IDIAP RESEARCH REPORT

Confidence Measures In Multiple Pronunciations Modeling For Speaker Verification

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

IDIAP–RR 03-53

October 2003

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 03-53

Confidence Measures In Multiple Pronunciations Modeling For Speaker Verification

Mohamed Faouzi BenZeghiba Hervé Bourlard

October 2003

Abstract. This paper investigates the use of multiple pronunciations modeling for User-Customized Password Speaker Verification (UCP-SV). The main characteristic of the UCP-SV is that the system does not have any *a priori* knowledge about the password used by the speaker. Our aim is to exploit the information about how the speaker pronounces a password in the decision process. This information is extracted automatically by using a speaker-independent speech recognizer. In this paper, we investigate and compare several techniques. Some of them are based on the combination of confidence scores estimated by different models. In this context, we propose a new confidence measure that uses acoustic information extracted during the speaker enrollment and based on *log likelihood ratio* measure. These techniques show significant improvement (15.7% relative improvement in terms of equal error rate) compared to a UCP-SV baseline system where the speaker is modeled by only one model (corresponding to one utterance).

1 Introduction

In most text-dependent speaker verification systems, the system has *a priori* knowledge about the text, such as the phonetic transcription (pronunciation) of the password which is given by an expert (phonetician) or extracted from a standard pronunciation dictionary. These systems achieve relatively good performance but at the cost of user constraints. In this paper, we are interested in what we referred to as User-Customized Password Speaker Verification (UCP-SV) [1], where the speaker is free to choose a password on which the verification will be performed. Typically, each speaker repeats a password a few times, then a speaker-independent speech recognizer infers the phonetic transcriptions where each utterance will be represented by a sequence of phonemes. The inferred phonetic transcriptions (pronunciations) can be considered as an information source of *how a speaker pronounces a password*.

In previous work [1], we have used the *best* (according to a posteriori probability criterion) phonetic transcription to build-up the speaker dependent Hidden Markov Model (HMM). In this work, we extend our investigation by allowing several alternative phonetic transcriptions to be used for speaker modeling. Hereafter, this will be referred to as multiple pronunciations modeling.

One of the factors that can degrade the performance of a speaker verification system is the intra-speaker variability, that is, the speaker can not pronounce the same word with the same manner even in the same session. By using multiple pronunciations, it could be possible to reduce this effect.

From these multiple models, the problem is then how to compute the confidence score in order to accept or reject a speaker? The goal of this paper is to investigate and compare several procedures. Some of them are based on the use of the confidence score of the best model, and others use some combination techniques. In this context, we propose a new confidence measure that uses acoustic information extracted using the train data. It normalizes the *log likelihood ratio* of the test data by the *log likelihood ratio* of the train data. All these techniques are *log likelihood ratio* based.

The rest of the paper is organized as follows; Section 2 describes briefly the databases we have used and the experimental set-up. Section 3 describes the speaker acoustic modeling during the enrollment step and Section 4 describes some of the techniques we have investigated and reports the obtained results.

2 Databases and Experimental Set-up

Two databases were used in this work. The Swiss French PolyPhone database [2] was used to train different speaker-independent speech recognizers. The speaker verification experiments were conducted using the PolyVar database [2]. 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 (composed of 3 to 12 phonemes each) common for all speakers. 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 are used as training data, between 18 and 22 utterances of the same word were used as client accesses. Each speaker has a subset of 19 speakers as impostors, each impostor has two accesses for each word. For acoustic features, 12 MFCCs coefficients with energy complemented by their first derivatives were calculated every 10 ms over 30 ms window, resulting in 26 coefficients.

Two speaker-independent speech recognizers were trained using PolyPhone database:

- A Multi-layer perceptron (MLP) with a set of parameter Θ. This SI-MLP has 234 input units with 9 consecutive 26 dimensional acoustic vectors, 600 hidden units and 36 outputs, each output associated with a specific phone.
- An HMM with a set of parameter λ. This HMM has 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 priori distribution for MAP (maximum *a posteriori*) adaptation [3].

3 Speaker Acoustic Modeling

In UCP-SV, the enrollment of a new speaker consists of two steps; (1) the inference of the phonetic transcriptions (pronunciations) and, (2) the creation of the speaker acoustic model (HMM).

3.1 Pronunciations inference

As mentioned above, we are interested in UCP-SV, where no *a priori* information is available about the possible pronunciations of the chosen password. These pronunciations should be determined automatically. A speaker-independent hybrid HMM/MLP [4] system is used to infer the pronunciation of each utterance in the enrollment data. To do this, we match (using Viterbi decoding) each of the utterances with an ergodic HMM model using local posterior probabilities estimated through the SI-MLP Θ. This results in L pronunciation models ($M_\ell,\ \ 1 \leq \ell \leq L$), in our case $L = 5$. The consistency of these pronunciations depends on the performance of the acoustic speech recognizer.

3.2 Models selection and parameter estimation

Once we infer the possible pronunciations, we then aim to create the speaker-dependent model that best represents the lexical content of the password and achieves the best performance. Two approaches are investigated.

3.2.1 Single model

In this approach, from all the inferred pronunciations, we chose the *best* one to build the speaker model. The best pronunciation $\widehat{M}$ is defined as the one produces the highest posterior probability over all the enrollment utterances, i.e:

$$\widehat{M} = \arg \max_{1 \leq l \leq L} \left[\sum_{i=1}^{I} \log P(M_\ell | X_i, \Theta) \right] \tag{1}$$

where L is the number of pronunciations,

$$\log P(M_\ell | X_i, \Theta) = \frac{1}{N_i} \sum_{n=1}^{N_i} \log p(q_k^{n,i} | x_{n,i}, \Theta) \tag{2}$$

and $p(q_k^{n,i} | x_{n,i}, \Theta)$ is the local posterior probability of the decoded phone q_k at time n associated with the frame $x_{n,i}$ of the i^{th} utterance and N_i is the length of the utterance X_i.

Once $\widehat{M}$ is selected, a MAP adaptation procedure is performed which consists of adapting the mean of the Gaussians of phone models of λ constituting $\widehat{M}$. This results in a speaker-dependent HMM model parametrized by λ_c.

3.2.2 Multiple models

The problem with the previous approach, is that the speaker can not pronounce exactly the same word in the same manner from one trial to another. So, if there is a mismatch between $\widehat{M}$ and the test utterance this will cause a degradation in the verification performance. In this approach, instead of selecting only one pronunciation, we keep all of them and we build-up an HMM model for each pronunciation using the same MAP adaptation procedure as explained in the previous section.

4 Confidence Measures based Speaker Verification

This section describes and reports results of several techniques that we have used to take the decision to accept or reject a speaker for both conditions explained in Sec (3.2). The decision can be expressed as follows:

$$S \;=\; S_k \;\; if \;\; CM \geq \Delta \tag{3}$$

where S represents the test speaker, S_k the claimed speaker, CM is the confidence measure based hypothesis testing criteria and Δ is a speaker-independent threshold. As we are using HMM models, all the confidence measures described bellow are *log likelihood ratio* (LLR) based. For all techniques, the performance is reported using a threshold determined *a posteriori* to minimize the equal error rate (EER).

4.1 Single model

In this case[1], the confidence measure is just the *log likelihood ratio* usually used in text-dependent speaker verification. Assuming that the simultaneous probability $P(M, S)$ of any speaker and any word is equal for all combinations of speakers and word, the CM in (3) can be expressed as:

$$CM_1 = \frac{1}{N}\left[\log P(X|\widehat{M}, \lambda_c) - \log P(X|\widehat{M}, \lambda)\right] \tag{4}$$

where N is the length of the test utterance after removing the silence segment, and $\widehat{M}$ is the inferred pronunciation according to (1). Table1 shows the performance of CM_1 compared to that obtained by the reference system where the correct pronunciation of the password is known.

	Posteriori threshold	EER (%)
Reference	1.36	5.65%
CM_1	1.45	7.03%

Table 1: *The performance of UCP-SV using CM_1 compared to the reference system.*

It is clear that the reference system achieves better performance than UCP-SV. It has also been found that this improvement is mainly due to the normalization model $(\widehat{M}, \lambda)$, which is automatically inferred in the case of UCP-SV. So, it did not reflect how all speakers pronounce the password but how the target speaker pronounces it, which reduced the competitiveness of this model.

4.2 Multiple models

There are two alternatives to compute the confidence measure from multiple models. The first solution is to dynamically (during the test) select the model with the best confidence measure. The second solution uses all the models and apply some combination techniques that will make use of all the individual model CM.

4.2.1 Dynamically choosing the best model

The criterion we have used to select the *best* model is based on the confidence measure estimated for each model using the test utterance. The CM in (3) can be defined as follows:

$$CM_2 = \max_{1 \leq \ell \leq L}\left[\frac{1}{N}\left[\log P(X|M_\ell, \lambda_{(c,\ell)}) - \log P(X|M_\ell, \lambda)\right]\right] \tag{5}$$

where L is the number of models, and $\lambda_{(c,\ell)}$ is the client model associated with the pronunciation M_ℓ. Table2 shows the performance of the system using the i^{th} $(1 \leq i \leq 5)$ best model.

[1]For more detail about the approach, see [1].

	Posteriori threshold	EER (%)
First best	2.35	6.95%
Second Best	1.86	6.36%
Third best	1.58	6.01%
Fourth best	1.28	5.97%
Fifth best	0.82	6.50%

Table 2: *The performance of UCP-SV using CM_2.*

From the table, we can observe that:

- Using multiple models always gives better results than the use of a single model.
- The best performance is not achieved by using the best model. In our case, surprisingly, the fourth model yields the best performance.

This can simply be explained by the fact that our multiple models decision strategy is probably not optimal. To analyze these results, we plot the false acceptance (FA) and the false rejection (FR) error rates for each system using a *priori* threshold $\Delta = 1.28$ (the threshold of the best system).

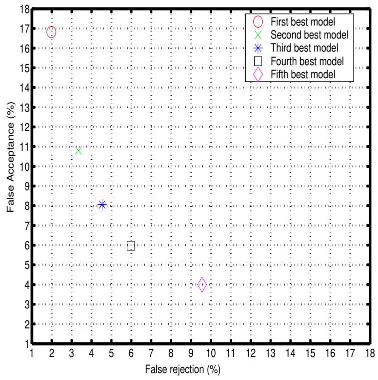

Figure 1: *False rejection and false acceptance error rates for each system using a priori threshold* $\Delta = 1.28$.

It can be observed from the Figure1 that using the first best model gave a low FR rate but increased considerably the FA rate, while the use of the worst model decreased the FA rate but increased the FR. So, the *optimal* model is somewhere between these two models. This is the weakness of this technique (in real application) as the selection of this *optimal* model is not obvious and depends on the data.

4.2.2 Averaging confidence measures

A simple way to do the combination is to take the average confidence measure. In this case, the CM in (3) can be defined as follows:

$$CM_3 = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{N}\left[\log P(X|M_\ell,\lambda_{(c,\ell)}) - \log P(X|M_\ell,\lambda)\right] \quad (6)$$

where L is the number of models.

Using this CM, the obtained EER was equal to 5.97% with a threshold $\Delta = 1.58$. This shows that taking the average over all the individual CM significantly improves the performance (15.14% relative improvement) compared to the system using a single model.

The obtained EER is equal to that obtained by the *optimal* system in section(4.2.1). This indicates that for a specific speaker, there is a model associated with a specific pronunciation that can be a *trade off* between a good client model and a good normalization model for the test utterance. So, all the discriminant information used to take the decision is represented by this model. It is worth mentioning that (for a specific speaker) all models are adapted with the same MAP adaptation procedure using the same data, but using different pronunciation models. So, if there is any complementary information that the combination could benefit from, it is most probably the pronunciation variation.

4.2.3 Normalized confidence measure

As suggested in [5], the confidence measure used here is similar to the *voting technique* where the final decision is based on the local decision taken by each subsystem (model). In this approach, the CM in (3) is then defined as follows:

$$CM_4 = \frac{1}{L}\sum_{\ell=1}^{L} f(cm_\ell) \tag{7}$$

where

$$f(cm_\ell) \quad = \quad \begin{cases} 1, & if \quad cm_\ell \geq \delta_{(c,\ell)}; \\ 0, & otherwise \end{cases} \tag{8}$$

$$\tag{9}$$

and cm_ℓ is the CM computed using (4) with the speaker model $(M_\ell, \lambda_{(c,\ell)})$. and $\delta_{(c,\ell)}$ is a local speaker and model dependent threshold. This CM_4, which belongs to the $[0,1]$ interval, can be interpreted as a percentage of times that the local confidence measure cm_ℓ exceeded its local threshold $\delta_{(c,\ell)}$.

One difficulty that can make the use of CM_4 undesirable in real application is the estimation of the local threshold $\delta_{(c,\ell)}$ for each speaker's model. It is desirable to have a local threshold which is;

- Speaker and model independent ($\delta_{(c,\ell)} = \delta$): So, it can be determined a *priori* on separate data.
- Interpretable and adjustable: so, it will be easy to adjust its value according to the application requirements.

The local confidence measure cm_ℓ has a wide dynamic range since, theoretically $cm_\ell \in]-\infty, +\infty[$. To satisfy the above two conditions, we propose a new confidence measure cm'_ℓ that transforms cm_ℓ to a more interpretable CM. The new cm'_ℓ uses the *log likelihood ratio* of the train data to normalize cm_ℓ. It is based on the following assumption:

$$CM(train) \geq CM(test) \tag{10}$$

$$\Rightarrow \qquad \frac{CM(test)}{CM(train)} \leq 1 \tag{11}$$

which says that the confidence measure computed using the train data is the best confidence measure we can get.

Using this assumption, the new confidence measure cm'_ℓ can be defined as follows[2]:

$$cm'_\ell = \frac{cm_\ell}{\frac{1}{I}\sum_{i=1}^{I}\left[\log P(X_i|M_\ell,\lambda_{(c,\ell)}) - \log P(X_i|M_\ell,\lambda)\right]} \tag{12}$$

where I is the number of training utterances for the speaker c. The denominator is the average *log likelihood ratio* of the training data computed for each model. By using (12):

[2]Actually, this confidence measure was introduced for utterance verification using hybrid HMM/ANN systems. The posterior probability of a phone is normalized by the average posterior probability of the same phone using the train data [6]

- The new confidence measure cm'_ℓ will have a limited dynamic range with an upper bound equal to 1. It indicates how much mismatch is there between the test data and the train data. Closer cm'_ℓ to 1, more probable the claimed identity is valid.

- The search for a local speaker-independent threshold δ will be in a fixed range $[0, 1]$ [3]. So, depending on the application requirements, we can fix *a prior* threshold without the need of a separate data.

Note that in this technique we have two thresholds, a local threshold δ and a global threshold Δ. In Figure2, we plot the variations of the FR and FA error rates for several values of Δ [4]. The EER was equal to 5.93% (15.7% relative improvement), obtained with global threshold $\Delta = 0.6$ and local threshold $\delta = 0.21$.

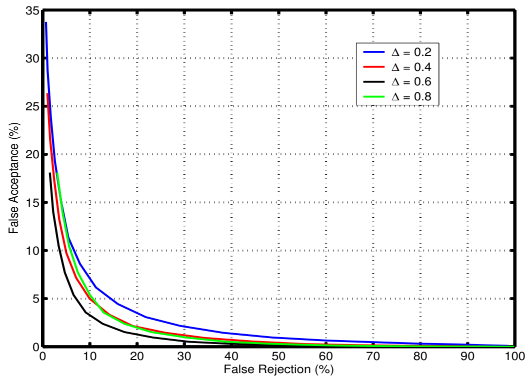

Figure 2: *FR and FA error rates using CM_4: For each global threshold Δ we have varied the local threshold δ from* 0 *to* 1 *and each step we compute the associated FR and FA error rates.*

This result is quite similar to that obtained using the average CM (6). The advantage of this technique, however, is that the threshold has a meaningful interpretation and is easily adjustable according to the application requirements.

5 Conclusion

Several techniques that exploit the use of multiple pronunciations modeling in user-customized password speaker verification have been investigated. Experiments carried out on PolyVar database showed that the use of multiple pronunciations significantly improved the performance (15.7% relative improvement in terms of EER) of the system compared to that obtained by the one allowing only one pronunciation.

A new confidence measure has also been presented. This CM normalizes the usual *log likelihood ratio* of the test utterance by the *log likelihood ratio* of the train data. This normalization makes the confidence score more meaningful and interpretable.

As a future work, we intend to further investigate the robustness of the new confidence measure in mismatch conditions, e.g., the effect of this normalization on the FA and FR error rates. We should also study how useful this normalization is for *a priori* threshold estimation and finally how to extend this approach to the sub-word (phone) level in text-dependent speaker verification.

[3] Theoretically, the threshold is equal to 0 (in log domain) and practically higher than 0.
[4] In our case, CM_4 has 6 possible values $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

6 Acknowledgment

The authors gratefully acknowledge the support of the Swiss National Science Foundation throught the project "MULTI :2000-068231.02/1. This work was also carried on in the framework of the SNSF National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)".

References

[1] M. F. BenZeghiba and H. Bourlard, "User-Customized Password HMM Based Speaker Verification" *Proceedings of the COST275 Workshop on the Advent of Biometrics on the Internet", 2002*, pp 103-106, Rome, Italy

[2] 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.

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

[4] 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.

[5] Q. li, B.-H. Juang, Q. Zhou, C.-H. Lee, "Automatic Verbal Information Verification for User Authentication", *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 5,2000.

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