IDIAP RESEARCH REPORT

User Customized Password HMM Based Speaker Verification

Mohamed Faouzi BenZeghiba [a]
Hervé Bourlard [a,b]

IDIAP–RR 02-35

October 30, 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny
[b] Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

IDIAP Research Report 02-35

User Customized Password HMM Based Speaker Verification

Mohamed Faouzi BenZeghiba Hervé Bourlard

October 30, 2002

Abstract. A speaker verification system which allows users to chose their own password is presented. The system has no *a priori* knowledge of passwords. A hybrid HMM/ANN system is used to infer the phonetic transcription of the password. The emission probabilities are then modeled by a multi-Gaussians HMM model. Evaluation experiments, conducted on PolyVar database, showed results comparable with a system where the correct phonetic transcription of the password is known *a priori*.

1 Introduction

In most text-dependent speaker verification systems, the password is constrained to be within a small vocabulary. So, the system has some *a priori* knowledge (e.g., the phonetic transcription of the word) about the password of the speaker. However, in speaker verification based on user customized password (SV-UCP), users can choose their own password from an unconstrained vocabulary. This enables better user friendliness and increased security. Nevertheless, the SV-UCP raises some issues in both enrollment and test phases:

1. We have to automatically find (infer) the topology (in terms of sub-word models like phonemes) of the password model from a few repetitions of the user password. The inferred model should be representative of the lexical content of the password. This step requires a good speaker independent speech recognition system.

2. We have to quickly adapt the parameters of the inferred model towards the characteristics of the speaker using only a small amount of the adaptation data.

3. We have to determine the *a priori* threshold (speaker dependent or speaker independent) for the decision to accept or reject a speaker.

4. We have to find an appropriate world model for score normalization.

The approach presented here exploits some of the advantages of the hybrid HMM/ANN systems [1] where an Artificial Neural Network (ANN) is used to estimate Hidden Markov Model (HMM) emission posterior probabilities (or scaled likelihoods). In this framework, HMM/ANN systems usually yield very good phonetic recognition rates, and are also well suited in estimating a confidence measure [2, 3], which makes them particularly amenable to perform HMM inference from acoustic data. The emission probabilities of the inferred HMM model are then modeled in terms of speaker adapted multi-Gaussians HMMs models. Some related works can be found in [4, 5].

The rest of the paper is organized as follows: Section 2 briefly introduces the similarity measure that we have used in experiments, while Section 3 describes the evaluation databases. Section 4 presents a detailed description of the method and Section 5 describes the experiments conducted and provides an analysis of the results obtained.

2 SV-UCP Decision Rules

In SV-UCP, we are interested in estimating the joint posterior probability $P(M_k, S_k|X)$ representing the probability that the correct speaker S_k has pronounced the correct password M_k given the observed acoustic vector sequence X. During verification, this probability is compared to the joint posterior probability that any other speaker (impostor) may have pronounced the correct password $P(M_k, \overline{S_k}|X)$ and to the joint posterior probability that any speaker (impostor or client) may have pronounced any other password (text) $P(\overline{M_k}, S|X)$. Hence, the decision rules can be formulated as follows:

$$S = S_k \quad \text{if} \quad P(M_k, S_k|X) \geq P(M_k, \overline{S_k}|X) \tag{1}$$

$$\text{and} \quad P(M_k, S_k|X) \geq P(\overline{M_k}, S|X) \tag{2}$$

Using Bayes rule, and assuming that the simultaneous probability of any speaker and any word is equal for all combinations of speakers and words, decision rules (1) and (2) can be rewritten as follows:

$$\frac{P(X|M_k, S_k)}{P(X|M_k, \overline{S}_k)} \geq \left[\frac{P(M_k, \overline{S}_k)}{P(M_k, S_k)} = \Delta_1\right] \tag{3}$$

$$\frac{P(X|M_k,S_k)}{P(X|\overline{M}_k,S)} \geq \left[\frac{P(\overline{M}_k,S)}{P(M_k,S_k)} = \Delta_2\right] \tag{4}$$

where Δ_1 and Δ_2 are the decision thresholds.

To estimate $P(X|\overline{M}_k,S)$ we have used a Gaussian Mixture Model (GMM). So $P(X|\overline{M}_k,S) \approx P(X|S)$.

Taking the logarithm of (3) and (4), and normalizing each probability by the number of frames in the test utterance, after having removed the silence frames, yields to the following *log-likelihood ratios:*

$$\left[LLR1 = \frac{1}{N_1}\log P(X|M_k,S_k) - \frac{1}{N_2}\log P(X|M_k,\bar{S}_k)\right] \geq \omega_1 \tag{5}$$

$$\left[LLR2 = \frac{1}{N_1}\log P(X|M_k,S_k) - \frac{1}{N_3}\log P(X|S)\right] \geq \omega_2 \tag{6}$$

The first decision rule (5) is used to verify if the speaker **who** pronounced the correct password is the true claimant or not. The verification is mainly based on the characteristics of the speaker, and hence, it represents the speaker verification part. The second decision rule (6) is used to verify if **what** is pronounced (text) is the correct password of the claimed identity or not. This is the utterance verification part, except that we use the speaker-dependent HMM password model instead of speaker-independent HMM model.

Usually, the decision to accept or reject a speaker is taken in two steps. First, we perform an utterance verification step, and if the score exceeds the threshold ω_2, then we perform a speaker verification step. So, the speaker is accepted if the two scores exceed their respective thresholds simultaneously. It has been found [6] that the combination of these two scores can significantly improve the performance of the system. In this paper, among different combination techniques, we have used a simple weighted combination technique. The final decision to accept a speaker can be defined as follows:

$$\alpha\ LLR1 + (1-\alpha)\ LLR2 \geq \delta \tag{7}$$

where $0 \leq \alpha \leq 1$. The values of the parameter α and the threshold δ are optimized using a development set.

3 Database, Protocol and Acoustic features

Two databases were used in this work. The **Swiss French Polyphone** database [7], was used to train different speaker-independent speech recognizers. The speaker verification experiments were conducted using the **PolyVar** database [7]. This database comprises telephone recordings from 143 speakers, each speaker recording between 1 and 229 sessions. Each session consists of one repetition of the same set of 17 words common for all speakers. This set of 17 words was divided into three subsets *data1*[1] and *data2*[2] with 7 words in each subset and *data3* with the remaining 3 words. A set of 38 speakers (24 males and 14 females) who have more than 26 sessions were selected. This set was also divided into two subsets *speakers1* and *speakers2* of 19 speakers. For each speaker in *speakers1* and *speakers2* and each word in *data1* and *data2*, the first 5 utterances of the same word are used as training data, and between 18 and 22 utterances were used as client accesses with the correct password. Each speaker in *speakers1* (respectively in *speakers2*) is considered as an impostor for each speaker in *speakers1* (respectively in *speakers2*). Two accesses with the correct password from each impostor (in *speakers1* and *speakers2*) and 3 accesses with a wrong password[3] from each speaker (client

[1] data1 = { exposition, message, mode d'emploi, musée, précédent, quitter, suivant}
[2] data2 = { annulation, casino, cinéma, concert, galerie du manoir, gainadda, louis moret }
[3] chosen from *data3*

and impostor in *speaker1*) are used as impostor accesses. Later we will refer to (*speakers1*, *data1*), (*speakers1*, *data2*), (*speakers2*, *data2*), (*speakers2*, *data1*) as *dev1*, *dev2*, *eva1* and *eva2* respectively. Unlike *dev1* and *dev2*, there is no impostor accesses in *eva1* and *eva2* with a wrong password; all the test accesses are done with the correct passwords.

For acoustic parameters, two kinds of parameters were used: 12 RASTA-PLP coefficients with their first temporal derivatives as well as the first and second derivative of the log energy were calculated every 10 ms over 30 ms window, resulting in 26 coefficients. These coefficients, which are more suitable for speech recognition, were used to train a speaker-independent multi-layer perceptron (SI-MLP) which is used to infer the password of the user. In order to keep the characteristic of the user, MFCCs were used for speaker adaptation. 12 MFCCs with energy complemented by their first derivatives were calculated every 10 ms over 30 ms window, resulting in 26 coefficients.

4 The approach

The approach presented here is based on hidden Markov models (HMM). We started from:

1. A well trained Speaker-Independent Multi-Layer Perceptron (referred to as SI-MLP) with parameters θ. This MLP was trained on Swiss French Polyphone database with RASTA-PLP features. The SI-MLP had 234 input units with 9 consecutive 26 dimensional acoustic vectors, 600 hidden units and 36 outputs, each output associated with a specific phone. This SI-MLP achieved 68% as a phonetic recognition rate. To this MLP, we associated an ergodic HMM model M with 36 states. Each state belongs to a specific output of the SI-MLP.

2. A HMM speaker-independent speech recognition (which we will refer to as λ) with 36 context-independent phone models. The phone models consisted of 3 states left-to-right HMM with 3 mixtures/state. This HMM model was used as a prior distribution for MAP (maximum *a posteriori*) adaptation [8] of a new client and to build the "world model" for score normalization. This HMM model was trained on Swiss French Polyphone database with MFCC.

3. A GMM with parameters Λ was modeled by 150 (diagonal covariance) Gaussians and trained on Swiss French Polyphone database with MFCC. This GMM was used for utterance score normalization.

4.1 HMM topology inference

For each new customer, we match (using Viterbi alignment) each of the utterances (5 repetitions) in the enrollment data with the ergodic HMM model M using local posterior probability $p(q_\ell|x_n,\theta)$ estimated by the SI-MLP (θ). The result is 5 phonetic transcriptions from which we select the best one to build up the user HMM password model. Two criteria were used to determine the best phonetic transcription:

1. The highest, time normalized accumulated log posterior probability (HNPP) defined as:

$$\widehat{M} = \arg\max_{1\leq i\leq 5}\left[\frac{1}{N_i}\sum_{n=1}^{N_i}\log P(q_\ell^{n,i}|x_{n,i},\theta)\right] \tag{8}$$

   where $q_\ell^{n,i}$ represents the phonetic symbol associated with the acoustic frame $x_{n,i}$ of the i^{th} repetition. N_i is the length of the utterance i.

2. The highest, global average time normalized accumulated log posterior probability (HANPP): we force align all the enrollment utterances on each phonetic transcription using local posterior probability estimated by the SI-MLP. The best phonetic transcription is the one which best match all the enrollment utterances.

The topology of the user customized HMM model M_k is then simply built-up by strictly concatenating left-to-right (with only loops and skips to the next state) HMM states corresponding to each of the phones in the above "optimal" phonetic sequence $\widehat{M}$.

4.2 HMM parameter adaptation

Once the user HMM model M_k is inferred, a MAP adaptation procedure using all the enrollment data is performed. This procedure consisted of adapting the mean of the Gaussians of the phone models of the speaker independent speech recognizer (λ) which constitute the inferred model M_k as well as the transition probabilities between these phone models. The result is a speaker-dependent HMM model (λ_k).

4.3 Score computation

Having created all models, the two *log-likelihood ratios LLR1* and *LLR2* in (5) and (6) are estimated as follows:

$$LLR1 = \frac{1}{N_1} \log P(X|M_k, \lambda_k) - \frac{1}{N_2} \log P(X|M_k, \lambda) \tag{9}$$

$$LLR2 = \frac{1}{N_1} \log P(X|M_k, \lambda_k) - \frac{1}{N_3} \log P(X|\Lambda) \tag{10}$$

During verification, a simple silence detector based on an unsupervised bi-Gaussian model [9] was used to remove the silence frames before GMM score computation, while in the HMM a silence model is applied in the beginning and the end of the password M_k.

4.4 Threshold determination

The performance of the method was optimized using a development set to minimize the equal error rate (EER) using a speaker-independent threshold. In our experiments, to assure that speakers in development (*dev1* and *dev2*) and evaluation (*eva1* and *eva2*) sets have different passwords, we have estimated the thresholds δ_1 and δ_2 using *dev1* and *dev2* respectively, and we have used them as a *priori* thresholds on *eva1* and *eva2* respectively.

4.5 Decision

To take the decision to accept or reject a speaker, first we normalize the score (7) (to get a common prior threshold for all speakers) of a test access from a speaker in *eva1* (respectively in *eva2*), by subtracting the threshold δ_1 (respectively δ_2) from the score of the test access. If the normalized score is positive, then the speaker is accepted, otherwise, the speaker is rejected.

5 Experiments and Results

All experiments reported here were conducted using the Torch library[4]. For comparison purposes, results with the correct phonetic transcription of the password are also given. This is the reference system.

5.1 Results

Figure 1 shows the performance of the reference (COR) and the SV-UCP systems with the two criteria used for HMM inference. The circles correspond to the performance of the systems with *a priori* threshold. Table 1 gives more details.

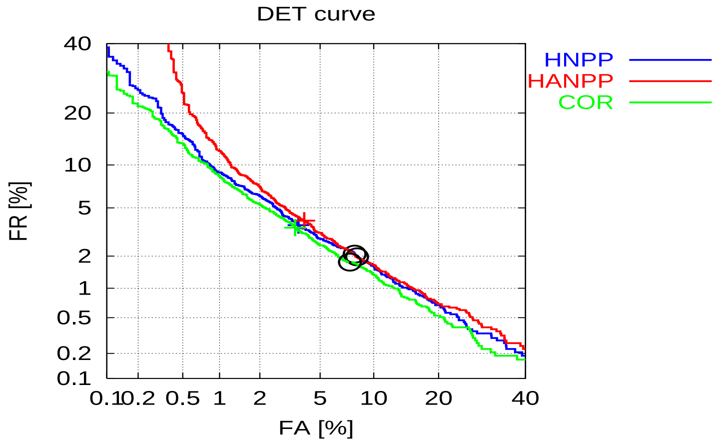

Figure 1: *DET curve comparing the performance of the reference system and the UCP systems on the evaluation set. FA [%] is the false acceptance rate and FR [%] is the false rejection rate.*

	Threshold	FA [%]	FR [%]	HTER[%]
HANPP	POSTERIOR	3.984	3.979	3.981
	PRIOR	**8.156**	**1.971**	**5.064**
HNPP	POSTERIOR	3.660	3.660	3.660
	PRIOR	**7.915**	**2.083**	**4.999**
COR	POSTERIOR	3.482	3.491	3.486
	PRIOR	**7.455**	**1.764**	**4.609**

Table 1: *The peformance of each system on the evaluation set: The first line is the EER estimated a postoriori and the second line is the HTER estimated using a priori threshold.*

From Figure 1, we can see that the circles are not close to the EER region (crosses). Indicating that both development and evaluation sets have two very distinct thresholds. This is probably due to the fact that both development and evaluation sets have different passwords and we do not have enough words in the development set to reliably estimate a speaker independent threshold. Figure 1 and Table 1 show that the two SV-UCP systems have comparable performance, but they perform slightly worse than the reference system.

5.2 Analysis

To determine the reasons for these results, and to evaluate how good the inferred HMM model is, we plot the variations of the EER on the development set and the associated HTER (Half Time Error Rate) estimated with *a priori* threshold on the evaluation set as a function of the combined parameter α. The results are shown in Figure 2.

There are two informative values of α which can help us to analyze the results obtained above. These values correspond to the performance of the system where $\alpha = 0$ and $\alpha = 1$.

- For $\alpha = 0$, the performance of the combined system becomes equal to the performance of the utterance verification part (*LLR2*). In this case, the reference and the SV-UCP systems will have the same normalization model (GMM) to estimate *LLR2*. So, if one of these systems performs better than the others, this should be attributed to the user HMM password model. Or

[4] http://www.Torch.ch

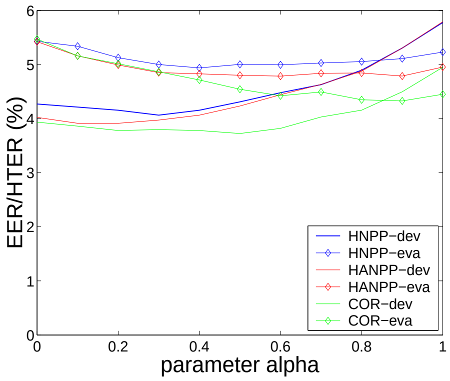

Figure 2: *EER variations on the development set and HTER variations on the evaluation set as a function of the combined parameter α.*

the HTERs on the evaluation set for the reference and the SV-UCP systems show no difference between these three systems. This indicates that the small improvement of the reference system can not be attributed to the fact that in this system we use the correct phonetic transcription of the password while for the other systems, we infer the phonetic transcription.

- For $\alpha = 1$, the performance of the combined system becomes equal to the performance of the speaker verification part (*LLR1*). To estimate *LLR1*, the client model (M_k,λ_k) and the normalization model (M_k,λ) have the same topology[5] within the same system, but it is different from one system to another. In this case, if one system performs better than the other, this performance can be attributed to either the user model (inferred HMM password) or the normalization (world) model. As we have just seen in the case for $\alpha = 0$, the client model performs comparably for all the systems. So, the improvement in the reference system is mainly caused by the normalization (world) model which is better than the one used in the SV-UCP.

One possible explanation is that the world model should reflect how speakers pronounce the password and not how the client pronounces it as used in this experiment for the SV-UCP. The simplest way to model this inter-variability in pronunciation is to choose the correct phonetic transcription of the password as a world model. This is another reason why we need good phonetic speaker independent speech recognition.

6 Conclusion

In this paper, user customized password speaker verification based on hidden Markov models is presented. A hybrid HMM/ANN was used to find phonetic transcriptions of the enrollment utterances, from which we built up the user HMM password model. The system gave comparable results compared to the reference system, where the correct phonetic transcription of the password is known *a priori*. The main conclusions are:

1. As we can expect from the results, randomly choosing one of the phonetic transcriptions to build the HMM model does not significantly affect the performance of the system. However, we have to make sure that the speaker effectively pronounces the same password.

2. In SV-UCP within the HMM framework, not only how to infer the HMM password is important, but which model we use as a "world" model for score normalization is also important.

[5] By the same topology, we mean the same states and the same connections between states

3. The technique used here for HMM inference is based on single pronunciation modeling. As a speaker can not pronounce exactly the same word in the same manner from one trial to another, choosing one phonetic transcription to model the password is not a good choice. It is clear that the word model allowing more than one pronunciation should perform better than word model allowing just one pronunciation. One solution is to use all the inferred phonetic transcriptions to model the HMM password by keeping them separately, or we can merge them to build one HMM model [10].

7 Acknowledgments

Mohamed BenZeghiba is supported by the Swiss National Science Foundation through the project "SV-UCP: Speaker Verification based on User-Customized Password" (2000-063721.00-1). The main author would like to thank Johnny Mariéthoz for helpful discussions, Conrad Sanderson and Andrzej Drygajlo for helpful comments on the paper.

References

[1] S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, "Connectionist Probability Estimators in HMM Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, Part II, 1994.

[2] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN Speech Recognition," *Proceedings of EUROSPEECH'97*, Rhodes, Greece, pp. 1955-1958, 1997.

[3] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems", *Proc. of Intl. Conf. on Spoken Language Processing* (Sydney), pp. 775-779, 1998.

[4] B. Jacob, J. Mariethoz, G. Gravier, F. Bimbot, "Robustesse de la Verification du Locuteur par Mot de Passe Paersonnalisé" in *XXIIIèmes Journées d'Etude sur la parole*, Aussois, pp. 357-360, 2000.

[5] J. Kharroubi and Gérard Chollet, "Utilisation de Mots de Passe Personnalisés pour la Vérification de Locuteur", in *XXIIIèmes Journées d'Etude sur la parole*, Aussois, pp. 361-364, 2000.

[6] L. Rodriguez-Linãres and C. Garcia-Mateo and J. L. Alba-Castro, "On the Use of Neural Networks to Combine Utterance and Speaker Verification Systems in a Text-Dependent Speaker Verification Task," *Proceedings of EUROSPEECH'99*, pp. 1003-1006, 1999, Budapest, Hungary.

[7] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais, "Swiss French Poly-Phone and PolyVar: telephone speech databases to model inter- and intra-speaker variability", *IDIAP Research Report*, IDIAP-RR-96-01, 1996.

[8] J. L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains", in *IEEE Transaction on Speech Audion Processing,* April 1994, Vol 2, pp. 291-298.

[9] S. Bengio, J. Mariéthoz and S. Marcel, "Evaluation of Biometric Technology On XM2VTS", *IDIAP Researcg Report*, IDIAP-RR-01-21, 2001, *Published in European Project BANCA Deliverable D71.*

[10] M. G. Thomason and E. Granum, "Dynamic Programing Inference of Markov Networks from Finite Sets of Sample Strings" in *IEEE Transaction on Pattern Analysis and Machine Intelligence,* Vol 8, No 4, pp. 491-501, 1986.