IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

# User Customized HMM/ANN-based Speaker Verification

Mohamed F. BenZeghiba [a]   Hervé Bourlard [a,b]

IDIAP–RR 01-32

October 23, 2001

submitted for publication

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 01-32

# User Customized HMM/ANN-based Speaker Verification

Mohamed F. BenZeghiba Hervé Bourlard

October 23, 2001

submitted for publication

**Abstract.** In this paper, we describe a new speaker verification approach, using a hybrid HMM/ANN system, and accommodating user customized passwords. This system is exploiting the high phonetic recognition rates usually achieved by HMM/ANN speaker independent systems to infer the HMM topology associated with the user specific password from a few utterances of that password. Different adaptation schemes are then compared to quickly adapt the speaker independent ANN parameters used for HMM inference into speaker dependent parameters used for speaker verification. Different scoring criteria, based on normalized accumulated posterior probabilities (previously used as confidence measures in speech recognition) are also compared. Based on these improvements, our best system achieved false acceptance and false rejection rates of 8.2% and 3.2%, respectively, corresponding to an a posteriori threshold set to the minimum of the HTER (half total error rate), and in the worse case where all customers are using the same password.

**Acknowledgements**: Mohamed Faouzi BenZeghiba is supported by the Swiss National Science Foundation through the project "SV-UCP: Speaker Verification based on User-Customized Password" (2100-054018.98-1).

# 1 Introduction

The approach presented in the present paper exploits some of the advantages of hybrid HMM/ANN systems [6] where 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 [11], which makes them particularly amenable to perform HMM inference from acoustic data [7].

Speaker verification based on user-customized password (SV-UCP) has to address two issues. The first problem consists in finding the topology of the HMM model which better represents the password chosen by the user, thus capturing/modeling the lexical content of the password. The second problem is to quickly adapt the ANN parameters towards the targeted speaker, based on a few utterances of the user specific password, thus capturing/modeling the speaker characteristics. Since adaptation data is thus very limited, and given that we want to capture very quickly the characteristics of the speaker, a new ANN adaptation scheme yielding significant improvements will also be discussed here.

Finally, the present paper also discusses and compares different scoring and decision strategies based on likelihood ratio or posterior probabilities (more appropriate to HMM/ANN systems). Indeed, most speaker verification approaches use likelihood ratio as hypothesis testing. However, following [9], we also tested different similarity measures based on a posteriori probabilities, and which were recently proposed to estimate confidence measures in HMM/ANN speech recognition systems.

# 2 User Customized HMM/ANN Speaker Verification

## 2.1 Generic Approach

As illustrated in Fig. 1, the user customized enrolment and speaker verification system described here is based upon:

1. A well-trained speaker independent ANN (in our case, a multilayer perceptron, MLP), of parameters $\Theta$, and which is known to perform very well on phonetic classification at the acoustic frame level.

2. A "world" HMM model $M$, defined as an ergodic (looped) HMM, of emission parameters $\Theta$. In our case (HMM/ANN hybrid), each phoneme was represented by a single state with a minimum duration constraint or with transition probabilities reflecting this duration constraint.

**Enrolment** of a new customer then consists in the following steps:

1. A new customer $S_k$ pronounces $J$ (typically, 2 or 3) times his/her password $X_k^j$, $j = 1, \dots, J$, where $X_k^j$ represents the sequence of acoustic vectors associated with the $j$-th utterance.

2. Match each of the enrolment utterances $X_k$ onto $M$, using the speaker independent parameters $\Theta$, to generate a phonetic transcription of each utterance, together with its associated accumulated posterior probability..

3. Choose the phonetic transcription yielding the highest accumulated posterior probability, and use it to build a reference HMM model $M_k$ representing the password of customer $S_k$.see section 4.

4. Match each of the enrolment utterances $X_k^j$, $j = 1, \dots, J$ on the speaker specific model $M_k$ to yield phonetic segmentation of these utterances.

5. As for the training of HMM/ANN systems, adapt the ANN parameters $\Theta$ by using the above segmentation to provide the ANN target outputs and to minimize (by the usual back-propagation algorithm) the square error between the observed output vector generated by each input vector of the enrolment utterances and the associated target output vector (obtained from the above

segmentation), yielding a *speaker specific* ANN model of parameter $\Theta_k$. Fast adaptation of the speaker independent MLP (SI-MLP) is not easy given the large number of parameters. A solution to this problem is discussed below.

6. Once the speaker-adapted ANN has been trained, possibly perform an additional Viterbi alignment on the HMM associated with the user-customized password (initially inferred with the speaker independent ANN) using the speaker specific ANN, thus generating new ANN targets used for further ANN training.

**Verification** of a speaker $S$, pronouncing $X$ and claiming to be $S_k$, then consists of the following steps:

1. Load the HMM model $M_k$ associated with the password of $S_k$, the speaker specific MLP model of parameters $\Theta_k$, the world HMM model $M$ with its associated MLP parameters $\Theta$.

2. Perform Viterbi matching of $X$ on model $M_k$ and parameters $\Theta_k$ to compute $P(X|M_k, \Theta_k)$, representing the likelihood that $X$ was actually produced by $S_k$ (since using $\Theta_k$) pronouncing $M_k$.

3. Perform Viterbi matching of $X$ on model $M$ and parameters $\Theta$ to compute $P(X|M, \Theta)$, representing the probability that $X$ was generated by another speaker (which, as for text-independent speaker verification, is estimated here from a speaker independent model) pronouncing any word (whence the looped world model). Another possibility, underestimating the likelihood of the world model, would be to use $P(X|M_k, \Theta)$, thus estimating the probability that the right password has been produced by a speaker different than $S_k$.

4. Compute the likelihood ratio and check whether or not it is above a speaker-specific threshold; see Section 3.

## 2.2 Improvements to the Generic Approach

As further discussed below, the above generic approach, if implemented as such, did not yield satisfactory results. Consequently, several improvements, representing the main contribution of the present paper, were required to achieve good results. Although further discussed below, the characteristics of the best system can be summarized as follows:

- Use different acoustic parametrization schemes for HMM inference and speaker adaptation. As illustrated in Figure 1, for each utterance, we extract RASTA-PLP parameters $X$ (more speaker independent) for HMM inference and mel frequency cepstral coefficients $Y$ (containing more speaker specific information) for ANN adaptation.

- Training two speaker independent ANNs:

  1. The large one already mentioned above and trained with $X$ on a large speaker independent database, resulting in the parameter set $\Theta$ and yielding high phonetic recognition rates. This ANN will be used for HMM inference.

  2. A small one (in our case, an MLP with no hidden units) trained on the same large database, but using acoustic features $Y$. Of course, this MLP performs quite poorly as a speaker independent model (given the limited number of parameters), but will be used for fast speaker adaptation during enrolment.

- Scoring: Using (normalized) accumulated posterior probabilities $P(M_k|Y_k, \Theta_k)$ as a scoring criteria, directly available from the output of the ANN, instead of the usual likelihood ratio ( 3).

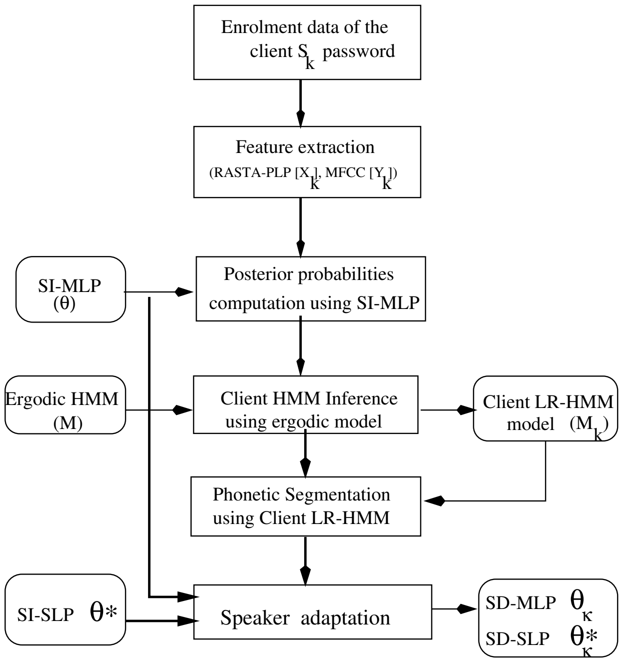


Figure 1: *SV-UCP block-diagram*

# 3 HMM Scoring

In the case of speaker verification based on user-customized password, the decision to validate a speaker $S$ claiming to be $S_k$, should be based on:

$$P(S = S_k) = P(M_k|X, \Theta_k) \tag{1}$$

resulting in the hypothesis test:

$$S = S_k \quad \text{if} \quad P(M_k|X, \Theta_k) \geq \delta_k \tag{2}$$

It is also known (see, e.g., [4]) that, assuming equal class priors, this criteria is also equivalent to the likelihood ratio test:

$$\mathcal{L}_k = \frac{P(X|M_k, \Theta_k)}{P(X|M, \Theta)} \geq \Gamma_k \tag{3}$$

usually used in speaker verification, and where the denominator is often referred to as the likelihood of the "world model".

In hybrid HMM/ANN systems, we only have access to estimates of local posterior probabilities $p(q_k|x_n)$ where $q_k$ is a particular HMM state (in our case, associated with a particular MLP output class and a specific phone) and $x_n$ (or $y_n$) the current acoustic vector at time $n$. Consequently, the global likelihood ratio (3) is estimated through the product along the optimal HMM path of scaled local likelihoods:

$$\frac{p(x_n|q_k)}{p(x_n)} = \frac{p(q_k|x_n)}{p(q_k)}$$

For posterior based scoring, and based on the recent work with HMM/ANN-based posteriors to estimate confidence measure [11], three scoring criteria have been tested.

- **TN**: Time normalized accumulated log-posteriors

$$\hat{P}_1(M_k|X,\Theta_k) = \frac{1}{N}\log P(M_k|X,\Theta_k) = \frac{1}{N}\sum_{n=1}^{N}\log P(q_\ell^n|x_n,\Theta_k) \quad (4)$$

  where $q_\ell^n$ represents the optimal state $q_\ell$ visited at time $n$ along the Viterbi path, and $N$ the length of $X$.

- **TNS**: Time normalized posteriors, after having extracted the password from the preceding and following silence frames

$$\hat{P}_2(M_k|X,\Theta_k) = \left[\frac{1}{N_u}\sum_{n=b}^{e}\log P(q_\ell^n|x_n,\Theta_k)\right] \quad (5)$$

  where $N_u = e - b + 1$ represents the length of the isolated password utterance.

- **DN**: Confidence measure based on double normalization [2]

$$\hat{P}_3(M_k|X,\Theta_k) = \left[\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{n=b_\ell}^{e_\ell}\log P(q_\ell^n|x_n,\Theta_k)\right]$$

  where $n_\ell = e_\ell - b_\ell + 1$ represents the length of phone $q_\ell$, and $L$ is the number of phones in the hypothesized HMM model $M_k$.

Since in our experiments likelihood ratio based systems (3) never performed as well as posterior based systems, only the latter criteria will be compared in the sequel.

# 4 HMM Inference

For enrolment, each new customer pronounces his/her password $J$ times (5 in our case). We then match each of the enrolment utterances with the ergodic HMM model $M$ using local posterior probabilities $P(q_\ell|x_n,\Theta)$ obtained from the large SI-MLP of parameters $\Theta$, and trained with RASTA-PLP parameters.

This results in 5 phonetic transcriptions, from which we select the one yielding the highest time normalized accumulated posterior probability (4) as the baseline phonetic transcription of the password. The other posterior-based criteria presented above were also tested, but did not result in significant speaker verification differences.

The topology of the user-customized HMM model $M_k$ is then simply built 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 "optimal" phonetic sequence.

# 5 Fast Speaker adaptation

Using the baseline model $M_k$, we then have to adapt on the few enrolment utterances the parameter set $\Theta$ to the customer. For the fast adaptation of SI-MLP parameters, different MLP adaptation schemes have been investigated, including:

1. Retraining all the MLP parameters
2. Linear Input Network (LIN) [10]: in this case, we add an additional input layer to the MLP, whose weights are adapted, while keeping the weights of the speaker independent ANN fixed.

In both cases, different cross-validation techniques (the first three repetitions used for adaptation, while the last two are used for testing the generalization properties) were used to avoid overtraining. Although many variants of these approaches were tested [1], none of them could achieve satisfactory results. The best results, referred to as **SD-MLP** in the sequel, were achieved by the LIN where the additional parameters were limited to a scaling of each input value.

In a much more successful approach, and as already mentioned in Section 2.2, we trained a small speaker independent ANN of parameters $\Theta^*$, simply consisting of Single Layer Perceptron, referred to as **SI-SLP** without input context (thus resulting, in our case, in a network with 26 input units, 36 output units, resulting in 973 parameters). While the big ANN of parameters $\Theta$ is simply used to perform HMM inference, this SLP (initially trained on the same data) is used during adaptation as the initial model to create the speaker dependent model, referred to as **SD-SLP**. Sine it has much few parameters and a simpler topology, the SLP network was found to converge much faster and to generalize much better.

As a further improvement to this approach, it was also found that training the SI-SLP and performing its adaptation towards the SD-SLP was also resulting in significant improvements when using more speaker specific features such as MFCC. The resulting network will then be referred to as **SDD-SLP** $\theta^*$.

In all cases, the SI-MLP or SI-SLP adaptation was done by using the initial segmentation resulting of a forced Viterbi alignment of the enrolment utterances to the inferred password-specific HMM, followed by one more Viterbi iteration (segmentation, followed by MLP training).

# 6 Experimental results

## 6.1 Experimental conditions

Two databases were used in this work. The (Swiss French) Polyphone database [3], a large telephone database containing prompted and natural sentences pronounced by a large number of different speakers, was used to train[1] the SI-MLP ($\Theta$) and the SI-SLP ($\Theta^*$). The SI-MLP consisted of 234 inputs (nine consecutive frames with 26 acoustic features each), 600 hidden units and 36 outputs (associated with 36 phones). Such an MLP typically achieved a frame-based phonetic recognition rate higher than 75% on the Polyphone test data, as well as on the data that has been used here for speaker verification.

To perform our speaker verification tests, we used the *PolyVar* database [3], from which we extracted 19 speakers (12 males and 7 females) repeating several times (between 26 and 229) the same set of 17 words. In our case, we used the *same password* (*"annulation"*), thus working in the most difficult conditions. For each speaker, the first 5 utterances were used as enrolment data, and between 15 and 26 utterances were used as true accesses. 36 utterances of the other speakers were used as impostor accesses.

For acoustic parameters, we used 12 RASTA-PLP ($X_k$) coefficients (to train the SI-MLP and the SI-SLP) and 12 cepstral coefficients (to train and adapt the SDD-SLP), always complemented by their first temporal derivatives, as well as the first and second temporal derivatives of the log energy.

The decision threshold was set a posteriori to minimize the HTER ($(FA + FR)/2$) error.

## 6.2 Speaker Verification Tests

In a first experiment, both the SI-MLP and the SI-SLP were trained with RASTA-PLP features, respectively achieving 75% and 45% frame-based phonetic recognition rates. The results obtained by the two systems are given in the first two columns of Table 1 for the three different scoring criteria.

In a second set of experiments, we improved the performance of the system by training and adapting the SD-SLP based on MFCC features, instead of RASTA-PLP, to preserve a bit more of

[1] Using the 10 phonetically rich sentences read by 400 speakers. The SI-MLP and SI-SLP were respectively trained on 5 and 1.5 hours of data.

speaker specific information. The resulting performance is given in the third column of Table 1. These results show a small (13% relative) improvement in false acceptance, but a very large (57.3% relative ) improvement in false rejection.

| Models | SD-MLP | | SD-SLP | | SDD-SLP | |
|---|---|---|---|---|---|---|
| Scoring | FR% | FA% | FR% | FA% | FR% | FA% |
| TN | 8.3 | 18.9 | 7.2 | 13.3 | 6.3 | 12.3 |
| TNS | 9.9 | 14.4 | 7.5 | 9.5 | **3.2** | **8.2** |
| DN | 11.1 | 15.8 | 10.4 | 10.7 | 6.5 | 8.0 |

Table 1: . *False Acceptance (FA) and False Rejection (FR) rates for the different adaptation schemes (SD-MLP: adapting the big speaker independent MLP; SD-SLP: adapting a small linear network; SDD-SLP: adapting a small linear network and using different features) and different scoring criteria (TN: time normalized; TNS: time normalized, without silence; DN: double normalisation).*

As a comparison point, the best HMM-based *text dependent* speaker verification [8] achieved an HTER of 2.9% in the same experimental conditions, except that each speaker was represented by 17 HMM models (one model for each repetition of the password). As reported in [5], an HMM-based user-customized speaker verification achieved 12.3% HTER, also in the same experimental conditions.

# 7 Conclusion

In this paper, we have described the modifications required to a baseline user customized HMM/ANN-based speaker verification to significantly improve verification performance. From the results presented in Table 1, we can draw the following conclusions:

1. It is usually better to remove non-informative silence frames of the test utterance in the score computation.

2. Using a smaller network (SD-SLP) yields significant improvements, as compared to the adaptation of the network used for HMM inference.

3. Using different features for HMM inference and speaker adaptation results in further improvements.

4. While double normalization (DN) has been shown to be a good confidence measure in speech recognition, it does not seem to be appropriate to speaker verification based on user-customized password. The reason may come from the fact that the automatically inferred HMM model is not a good model of the password.

5. The SD-SLP was adapted to discriminate between the phones in the speaker adaptation data. This means that the SD-SLP has also learned the lexical content of the password.

# References

[1] M.F. BenZeghiba, H. Bourlard and J. Mariéthoz, “Speaker Verification Based On User-Customized Password”, *IDIAP Research Report*, IDIAP-RR-13-01, 2001.

[2] 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.

[3] 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.

[4] B. Gold and N. Morgan, *Speech and Audio Processing*, Wiley, 2000.

[5] B. Jacob, J. Mariéthoz, G. Gravier, and F. Bimbot, "Robustesse de la vérification du locuteur par mot de passe personnalisé", *Proceedings of JEP'2000* (Aussois, France), pp. 357-360, 2000.

[6] 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.

[7] M. Magimai Doss and H. Bourlard, "Pronunciation models and their evaluation using confidence measures" *IDIAP research Report*, http://www.idiap.ch, IDIAP-RR-01-29, 2001.

[8] J. Mariéthoz and F. Bimbot, "Adaptation robuste de modèles HMM pour la vérification du locuteur dépendente du text", *Proceedings of JEP'2000* (Aussois, France), pp. 349-352, 2000.

[9] T. Matsui and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (martigny, Switzerland), pp. 59-62, 1994.

[10] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM/ANN continuous speech recognition system", *Proceedings of Eurospeech* (Madrid), pp. 2171-2174, 1995.

[11] G. Williams and S. Renals, "Confidence measures for hybrid HMM/ANN speech recognition," *Proceedings of Eurospeech'97*, pp. 1955-1958, 1997.