A MULTI-BOOSTED HMM APPROACH TO LIP PASSWORD BASED SPEAKER VERIFICATION

Xin Liu and Yiu-ming Cheung

Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
{xliu,ymc}@comp.hkbu.edu.hk

ABSTRACT

This paper presents a multi-boosted Hidden Markov Model (HMM) approach to lip password (i.e. the password embedded in the lip motion) based speaker verification, where the speaker is verified by both of lip password and the underlying characteristics of lip motions. That is, the target speaker saying the wrong password or an impostor even knowing the correct password will be detected as well. To this end, we firstly propose an effective lip motion segmentation algorithm to segment the password sequence into a small set of discrete subunits. Then, we integrate HMMs with boosting learning framework associated with the random subspace method (RSM) and data sharing scheme (DSS) to model the segmental sequence of the input subunit discriminatively so that a precise decision boundary is formulated for these subunits verification. Finally, the speaker is verified based on all verification results of the subunits learned from multi-boosted HMMs. Experimental results show the promising results.

Index Terms— Lip Password, Lip Motion, Speaker Verification, Multi-boosted HMMs.

1. INTRODUCTION

Speaker verification (SPV) has received considerable attention in the community because of its potential applications in financial transaction, secure access, human-computer interfaces and so forth [1]. In general, speech not only conveys the linguistic information, but also characterizes the speaker's identity, which can therefore be utilized for SPV [2]. Face and acoustic speech signal may be the most natural modalities to achieve SPV, but which, unfortunately, suffers from some limitations. In the former modality, the SPV system utilizing the still face image is very susceptible to the poor picture quality, variations in pose or facial expressions, and also easily deceived by a face image placed in front of the camera. In the latter, the SPV system will be quite sensitive to the environment. Evidently, such a system will always degrade its performance dramatically in the noisy or multi-speaker environment. Under the circumstances, the SPV fused with lip motions has shown an improved performance over pure acoustic systems [3]. Actually, as a kind of behavior characteristics, the lip motions accompanying with the lip movements, tongue and teeth visibility contain extremely rich information for speaker verification. Hence, it is feasible to develop a lip-motion based approach to speaker verification. Compared to the acoustic speech based SPV, the advantages of lip-motion based one are at least three-fold: (1) Lip-motion based SPV system is completely insensitive to the background noise; (2) Such a system can be utilized in a moderate distance; (3) Such an SPV system is easily applicable to a dumb person. In the literature, traditional lip-motion based SPV systems just adopt a fixed scheme of utilizing a single Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) for lip-motion modeling and similarity measurement, thus their discrimination power is very limited. To the best of our knowledge, the performance of the current lip-motion based SPV systems is far behind our expectations.

In this paper, we shall concentrate on digital lip-motion based SPV, in which a lip password (i.e. the password embedded in the lip motion) based SPV is presented. Definitely, password protected SPV system will hold a double security to the system, where a speaker is verified by both of lip password and the underlying characteristics of lip motions simultaneously. That is, the target speaker saying the wrong password or an impostor even knowing the correct password will be detected and rejected as well. In general, the password utterance comprises several visibly distinguishable units (*i.e.,* subunit). Each subunit indicates a short period of lip motion and always has diverse styles between different elements. To investigate the lip password in detail, these subunits should be considered individually instead of taking into account the whole utterance as the basic processing unit. To this end, we firstly present an effective lip-motion segmentation algorithm to segment the password sequence into several subunits. Then, we integrate HMMs with boosting learning framework associated with the random subspace method (RSM) and data sharing scheme (DSS) [4] to model the input subunit sequence discriminatively so that a precise decision boundary is formulated for these subunits verification. Finally, the

The work was supported by the Research Grant Council of Hong Kong SAR with Project No: HKBU 210309 and a Faculty Research Grant of Hong Kong Baptist University with Project No: FRG2/09-10/098 and FRG2/10-11/056. Yiu-ming Cheung is the corresponding author.

speaker is verified based on all verification results of the subunits learned from multi-boosted HMMs. The experimental results have shown the promising results.

2. OVERVIEW OF HMM-BASED SPV

In general, the modality for HMM-based SPV can be regarded as a binary classification between the target-speaker $\lambda(T)$ and impostor $\lambda(I)$, which can be formulated as either the closed-set or open-set learning problem. Specifically, let $\mathcal{O}_s = \{o_1^s, o_2^s, \cdots, o_{l_s}^s\}$ be a test observed sequence. In the closed-set problem, the testing utterances of the speakers are recorded to be known, and the models of both the target-speakers and imposters can be trained during the training phase. The classification for SPV is performed based on the log likelihood ratio (LLR):

$$LLR(\mathcal{O}_s) = \sum_{t=1}^{l_s} [\log P(o_t^s|\lambda(T)) - \log P(o_t^s|\lambda(I))]$$
$$if \;\; LLR(\mathcal{O}_s) \geq \tau : \; accepted. \qquad (1)$$
$$Otherwise : \; reject.$$

In the open-set problem, the imposter cannot be trained due to its arbitrariness. The task is to find whether the test sequence belongs to the target speaker registered in the database or not. Since the frame length of the utterance often changes even among the same phrase uttered by the same speaker, the following normalized log likelihood (NLL) is therefore often employed:

$$NLL(\mathcal{O}_s) = \frac{1}{l_s} \sum_{t=1}^{l_s} \log P(o_t^s|\lambda(T)).$$
$$if \;\; NLL(\mathcal{O}_s) \geq \tau : \; accepted. \qquad (2)$$
$$Otherwise : \; reject.$$

3. THE PROPOSED APPROACH

3.1. Lip Motion Segmentation

Lip motion segmentation aims at detecting the start and stop frames of subunit utterance from a group of lip sequence. In general, the mouth areas change significantly as the frame length increases. The point position with the minimum mouth area always represents the status of mouth closing or intersection point between subunit utterances. Accordingly, the proposed lip motion segmentation approach is comprised of three phases: First, we obtain the signal $\mathcal{A}_c$ in terms of the mouth area variations via lip tracking [5]. Next, we utilize a forward-backward filtering [6] to process the input area signal $\mathcal{A}_c$ in both the forward and reverse directions such that the filtered signal $\mathcal{A}_c^f$ is obtained. Finally, we can easily obtain the positions of the valley points via the filtered signal $\mathcal{A}_c^f$. In general, speakers usually keep the same speaking pace during the utterance. Therefore, the frame length of each subunit differs not quite much from each other. If the frame length of the whole utterance and the number of subunit elements that are recorded are known, the intersection points can be computed within a pre-defined threshold ΔT as follows:

$$\begin{cases} T_{left} \leq P_e^1 \leq T_{right} \\ P_e^{i-1} + T_{left} \leq P_e^i \leq P_e^{i-1} + T_{right} \end{cases} \qquad (3)$$

where $T_{left} = \frac{N_{frame}}{N_{element}} - \Delta T$ and $T_{right} = \frac{N_{frame}}{N_{element}} + \Delta T$ are the left and right threshold value, respectively.

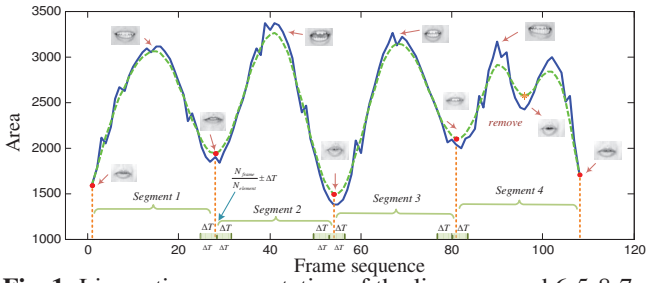

Fig. 1. Lip motion segmentation of the lip-password 6-5-8-7.

Fig. 1 shows an example, in which the solid curve representing the area variations of the password utterance 6-5-8-7, has many unstable peak or valley points. In contrast, the dotted curve describing the processed signal performed by the forward-backward filtering only has some major peak or valley points. According to the constrains of Eq. (3), the proposed valley point searching method can successfully find the intersection points between the subunits. Meanwhile, the valley point that does not belong to intersection points can be removed.

3.2. Discriminative Learning

In general, discriminative learning in the existing HMM-based systems mainly includes discriminative feature selection and discriminative model learning [7]. Discriminative feature selection aiming at minimizing the classification loss will not only emphasize the informative features, but also filter out the irrelevant ones. However, the features in each category are not statistically independent. It is very difficult to determine which single feature component has more discrimination power. Discriminative model learning approaches featuring on parameter optimizations always achieve a better result than non-discriminative learning approaches, *e.g.,* Maximum Mutual Information (MMI), conditional maximum likelihood (CML) and minimum classification error (MCE) [7]. These methods aiming at maximizing the conditional likelihood or minimizing the classification error rate are usually superior than Maximum Likelihood Estimation (MLE) approaches, but they are applicable to some special tasks only.

Recently, researchers have found that classifier ensemble approaches trained on different data subsets or feature subset are capable of generating more discrimination power in both modeling and classification. The most popular one, AdaBoost, aims at building a strong classifier by sequentially training and combining a group of weak classifiers such that the ensemble classifier has an improved correct classification rate. Recently, GMM and HMM have been successfully in-

tegrated with boosting learning framework to form a strong learning approach [8].

3.3. The Proposed Multi-boosted HMMs Approach

By integrating the superiority of segmental scheme and boosting learning ability, the whole utterance can be verified via multi-boosted HMMs jointly, which therefore generates more discrimination power than single HMM or boosted HMMs performed on the whole utterance. Nevertheless, simply utilization of the whole feature vectors may lead to the curse of dimensionality. We therefore adopt the RSM to circumvent this problem and utilize the DSS to form a train data set, which can handle the small sample size problem. Given a set of positive examples $A = \{x_1^a, x_2^a, \cdots, x_{N_a}^a\}$ of the target speaker and a set of negative examples $B = \{x_1^b, x_2^b, \cdots, x_{N_b}^b\}$ of imposters. From A and B, we form a novel 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 pairs of examples that are from A and B, respectively. As the imposters may have many different categories, it is very difficult to utilize one single model to represent all the imposters. Hence, we prefer not to train the imposter models.

Let λ be an HMM trained from A, the NLL of $x_i^a \in A$ conditioned on λ should be larger than the NLL of $x \in B$ conditioned on λ. We learn a similarity score $h(x_i^a, x, \lambda)$:

$$h(x_i^a, x, \lambda) = |NLL(x_i^a, \lambda) - NLL(x, \lambda)|. \quad (4)$$

By setting an appropriate threshold τ, the similarity between the testing example x and data set A is computed as:

$$\hat{h}_{\min} = \min_{x_i^a \in A} h(x_i^a, x, \lambda), \quad (5)$$

where x belongs to the target speaker if $\hat{h}_{\min} \leq \tau$, and imposter otherwise, *i.e.,* we compare the test example with all the positive samples and take the highest score (*i.e.,* minimum value) to make the decision. As shown in Algorithm 1, the password utterance which belongs to the target speaker or not is determined via all the subunit verification results.

3.4. Feature Extraction

In general, the combination of contour-based features and area-based features will deliver a better performance in lip motion analysis [2]. Hence, we first compute nine geometric shape parameters, *i.e.,* maximum horizontal distance, seven vertical distances [2] and mouth area $\{L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8, \mathcal{A}_c\}$ to model the contour-based features, denoted as F_{cf}. The geometric shape parameters are normalized with respect to the corresponding values of the first lip frame. Next, the located ROIs of lip images are convert to gray level case and normalized to have a similar distribution characteristic. Subsequently, mean subtraction is performed to remove the basis effect across each utterance. Then, the principal components of top N_{pca} numbers are chosen as PCA features F_{pca}, while the first M 2D-DCT coefficients along the Zig-zag

Algorithm 1: Multi-boosted HMMs for SPV.

Input:

Lip motion sequences of the training data set, D.

The password number: p, RSM percentage: $P_{rsm}\%$.

Preprocessing:

Frame feature extraction, lip motion segmentation.

Multi-boosted HMMs: ($m = 1, \cdots, p$)

Form a new training set via DSS [4] from subunit data: $D_m^T = \{X_1^T, \cdots, X_{N_a}^T\}$, $D_m^I = \{X_1^I, \cdots, X_{N_b}^I\}$.

Initialize weights $w_{i,j}^T = \frac{2}{N_a(N_a-1)}$, $1 \leq i \leq j \leq N_a$; $w_{i,j}^I = \frac{1}{N_a N_b}$, $1 \leq i \leq N_a, 1 \leq j \leq N_b$. $r = 0, \varepsilon^0 = 0$;

while $r \leq R$ and $\varepsilon^r < 0.5$ **do**

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

  RSM sampling D_m^T of $P_{rsm}\%$, build a HMM $\lambda_m^r(T)$.

  Call WeakLearner with respect to Eq. (4).

  Train a threshold τ_m, minimizing error:
  $\varepsilon^r = \sum_{i,j} w_{i,j}^T e_{r,i,j}^T + \sum_{i,j} w_{i,j}^I e_{r,i,j}^I$, where $e_{r,i,j}^T = 1$ if $h_m^r(X_i^T, X_j^T, \lambda_m^r(T)) \geq \tau_m$, $e_{r,i,j}^I = 1$ if $h_m^r(X_i^T, X_j^I, \lambda_m^r(T)) < \tau_m$, and 0 otherwise.

  Set $\alpha_m^r = \frac{1}{2}\log[(1-\varepsilon^r)/\varepsilon^r]$.

  Update the weights to be:
  $w_{r+1,i,j}^T = w_{r,i,j}^T \cdot \exp(2\alpha_m^r e_{r,i,j}^T)$,
  $w_{r+1,i,j}^I = w_{r,i,j}^I \cdot \exp(2\alpha_m^r e_{r,i,j}^I)$.

  $r = r + 1$.

end while

Obtain similarity score between X_p^T and X_q via Eq. (5):
$\hat{h}_m(X_p^T, X_q) = \sum_{w=1}^r \alpha_m^w h_m^w(X_p^T, X_q, \lambda_m^w(T))$.

Output: $\hat{h}_{\min}^m = \min_{X_i^T \in D_m^T} \hat{h}_m(X_i^T, X_q)$, $m = 1, \cdots, p$

Scan order are selected as the 2D-DCT features F_{dct}. Finally, we obtain the joint features, i.e. $\{F_{cf}, F_{pca}, F_{dct}\}$.

4. EXPERIMENTAL RESULTS

A database consisting of 46 speakers (28 males, 18 females) repeating the fixed digit password 3175 for 20 times ($\mathcal{D}_p$) and randomly uttering another 10 different four-digit password ($\mathcal{D}_r$) is established. All the password phrases are uttered with same speaking pace during approximate 4-second recording in 30 fps. The located ROIs of the lip images are of size 112×76. A left to right HMM with six hidden states incorporating two continuous density Gaussian mixtures output is employed. The biased Baum-Welch estimation [8] was utilized for HMM parameters learning. Equal error rate (EER) [2] was adopted as the evaluation metric.

The database $\mathcal{D}_p$ is divided into two disjoint data sets with the equal size, i.e. $\mathcal{D}_{p_1}$ for training and $\mathcal{D}_{p_2}$ for testing. The experiments are conducted with two cases: (1) speaker-dependent case, and (2) speaker-independent case. In the former, the utterances differ from the registered one (i.e. 3175)

Table 1. The verification results of different approaches.

Feature set	Equal Error Rate [EER %] (The operating point where the FAR equals to FRR)						
	GMM	HMM	Segmental+GMM	Segmental+HMM	boosted GMM	boosted HMM	Multi-boosted HMM
Speaker-dependent	12.17	12.39	11.73	11.3	13.91	11.08	3.91
Speaker-independent	16.88	15.74	13.78	10.15	10.58	11.16	4.06

are considered as the imposters. The subunit imposters are generated via leave-one-out scheme [2], where each segmental unit that does not belong to the subunit of the fixed order of the password is selected as imposter. We randomly selected one segmental unit of each digit "0-9" from $\mathcal{D}_{p_1}$ and $\mathcal{D}_r$ as the subunit imposter data. In the latter case, as the password utterances differing from the registered one and uttered by different speakers can be easily distinguished using the existing methods, we therefore focus on the scenario provided that an imposter knows the password in advance, i.e. each imposter utters the same password. Subsequently, by specifying a speaker as the target one, the other speakers becomes the imposters. We randomly selected two examples of each speaker excluding the target speaker from subset $\mathcal{D}_{p_1}$ to form the imposter training data. The DSS [4] was employed to form the training samples in pairs.

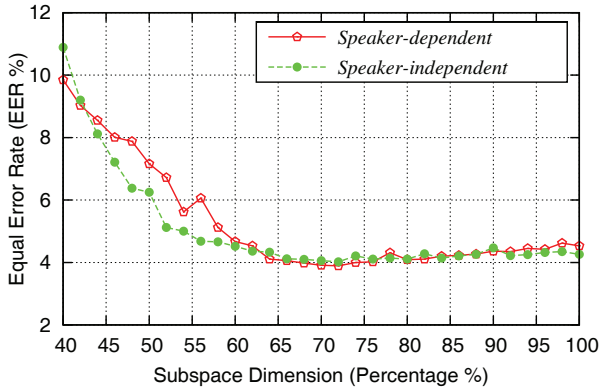

Fig. 2. SPV performance via different subspace dimensions.

Table 1 shows the comparative results between the proposed approach and the existing GMM and HMM methods. It can be seen that, with the segmental modeling or boosted learning, the HMM outperforms the GMM. Further, the proposed approach has made the significant performance improvement in comparison with the existing counterparts. In addition, the values of EER performed via various subspace dimensions are shown in Figure 2. It can be observed that the feature sets with subspace dimension of 65-75% have the lowest EER values.

5. CONCLUSION

In this paper, we have proposed an effective lip motion segmentation method and addressed a multi-boosted HMMs approach incorporating the RSM and DSS to realize the lip password based speaker verification. The proposed approach is capable of detecting an imposter even knowing the password, as well as the target speaker with the wrong password. Experiments have shown the promising result.

References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, 2007.

[2] 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, 2006.

[3] 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, 2005.

[4] X.G. Wang, C. Zhang, and Z.Y. Zhang, "Boosted multi-task learning for face verification with applications to web image and video search," in *Proc. IEEE CVPR*, 2009, pp. 142–149.

[5] X. Liu and Y.M. Cheung, "A robust lip tracking algorithm using localized color active contours and deformable models," in *Proc. IEEE ICASSP*, 2011, pp. 1197–1200.

[6] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 988–992, 1996.

[7] 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 ICASSP*, 2007, vol. 4, pp. IV313–316.

[8] 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, 2004.