IDIAP RESEARCH REPORT

Hybrid HMM/ANN and GMM combination for User-Customized Password Speaker Verification

Mohamed Faouzi BenZeghiba [a]

Hervé Bourlard [a,b]

IDIAP–RR 02-45

November 19, 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 02-45

Hybrid HMM/ANN and GMM combination for User-Customized Password Speaker Verification

Mohamed Faouzi BenZeghiba Hervé Bourlard

November 19, 2002

Abstract. Recently we have proposed an approach for user-customized password speaker verification; in this approach, we combined a hybrid HMM/ANN model (used for utterance verification) and a GMM model (used for speaker verification). In this paper, we extend our investigations. First, we propose a new similarity measure that uses confidence measures developed in the HMM/ANN framework. Secondly, we analyze the contribution of each model using a weighted sum combination technique. Experiments conducted on a subset of the PolyVar database show that for a short password the performance of the combined system did not improve significantly compared to the performance using the GMM model alone, and that the HMM/ANN did not contribute much in the combined system. We discuss possible reasons for this.

1 Introduction

This paper addresses the problem of the User-Customized Password Speaker Verification (SV-UCP) where the user has the possibility to chose his/her password from an unconstrained vocabulary. This rises some difficulties. First, we have to infer the hidden Markov model of the password. Second, we have to create (or adapt) a speaker dependent model which captures/models both the lexical content of the password and the speaker characteristics using a small amount of enrollment data.

Recently [1], we have proposed an approach that combined the hybrid HMM/ANN [2] and GMM [3] models in the same statistical framework. In the HMM/ANN model, the artificial neural network (ANN) is used to estimate the emission posterior probabilities of the inferred Hidden Markov Model (HMM). The advantages of the hybrid HMM/ANN system is that the estimated posterior probabilities can be used to derive some confidence measure to tell us how well the pronounced utterance matches a word model [4, 5]. The GMM model is used as usually done in text-independent speaker verification to model the characteristics of the speaker. We have shown [1] that the adapted speaker-dependent ANN with its associated maximum a posteriori probability mainly modeled the lexical content of the password. To accept a speaker, we have used the following decision rule:

$$P(M_k, S_k|X) \geq P(M, S|X) \tag{1}$$

where $P(M_k, S_k|X)$ is the joint posterior probability that the correct speaker pronounced the correct password and $P(M, S|X)$ is the joint posterior probability that any speaker pronounced any text. By developing (1) and using Bayes rule with assumptions that all speakers have the same *a priori* probability and $P(M|S, X) = 1$[1], the final decision was rewritten as follows:

$$[P(M_k|S_k, X)] \left[\frac{P(X|S_k)}{P(X|S)}\right] \geq \Delta \tag{2}$$

where Δ is the threshold.

In this paper, we extend our investigation by processing separately impostors pronouncing the correct password of the claimant and impostors pronouncing the wrong password. Usually, we make a utterance verification step, so, if the pronounced utterance matches the password model of the claimant then we verify the identity of the speaker. In this work, we have used a weighted sum combination technique to combine the utterance verification and the speaker verification scores. We also investigated and analyzed the contribution of each model (HMM/ANN and GMM) to the final decision.

2 Decision rule

In SV-UCP, speaker S pronouncing the word M and claiming the identity of the user S_k is accepted if:

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

$$P(M_k, S_k|X) \geq P(\overline{M}_k, S|X) \tag{4}$$

where $P(M_k, \overline{S}_k|X)$ is the joint posterior probability that an impostor pronounced the correct password and $P(\overline{M}_k, S|X)$ is the joint posterior probability that any speaker (client or impostor) pronounced any other password (text).

Developing these two Equations and using Bayes rule with the assumption that all speakers have the same *a priori* probability, decision rules (3) and (4) can be rewritten as follows:

$$\left[\frac{P(M_k|S_k, X)}{P(M_k|\overline{S}_k, X)}\right] \left[\frac{P(X|S_k)}{P(X|\overline{S}_k)}\right] \geq \Delta_1 \tag{5}$$

[1] Which is true if we use Baum-Welch algorithm instead of Viterbi decoding without taking into account the transition probabilities.

$$\left[\frac{P(M_k|S_k,X)}{P(\overline{M}_k|S,X)}\right]\left[\frac{P(X|S_k)}{P(X|S)}\right] \geq \Delta_2 \tag{6}$$

where Δ_1 and Δ_2 are the thresholds.

In these Equations, there are two kinds of scores. The likelihood ratio scores used for *speaker verification* and the posterior probability scores used for *utterance verification*. The likelihood ratios are estimated as usually done in text-independent speaker verification. The model used to estimate the denominator (referred to as world model) has a discriminative role: it discriminates between true speakers and impostors.

The posterior probabilities are estimated through a neural network, which is trained (or adapted) in a discriminative way (unlike the maximum likelihood). It has been found (for speech recognition) that these posterior probabilities are equivalent to the likelihood ratio used for utterance verification [6]. So, taking the ratio of two posterior probabilities estimated by two different neural network is not useful. So, all the posterior probabilities in (5) and (6) should be estimated using the same neural network (the adapted speaker dependent neural network in occurrence). This yields the following simplifications:

$$\frac{P(M_k|S_k,X)}{P(M_k|\overline{S}_k,X)} = 1 \tag{7}$$

and

$$\frac{P(M_k|S_k,X)}{P(\overline{M}_k|S,X)} = \prod_{n=1}^{N}\frac{p(q_k^n|x_n)}{p(q_{best}^n|x_n)} \tag{8}$$

where N is the length of the utterance X. Assuming that the transition probabilities are equal[2], $\frac{p(q_k^n|x_n)}{p(q_{best}^n|x_n)}$ represents the posterior probability of being in the decoded (according to the forced Viterbi alignment) state q_k at time n given the frame x_n divided by the best posterior probability of that frame at the time n. This Confidence Measure (CM) which is called Relative Posterior Confidence Measure (RPCM) [5], tells us how close the score of the decoded word is compared to the best acoustic score of the utterance. If a word is correctly recognized [3], this RPCM will be equal to 1. Substituting (7) and (8) into (5) and (6) respectively, and taking the logarithm, decision rules (5) and (6) can be rewritten as follows:

$$\log P(X|S_k) - \log P(X|\overline{S}_k) \geq \delta_1 \tag{9}$$

$$\sum_{n=1}^{N}\log\left[\frac{p(q_k^n|x_n)}{p(q_{best}^n|x_n)}\right] + [\log P(X|S_k) - \log P(X|S)] \geq \delta_2 \tag{10}$$

As explained in the introduction, a weighted sum combination technique is used. If we refer to the scores in (9) and (10) as s_1 and s_2 respectively, the combined score can be written as follows:

$$s_{com} = \alpha s_1 + (1-\alpha)s_2 \tag{11}$$

In this work, we have represented S and $\overline{S}_k$ by the same model. By expending Equation (11), we obtain the following decision rule to accept a speaker:

$$\alpha\left(\sum_{n=1}^{N}\log\left[\frac{p(q_k^n|x_n)}{p(q_{best}^n|x_n)}\right]\right) + \left[\log P(X|S_k) + \log P(X|\overline{S}_k)\right] \geq \delta \tag{12}$$

The parameter α ($0 \leq \alpha \leq 1$) indicates how much the contribution of the posterior probability score (related to the utterance verification) is in the final decision. As we can see, the weight of the likelihood ratio (related to the speaker verification) is equal to 1, indicating the importance of the

[2] Which is generally the case in HMM/ANN speech recognition systems.

[3] Which means that the decoded phone at each time has the best local posterior probability, even if it is not high.

GMM score in the final decision. In this paper, we compare the use of RPCM criterion to estimate the confidence measure of the utterance verification with the Standard Posterior Confidence measure (SPCM) criterion where $\left(\sum_{n=1}^{N}\log\left[\frac{p(q_k^n|x_n)}{p(q_{best}^n|x_n)}\right]\right)$ is replaced by $\left(\sum_{n=1}^{N}\log p(q_k^n|x_n)\right)$.

3 Databases and Experiment setup

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 words was divided into two subset *data1* and *data2* with 14 and 3 words respectively. A set of 38 speakers (24 males and 14 female) 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 in *data1* are used as training data, between 18 and 22 utterances of the same word were used as client accesses with the correct password. Each speaker has a subset of 19 speakers as an impostors. Two accesses with the correct word from each impostor and, *three* accesses with wrong password (form *data2*) for each speaker (client or impostor) were added to the test accesses.

For acoustic features, two kinds of features 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 which is used for HMM inference. In order to keep the characteristics of the user, MFCCs were used for speaker adaptation. 12 coefficients with energy complemented by their first and second derivatives were calculated every 10 ms over 30 ms window, resulting in 26 coefficients.

4 The Approach

As we have seen, there are two problems that we have to solve: the HMM inference and the speaker adaptation. In this section, we will describe briefly our approach. More details can be found in [1].

4.1 HMM Inference

We match (using Viterbi alignment) each of the enrollment utterances with an ergodic HMM model using local posterior probabilities estimated by a large Speaker-Independent Multi-Layer Perceptron (SI-MLP). This SI-MLP is trained on PolyPhone database with RASTA-PLP. We then chose from the inferred phonetic transcriptions the one with the highest normalized posterior probability to build the user HMM model M_k as explained in [1].

4.2 Speaker adaptation

Two models were adapted for each new speaker:

4.2.1 GMM adaptation

The GMM adaptation consisted of adapting the mean of Gaussians of a speaker independent GMM model (referred to as "world model") with 120 diagonal covariance Gaussians. The world model is trained on PolyPhone database with MFCC coefficients. The adaptation is performed using a simplified version of MAP (maximum *a posteriori*) adaptation algorithm [3].

4.2.2 Neural network adaptation

As the amount of adaptation data is very limited, the neural network adaptation consisted of adjusting the weights of a small Speaker-Independent Single-Layer Perceptron (SI-SLP) in a supervised manner. This SI-SLP is trained on PolyPhone database with MFCC coefficients. The segmentation was obtained by matching each of the enrollment utterances on the inferred HMM model M_k using local posterior probabilities estimated by the SI-MLP. One difficulty is that the adaptation data contains a small number of phonemes (those constituting the user HMM model M_k). During adaptation, the neural network will be biased to the outputs belonging to those phonemes, and may destroyed the performance on the other outputs and thus on the other words. To alleviate this problem, we added some examples (from PolyPhone) of phonemes that did not appear in the segmentation. The number of the added examples for each phoneme is equal to the average number of examples per phoneme in the segmentation.

5 Experiments and Results

All experiments reported here were conducted using the Torch library[4]. To compensate the difference in the dynamic range and to make the scores (posterior probability and likelihood ratio) more mathematically convenient, we mapped them to the [0,1] interval using sigmoid function [8, 9]. A speaker-independent threshold was set *a posteriori* to equalize the probability of false acceptance rate (FAR) and false rejection rate (FRR). For comparison purposes, results with the *a priori* knowledge of the correct phonetic transcription of the password are also reported.

5.1 Results

Figure 1 shows the variations of the equal error rate (EER) as a function of α and Table 1 gives the performance of each system using the corresponding optimal value of α. It can be observed that:

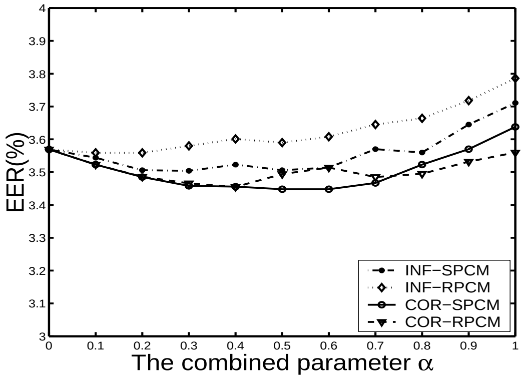

Figure 1: *EER variations as a function of the combined parameter α: INF (respectively COR) for systems with the inferred (respectively the correct) phonetic transcription. SPCM (respectively RPCM) for systems using the SPCM (respectively RPCM) criterion to estimate the HMM/ANN score.*

- All systems perform comparably and the RPCM criterion does not improve the result compared to SPCM.
- The value of α in systems with the inferred phonetic transcription is very small (0.2 and 0.3), indicating that the HMM/ANN score did not contribute much in the combined score. While in systems with the correct phonetic transcription, this value is higher (0.5 and 0.4) but still small.

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

	INF-SPCM	INF-RPCM	COR-SPCM	COR-RPCM
α	0.3	0.2	0.5	0.4
EER	3.51%	3.56%	3.45%	3.46%

Table 1: *The performance of different systems with optimal α.*

- Compared to a GMM only approach (Equation (12) with $\alpha = 0$, the corresponding EER is equal to 3.57%), the combined system shows no significant improvement in performance

- In Equation (12), if we chose SPCM instead of RPCM criterion and we put $\alpha = 1$, we will get the decision rule that has been used in [1]. The corresponding EER is equal to 3.72%, which is a little worse compared to the optimal EER.

5.2 Analysis and discussion of the results

The distribution (after mapping) of the HMM/ANN scores against the GMM scores for the system with the inferred phonetic transcription and the SPCM criterion (INF-SPCM) is shown in Figure 2. To make the Figure clear, we did not plot the distribution of impostors accesses with wrong passwords, since, we have found that the system is very robust in this situation (as explained later). The vertical and the horizontal lines correspond to the individual HMM/ANN and GMM thresholds respectively. The diagonal line corresponds to the decision boundary as found by the combination technique. From this Figure, we can conclude that:

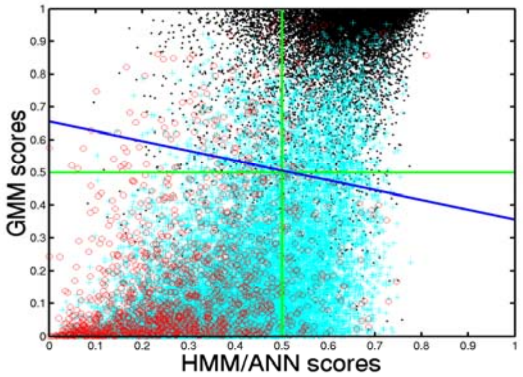

Figure 2: *The distribution of scores: The black point (respectively the cyan plus) corresponds to the true client (respectively to impostors) pronouncing the correct password. The red circle corresponds to the client pronouncing other password. The diagonal line shows the decision boundary. The distribution upper the decision line corresponds to the accepted speaker.*

1. Many impostor accesses with the correct password (cyan plus) obtained a good HMM/ANN score, confirming that the neural network mainly modeled the lexical content of the password.

2. As the decision boundary shows, the combined system uses information mainly given by the GMM score. This explains why the value of the parameter α is very small and why the performance of the combined system did not improve significantly compared to the use of only GMM scores.

3. Most of the client accesses with wrong passwords (circles) obtained a low GMM score, indicating, that the GMM model did not capture properly the characteristics of the client. One possible

explanation is that the adaptation data contains only a few phonemes (short password). So, the GMM model keeps only the speaker characteristics that are extracted from those phonemes, which are not sufficient to properly model all the speaker characteristics. In reality, it could happen that the client can not remember exactly his/her password. So, for applications with low level of security, we can use only the decision of the GMM model.

For more analysis, Table 2 shows different acceptance (true and false) rates related to different situations that may happen in real life. The first column gives the true acceptance rate of Client pronouncing the Correct Password (CCP). The second column gives the FAR rate of the true Client pronouncing Other Password (COP). The third column corresponds to the FAR of Impostors pronouncing the Correct Password (ICP) and the last column corresponds to FAR of Impostors pronouncing Other Password (IOP). For each situation, the rate corresponds to the ratio of the number of accepted accesses to the total number of accesses in that situation. We can make the following observations:

1. This approach is very robust to impostor accesses with wrong passwords, making the SV-UCP system more secure. Indeed, the fact that the password is chosen from an unconstrained vocabulary, makes it more difficult to an impostor to guess the user password.

2. The false acceptance accesses are mainly caused by impostors pronouncing the correct password or true clients pronouncing other passwords, making the set up of the parameter α more difficult. Indeed, in this case, the client will get (probably) a good GMM score (as it supposed to keep the characteristics of the speaker) and a low HMM/ANN score (as it modeled the lexical content of the password). In contrast, the impostor will get a low GMM score and a good HMM/ANN score. Depending on the value of the parameter α, we can distinguish two cases:

   - α has a small value: The combined score of client pronouncing the wrong password will be good enough and the client will be accepted, while the combined score of the impostor pronouncing the correct password will be low and the impostor will be rejected.
   - α has a high value: The combined score of the client access will be low and the client will be rejected, while the combined score of the impostor access will be good and the impostor will be accepted.

   So, a small value of parameter α penalizes the impostor pronouncing the correct password, while a high value penalizes the client pronouncing the wrong password. An optimal value of parameter α should make a compromise between these two types of FA. This corresponds to the value which minimize the sum of this two FAR. As the GMM model does some work that the HMM/ANN is supposed to do, by giving a low score to client accesses with wrong password, the optimal value of α will be small.

Models	CCP	COP	ICP	IOP
INF-SPCM	96.50%	8.75%	8.12%	0.15%
INF-RPCM	96.44%	9.37%	8.12%	0.17%
COR-SPCM	96.55%	7.04%	8.23%	0.07%
COR-RPCM	96.56%	7.74%	8.12%	0.09%

Table 2: *Different false and true acceptance rates corresponding to different situations for each system.*

6 Conclusion

In this paper, we investigated and analyzed the combination of the hybrid HMM/ANN and GMM models for user-customized password speaker verification. For a short password, we have found that the GMM model did not properly keep the speaker characteristics and it did some work that the HMM/ANN was supposed to do, explaining why the GMM model has much more contribution in the combined system than the HMM/ANN model.

7 Acknowledgment

Mohamed F. BenZeghiba is supported by the Swiss National Science Foundation thought the project "SV-UCP: Speaker Verification based on User-Customized password" (2000-063721.00-1). The main author would like to thank Conrad Sanderson for fruitful discussion and helpful comments on the paper.

References

[1] M. F. BenZeghiba and H. Bourlard, "User-Customized Password Speaker Verification based on HMM/ANN and GMM models", *Proceedings of ICSLP 2002*, pp 1325-1328, 2002.

[2] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A hybrid approach", Kluwer Academic Publisher,1994.

[3] D. A. Reynolds, T. F. Quatieri and R.B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol.10, N 1-3, 2000.

[4] G. Williams and S. Renals, “Confidence Measures for hybrid HMM/ANN speech recognition”, *Proceedings of Eurospeech’97*, pp. 1955-1958, 1997.

[5] E. Mengusoglu and C. Ris, “Use of Acoustic Prior Information for Confidence Measure in ASR Applications", *Proceedings of Eurospeech 2001*, pp 2557-2560, 2001.

[6] B. Gold and N. Morgan, “Speech and Audio processing”, Wiley, 2000.

[7] 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-RR96-01, 1996.

[8] P. Jourlin, J. luetten, D. Genoud and H. Wassner, “Acoustic-labial speaker verification”, *Pattern Recognition letters*, Vol. 18, No 9, 1997, pp. 853-858.

[9] C. Sanderson, “Information Fusion and Person Verification Using Speech and Face Information”, *IDIAP Research Report*, RR-02-33, Martigny, Switzerland, 2002.