IDIAP RESEARCH REPORT

User-Customized Password Speaker Verification based on HMM/ANN and GMM Models

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

IDIAP–RR 02-10

April 11, 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 02-10

User-Customized Password Speaker Verification based on HMM/ANN and GMM Models

Mohamed Faouzi BenZeghiba   Hervé Bourlard

April 11, 2002

Abstract. In this paper, we present a new approach towards user-customized password speaker verification combining the advantages of hybrid HMM/ANN systems, using Artificial Neural Networks (ANN) to estimate emission probabilities of Hidden Markov Models , and Gaussian Mixture Models. In the approach presented here, we indeed exploit the properties of hybrid HMM/ANN systems, usually resulting in high phonetic recognition rates, to automatically infer the baseline phonetic transcription (HMM topology) associated with the user customized password from a few enrollment utterances and using a large, speaker independent, ANN. The emission probabilities of the resulting HMMs are then modeled in terms of speaker specific/adapted multi-Gaussian HMMs or speaker specific/adapted ANN. In the proposed approach, the hybrid HMM/ANN system is used as a model for utterance (password) verification, while still using a speaker independent GMM for speaker verification. Results (EER) are compared to a state-of-the-art text-dependent approach, using multi-Gaussian HMMs only.

Contents

1 Introduction

In most text-dependent speaker verification systems, the password is constrained to be within a small vocabulary. However, in speaker verification based on user-customized password (SV-UCP), users can choose their own password without any constraint on the vocabulary size. This characteristic raises new issues. First, we have to find (infer) the topology of the password model which represent best the lexical content of the password. Second, the parameters of the inferred model have to be quickly adapted towards the acoustic characteristics of the user. An additional problem related to SV-UCP is the dilemma/compromise between a good utterance verification system and a good speaker verification system, which will also be addressed here in a same framework, using utterance posterior probabilities and speaker likelihoods.

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 are usually yielding very good phonetic recognition rates, and are also well suited to estimate confidence measure [2], which makes them particularly amenable to perform HMM inference from acoustic data. In some previous work [4], a preliminary HMM/ANN speaker verification approach was proposed, and was shown yielding good SV-UCP performance. The main idea of this approach was to use (normalized) posterior-based confidence measures towards SV-UCP. User-customized password HMM was inferred by using a large, speaker-independent, ANN. A smaller speaker independent ANN (to limit the number of parameters) was then adapted to the speaker specific enrollment utterances to model the emission probabilities of the inferred HMM. In this work, we were thus using exclusively ANN posterior probabilities, which resulted in some SV weaknesses. Indeed, confirming the conclusions reported in [5], it was observed that the hybrid HMM/ANN system, and its associated maximum a posteriori (MAP) training, was mainly modeling the lexical content of the password and was not properly capturing the speaker characteristics.

In this paper, we thus present a new approach alleviating this problem, combining the use of hybrid HMM/ANN models (mainly performing utterance verification) and Gaussian mixture models (mainly performing speaker verification) in a same framework. In this method, the hybrid HMM/ANN model was adapted to learn the lexical content of the user password, while a text-independent state-of-art GMM model was used to capture the characteristics of the user. Results on large evaluation databases in different experimental conditions (including the pessimistic case were all costumers are using the same password) show the effectiveness of the proposed method compared to previous SV-UCP approaches or state-of-the-art text independent

In the following, Section 2 briefly introduces the similarity measures that will be used here, while Section 3 describes the evaluation databases. Section 4 presents a detailed description of the two methods, and Section 5 describes the experiments conducted and the results obtained. A discussion is given in Section 6.

2 SV-UCP Decision Rules

In SV-UCP, we are interested in estimating the join 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, we thus want to compare this join posterior probability to the probability that any other speaker (impostor) may have pronounced the correct password or anything else. Based on this, we can define different SV-UCP decision rules, depending on whether we compare this join posterior probability to:

1. The posterior probability of any speaker S pronouncing any word M:

$$P(M_k, S_k|X) \geq P(M, S|X) \quad (1)$$

where M is represented by an ergodic HMM. Similarly, we could also use the following criterion:

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

2. The posterior probability of any speaker S pronouncing the correct password M_k:

$$P(M_k, S_k|X) \geq P(M_k, S|X) \tag{3}$$

Using Bayes rule, decision rule (1) can be rewritten as:

$$\frac{P(X|M_k, S_k)}{P(X|M, S)} \geq \frac{P(M, S)}{P(M_k, S_k)} = \Delta_1 \tag{4}$$

and decision rule (3) as:

$$\frac{P(X|M_k, S_k)}{P(X|M_k, S)} \geq \frac{P(M_k, S)}{P(M_k, S_k)} = \Delta_2 \tag{5}$$

where Δ_1 and Δ_2 are the decision thresholds. Threshold of decision rule (5) is based on the assumption that the impostor could pronounce the right password, and is thus more competitive than (4).

To estimate the denominator probability $P(X|M, S)$ in (4), an ergodic HMM model was used as the "world" model. This method will be referred to as *unconstrained HMM*. The denominator probability $P(X|M_k, S)$ in (5) is estimated by a forced Viterbi alignment performed on the password HMM model M_k. This method will be referred to *constrained HMM*. Both approaches will be further described in Section 4.3.

As an alternative, we can also start from inequality (2) and develop the first term, yield the following decision rule:

$$P(M_k|S_k, X)P(S_k|X) \geq P(S|X) \tag{6}$$

which, using Bayes rule, can be rewritten as:

$$P(M_k|S_k, X)\frac{P(X|S_k)}{P(X|S)} \geq \frac{P(S)}{P(S_k)} = \Delta_3 \tag{7}$$

The **first term** of the left hand side $P(M_k|S_k, X)$ is text-dependent and corresponds to the posterior probabilities estimated through an ANN, as used in [4]. However, as already mentioned above, this probability was found to be mainly modeling the lexical characteristics and not necessarily the speaker characteristics. The **second term** $\frac{P(X|S_k)}{P(X|S)}$ is the **likelihood ratio**, usually used for text-independent SV and actually represents the contribution of the speaker characteristics. This contribution will be estimated through usual GMMs, the speaker independent GMM (world model, estimating the denominator) and the speaker adapted GMM (estimating the numerator). Together, these two probabilities give us the information on which we can take the decision to accept or reject a speaker pronouncing a specific password. In the following, this method will be referred to as combined HMM/ANN-GMM, and will be described in Section 4.2.

3 Database and Acoustic features

Two databases were used in this work. The **Swiss French Polyphone** database [6], a large telephone database containing prompted and natural sentences pronounced by a large number of speakers, was used to train[1] different HMM and HMM/ANN (thus including a large, speaker independent, ANN) speaker-independent speech recognizers.

[1] Using the 10 phonetically rich sentences read by 400 speakers.

Our speaker verification experiments were conducted using the **PolyVar** database [6], which is designed to address inter-speaker variability issues. 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 the speakers, which makes this database particularly well suited to test SV-UCP. A client subset of 19 speakers (12 males and 7 females) who have more than 26 sessions were selected. For each of these speakers, the first 5 utterances (corresponding to the first 5 sessions) of the same word are used as training data, and around 22 utterances were used as true accesses. Another subset of 19 speakers different from the client subset were used as impostors.

For acoustic parameters, two kinds of parameters were used: RASTA-PLP coefficients (more speaker independent) for HMM inference, while mel-frequency cepstral coefficients (MFCC) were used for speaker adaptation and verification. 12 RASTA-PLP coefficients with their first temporal derivatives as well as the first and second derivative of the log energy were thus calculated every 10 ms over 30 ms windows, 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 coefficient with energy complemented by their first derivative were calculated every 10 ms over 30 ms windows, resulting in 26 coefficients.

4 Approaches

As hybrid HMM/ANN systems were known to yield very good phonetic recognition rates, only hybrid HMM/ANN system were used here to perform HMM inference. Results using HMM inference from standard ergodic multi-Gaussian HMMs are not reported here since they were consistently yielding much poorer performance.

For all methods described below, the same HMM inference procedure (using HMM/ANN) was used, as described below.

4.1 HMM Inference

We start from a well trained Speaker-Independent Multi-Layer Perceptron (referred to as SI-MLP) with parameters (θ), this ANN was trained on Polyphone with RASTA-PLP features. This SI-MLP has 234 input units with 9 consecutive 26 dimensional acoustic frames, 600 hidden units and 36 outputs, each output associated with a specific phone. This SI-MLP achieved 68% as a phonetic recognition rate.

Each new customer pronounces 5 times his/her password. The first three utterances constitute the inference data while the last two were used as cross-validation set for speaker adaptation process (Section 4.2). We match each of the utterances in the inference data with the ergodic HMM model M using local posterior probability $p(q_\ell|x_n,\theta)$ estimated by the SI-MLP, resulting in 3 phonetic transcriptions from which we selected the one yielding the highest time normalized accumulated log posterior probability (along the line of confidence measure used in [7]), defined as (for an utterance $X = \{x_1, \dots, x_n, \dots, x_N\}$ and a phonetic transcription Q):

$$\frac{1}{N}\sum_{n=1}^{N} \log P(q_\ell^n|x_n,\Theta) \tag{8}$$

where q_ℓ^n represents the phonetic symbol associated with x_n.

The topology of the resulting user-customized HMM model (M_k) is then simply built-up by concatenating strictly 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.

4.2 Combined HMM/ANN-GMM approach

As given in (7), hybrid HMM/ANN-GMM user-customized speaker verification can be expressed in terms of utterance verification and speaker verification parts.

Utterance verification, represented by $P(M_k|S_k, X)$ in (7), consists in adapting for each new client the SI-MLP of parameters θ. As the amount of adaptation data is very limited, different MLP adaptation schemes were tested in [3]. Finally, it was found that the best solution consists in adapting a smaller, speaker independent, single-layer perceptron, referred to as SLP [4], of parameters (θ^*), to yield the speaker specific parameters θ_k^* that will be used to estimate $P(M_k|\theta_k^*, X)$. The speaker independent SLP used as the initial network for speaker adaptation was also trained using the Polyphone database, using MFCCs coefficients (more robust to speaker characteristics). The SLP adaptation is then performed by matching each enrollment utterance on the inferred speaker-specific word model M_k, thus providing targets for the SLP training. To avoid over-training, cross-validation was used to stop this adaptation process. In our case, the enrollment data (5 utterances) was thus divided into two parts. The first three utterances which were used as HMM inference data were also used as adaptation data, while the last two utterances were used for cross-validation.

Speaker characteristics are captured by the second term $\frac{P(X|S_k)}{P(X|S)}$ in (7), representing the **likelihood ratio** classically used for text-independent SV. This contribution is estimated through usual GMMs, the speaker independent GMM (world model, estimating the denominator) and the speaker adapted GMM (estimating the numerator). The world model GMM (of parameters Λ) was modeled by 150 (diagonal covariance) Gaussian mixtures trained using Polyphone database with MFCCs coefficients. This world model GMM was then used as a priori information for MAP adaptation for a new client, using all the enrollment data to yield speaker specific GMM parameterized by Λ_k. For each new client, a simplified version of the MAP algorithm [8, 9] was used. This version consists of adapting only the Gaussian means:

$$\hat{\mu}_{j_k} = \alpha\mu_{j_\Lambda} + (1-\alpha)\frac{\sum_{n=1}^{N} P(j|x_n)x_n}{\sum_{n=1}^{N} P(j|x_n)} \tag{9}$$

where $\hat{\mu}_{j_k}$ is the new mean of the j-th Gaussian for client k, μ_{j_Λ} is the corresponding mean in the world model (Λ), and α is the adaptation rate.

At the end of the enrollment process, each client is thus modeled by the set of parameters $\{M_k, \theta_k^*, \Lambda_k\}$. During verification, we estimate the normalized $\log P(M_k|\theta_k^*, X)$ by performing a forced Viterbi algorithm between the test utterance X and the inferred model M_k using local posterior probability estimated by the SD-SLP (θ_k^*). We then estimate the normalized log likelihood ratio as usually done in text-independent speaker verification, and compute the final score which is compared to a speaker-independent threshold:

$$\frac{1}{N}\left[\log P(M_k|\theta_k^*, X) + \log P(X|\Lambda_k) - \log P(X|\Lambda)\right] \geq \delta_1 \tag{10}$$

Where N is the length of the test access after silence frames have been removed.

4.3 Unconstrained and constrained HMM approach

The speaker verification systems with hidden Markov models (HMM)are built up by training two speaker independent speech recognizers, each one with 36 context-independent phone models. The phone models of the first HMM speech recognizer (λ) consist of 3 states left-to-right HMM with 24 mixtures/state. This HMM model was used as a "world model" to estimate the probability of the denominator in (4) in the unconstrained HMM method. The phone models of the second HMM speech recognizer $(\hat{\lambda})$ consist of 3 states left-to-right HMM with 3 mixtures/state. This HMM model was used as a prior distribution for MAP adaptation of the new client in both constrained and unconstrained HMM methods and as "world model" to estimate the probability of the denominator in (5) in the constrained HMM method. Both HMM models are trained using Polyphone database and MFCC coefficients. Once the user HMM model (M_k) is inferred, a MAP adaptation procedure using all the enrollment utterances is performed. This procedure consists of adapting the mean of the Gaussians of the phone models of the model ($\hat{\lambda}$) which constitute the inferred model (M_k) using (9). The result is a SD-HMM model ($\hat{\lambda}_k$). For the verification process, we have used the normalized log likelihood ratio which is compared to a speaker independent threshold:

$$\frac{1}{N}[\log P(X|M_k,\hat{\lambda}_k) - \log P(X|M,\lambda)] \geq \delta_2 \tag{11}$$

for **unconstrained HMM** method or

$$\frac{1}{N}[\log P(X|M_k,\hat{\lambda}_k) - \log P(X|M_k,\hat{\lambda})] \geq \delta_3 \tag{12}$$

for **constrained HMM** method, and N is the length of the test access after the silence frames have been removed.

5 Experiments and Results

All experiments reported here were conducted using the Torch library recently developed at IDIAP[2]. To thoroughly test the proposed approach, we investigated the worse scenario where all customers would be using the same password, as well as a more realistic scenario where customers would have different passwords (and with impostors using the right and/or wrong password). For comparison purpose, results of each method with the a priori knowledge of the correct phonetic transcription of the password is also given.

5.1 Clients with the same password

The first scenario of the experiments consists of the evaluation of the performance of the system when all the clients choose the same password, in our case, the word *'annulation'*. For this purpose, each client (19 clients) pronounces about 27 utterances for testing, 5 of them being wrong accesses (i.e., words different than the correct password), while the others being true accesses. All impostors are from outside the set of registered clients, which is more likely in practical applications. 19 impostors were selected from PolyVar database. Each impostor has two accesses for testing, one access with the correct password of the claimed identity and one with the wrong password. This makes up a total of 420 true client accesses, and 779 false impostor accesses (including the true client pronouncing wrong words). A speaker-independent threshold was set a posteriori to equalize the probability of false acceptance and false rejection. The results of this experiment are given in Table 1.

Models	Correct password	Inferred password
Constrained HMM	1.5%	2.08%
Unconstrained HMM	2.34%	2.4%
HMM/ANN-GMM	2.14%	2.14%

Table 1: *Equal error rates for constrained and unconstrained HMM and combined HMM/ANN-GMM methods, with the correct and the inferred phonetic transcription of the password.*

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

5.2 Clients with different password

In practical applications, clients will probably choose different passwords. In order to investigate this case, we conducted a second experiment, involving the same 19 clients with 17 of them with a different password. The experimental set up was identical to the one used in the previous experiment, resulting in a total of 417 true client accesses, and 779 false impostor accesses. Table 2 gives the results of this experiment.

Models	Correct password	Inferred password
Constrained HMM	4.1%	2.6%
Unconstrained HMM	5.27%	4.5%
HMM/ANN-GMM	3.9%	3.9%

Table 2: *Equal error rates for the three methods with different password for each client.*

6 Discussion

From the results reported in Table 1 and Table 2, we can conclude that: The proposed method performs better than the unconstrained HMM method, but, compared to the constrained HMM, it performs slightly worse except with the correct phonetic transcription of the password in the second experiment. This is probably because, the "world model" used for score normalization in the constrained HMM is more competitive and reduces the impostor scores. In the proposed method, this competitivity can be improved by using other confidence measures [7] [2] for utterance verification, which are developed in the hybrid HMM/ANN framework. Surprisingly, the results in the first experiment are better than the results in the second experiment. The reason is that, in the second experiment, the length of the passwords varied between 3 and 12 phonemes. So, for the clients who chose a short password, we did not have enough data to properly model their characteristics.

For a comparison purpose with text-independent speaker verification, the equal error rate obtained in the second experiment using only GMMs is 4.1%, which is quite similar to the performance of the proposed method. The advantage of the SV-UCP system is that they should be more secure than text-independent SV system. Given that the password is chosen from an unconstrained vocabulary, it will be difficult to an impostor to guess the password of the user.

Finally, for further analysis, we found that for the hybrid HMM/ANN, the false acceptance rate of impostors or clients who pronounce a different word than the claimed identity's password is 2.6%, while the false acceptance rate of impostors who pronounce the correct password is 30.4%. This shows that the hybrid HMM/ANN mainly discriminates between speakers based on the lexical content of the password. For the GMMs model, we found that the false acceptance rate of impostors who pronounce the correct password is only 6.1%.

7 Conclusion

A new method for speaker verification based on user-customized password is proposed. This method combines the advantages of the hybrid HMM/ANN systems used for utterance verification and GMM models used for speaker verification. Results on different application scenarios show the effectiveness of this method. In future work, we intend to use other HMM inference techniques with the hybrid HMM/ANN systems by using other confidence measures. These confidence measures will be used also for utterance verification. In the proposed method, the final score which is used to take the decision is a combination with equal weight between the scores of two models (utterance verification model and speaker verification model). This technique is optimal if the probabilities of both models are perfectly estimated. As this is not the case, the results can be further improved by using other fusion techniques (Support Vector Machines, MLP,...).

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*, pp. 1955-1958, 1997.

[3] M.F. BenZeghiba, H. Bourlard, J. Mariéthoz, "Speaker Verification based on User-Customized password" *IDIAP Research Report*, IDIAP-RR-13,2001

[4] M.F. BenZeghiba, H. Bourlard, "User-Customized HMM/ANN based Speaker Verification', *IDIAP Research Report*, IDIAP-RR-32, 2001.

[5] D. Genoud, D. Ellis and N. Morgan, "Combined speech and speaker recognition with speaker-adapted connectionist models", *Proc. Auto. Speech recog. and Understanding Workshop,* keystone

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

[7] 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.

[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] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models", *Digital Signal Processing,* vol. 10,n0 1-3,2000.