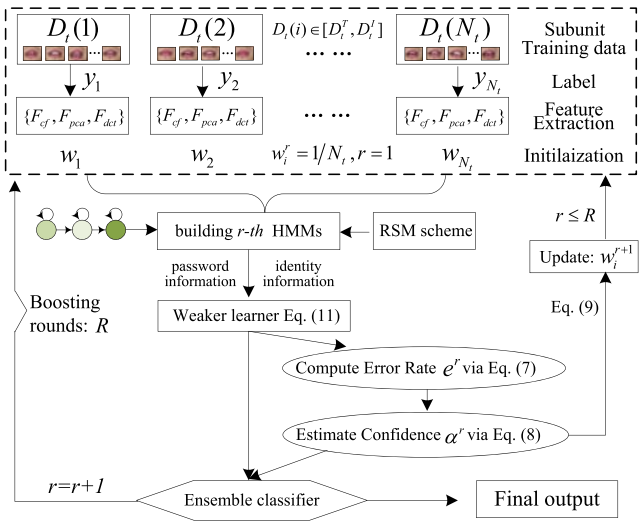


Fig. 4. The diagram of the boosted HMM learning for subunit verification.

learned discriminatively. As for the combined feature vector $F$, the feature components of different categories always share the distinct power for lip motion analysis. In general, the contour-based features are effective in verifying the different password elements while the appearance-based features would be useful in characterizing the identity information. In RSM, if one feature category is not included in the random subspace, the current base classifier may degrade the accuracy in verifying the lip-password information. Therefore, the selected feature subset should cover different kinds of feature components to characterize both of the password and identify information sufficiently. Therefore, Algorithm 2 is particularly utilized for the feature subset selection.

Moreover, as reported in [22], boosting method is especially utilized for large training sample size while the RSM is susceptible to the inadequate training samples. Nevertheless, a small number of training samples are usually available for learning because it is unamiable to ask the speakers to repeat their password phrases many times. Under the circumstances, we employ the data sharing scheme (DSS) proposed by Wang et al. [42] to form a novel train data set in pairs, which can generate more examples to reduce the impact of small sample size problem. Specifically, suppose there are a set of positive examples $A = \{x_1^a, x_2^a, \ldots, x_{N_a}^a\}$ of the target speaker and a set of negative examples $B = \{x_1^b, x_2^b, \ldots, x_{N_b}^b\}$ of imposter excluding the target speaker. From $A$ and $B$, we construct a new training set, where the positive examples are the pairs of the ones that are both from $A$, i.e. $\{(x_i^a, x_j^a)\}$, and negative examples $\{(x_i^a, x_j^b)\}$ are the pairs of examples that are from $A$ and $B$, respectively.

As introduced in Section II-B, the verification problem can be grouped into close-set and open-set cases. For the close-set case, the imposter model can be learned through the training data and the verification problem is performed based on the LLR. Nevertheless, as the imposters may have many different categories, it is very difficult to utilize one single model to represent all imposter modalities. Hence, we prefer to utilize the open-set case in our proposed approach. That is, each test sequence can generate an acceptance or rejection result via Eq. (3) by setting a decision threshold $\tau$. Let $\lambda$ be an HMM trained via data set $A$ of the target speaker, it can be concluded that the NLL of the target speaker conditioned on $\lambda$ should be larger than the NLL of the imposter conditioned on $\lambda$. Thereupon, we define a similarity score $h(x_i^a, x, \lambda)$ between $x_i^a$ and the testing sample $x$ as follows:

$$h(x_i^a, x, \lambda) = \left| NLL(x_i^a, \lambda) - NLL(x, \lambda) \right|. \tag{12}$$

Therefore, the similarity between the testing example $x$ and the whole positive data set $A$ can be measured as:

$$\hat{h}_{\min} = \min_{x_i^a \in A} h(x_i^a, x, \lambda), \tag{13}$$

where $x$ belongs to the target speaker if $\hat{h}_{\min} \le \tau$, and imposter otherwise. In other words, we compare the testing example with all the examples of the positive data set $A$ and take the highest score (i.e. minimum value) to make the decision. Since a number of HMMs are trained individually in ensemble learning approaches, the reduction of the computational load per HMM is also an important issue to be considered. Therefore, the Baum-Welch algorithm is selected to estimate the HMM parameters due to its less computations. As investigated in [27], the hard-to-classify samples should be treated differently for optimal parameter estimation. Therefore, the biased Baum-Welch method is adopted. Given an $N$-state-$M$-symbol HMM $\lambda = (\pi, A, B)$, we denote the training set consisting of $K$ observations as

$$\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_K\} \tag{14}$$

where $\mathcal{O}_k = \{o_1^k, o_2^k, \ldots, o_{l_k}^k\}$ is the $k^{th}$ sequence with $l_k$ observation frames and each observation is independent of each other. The Baum-Welch algorithm aiming at adjusting the parameters of the model $\lambda$ is to maximize:

$$P(\mathcal{O} \,|\lambda_i) = \prod_{k=1}^{K} P(\mathcal{O}_k \,|\lambda) = \prod_{k=1}^{K} P_k. \tag{15}$$

As shown in [28], we define the forward variables $\alpha_t^k(i) = P(o_1^k, o_2^k, \ldots o_t^k, s_t = S_i \,|\lambda)$ and backward variables $\beta_t^k(i) = P(o_{t+1}^k, o_{t+2}^k, \ldots o_{l_k}^k \,|s_t = S_i, \lambda)$ for observation $\mathcal{O}_k$. The parameters of HMM are estimated as follows:

$$\bar{a}_{i,j} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{l_k-1} \alpha_t^k(i) a_{i,j} b_j(\mathcal{O}_{t+1}^k) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{l_k-1} \alpha_t^k(i) \beta_t^k(j)} \tag{16}$$

$$\bar{b}_j(\ell) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1, \; s.t.\; O_t = v_\ell}^{l_k-1} \alpha_t^k(i) \beta_t^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{l_k-1} \alpha_t^k(i) \beta_t^k(j)} \tag{17}$$

where $v_\ell$ is the $\ell^{th} (1 \le \ell \le M)$ symbol output. In this strategy, all the samples are treated equally. As for the biased Baum-Welch estimation [27], the sample weights obtained from the boosting learning framework are employed.

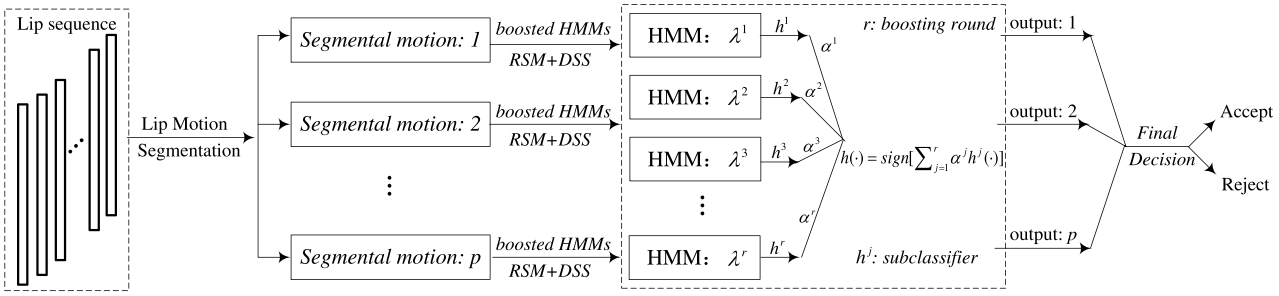


Fig. 5. The schematic view of the proposed multi-boosted HMMs learning approach to digital lip-password based speaker verification.

**Algorithm 2** Subset Selection via RSM for combined features

**Input:** Feature vector, $F = \{F_{cf}, F_{pca}, F_{dct}\}$, RSM sampling proportion $\rho \in (0, 1]$;

1: Obtained feature dimension $d_F = \{d_{cf}, d_{pca}, d_{dct}\}$.
2: Computed $k_{cf}=\lfloor \rho \cdot d_{cf} \rfloor$, $k_{pca}=\lfloor \rho \cdot d_{pca} \rfloor$, $k_{dct}=\lfloor \rho \cdot d_{dct} \rfloor$
3: $F^*_{cf} = RSM\_Projection(F_{cf}, k_{cf})$.
4: $F^*_{pca} = RSM\_Projection(F_{pca}, k_{pca})$.
5: $F^*_{dct} = RSM\_Projection(F_{dct}, k_{dct})$.
6: $F^*=\{F^*_{cf}, F^*_{pca}, F^*_{dct}\}$, $k=[k_{cf}, k_{pca}, k_{dct}]$.

**Output:** Feature subset $F^*$ of $k$ dimensionality.

In our proposed framework, the training samples are formulated in pairs. For the target speaker incorporates $K$ samples, the number of positive training data set is equal to $\frac{K(K-1)}{2}$. Let $w^T_{i,j} (1 \le i < j \le K)$ denote the weight of the coupled training sample $\{\mathcal{O}_i, \mathcal{O}_j\}$, the normalized weight for original target sample $\mathcal{O}_k (1 \le k \le K)$ is computed as:

$$\omega_k = \frac{\sum_{i=k \text{ or } j=k} w^T_{i,j}}{2 \cdot \sum_{i,j} w^T_{i,j}}. \tag{18}$$

By assigning this weight to the sample $\mathcal{O}_k$, the re-estimated parameters can be computed as:

$$\hat{a}_{i,j} = \frac{\sum_{k=1}^{K} \frac{\omega_k}{P_k} \sum_{t=1}^{l_k-1} \alpha^k_t(i) a_{i,j} b_j(O^k_{t+1}) \beta^k_{t+1}(j)}{\sum_{k=1}^{K} \frac{\omega_k}{P_k} \sum_{t=1}^{l_k-1} \alpha^k_t(i) \beta^k_t(j)} \tag{19}$$

$$\hat{b}_j(\ell) = \frac{\sum_{k=1}^{K} \frac{\omega_k}{P_k} \sum_{t=1, \; s.t.\; O_t=v_\ell}^{l_k-1} \alpha^k_t(i) \beta^k_t(j)}{\sum_{k=1}^{K} \frac{\omega_k}{P_k} \sum_{t=1}^{l_k-1} \alpha^k_t(i) \beta^k_t(j)}. \tag{20}$$

These estimated parameters are able to discriminatively model the lip motions such that some hard-to-classify samples can be verified. Therefore, as shown in Fig. 5, we integrate the HMMs with boosting learning framework associated with the RSM and DSS to precisely formulate a decision boundary for each subunit verification. Finally, the whole utterance whether spoken by the target speaker or not is verified according to all subunit verification results learned from multi-boosted HMMs. As summarized in Algorithm 3, if all subunit motions meet the accepted condition, the testing lip-password will be considered as the target-speaker saying the pre-registered password; otherwise, it will be an imposter.

## IV. Experimental Results and Discussion

In the visual speech domain, existing publicly available digital databases such as XM2VTS [8] and MVGL [14], have been widely utilized for multi-modal speech recognition and speaker identification. However, these databases are more or less incompetent for the evaluation of the lip-password based speaker verification problem. Therefore, a database consisting of digital password phrase uttered by 46 speakers (i.e. 28 males, 18 females) has been established. All the speakers were asked to phonically repeat the English phrase three-one-seven-five (3175) for twenty times ($\mathcal{D}_p$) and randomly say another ten different phrases ($\mathcal{D}_r$) of 4-digit length, which covered all "0"-"9" elements. All the utterances were captured by a high-quality digital video camera at 30 frames per second (fps) and recorded with almost the same speaking pace in a natural way. Specifically, as shown in Fig. 6, the recording was done in an office environment with almost uniform lighting conditions. The starting (ending) point of an utterance can be easily assigned to the earliest (latest) frame that has significant changes with respect to the first (last) frame in the feature sequence. The located and resized ROIs of lip images are of frontal view with a resolution of $112 \times 76$ pixels.

In the experiments, $\Delta T$ was empirically set at 5 for lip motion segmentation and the frame length of each subunit motion was aligned at 30. The boosting round $R$ was set at 30. The selected dimensionality of the PCA feature vector was fixed to be 80 and a $13 \times 13$ triangular mask was utilized to extract 2D-DCT coefficients of 91 dimensionality. The equal error rate (EER) [13] was utilized for performance evaluation and the value of $\tau$ in Algorithm 3 was deterred by the achievement of EER (i.e. the proportion of false acceptances was equal to the proportion of false rejections). To make a comparative evaluation, the configurations of GMM [12] and HMM [15] approaches were utilized for lip motion analysis. Meanwhile, these approaches associated with the proposed segmental scheme were denoted as "S-GMM" and "S-HMM", respectively. In addition, with the half of the data collections for training and testing, we followed the procedures of boosted GMM [25] and boosted HMM [27] to handle the speaker verification as well. Heuristically, we utilized a left to right HMM to train each subunit motion model and tested various

**Algorithm 3** Learning multi-boosted HMMs for lip-password based speaker verification

**Input:**

Training data set of lip-password sequences: $D$.
Password component number $p$, boosting rounds $R$.
The subspace dimension $k$ utilized in RSM.

**Multi-boosted HMMs:**

Feature extraction for each lip frame: $F$.
Lip motion segmentation, $D = \{D_1, D_2, \cdots, D_p\}$.
**for** $m = 1, \cdots, p$ **do**
  Get the training set $D_m^T = \{X_1^T, X_2^T, \cdots, X_{N_a}^T\}$ of the target speaker and $D_m^I = \{X_1^I, X_2^I, \cdots, X_{N_b}^I\}$ of the imposter, and form a novel training set via DSS [42].
  Initialize weights $w_{i,j}^T = \frac{2}{N_a(N_a-1)}$ with $1 \le i \le j \le N_a$, and $w_{i,j}^I = \frac{1}{N_a N_b}$ with $1 \le i \le N_a, 1 \le j \le N_b$. Set $r = 0, \varepsilon^r = 0$;
  **while** $r \le R$ and $\varepsilon^r < 0.5$ **do**
    Normalize the weights:

$$w_{r,i,j}^T = \frac{w_{r,i,j}^T}{\sum_{i',j'} w_{i',j'}^T + \sum_{i',j'} w_{i',j'}^I}$$

$$w_{r,i,j}^I = \frac{w_{r,i,j}^I}{\sum_{i',j'} w_{i',j'}^T + \sum_{i',j'} w_{i',j'}^I}$$

    Employ Algorithm 2 to obtain the $k$ dimensional feature subset in positive data set $D_m^T$ and build an HMM $\lambda_m^r(T)$ via Eq. (19) and Eq. (20).
    Call WeakLearner learning with respect to Eq. (13).
    Train a threshold $\tau_m$ to minimize the weighted classification error:

$$\varepsilon^r = \sum\nolimits_{i,j} w_{i,j}^T e_{r,i,j}^T + \sum\nolimits_{i,j} w_{i,j}^I e_{r,i,j}^I,$$

    where $e_{r,i,j}^T = 1$ if $h_m^r(X_i^T, X_j^T, \lambda_m^r(T)) \ge \tau_m$ and 0 otherwise, where $e_{r,i,j}^I = 1$ if $h_m^r(X_i^T, X_j^I, \lambda_m^r(T)) < \tau_m$ and 0 otherwise.
    Set $\alpha_m^r = \frac{1}{2}\log[(1-\varepsilon^r)/\varepsilon^r]$.
    Let $r = r + 1$ and update the weights to be:

$$w_{r+1,i,j}^T = w_{r,i,j}^T \cdot \exp(2\alpha_m^r e_{r,i,j}^T)$$
$$w_{r+1,i,j}^I = w_{r,i,j}^I \cdot \exp(2\alpha_m^r e_{r,i,j}^I)$$

  **end while**
  Obtain the similarity score between $X_p^T$ and $X_q$, where $X_p^T$ is from the data set of the target speaker:

$$\hat{h}_m(X_p^T, X_q) = \sum\nolimits_{w=1}^{r} \alpha_m^w h_m^w(X_p^T, X_q, \lambda_m^w(T)).$$

**end for**

**Output:**

Given the testing lip-password sequence $V = \{\nu_1, \nu_2, \cdots, \nu_p\}$, each subunit is verified via Eq. (13):

$$\hat{h}_{\min}^m = \min_{X_i^T \in D_m^T} \hat{h}_m(X_i^T, \nu_m).$$

If $\hat{h}_{\min}^m \le \tau$ for $m = 1, \cdots, p$, lip-password $V$ is verified to be uttered by target speaker, and otherwise not.

number of hidden states (i.e. 3-6 ) and Gaussian mixture components (i.e. 1-5) in the experiments. Due to the limited subunit motion frames, experimental results showed that the left to right HMM with three states, two continuous density

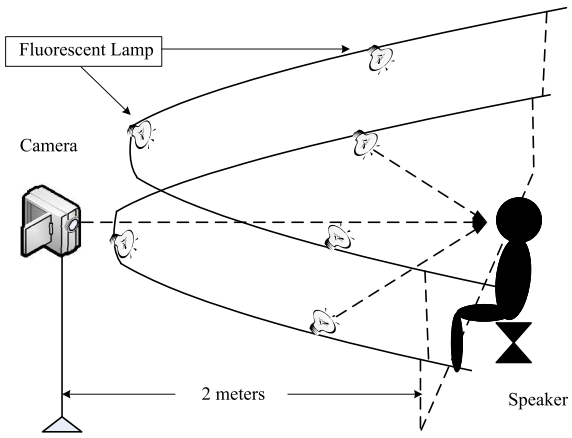


Fig. 6. The simulative circumstance for database construction.

Gaussian mixtures with diagonal covariance matrix output delivers the best performance. Specifically, this kind of HMM parameter setting is selected for analysis. With the same collections of the target speaker saying the registered password, we shall concentrate on verifying the following three types of imposters: (1) The target speaker saying the incorrect passwords; (2) The imposter saying the correct password; (3) The imposter saying the incorrect passwords.

### *A. The Target Speaker Saying the Incorrect Passwords*

In this case, the lip-passwords differing from the registered one (i.e. 3175) are considered as the imposters. The database $\mathcal{D}_p$ is divided into two disjoint data sets: $\mathcal{D}_{p_1}$ and $\mathcal{D}_{p_2}$, each of which has ten repetitions of the same utterance from each speaker. The subset $\mathcal{D}_{p_1}$ is utilized for training, while $\mathcal{D}_{p_2}$ and $\mathcal{D}_r$ are utilized for testing. The model of the target speaker saying the registered password can be trained through the data set $\mathcal{D}_{p_1}$. Since the utterances of different lip-passwords are selected to be imposters, the imposter category cannot be well determined due to its arbitrariness. Fortunately, the proposed lip motion segmentation scheme has the ability to make each subunit imposter category determined (i.e. the digits only from "0" to "9"), while the whole utterance fails. We employed the leave-one-out scheme [14] to generate imposters, where each segmental unit not belonging to the current subunit was selected as an imposter. For instance, all the subunit motions differing from the target subunit "3" are considered to be the imposters, i.e. the segmental motions of digits "0-2, 4-9". Under such circumstances, the number of imposter category for each subunit was equal to 9. We randomly selected one segmental motion of each digit "0-9" from $\mathcal{D}_{p_1}$ and $\mathcal{D}_r$ to form the imposter data. The RSM sampling proportion $\rho$ was fixed to 0.7 and the DSS was employed to form the training data set in pairs. For each speaker, the total numbers for positive training examples and negative training examples were equal to 45 and 90, respectively.

The experimental results are shown in Table II. It can be observed that the performance of S-GMM and S-HMM methods each outperforms the non-segmental approaches, i.e. single GMM [12] and HMM [15]. The EER value obtained

TABLE II
THE VERIFICATION RESULTS OF THE TARGET SPEAKER SAYING THE INCORRECT PASSWORDS

| Feature set | Equal Error Rate [EER %] (The operating point where the FAR equals to FRR) | | | | | | |
|---|---|---|---|---|---|---|---|
| | GMM [12] | HMM [15] | S-GMM | S-HMM | boosted GMM [25] | boosted HMM [27] | M-boosted HMM+RSM (70%) |
| $F_{cf}$ | 17.82 | 14.56 | 14.13 | 12.39 | 14.56 | 13.26 | 7.39 |
| $F_{pca}$ | 19.13 | 16.95 | 15.21 | 14.34 | 16.3 | 14.13 | 8.04 |
| $F_{dct}$ | 18.47 | 16.08 | 14.56 | 14.13 | 13.26 | 13.47 | 7.82 |
| $F_{cf} + F_{pca}$ | 13.47 | 11.52 | 10.21 | 10.43 | 12.39 | 10.86 | 4.34 |
| $F_{cf} + F_{dct}$ | 13.04 | 11.95 | 11.52 | 9.78 | 12.17 | 10.43 | 4.78 |
| $F_{pca} + F_{dct}$ | 13.91 | 12.6 | 11.08 | 10.65 | 12.82 | 11.73 | 5.21 |
| $F_{cf} + F_{pca} + F_{dct}$ | 12.17 | 12.39 | 11.73 | 11.3 | 13.91 | 11.08 | 3.91 |

by S-HMM method associated with the $F_{cf} + F_{dct}$ was less than 10%. The reason lies that the segmental scheme is capable of providing more detailed information that is not easily revealed in the whole lip-password sequence. For example, the incorrect password "3178" just has one different element which is so similar to the registered one such that this utterance may fail to be distinguished under non-segmental methods. Meanwhile, some imposter motion uttered by the same speaker often differs slightly from the registered one. For example, some imposter subunit motions of "0" and "8" were somewhat similar to the motions of "7" and "5", respectively, uttered by some speakers, which often failed to be verified under the segmental scheme. Under such circumstances, the boosted learning framework aiming at paying more attention on hard-to-classify samples would hold the promise of verifying these similar examples. However, the boosted GMM [25] and boosted HMM [27] approaches taking the whole utterance as the basic processing unit may not always deliver a better result than non-boosted methods. For example, the imposter utterances, e.g. "3715", "3157" of some speaker, often failed to be verified along this way. The main reason lies that these similar utterance associated with the fixed feature vector cannot be verified within the very limited training samples. In contrast, the proposed multi-boosted HMMs learning approach was able to detect these imposters, meanwhile boosting the performance. The main reasons are two-fold: (1) The segmental scheme has a ability to make each imposter category determined; (2) The utilization of RSM and DSS can solve the feature overfitting and small training sample size problem. As a result, the EER values incorporating the different kinds of feature vectors were all less than 10%.

To reveal the ambiguity between different digital motions, we extracted 10 subunit motions of all "0"-"9" digits and selected half size of the collected samples for training and testing. For each digit uttered by the same speaker, the DSS was employed to construct the training samples in pairs. The total number for the testing samples was equal to 230. Under the segmental scheme, we utilized the fixed HMM parameter settings with the whole extracted features and extended the booted HMM [27] with the utilization of DSS and RSM to classify each digital lip motion (simply called boosted S-HMM hereinafter), in which the one generating the maximum probability was chosen as the identification

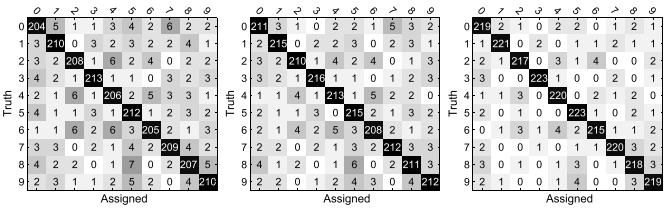


Fig. 7. Confusion Matrix on segmental digits classification. (a) The classification results obtained by S-GMM; (b) the classification results obtained by S-HMM; (c) the classification results obtained by boosted S-HMM.

result. The classification results in terms of the Confusion Matrix are shown in Fig. 7, in which the entry in the $i^{th}$ row and $j^{th}$ column is the number of digit from class $i$ that is misidentified as class $j$. It can be found that the S-GMM and S-HMM approaches often failed to well identify some digital lip motions due to their limited discrimination power. The reason lies that some segmental utterances corresponding to different digits often produce ambiguous lip motions visually, which often make it difficult to be recognized via a single classifier. In contrast, the boosted S-HMM has achieved better classification performance. For example, the segmental lip motion corresponding to digit "0" was just misclassified as digit "7" for only once, which was greatly decreased in comparison with S-GMM (i.e. six times) and S-HMM (i.e. five times) approaches. That is, the boosted learning strategy incorporating more discrimination power is able to verify most of the ambiguous samples, which can be utilized to verify the wrong password information effectively within the proposed lip-password protected speaker verification system.

### B. The Imposter Saying the Correct Password

In this case, the subset $\mathcal{D}_{p_1}$ was utilized for training while $\mathcal{D}_{p_2}$ was adopted for testing. We followed the leave-one-out scheme to generate the imposter samples, i.e. each speaker became an imposter datum of the remaining speakers. Note that each testing sample can be applied as an imposter for different target speakers simultaneously. Given a pre-defined target speaker within the data set $\mathcal{D}_{p_2}$, the resultant number of the imposter data excluding the target speaker for the true rejection was 450. We randomly selected two examples of each speaker excluding the target speaker from subset $\mathcal{D}_{p_1}$ to

TABLE III
THE VERIFICATION RESULTS OF THE IMPOSTER SAYING THE CORRECT PASSWORD

| Feature set | Equal Error Rate [EER %] (The operating point where the FAR equals to FRR) | | | | | | |
|---|---|---|---|---|---|---|---|
| | GMM [12] | HMM [15] | S-GMM | S-HMM | boosted GMM [25] | boosted HMM [27] | M-boosted HMM+RSM (70%) |
| $F_{cf}$ | 23.29 | 19.14 | 17.44 | 14.61 | 15.69 | 13.31 | 9.78 |
| $F_{pca}$ | 22.23 | 18.24 | 17.18 | 14.47 | 16.12 | 13.74 | 8.58 |
| $F_{dct}$ | 21.11 | 17.55 | 17.07 | 13.94 | 15.32 | 11.98 | 7.63 |
| $F_{cf} + F_{pca}$ | 19.77 | 17.34 | 16.17 | 12.27 | 14.52 | 10.56 | 5.69 |
| $F_{cf} + F_{dct}$ | 19.33 | 16.22 | 15.27 | 12.76 | 14.26 | 11.02 | 5.37 |
| $F_{pca} + F_{dct}$ | 18.88 | 16.86 | 14.31 | 11.31 | 12.21 | 10.79 | 4.87 |
| $F_{cf} + F_{pca} + F_{dct}$ | 16.88 | 15.74 | 13.78 | 10.15 | 10.58 | 11.16 | 4.06 |

TABLE IV
THE VERIFICATION RESULTS OF THE IMPOSTER SAYING THE INCORRECT PASSWORDS

| Feature set | Equal Error Rate [EER %] (The operating point where the FAR equals to FRR) | | | | | | |
|---|---|---|---|---|---|---|---|
| | GMM [12] | HMM [15] | S-GMM | S-HMM | boosted GMM [25] | boosted HMM [27] | M-boosted HMM+RSM (70%) |
| $F_{cf}$ | 13.31 | 11.7 | 8.35 | 8.21 | 7.33 | 7.57 | 5.69 |
| $F_{pca}$ | 10.23 | 8.32 | 9.67 | 8.09 | 6.18 | 7.28 | 5.11 |
| $F_{dct}$ | 9.61 | 8.58 | 9.31 | 8.76 | 6.38 | 6.21 | 4.15 |
| $F_{cf} + F_{pca}$ | 8.23 | 6.69 | 6.47 | 6.93 | 5.98 | 5.43 | 4.16 |
| $F_{cf} + F_{dct}$ | 8.61 | 7.53 | 6.91 | 7.72 | 5.43 | 5.49 | 3.77 |
| $F_{pca} + F_{dct}$ | 7.88 | 6.3 | 6.45 | 6.12 | 5.25 | 5.05 | 3.48 |
| $F_{cf} + F_{pca} + F_{dct}$ | 7.35 | 6.47 | 6.35 | 5.77 | 4.87 | 4.53 | 3.37 |

construct the imposter training data. For each target speaker, the total numbers for positive training examples and negative training examples were equal to 45 and 900, respectively.

The experimental results are listed in Table III. For the same utterance, the motion modeling approaches through the whole utterance were able to characterize the temporal pattern over the segments, which were expected to obtain a better performance. Nevertheless, the EER values obtained by a single GMM or HMM based approaches were all greater than 15%, which always failed to verify most samples within the large imposter categories due to their limited discrimination power. In contrast, the S-GMM and S-HMM methods can improve the verification performance to a certain degree. The reason lies that the segmental scheme would obtain more detailed information within a short period of motions to verify some similar speakers. Although the boosted learning methods have been demonstrated to be successful in increasing the robustness of the verification performance, the boosted GMMs [25] and boosted HMMs [27] taking the whole utterance as the basic processing unit also degraded their performance due to the very limited training samples. Subsequently, the EER values were all higher than 10%. Comparatively speaking, the proposed multi-boosted HMMs learning approach integrating the advantages of the segmental scheme and boosted learning ability was able to formulate a precise decision boundary discriminatively so that some hard-to-classify speakers can be verified. Although the proposed approach does not model the temporal pattern over the segments, the concatenation of each segmental motion modeling was able to characterize the significant information as well within the whole sequence for speaker verification. Accordingly, the promising verification results with all EER values less than 10% were obtained. In particular, the feature vector $F_{cf} + F_{pca} + F_{dct}$ with 70% subspaces has yielded much better performance, with the EER value equal to 4.06% only, in comparison with the other kinds of feature vectors.

### C. The Imposter Saying the Incorrect Passwords

In this case, the subset $\mathcal{D}_{p_1}$ was utilized for training while $\mathcal{D}_{p_2}$ and $\mathcal{D}_r$ were adopted for testing. The imposter model cannot be determined due to its diversity and arbitrariness. Fortunately, the segmental scheme can make the imposter categories of subunit element determined, i.e. the imposters with all subunit sequences can be selected as the imposter datum. Therefore, the maximum number of the imposter categories for each subunit was equal to 450. According to the collections in the previous two experiments, we randomly selected two target speakers and one subunit sample of each imposter category to form the negative training examples. Accordingly, the resultant numbers for positive training samples and negative training samples were equal to 45 and 900, respectively.

The experimental results are listed in Table IV. It can be found that the majority of the EER values obtained by different kinds of methods are less than 10%. The reason lies that the lip-password sequences differing from the registered one and uttered by the different speakers are significantly distinct from the sequence of the target speaker saying the registered password. That is, the imposters saying the incorrect passwords always encode the significantly valuable information to be easily identified. In this case, the approaches associated with the segmental scheme, i.e. S-GMM, S-HMM, may not

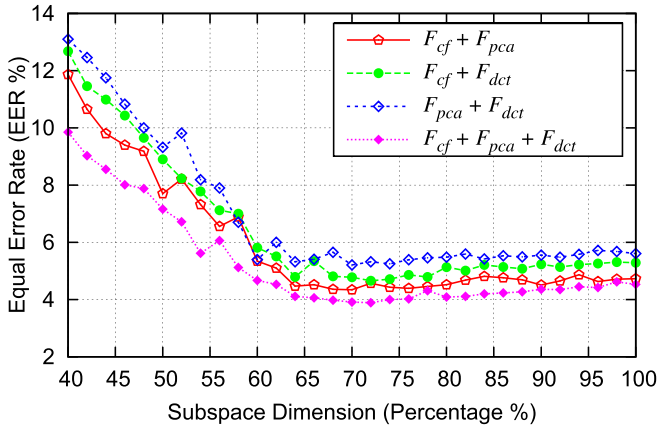


Fig. 8. The verification performance of the target speaker saying the incorrect passwords via different subspace dimensions.

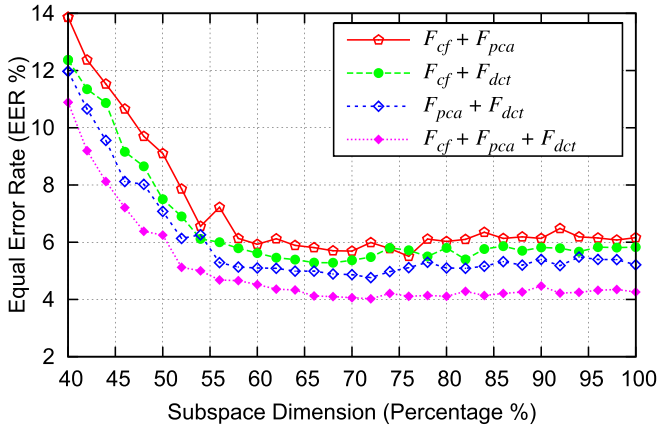


Fig. 9. The verification performance of the imposter saying the correct password via different subspace dimensions.

always deliver the better results than non-segmental methods. Within the limited training samples, the boosted GMM [25] and boosted HMM [27] approaches also failed to achieve a significant improvement of verification performance by taking the whole lip-password sequence as the basic processing unit. In contrast, the proposed approach kept achieving a better verification result in terms of the lower EER values. The reasons are two-fold: (1) The segmental scheme is capable of making each imposter category of subunit determined so that the establishment of the negative training samples can be succeeded; (2) The utilization of boosting learning framework associated with the RSM and DSS, can not only solve the feature overfitting and small training sample size problem, but also significantly increase the discrimination power to verify some hard-to-classify examples, e.g. some failed examples during the training phase.

Next, the EER values performed on different feature combinations with various subspace dimensions are shown in Figs. 8–10, respectively. For the first case, it can be found that the contour-based features $F_{cf}$ associated with $F_{pca}$ or $F_{dct}$ feature vector generally yield better performance than the feature vector $F_{pca}+F_{dct}$ in terms of EER values. This implies that the contour-based features are of crucial importance to the

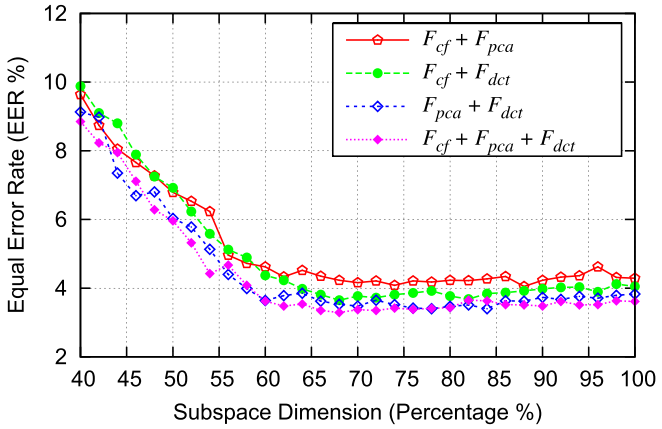


Fig. 10. The verification performance of the imposter saying the incorrect passwords via different subspace dimensions.

verification of different password subunits because the lip contours always have significantly different moving trajectories between different password elements. Comparatively speaking, the texture features serve as an important discrimination information especially in identifying the imposters saying the correct password. The main reason lies that the appearances of teeth and tongue are always diverse between the different speakers such that the utilization of texture biometrics of lip motions can well verify the imposter speakers. Moreover, it can be seen that the subspace dimension with 65-75% of original feature vectors always reports the lower EER values while the direct utilization of all the extracted feature vectors may not always generate the best performance. The reason lies that the utilization of RSM resampling the feature vector into different kinds of low dimensional subsets not only has a strong ability to solve the feature overfitting problem, but also would increase the discrimination power to improve the verification performance. Meanwhile, it would not obtain a good verification result when the size of subspace dimensionality is too small because the weaker learners in boosting learning framework are not able to learn well when the data feature vectors are too uninformative, e.g. the subspace dimensionality with 40-50% of the original feature vectors have delivered the unsatisfied results. As discussed in paper [43], diversity has been recognized as an important factor to the success of classifier ensemble approaches. Within the proposed learning framework, the sampling distribution is generally employed to resample the training data sets for subsequent component classifier learning. As a result, the likelihood for those samples which have been misclassified by the previous component classifier is increased so that the classifier ensemble becomes progressively diverse. In addition, the feature subset obtained by RSM is also capable of making each training motion model diverse synchronously. That is, the predictions obtained from each component classifier are not equal such that the learned ensemble classifier would become diverse as well.

Further, we maintained the segmental scheme and conducted the above experiments on the unpaired training samples. That is, the DSS was not employed. The EER values performed on two training patterns are listed in Table V, in which only

TABLE V
THE VERIFICATION RESULTS WITHIN THE DIFFERENT TRAINING PATTERNS

| Training Patterns | Equal Error Rate [EER %] | | |
|---|---|---|---|
| | Case A | Case B | Case C |
| Non-DSS + RSM (70%) | 9.13 | 8.25 | 4.93 |
| DSS + RSM (70%) | 3.91 | 4.06 | 3.37 |

the feature set $F_{cf}+F_{pca}+F_{dct}$ associated with RSM (70%) was employed. It can be found that the EER values obtained through the unpaired training samples (i.e. Non-DSS + RSM) were all larger than the results generated by the proposed framework (i.e. DSS + RSM). A plausible reason lies that overfitting is inevitable when the limited training sets are available for each component classifier learning. To avoid the variability caused by the small training sets, the size of the positive and negative training sets should be relatively large. Under such circumstances, the DSS aiming to train a generic classifier that determines any two examples coming from the same target or not, would construct more training samples for learning. Accordingly, the proposed boosted HMMs learning framework can effectively reduce the chance of the overfitting occurrence. As a result, the satisfactory verification performance is achieved. To the best of our knowledge, the proposed multi-boosted HMM learning framework is the first one that inherently incorporates the RSM and DSS to overcome potential overfitting issues due to the features redundant and the lack of training samples. The experimental results have demonstrated the efficacy of the proposed approach in comparison with the state-of-the-art methods.

## V. CONCLUDING REMARKS

In this paper, we have proposed the concept of lip-password, which has provided a double security to the speaker verification system. That is, a speaker will be verified by both of the password embedded in the lip motion and the underlying behavioral biometrics of lip motions simultaneously. To this end, we have presented a multi-boosted HMMs learning approach to solving such lip-password based speaker verification problem. Within the presented approach, an effective lip motion segmentation approach is addressed to segment the lip-password sequence into a small set of distinguishable subunits so that the more detailed motion information can be obtained. Further, the utilization of RSM can not only circumvent the occurrence of feature overfitting problem, but be also capable of making each component classifier diverse and increasing the discrimination power of the learning framework. Moreover, the adoption of DSS reorganizing the training samples in pairs, is able to solve the small training sample size problem. The experiments have shown the efficiency of the proposed approach in comparison with the existing counterparts.

Along the line of the present work, there still exist some open problems for further studies. For example, how to adaptively learn the optimal parameters for lip motion analysis and how to effectively verify the non-digital lip-password have yet to be studied. Further, from a practical viewpoint, it would also be useful to extend the algorithm to the less constrained conditions, especially for the large variations in speaking pace, facial expressions and illumination conditions. We shall leave them somewhere in our future work.

## REFERENCES


[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[2] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 802–809, Dec. 2010.

[3] A. Roy, M. Magimai-Doss, and S. Marcel, "A fast parts-based approach to speaker verification using boosted slice classifiers," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 1, pp. 241–254, Feb. 2012.

[4] E. Engin, 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] G. Chetty and M. Wagner, "Robust face-voice based speaker identity verification using multilevel fusion," *Image Vis. Comput.*, vol. 26, no. 9, pp. 1249–1260, Sep. 2008.

[6] A. K. Sao and B. Yegnanarayana, "Face verification using template matching," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 636–641, Sep. 2007.

[7] C. Chi Ho, B. Goswami, J. Kittler, and W. Christmas, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 602–612, Apr. 2012.

[8] M. I. Faraj and J. Bigun, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1169–1175, Sep. 2007.

[9] M. N. Kaynak, Z. Qi, A. D. Cheok, K. Sengupta, J. Zhang, and C. Ko Chi, "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 4, pp. 564–570, Jul. 2004.

[10] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," in *Proc. IEEE 4th Int. Conf. Spoken Lang.*, vol. 1. Oct. 1996, pp. 62–65.

[11] T. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *Proc. IEEE Int. Conf. Pattern Recognit.*, vol. 1. Aug. 1998, pp. 123–125.

[12] F. Shafait, R. Kricke, I. Shdaifat, and R. R. Grigat, "Real time lip motion analysis for a person authentication system using near infrared illumination," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 1957–1960.

[13] M. I. Faraj and J. Bigun, "Person verification by lip-motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2006, pp. 37–44.

[14] H. E. Cetingul, Y. Yemez, E. Engin, 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. 2879–2891, Oct. 2006.

[15] 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. Acoust., Speech, Signal Process.*, vol. 3. May 2004, pp. 397–400.

[16] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 701–714, Jun. 2007.

[17] X. Liu and Y. M. Cheung, "A multi-boosted HMM approach to lip password based speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 2197–2200.

[18] C. T. Lin, H. W. Nein, and W. C. Lin, "A space-time delay neural network for motion recognition and its application to lipreading," *Int. J. Neural Syst.*, vol. 9, no. 4, pp. 311–334, 1999.

[19] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[20] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 11. Apr. 1986, pp. 49–52.

[21] S. Fei and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4. Apr. 2007, pp. 313–316.
[22] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.
[23] T. Hao and T. S. Huang, "Boosting Gaussian mixture models via discriminant analysis," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.
[24] Y. Pei, I. Essa, and J. M. Rehg, "Asymmetrically boosted HMM for speech reading," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jul. 2004, pp. 755–761.
[25] S. Z. Li, D. Zhang, C. Ma, H. Y. Shum, and E. Chang, "Learning to boost GMM based speaker verification," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 1677–1680.
[26] S. Man Hung, Y. Xi, and G. Herbert, "Discriminatively trained GMMs for language classification using boosting methods," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 1, pp. 187–197, Jan. 2009.
[27] S. W. Foo, Y. Lian, and L. Dong, "Recognition of visual speech elements using adaptively boosted hidden Markov models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 693–705, May 2004.
[28] 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.
[29] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, 1999.
[30] S. L. Wang and A. W. C. Liew, "ICA-based lip feature representation for speaker authentication," in *Proc. 3rd Int. Conf. Signal, Image Technol. Internet, Based Syst.*, 2007, pp. 763–767.
[31] A. Mehra, M. Kumawat, R. Ranjan, B. Pandey, S. Ranjan, A. Shukla, *et al.*, "Expert system for speaker identification using lip features with PCA," in *Proc. 2nd Int. Workshop Intell. Syst. Appl.*, May 2010, pp. 1–4.
[32] X. Liu and Y. M. Cheung, "A robust lip tracking algorithm using localized color active contours and deformable models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 1197–1200.
[33] S. Wang, W. Lau, and S. Leung, "Automatic lip contour extraction from color images," *Pattern Recognit.*, vol. 37, no. 12, pp. 2375–2387, 2004.
[34] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Lip contour extraction from color images using a deformable model," *Pattern Recognit.*, vol. 35, no. 12, pp. 2949–2962, 2002.
[35] J. S. Lee and C. H. Park, "Hybrid simulated annealing and its application to optimization of hidden Markov models for visual speech recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1188–1196, Aug. 2010.
[36] M. W. Mak and W. G. Allen, "Lip-motion analysis for speech segmentation in noise," *Speech Commun.*, vol. 14, no. 3, pp. 279–296, 1994.
[37] W. C. Yau, H. Weghorn, and D. K. Kumar, "Visual speech recognition and utterance segmentation based on mouth movement," in *Proc. Biennial Conf. Austral. Pattern Recognit. Soc. Digital Image Comput. Tech. Appl.*, vol. 8. 2007, pp. 7–14.
[38] H. Talea and K. Yaghmaie, "Automatic visual speech segmentation," in *Proc. IEEE 3rd Int. Conf. Commun. Softw. Netw.*, May 2011, pp. 184–188.
[39] A. A. Shaikh, D. K. Kumar, and J. Gubbi, "Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments," *Vis. Comput.*, vol. 29, no. 10, pp. 1–14, 2012.
[40] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 988–992, Apr. 1996.
[41] H. Tin Kam, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
[42] X. Wang, C. Zhang, and Z. Zhang, "Boosted multi-task learning for face verification with applications to web image and video search," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Feb. 2009, pp. 142–149.
[43] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.

**Xin Liu** (M'08) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2013. He is currently with the Department of the Computer Science and Technology, Huaqiao University, Xiamen, China. His research interests include image processing, computer vision, pattern recognition, and medical image analysis.

**Yiu-ming Cheung** (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, in 2000. Currently, he is a Professor with the Department of Computer Science, Hong Kong Baptist University. His research interests include machine learning, information security, signal processing, pattern recognition, and data mining. He is the Founding Chair of the Computational Intelligence Chapter of IEEE Hong Kong Section. He is a senior member of ACM.