IDIAP RESEARCH REPORT

User-Customized Password Speaker Verification Using Multiple Reference and Background Models

Mohamed Faouzi BenZeghiba [a,b]

Hervé Bourlard [a,b]

IDIAP–RR 04-41

July 30th, 2004

revised in April 2005

published in

Speech Communication, 2006, Volume 48

[a] Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny

[b] Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

IDIAP Research Institute **www.idiap.ch**

Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland

Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-41

User-Customized Password Speaker Verification Using Multiple Reference and Background Models

Mohamed Faouzi BenZeghiba   Hervé Bourlard

July 30th, 2004

revised in April 2005

published in
Speech Communication, 2006, Volume 48

Abstract. This paper discusses and optimizes an HMM/GMM based User-Customized Password Speaker Verification (UCP-SV) system. Unlike text-dependent speaker verification, in UCP-SV systems, customers can choose their own passwords with no lexical constraints. The password has to be pronounced a few times during the enrollment step to create a customer dependent model. Although potentially more "user-friendly", such systems are less understood and actually exhibit several practical issues, including automatic HMM inference, speaker adaptation, and efficient likelihood normalization. In our case, HMM inference (HMM topology) is performed using hybrid HMM/MLP systems, while the parameters of the inferred model, as well as their adaptation, will use GMMs. However, the evaluation of a UCP-SV baseline system shows that the background model used for likelihood normalization is the main difficulty. Therefore, to circumvent this problem, the main contribution of the paper is to investigate the use of multiple reference models for customer acoustic modeling and multiple background models for likelihood normalization. In this framework, several scoring techniques are investigated, such as Dynamic Model Selection (DMS) and fusion techniques. Results on two different experimental protocols show that an appropriate selection criteria for customer and background models can improve significantly the UCP-SV performance, making the UCP-SV system quite competitive with a text-dependent SV system. Finally, as customers' passwords are short, a comparative experiment using the conventional GMM-UBM text-independent approach is also conducted.

Contents

1 Introduction

Speaker Verification (SV) is the task of automatically accepting or rejecting a claimed identity based on the voice characteristics of a speaker [1]. Speaker verification can be divided into text-dependent and text-independent. In Text-Dependent Speaker Verification (TD-SV), the text is constrained to be a known phrase, which can be fixed or randomly prompted from a small vocabulary (usually digits) [2]. Hence, the system has *a priori* knowledge about the text. In Text-Independent Speaker Verification (TI-SV), there is no constraint on the text during enrollment and verification steps. Test utterances can be completely different from enrollment utterances. Consequently, TI-SV systems need a large and rich training data to model properly the characteristics of the speaker's voice. Because in TD-SV systems, the speaker's model encodes both the lexical properties and the speaker's voice characteristics, these systems usually achieve better performance compared to TI-SV systems. However, because user has no freedom to choose the predefined password, TD-SV systems are less user-friendly and not fully appreciated by users.

This paper studies another kind of TD-SV systems which are more user-friendly. That is customer can choose easily his/her own password without any lexical constraints. The password has to be pronounced a few times during a short enrollment step to create a customer specific model that will be subsequently used for verification. Such a system is referred to as User-Customized Password Speaker Verification (UCP-SV). Given that the password is chosen from an unconstrained lexicon, it makes it more difficult for an impostor to guess the customer's password. The main assumption in a UCP-SV system is that no *a priori* knowledge about the password is available to the system. The goal of this paper is to study how does this assumption affect the performance of the UCP-SV compared to a TD-SV.

However, UCP-SV systems present new challenges. First, the system has to automatically infer the topology of the Hidden Markov Model (HMM) associated with the password simply based on a few utterances. The inferred model has then to be parameterized in terms of speaker-independent parameters (in our case, Gaussian Mixture Models (GMM)) that can easily be adapted to the customer characteristics. Finally, we have to design an appropriate likelihood normalization model that best competes with the customer model with respect to the test utterance. The likelihood normalization model will be used in the usual log likelihood ratio test. As we will see, this is considered as the main problem of an HMM/GMM based UCP-SV system and it will be the focus of this paper.

2 HMM/GMM based UCP-SV system

Figure 1 illustrates the enrollment process of a new customer in the HMM/GMM based UCP-SV system we have implemented. Each step will be discussed in more detail in the sequel of the paper.

1. *HMM inference:* The main assumption in UCP-SV systems is that no *a priori* information about the lexical content of the password is available. This information should be inferred automatically in terms of sub-word units like phonemes. A speech recognizer is used to transcribe each utterance of the customer's password to a sequence of phonemes. The inferred Phonetic Transcriptions (PTs) should be representative of the lexical content of the password. The accuracy of the inferred PTs depends on the accuracy and the consistency of the speech recognizer. Ideally, the inferred PTs should almost be the same for all utterances of the same password, but in practice, this is never the case. Hence, we have to pick the one that best represents the enrollment data or find ways to consolidate all resulting PTs. In our case, the HMM inference is performed using a speaker-independent hybrid HMM/MLP system [3]. In this system, the Multi-Layer Perceptron (MLP) is used to estimate HMM states posterior probabilities (or scaled likelihoods) of an ergodic lexical model to map each of the enrollment utterance into a phonetic sequence. The HMM/MLP systems are successfully used in speech recognition and are known to usually yield better performance at the frame level compared to other systems like HMM/GMM, thus being better suited to recognize utterances in terms of phone sequences.

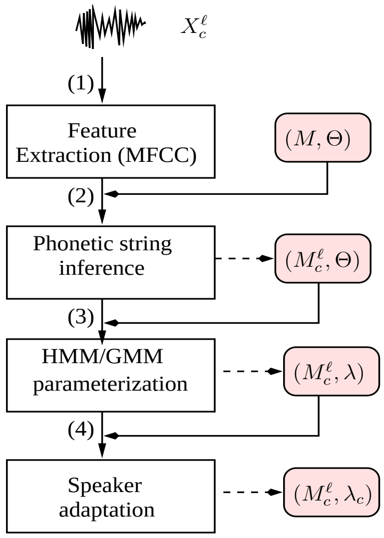

Figure 1: *Block-diagram of the enrollment process in an HMM/GMM based UCP-SV system: For each enrollment utterance* X_c^ℓ*, corresponding to the* ℓ^{th} *repetition of the password pronounced by the customer (c), first (1) we extract MFCC features, which will be used in an ergodic HMM/MLP model* (M,Θ) *to (2) generate the most probable phonetic string* M_c^ℓ *(still parameterized by* Θ*) associated with each utterance. We then (3) parametrize the resulting HMM topologies* M_c^ℓ *with the parameter set* λ*, corresponding to speaker-independent GMMs, resulting in left-to-right speaker-independent HMM/GMM models* (M_c^ℓ,λ) *which will be used as background models for likelihood normalization. Finally, (4) a MAP adaptation procedure is applied to* (M_c^ℓ,λ) *to create the speaker-dependent HMM/GMM model* (M_c^ℓ,λ_c)*.*

2. *Speaker adaptation:* Having inferred the HMM topology associated with the password, the next step is to create the customer specific acoustic model. Since in the UCP-SV the amount of enrollment data is very limited, a *model adaptation techniques* are used. The adaptation process requires us to consider an appropriate parameter adaptation scheme as well as the number of parameters to be adapted. If we stick to hybrid HMM/MLP model, we have to adapt the MLP parameters. In [4], we reported all the attempts we made to adapt the parameters of the SI-MLP for each new customer, where we tried different adaptation techniques based on the training of a small linear input transformation [5]. However, none of the tested techniques were satisfactory enough for such small amounts of adaptation data. Consequently, we decided to parametrize the resulting HMM models in terms of speaker-independent GMMs, which are then used as *a priori* distribution for *Maximum A Posteriori* (MAP) adaptation [6], yielding the customer and password specific acoustic model.

3. *Likelihood normalization:* Speaker verification is an hypothesis testing, usually casted in terms of a *likelihood ratio* test. That is the likelihood estimated by the speaker model is normalized by the likelihood estimated by a background model (a model representing all other possible speakers). This is still the case with UCP-SV systems, except that the hypothesis to test is whether we have the right speaker pronouncing the right password. The background model should be determined in such a way that the discriminant capability of the system against impostor accesses will improve. Empirical studies, suggest that a background model which is close to the target speaker is a reasonable choice [7]. This statement makes the design of such a model in UCP-SV systems more difficult, since there is no *a priori* knowledge available about the phonetic coverage of the customer password. This will be the main investigation of this paper.

In this paper, we first describe and evaluate a baseline UCP-SV system similar to the one presented in [8] for an open set speaker identification. This baseline system uses the best inferred PT to create the customer password HMM model. A comparison with a *reference* TD-SV system that uses the *correct* PT of the password (given by a dictionary) shows that the main difficulty in UCP-SV system lies in the background model.

The main contribution of this paper is then to improve the performance of this baseline system using multiple reference models for customer acoustic modeling and multiple background models for likelihood normalization. That is for each inferred PT, a customer and a background HMM models are created. The different inferred PTs provide us with information about how customer pronounces his/her password. This information can be used to improve the performance of the UCP-SV system[1]. This paper will demonstrate that the use of multiple reference and background models allows us to select an appropriate customer and background models in terms of customer and password characteristics with respect to the test utterance.

3 HMM inference

The goal of this step is to infer the speaker-independent password HMM. For this purpose, we have used an ergodic speaker-independent HMM/MLP system with a set of MLP parameters Θ to map each of the enrollment utterances X_c^ℓ into a phonetic sequence. More precisely, for each acoustic sequence $X_c^\ell = \left\{x_{(1,c)}^\ell, x_{(2,c)}^\ell, ..., x_{(T,c)}^\ell\right\}$ associated with each utterance of the customer password, the MLP outputs provide, for each acoustic frame $x_{(t,c)}^\ell$ at its input, an estimate of the posterior probabilities $p(q_k^t|x_{(t,c)}^\ell, \Theta)$ of phones q_k, with $k = 1, ..., K$ and, where K is the total number of phones. These posterior probabilities are then converted to scaled likelihood using Bayes rule[2]. Using these phone

[1] SuperSID project at the JHU summer workshop, 2002, http://www.clsp.jhu.edu/ws2002/groups/supersid

[2] $\frac{P(q_k^t|x_t)}{P(q_k)} = \frac{p(x_t|q_k^t)}{P(x_t)}$.

likelihoods and an ergodic HMM model M with minimum state duration constraints equal to 3 and phone transition probability[3] set to 0.5, a simple dynamic programming algorithm [9] is applied to estimate the best phonetic sequence. This results in L phonetic transcriptions M_c^ℓ (with $1 \leq \ell \leq L$ and L is the number of enrollment utterances) which are still parameterized with Θ. Once the PTs have been inferred, we then aim to create the customer-dependent password HMM/GMM, that best represents the lexical content of the password. In this work, we have investigated two approaches: single reference modeling approach and multiple reference modeling approach.

4 Single reference approach

4.1 HMM/GMM parameterization

In this case, we simply select the phonetic transcription $\widehat{M_c}$ yielding the highest normalized (by the number of frames) likelihood over all the enrollment utterances using forced alignment technique, i.e.:

$$\widehat{M_c} = \underset{1 \leq \ell \leq L}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log P(X_c^i | M_c^\ell, \Theta) \right] \tag{1}$$

where $I = L$, the number of enrollment utterances (hence phonetic transcriptions), and $\log P(X_c^i | M_c^\ell, \Theta)$ is defined as follows:

$$\log P(X_c^i | M_c^\ell, \Theta) = \frac{1}{T_i} \sum_{t=1}^{T_i} \log \left(\frac{P(q_k^{(t,\ell)} | x_{(t,c)}^i, \Theta)}{P(q_k)} \right) \tag{2}$$

where $\frac{P(q_k^{(t,\ell)} | x_{(t,c)}^i, \Theta)}{P(q_k)}$ is the local scaled likelihood of the decoded phone $q_k^{(t,\ell)}$ using forced Viterbi alignment on the inferred model M_c^ℓ at time t associated with the frame $x_{(t,c)}^i$ of the i^{th} enrollment utterance, and T_i is the length of the utterance X_i without silence frames.

The HMM password model is then built-up by strictly concatenating left-to-right (with only loops and skips to the next state) HMM phone models from λ [4] corresponding to each of the phones in the above "optimal" phonetic sequence $\widehat{M_c}$. This results in an HMM model $(\widehat{M_c}, \lambda)$ which is *lexically* customer-dependent but *acoustically* speaker-independent. The HMM model $(\widehat{M_c}, \lambda)$ will be used as *background* model for likelihood normalization.

4.2 Speaker adaptation

Once the speaker-independent password models $(\widehat{M_c}, \lambda)$ has been inferred, a MAP adaptation procedure [6] is then performed using the enrollment data to estimate $(\widehat{M_c}, \lambda_c)$, where λ_c represents the set of speaker adapted phonetic HMM/GMM parameters. This procedure consists of adapting the mean of state Gaussians of $(\widehat{M_c}, \lambda)$ models. We have used a modified version of the adaptation formula:

$$\hat{\mu}_{j_c}^{q_i} = \alpha \mu_{j_\lambda}^{q_i} + (1-\alpha) \frac{\sum_{t=1}^{T} P(j, q_i | x_t) x_t}{\sum_{t=1}^{T} P(j, q_i | x_t)} \tag{3}$$

where $\hat{\mu}_{j_c}^{q_i}$ is the new mean of the j-th Gaussian in the state q_i for client S_c, $\mu_{j_\lambda}^{q_i}$ is the corresponding mean in the model $(\hat{M}_c, \lambda)$, $P(j, q_i | x_t)$ is the joint posteriori probability of the state q_i and the Gaussian j and α is the adaptation rate.

[3] Several of the values have been tested, including $\frac{1}{K}$ (K is the number of phones). We have observed that this probability has no significant effect on the topology of the model. We thus chose 0.5 as a uniform value for transition probabilities.

[4] A context-independent and speaker-independent phonemes HMM speech recognizer (see Section 8.2 for more details).

5 Multiple reference approach

Using the above selection criterion (1), the resulted customer dependent password HMM $(\widehat{M_c}, \lambda_c)$ might match well with the speaker enrollment data, but it does not mean that during the access to the system (test): (1) this model will be *lexically* the most likely during verification and (2) the associated *background* model $(\widehat{M_c}, \lambda)$ will be *lexically*[5] the appropriate model for likelihood normalization. To alleviate these two problems, we propose the use of multiple reference and background modeling approach. In this approach, for each phonetic transcription, we create a customer password and a background models, using the same procedure described above in single reference approach. This results in a set of L customer dependent password HMMs (M_c^ℓ, λ_c) and a set of L background models (M_c^ℓ, λ).

6 Decision Rules

In speaker verification, the decision that a test speaker S is indeed verified as the claimed identity S_c can be expressed as follows:

$$S \;=\; S_c \;\text{ if }\; CS \geq \Delta \tag{4}$$

where CS is the estimated confidence score representing the reliability that the speech segment comes from the claimed identity and Δ is a speaker-independent threshold.

In UCP-SV system, we should verify both the identity of the speaker, as well as the validity of the pronounced password. Formally, we are interested in estimating $P(M_c, S_c|X)$, representing the joint posteriori probability that the customer S_c has pronounced the expected password M_c given the observed acoustic vector X. During verification, this probability is compared to (1) $P(M_c, \overline{S_c}|X)$, representing the joint posterior probability that any other speaker (impostor) $\overline{S_c}$ may have pronounced the expected password M_c, and (2) $P(\overline{M_c}, S|X)$, representing the joint posterior probability that any speaker (impostor or customer) S may have pronounced any other password $\overline{M_c}$. Hence, the decision rules can be formulated as follows:

$$(S, M) = (S_c, M_c) \quad \text{if} \quad P(M_c, S_c|X) \geq P(M_c, \overline{S_c}|X) \tag{5}$$

$$\text{and} \quad P(M_c, S_c|X) \geq P(\overline{M_c}, S|X) \tag{6}$$

Using Bayes' rule, and assuming that the joint *a priori* probability of any speaker and any word is equal for all combinations of speakers and words, decision rules (5) and (6) can be rewritten as follows:

$$\frac{p(X|M_c, S_c)}{p(X|M_c, \overline{S}_c)} \geq \frac{P(M_c, \overline{S}_c)}{P(M_c, S_c)} = \Delta_1 \tag{7}$$

$$\frac{p(X|M_c, S_c)}{p(X|\overline{M}_c, S)} \geq \frac{P(\overline{M}_c, S)}{P(M_c, S_c)} = \Delta_2 \tag{8}$$

where Δ_1 and Δ_2 are the decision thresholds. The normalization models $(M_c, \overline{S}_c)$ and $(\overline{M}_c, S)$ in (7) and (8) used to estimate the normalization scores $p(X|M_c, \overline{S}_c)$ and $p(X|\overline{M}_c, S)$, respectively, have two different roles. The first normalization model $(M_c, \overline{S}_c)$ is supposed to represent the correctly pronounced password. So, it is used to discriminate between the customer and impostors pronouncing the expected password. This likelihood normalization model will be referred to as *background model* and the decision using (7) to as *speaker verification* decision. If the speech content of the test utterance is different from the expected password, both customer and background models in (7) will have a poor individual likelihood which might result in a good likelihood ratio and leads to the acceptance of an impostor. A solution to this problem is to make a speech recognition or utterance verification step to recognize or to verify the lexical content of the pronounced word. This is the role of the second

[5] In text-dependent speaker verification, the lexical content of the background model is important.

likelihood normalization model in (8). This model is supposed to represent the incorrectly pronounced password. This likelihood normalization model will be referred to as *world model* and the decision using (8) to as *utterance verification* decision. The speaker is then, accepted if the two scores in (7) and (8) exceed their respective thresholds Δ_1 and Δ_2 simultaneously. It has been found [10] that the combination of these two scores can significantly improves the performance of the system. In this paper, a weighted sum combination technique is used. The confidence score CS in (4) is then defined as follows:

$$CS = \alpha \ LLR_s + (1 - \alpha) \ LLR_u \tag{9}$$

with $0 \leq \alpha \leq 1$. LLR_s is the normalized *speaker verification* log likelihood ratio, estimated as:

$$LLR_s = \frac{1}{T} \log \left[\frac{p(X|M_c, S_c)}{p(X|M_c, \overline{S}_c)} \right] \tag{10}$$

and LLR_u is the normalized *utterance verification* log likelihood ratio, estimated as:

$$LLR_u = \frac{1}{T} \log \left[\frac{p(X|M_c, S_c)}{p(X|\overline{M}_c, S)} \right] \tag{11}$$

We used $\frac{1}{T}$ to normalize the two *log likelihood ratio* for test utterance duration, where T is the length of the test utterance after having removed the silence frames.

7 Speaker verification

7.1 Score normalization

To verify the identity of the speaker, we need to define the world model $(\overline{M}_c, S)$ and the background model $(M_c, \overline{S}_c)$ to estimate LLR_s and LLR_u. If we have some *a priori* knowledge about the content of the password, this can help us in designing effective score normalization models for both speaker and utterance verification parts. Unfortunately, in UCP-SV, such information is not available.

For the *utterance verification* part, the *world* model should represent all the words but the customer's password. Training a model satisfying this condition is a very difficult task (actually impossible). In this work, we have used a general speech Gaussian mixture model (GMM) with a set of parameters Λ.

For the speaker verification part, A straightforward way to define a *background* model in UCP-SV system is to use the inferred speaker-independent password HMM. However, this model might not be optimal. To improve the competitiveness of this model, a previous study [11] proposed the use of a modified normalizing model (MNM) determined by perturbing the inferred background model using the enrollment data to reflect the lexical content of the speaker's password. In another study [12], the authors proposed the use of the speaker enrollment data to (1) train a background model with fewer number of parameters compared to the speaker model or (2) perturbing the temporal information by reversing the state order of the previously trained background model. In the work reported here, we will demonstrate that the use of multiple background models, corresponding to the inferred speaker-independent password HMM models can improve the performance of the UCP-SV system.

7.2 Single reference approach

In single reference modeling approach, the two *log likelihood ratios* LLR_s in (10) and LLR_u in (11) are estimated as follows:

$$LLR_s = \frac{1}{T} \log \left[\frac{p(X|\widehat{M}_c, \lambda_c)}{p(X|\widehat{M}_c, \lambda)} \right] \tag{12}$$

$$LLR_u = \frac{1}{T} \log \left[\frac{p(X|\widehat{M}_c, \lambda_c)}{p(X|\Lambda)} \right] \tag{13}$$

During the forced Viterbi decoding [13], a silence phone model is applied at the beginning and the end of the customer model ($\widehat{M}_c$,λ_c) to detect the silence frames and ignore them during the log likelihood estimation. The resulted speech/silence segmentation is used to estimate the log likelihoods $p(X|\widehat{M}_c, \lambda)$ and $p(X|\Lambda)$.

7.3 Multiple reference approach

Given a set of customer specific HMM models and a set of background models, the CS is now estimated by selecting (1) the customer model that best represents the test utterance, and (2) the background model that best competes with the customer model.

There are two possible solutions to that problem. The first solution consists in dynamically selecting during the access to the system the customer and the background models that satisfy some "optimal" criteria. Such techniques will be referred hereafter to as *dynamic model selection* (DMS) techniques. The second solution is to fuse the confidence scores or the partial decisions estimated/made by each individual subsystem [6] to derive the final score. Such techniques will be referred hereafter to as *confidence score fusion* and *partial decision fusion*, respectively.

7.3.1 Dynamic Model Selection techniques

In multiple reference modeling approach, the confidence score in (4) will be estimated as follows:

$$CS = \alpha CS_s + (1 - \alpha) CS_u \tag{14}$$

where CS_s and CS_u are the confidence scores of the speaker verification and the utterance verification parts, respectively.

- *Utterance verification part:*
  The performance of the *utterance verification* part will largely depend on how good the customer model matches the test utterance, since the estimation of CS_u uses a GMM as a score normalization model. The optimal criterion, with respect to the role of the *world* model, is probably to select the most likely customer model $(\widehat{M}_c^\ell, \lambda_c)$. That is:

  $$(\widehat{M}_c^\ell, \lambda_c) = \underset{1 \le \ell \le L}{\text{argmax}} \ \frac{1}{T} \log p(X|M_c^\ell, \lambda_c) \tag{15}$$

  The CS_u in (14) is then estimated as follows:

  $$CS_u = \frac{1}{T} \log \left[\frac{p(X|\widehat{M}_c^\ell, \lambda_c)}{p(X|\Lambda)} \right] \tag{16}$$

- *Speaker verification part:*
  The performance of the *speaker verification* part does not depend only on how good the customer model matches the test utterance, but also on how well the background model competes *lexically* (as all of them are speaker-independent) with the customer model. Consequently:

  – If we assume that (M_c^ℓ, λ_c) and (M_c^ℓ, λ) are statistically independent, then both customer and background models may have different model selection criteria to optimize CS_s.

  – If we assume that (M_c^ℓ, λ_c) and (M_c^ℓ, λ) are statistically dependent[7], then the selection criterion might depend on some statistics applied directly to the LLR_s estimated by each subsystem.

  In this work, three different criteria are tested. They are presented below according to the competitiveness of the background model to the customer model from low to high level:

[6] In a subsystem, both the customer and background models have the same pronunciation model.

[7] At least they have the same topology and (M_c^ℓ, λ_c) is adapted from (M_c^ℓ, λ).

1. **Maximizing** $p(X|M_c^\ell, \lambda_c)$: Using this criterion, the background model associated with the best customer model selected according to (15), is used for likelihood normalization. The CS_s in (14) is then estimated as follows:

$$CS_s = LLR_s^{\widehat{M}_c^\ell} = \frac{1}{T}\log\left[\frac{p(X|\widehat{M}_c^\ell, \lambda_c)}{p(X|\widehat{M}_c^\ell, \lambda)}\right] \tag{17}$$

   However, this criterion might not be a good criterion for the *speaker verification* part with respect to the competitiveness constraint. As we will see in the results, a good customer model might have a poor associated background model.

2. **Maximizing** $p(X|M_c^\ell, \lambda)$: While keeping the same customer HMM model selection criterion (15) as before, maximizing $p(X|M_c^\ell, \lambda)$ aims to make the background model more competitive by selecting the one that best matches the test utterance X as follows:

$$(\widehat{M}_c^{\ell'}, \lambda) = \underset{1 \leq \ell \leq L}{\text{argmax}} \ \frac{1}{T}\log p(X|M_c^\ell, \lambda) \tag{18}$$

   Thus the CS_s in (14) will be estimated as follows:

$$CS_s = \frac{1}{T}\log\left[\frac{p(X|\widehat{M}_c^\ell, \lambda_c)}{p(X|\widehat{M}_c^{\ell'}, \lambda)}\right] \tag{19}$$

   It might happen that both customer and background models will have the same topology (i.e., $\ell = \ell'$). In this case, this criterion will be equivalent to the previous one.

3. **Minimizing** $LLR_s^{M_c^\ell}$: Because (M_c^ℓ, λ_c) is derived from (M_c^ℓ, λ) by adapting only the mean of state GMMs, hence (M_c^ℓ, λ) is probably, the most appropriate background model for the customer model (M_c^ℓ, λ_c). Consequently, it might be better if the model selection criterion for the speaker verification part will be applied directly to the $LLR_s^{M_c^\ell}$, with respect to the competitiveness constraint. That is the background model should be close to the customer model. The criterion proposed here, selects the phonetic transcription M_c^ℓ that minimizes the $LLR_s^{M_c^\ell}$. Hence, the CS_s in (14) will be estimated as follows:

$$CS_s = \min_{1\leq\ell\leq L}\left(\frac{1}{T}\log\left[\frac{p(X|M_c^\ell, \lambda_c)}{p(X|M_c^\ell, \lambda)}\right]\right) \tag{20}$$

The drawback of dynamic model selection criteria though is that there is no guarantee that the selected set of parameters (customer and background models) are "optimal" in the sense of yielding the optimal EER.

7.3.2 Confidence score fusion

In confidence score fusion, the inputs to the fusion system are the *individual* confidence scores estimated by each subsystem, and the outputs are the average of $LLR_s^{M_c^\ell}$ and $LLR_u^{M_c^\ell}$ over all subsystems. The confidence scores CS_s and CS_u are then estimated as follows:

$$CS_s = \frac{1}{L}\left[\sum_{\ell=1}^{L} LLR_s^{M_c^\ell}\right] \tag{21}$$

and

$$CS_u = \frac{1}{L}\left[\sum_{\ell=1}^{L} LLR_u^{M_c^\ell}\right] \tag{22}$$

where L is the number of subsystems. The final confidence score CS is then a weighted sum of CS_s and CS_u:

$$CS = \alpha CS_s + (1-\alpha)CS_u \tag{23}$$

The use of the average criterion prevents us from the choose of a poor set of parameters (subsystem) to estimate CS_u and CS_s.

7.3.3 Partial decision fusion

In partial decision fusion, the inputs to the fusion system are the *individual* decisions made by each subsystem and the output is the final confidence score. The fusion system uses a *majority voting* technique, as suggested in [14]. The CS in (4) is then defined as follows:

$$CS = \frac{1}{L}\sum_{\ell=1}^{L} f(cs_\ell) \tag{24}$$

where

$$f(cs_\ell) = \begin{cases} 1, & \text{if} \quad cs_\ell \geq \delta_{(c,\ell)} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

$$\tag{26}$$

with cs_ℓ being the combined individual confidence score estimated using the phonetic transcription M_c^ℓ, and $\delta_{(c,\ell)}$ being a local speaker and model dependent threshold. This CS, which belongs to the $[0,1]$ interval, can be interpreted as a percentage of times that the local confidence score cs_ℓ exceeded its local threshold $\delta_{(c,\ell)}$.

One difficulty that can make the use of this technique undesirable in real application is the estimation of the local threshold $\delta_{(c,\ell)}$ for each speaker's subsystem. Indeed, it is desirable to have a local threshold that:

1. Is customer and model independent ($\delta_{(c,\ell)} = \delta$), hence, it can be determined *a priori* on separate data.
2. Is interpretable and adjustable, so it can easily be adjusted according to the application requirements.
3. Allows the parameter α to be optimized independently on the subsystem.

$LLR_s^{M_c^\ell}$ and $LLR_u^{M_c^\ell}$ have a large dynamic range, theoretically belonging to $]-\infty, +\infty[$ interval. To satisfy the above conditions, we introduce the *normalized log likelihood ratio* (NLLR) that transforms $LLR_s^{M_c^\ell}$ and $LLR_u^{M_c^\ell}$ into more interpretable scores. The *normalized log likelihood ratio* uses the *log likelihood ratio* of the train data to normalize the *log likelihood ratio* of the test data, and it is based on the following assumption:

$$\frac{LLR(test)}{LLR(train)} \leq 1 \tag{27}$$

which states that the *log likelihood ratio* estimated using the train data is the best *log likelihood ratio* we can get. We have used this assumption to normalize the $LLR_s^{M_c^\ell}$ and $LLR_u^{M_c^\ell}$. Given a customer model (M_c^ℓ, λ_c), the $NLLR_s^{M_c^\ell}$ can be defined as:

$$NLLR_s^{M_c^\ell} = \frac{LLR_s^{M_c^\ell}}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{T_i}\log\left[\frac{p(X_c^i|M_c^\ell,\lambda_c)}{p(X_c^i|M_c^\ell,\lambda)}\right]} \tag{28}$$

and $NLLR_u^{M_c^\ell}$ as:

$$NLLR_u^{M_c^\ell} = \frac{LLR_u^{M_c^\ell}}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{T_i}\log\left[\frac{p(X_c^i|M_c^\ell,\lambda_c)}{p(X_c^i|\Lambda)}\right]} \tag{29}$$

where I is the number of enrollment utterances for the speaker S_c. The denominators in (28) and (29) are the average *log likelihood ratio* estimated over all the enrollment data.

The new confidence score cs_ℓ in (25) will be estimated as:

$$cs_\ell = \alpha\ NLLR_s^{M_c^\ell} + (1-\alpha)\ NLLR_u^{M_c^\ell} \tag{30}$$

Using (28) and (29), together with the assumption (27):

- The $NLLR_u^{M_c^\ell}$ and $NLLR_s^{M_c^\ell}$ will have, theoretically, a limited dynamic range with an upper bound equal to 1. Consequently, the cs_ℓ in (30) will be bounded by 1. The value $NLLR_s^{M_c^\ell}$ and $NLLR_u^{M_c^\ell}$ indicate how probable the test utterance belongs to the claimed identity. Closer are $NLLR_u^{M_c^\ell}$ and $NLLR_s^{M_c^\ell}$ to 1, more probable the claimed identity is to be valid.
- The search for a local speaker and subsystem independent threshold δ will be in a fixed range $[0, 1]$. So, depending on the application requirements, we can adjust the threshold without difficulty.

Note that in this approach we now have two thresholds, a local threshold δ and a global threshold Δ.

8 Databases and Experimental Set-up

In this work, we have used two databases, the *PolyPhone* [15] database to train different speaker-independent models, and the *PolyVar* [15] database to perform customer enrollment and verification test.

8.1 PolyPhone database

The Swiss-French PolyPhone databases [15] contains telephone calls from about $4,500$ speakers recorded over the Swiss telephone network. The calling sheets were made up of 38 prompted items and questions. Among other items, each speaker was invited to read 10 sentences selected from different corpora to ensure good phonetic coverage for the resulting database. Different kinds of irregularities (i.e; noise in the recording, strange utterances) were discovered, and the training set was finally reduced to $3,272$ sentences, corresponding to approximately 5 hours of speech.

8.2 PolyVar database

The PolyVar database was recorded and designed at IDIAP as a complement to the PolyPhone database to address the intra-speaker variability. This database comprises telephone recordings from about 143 speakers (85 male and 58 female speakers). Each speaker recorded between 1 and 229 sessions. Several speakers pronounced the same set of 17 words several times, which makes this database particularly well suited to test UCP-SV systems. i.e.,

- Assigning each of the words to one specific customer
- Providing enrollment utterances for each of those words, test utterances, as well as many impostor utterances pronouncing the right password.
- Providing several utterances associated with words different than the chosen password, from both the customer and potential impostors.

A set of 38 speakers (24 males and 14 female) who have more than 26 sessions were selected. The set of 17 words is divided into *data1* and *data2* with 14 and 3 words, respectively. For each speaker and each word in *data1*, the first 5 utterances (corresponding to the first 5 sessions) of the word are used for training, to create the customer-dependent password HMM. For testing, two protocols are defined:

1. *Protocol P1:*
   In this protocol (summarized in Table 1), between 18 and 22 utterances of the same word are used as a customer accesses with the expected password. Each speaker has a subset of 18 speakers as impostors (11 males and 7 females if the user is a male and 6 females and 12 males if the customer is a female). Each impostor has two accesses with the expected password. Each speaker, including customers and impostors, has 3 accesses with three different words taken from *data2* to simulate the case where speaker pronounces wrong password.

Type of access	#nb of accesses
Training	5
Testing: C-EP	18-22
Testing: C-IP	3
Testing: I-EP	36
Testing: I-IP	54

Table 1: *Distribution of customer (C) and impostor (I) accesses with expected password (EP) and invalid passwords (IP).*

2. *Protocol P2:*
   To evaluate our approach on more difficult conditions, a second protocol $P2$ where customers and impostors pronounce only the expected password was defined. There were 12930 customers' accesses and 23256 impostors' accesses.

8.3 Experimental set-up

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 every 10 ms.

The approach studied here assumes the availability of some *a priori* speaker-independent acoustic models for HMM inference, speaker adaptation and score normalization. Three speaker-independent speech recognizers are trained using *PolyPhone* databases:

1. *A Hybrid HMM/MLP system:* The speaker independent MLP (SI-MLP) used for HMM inference consisted of 234 input units with 9 consecutive 26 dimensional acoustic vectors, 600 hidden units and 36 outputs, such that each output is associated with a specific phone. The phone level accuracy obtained by this system on *PolyVar* using customer enrollment data is 56.6%.

2. *A Hidden Markov model* with a set of parameter Λ is trained using the segmental K-means algorithm [16] followed by EM algorithm. 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 is used as a priori distribution for *maximum a posteriori* (MAP) adaptation.

3. *A Gaussian mixture model* with a set of parameters Λ is modeled by 240 (diagonal covariance) Gaussian and trained using the segmental K-means algorithm followed by EM algorithm. This GMM is used for utterance verification score normalization.

9 Experiments

All experiments described here were conducted using the Torch library [17]. For comparison purposes, results for a SV system uses the correct phonetic transcription of the password given by a dictionary are also reported. This will be referred to as TD-SV system. The combined parameter α as well as the speaker-independent decision threshold are determined *a posteriori* to optimize the Equal Error Rate (EER). This is not realistic, but it gives a good way to evaluate the discrimination capabilities of the modeling approach. Note that for the second protocol $P2$, the evaluation is made using only the speaker verification part based on LLR_s.

9.1 Single reference approach evaluation

The goal of this experiment is to evaluate and analyze the performance of the UCP-SV system using single reference approach. This system will be referred to as baseline system. The obtained EER is compared to that obtained by a TD-SV system.

Figure 2 shows the EER variations of the UCP-SV and the TD-SV systems as a function of the combined parameter α using the first protocol $P1$. Table 2 shows the EER for both systems using $P1$ and $P2$. It is clear that the use of *a priori* information about the lexical content of the password helps in improving the verification performance of a TD-SV system. This performance becomes significant as the parameter α increases.

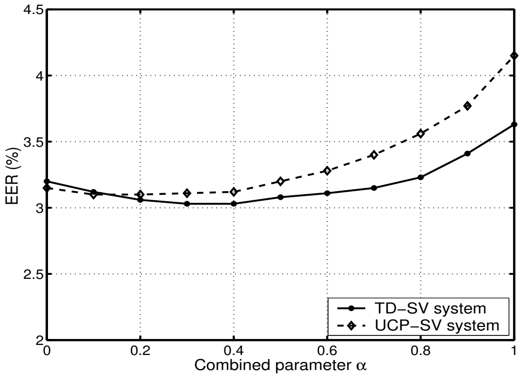

Figure 2: *Equal error rate variations of the reference TD-SV and the baseline UCP-SV systems as a function of the combined parameter α using the first protocol $P1$.*

Protocol	System	LLR_s $\alpha = 1$	LLR_u $\alpha = 0$	EER(α)
$P1$	TD-SV	3.6	3.2	3.0 (0.3)
	UCP-SV	4.2	3.2	3.1 (0.2)
$P2$	TD-SV	-	-	5.3
	UCP-SV	-	-	5.8

Table 2: *EER of the UCP-SV and TD-SV systems using both first ($P1$) and second ($P2$) protocols.*

9.1.1 Analysis

There are two informative values that can help us to analyze the results. These values correspond to the performance of the two systems for $\alpha = 0$ and $\alpha = 1$.

- $\alpha = 0$: The performance of both TD-SV and UCP-SV systems using the combined confidence score (9) becomes equal to the performance using only the *utterance verification* part (LLR_u). In this case, the TD-SV and the UCP-SV systems have the same world model (GMM) for likelihood normalization, but they use a customer HMM model created from two different PTs. So, if one of these systems performs better than the other, this should be attributed to the customer HMM model. The EERs associated with $\alpha = 0$ show comparable performance for both TD-SV and UCP-SV systems. This indicates that the improvement of the TD-SV system cannot be completely attributed to the fact that this system uses the correct PT to create the customer HMM model while the UCP-SV uses the inferred PT.

- $\alpha = 1$: The performance of both TD-SV and UCP-SV systems becomes equal to the performance using only the *speaker verification* part (LLR_s). Both UCP-SV and TD-SV systems have two different customer HMM models and two different background models. Within the same system, the customer and the background models have the same topology (i.e.; the same states and the same connections between states). In this case, if one system performs better than the other, this improvement can be attributed either to the customer HMM model or to the *background* model. As we have seen in the case of $\alpha = 0$, the customer model performs comparably in both TD-SV and UCP-SV systems. Hence, the improvement in the TD-SV system is in great part due to the background model which -in the case of TD-SV system- is more competitive than the one used in the UCP-SV system. This explains why the difference between the two EERs obtained by the TD-SV and the UCP-SV systems increases as the weight given to the *speaker verification* part increases and why the TD-SV system performs better than the UCP-SV system using $P2$.

This is consistent with what has been found in [8]. One possible explanation is that the *background* model should cover as much as possible the acoustic space of how other speakers pronounce the expected password, and not only how a specific speaker (customer) pronounces it.

9.2 Multiple reference approach evaluation

9.2.1 Dynamic Model Selection techniques

Tables 3 reports EERs of the *speaker verification* part (2 column), the *utterance verification* part (3 column) and the UCP-SV system using DMS criteria and $P1$. Table 4 reports EERs of the UCP-SV system using DMS criteria and $P2$.

DMS criterion	CS_s	CS_u	EER(α)
Baseline UCP-SV	4.2	3.2	3.1 (0.2)
Max $p(X\|M_c^\ell, \lambda_c)$	5.0	3.3	3.3 (0.1)
Max $p(X\|M_c^\ell, \lambda)$	4.5	3.3	3.3 (0.2)
Min $LLR_s^{M_c^\ell}$	3.5	3.3	3.1 (0.5)

Table 3: *EER of the UCP-SV system using $P1$ with different dynamic model selection criteria. The second raw corresponds to the EER of the UCP-SV Baseline system using single reference model.*

DMS criterion	*EER*
Baseline UCP-SV	5.8
Max $p(X\|M_c^\ell, \lambda_c)$	6.2
Max $p(X\|M_c^\ell, \lambda)$	6.0
Min $LLR_s^{M_c^\ell}$	5.5

Table 4: *EER of the UCP-SV system using P2 with different dynamic model selection criteria, The second raw corresponds to the EER of the UCP-SV Baseline system using single reference model.*

9.2.2 Discussion

Several observations can be made from these results:

1. *Second protocol evaluation:*

   - The performance using the *background* model associated with the best customer model (17) is worse than that obtained with the baseline system. A possible reason is that the $(\widehat{M}_c^\ell, \lambda_c)$ is selected dynamically according to the maximum likelihood criterion. For many impostors, the alignment of the test utterance against $(\widehat{M}_c^\ell, \lambda_c)$ results in a good likelihood score, and because $(\widehat{M}_c^\ell, \lambda)$ is not necessarily an appropriate *background* model, many impostor accesses will get accepted[8].

- The selection of $(\widehat{M}_c^\ell, \lambda_c)$ and $(\widehat{M}_c^{\ell'}, \lambda)$ separately, according to the maximum likelihood criterion (15) and (18), improved the performance compared to the use of (17). But the obtained performance is still worse than the baseline system. This indicates that the appropriate selection criterion might well applied to the $LLR_s^{M_c^\ell}$.
- Significant improvement is obtained using the $Minimum\ LLR_s^{M_c^\ell}$ as a selection criterion (20). The EER dropped from 5.8% to 5.5%. As we can see, the performance of the UCP-SV system is quite competitive with the TD-SV system. We should mentioned here, that the use of (20) might not be optimal and depends on the experimental set-up, making the selection of the optimal model not obvious [18].

2. *First protocol evaluation:*
   - The use of (15) to select the customer HMM model did not improve the performance of the *utterance verification* part. Taking into account our acoustic modeling approach, it seems that the value of 3.2% is the best we can achieve.
   - Surprisingly, and despite the significant improvement in the *speaker verification* part (see Table 3, column 2), no improvement in the performance of the UCP-SV system is obtained (column 4). A possible reason is that the GMM world model (Λ) is trained with general speech data from a large set of speakers. It covers the general acoustic space including the customer password. Hence, it has some acoustic characteristics of the *background* model (M_c^ℓ, λ), making the amount of new (complementary) information given by the *speaker verification* part very low. The *correlation coefficients* between CS_u and CS_s for customer and impostor accesses were found to be 0.90 and 0.80, respectively. This indicates that these two scores are highly correlated.

9.2.3 Confidence scores fusion

Figure 3 shows the EER variations of the UCP-SV system using (21) and (22) with the first protocol $P1$ as a function of the combined parameter α. Results of the reference TD-SV and the baseline UCP-SV systems are also shown. Table 5 reports the EER of the UCP-SV system using $P1$ and $P2$.
It can be seen from Table 5 that the use of average confidence score criterion improves the UCP-SV

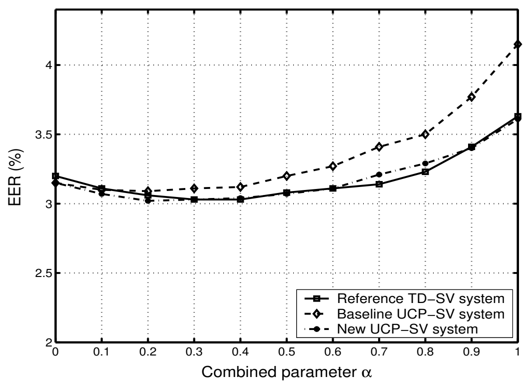

Figure 3: *ERR variations of the UCP-SV systems as a function of the combined parameter α using the confidence score fusion technique (21), (22) and (23) with the first protocol $P1$. Results of the reference TD-SV and the Baseline UCP-SV systems are also shown.*

[8]It is worth mentioning that an appropriate background model is useful in reducing the false acceptance rate.

Protocol	CS_s	CS_u	EER(α)
$P1$	3.6	3.2	3.0(0.2)
$P2$	-	-	5.6

Table 5: *EER of the UCP-SV system using confidence score fusion technique (21), (22) and (23) with* $P1$ *and* $P2$*. For* $P2$*, only* CS_s *are used.*

performance. Figure 3 shows that both systems (i.e., the UCP-SV and the TD-SV) have the same performance for all the values of the parameter α. Also, It can be noted that the use of average score criterion gives comparable results with those obtained using the best dynamic model selection criteria, i.e., the *min* function (20) for CS_s estimation and the maximum likelihood criterion (17) for CS_u estimation. This indicates that (20) and (17) are a good criteria.

We should note here, that, for a given customer, the verification score estimated by each subsystem are not statistically independent. Indeed, all subsystems are trained using the same adaptation data and the same adaptation procedure, only phonetic transcriptions are different. Consequently, given a test utterance X, there is a set of *optimal* parameters corresponding to only one customer HMM model (M_c^ℓ, λ_c) that gives the best $LLR_u^{M_c^\ell}$ and a set of *optimal* parameters corresponding to only one phonetic transcription $M_c^{\ell'}$ that gives the best $LLR_s^{M_c^{\ell'}}$. The combination of these two scores will give the best performance. The use of the other models will be useless as they do not carry any complementary information. Because the search for these *optimal* models is not obvious, using the average score will prevent us from the choose of the poor parameters.

9.2.4 Partial decisions fusion

Table 6 shows the EER of the UCP-SV for both protocols. The global threshold Δ is set to 0.6. This

Protocol	*Loc.thrd* δ	*Glob.thrd* Δ	EER(α)
$P1$	0.28	0.6	3.1 (0.2)
$P2$	0.25	0.6	5.6

Table 6: *EER of the UCP-SV system with its optimal local and global thresholds using partial decision fusion technique (24) with the first and second protocols.*

is mean that the speaker is accepted if $3/5-th$ of local confidence scores exceeded the local threshold δ. We can see that we have got an improvement in the performance compared to the baseline system, but not as significant as in the two previous techniques. For comparison purposes, we also used the original $LLR_u^{M_c^\ell}$ and $LLR_s^{M_c^\ell}$ to estimate the combined score, and we have got the same performance with $P1$ and $P2$. The advantage of the NLLR, however, is that the threshold has a meaningful interpretation and is easily adjustable according to the application requirements. Another advantage is that the NLLR can be used as a criterion to select the utterance test that has a high NLLR for incremental adaptation.

9.3 GMM-UBM approach

It can be argued that when a GMM is trained with a limited data that covers a few phonemes (actually this is our case), the GMM becomes speaker and text dependent. For the sake of comparison, we have conducted an experiment using the conventional Gaussian Mixture Model-Universal Background Model (GMM-UBM) text-independent speaker verification approach [19], where a speaker-independent GMM referred to as Universal Background Model (UBM) is used as *a priori* distribution for speaker adaptation. In the approach tested here, the enrollment step consists of using the 5 repetitions of the customer's password for adaptation using MAP adaptation. The UBM consists of 120 mixtures trained on *PolyPhone* database. For testing, the usual *log likelihood ratio* test is used.

Protocol	EER
$P1$	3.5
$P2$	5.5

Table 7: *EER of the UCP-SV system using GMM-UBM approach on both protocols* $P1$ *and* $P2$

Table 7 reports the obtained results on both protocols $P1$ and $P2$. Compared to the multiple reference approach, the GMM-UBM approach performs comparably on $P2$ but poorly on $P1$. A possible reason is that, the GMM-UBM system does only speaker verification. In contract, the HMM/GMM system performs both utterance and speaker verification. Because the protocol $P2$ contains only accesses with expected password, we are only interested in verifying the claimed identity, hence, both systems become equivalent. However, the protocol $P1$ contains some customer's accesses with invalid passwords. Hence, there could be an overlap in the acoustic space between the customer password and customer test accesses with invalid password. In the case of GMM-UBM system, this might result in a good *log likelihood ratio*. Consequently, the customer gets accepted even if the pronounced word is not correct, which is not the case in HMM/GMM approach.

10 Conclusion

This paper has developed and compared HMM/GMM based User-customized password speaker verification (UCP-SV) systems using single reference and multiple reference approaches. A speaker-independent HMM/MLP is used to infer the phonetic transcriptions associated with the enrollment utterances. These phonetic transcriptions are then used to create customer-dependent password HMMs and background models.

First, a UCP-SV system using single reference approach is developed. This system used the best phonetic transcription determined during the enrollment step to create the background model which is then adapted towards customer voice's characteristics. This system achieved acceptable performances but not competitive with TD-SV system using the correct phonetic transcription of the password. Our analysis has revealed that the main reason of this limitation lies in the background model.

Second, to improve the performance of the baseline UCP-SV system, we have investigated the use of multiple reference approach. Different scoring criteria are proposed and evaluated. Results showed that significant improvement could be achieved if an appropriate selection criterion of the customer and background models is used. However, comparable improvement could be obtained by taking the average log likelihood ratios estimated by each subsystem or fusing partial decisions made by each subsystem.

Finally, a comparative experiments using the conventional GMM-UBM text-independent speaker verification is conducted. Results showed that under certain conditions, the GMM-UBM approach performs comparably with multiple reference approach. This indicates that when a GMM is trained (adapted) with data associated with a short password, the GMM becomes text and speaker dependent, although the temporal phonetic structure is not preserved.

11 Acknowledgment

The authors gratefully acknowledge the support of the Swiss National Science Foundation through 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] S. Furui, "An Overview of Speaker Recognition Technology" *ESCA workshop on Automatic speaker Recognition, Identification and Verifica tion*, pp. 1-9, 1994, Martigny, Switzerland

[2] J. de Veth, G. Gallopyn, and H. Bourlard, "Limited Parameter Hidden Markov Models for Random Connected Digit Speaker Verification over Telephone Channels," *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (Minneapolis, MI)*, pp. II:247-250, 1993.

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

[4] M.F. BenZeghiba, H. Bourlard and J. Mariethoz "Speaker Verification based on User-Customized Password " *IDIAP Research report*, IDIAP-RR-32, 2001.

[5] J. Neto and L. Almeida and M. hochberg and C. Martins and L. Nunes and S. Renals and T. Robinson, "Speaker-Adaptation for Hybrid HMM/ANN Continous Speech Recognitio n System", *Proceeding of Eurospeech'95*, pp. 2171-2174, 1995.

[6] 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.

[7] A. E. Rosenberg and P. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Verification", *Proc. of Inter. Conf. on Speech and Signal Processing*,pp. 81-84, 1996.

[8] A. E. Rosenberg and P. Parthasarathy, "Speaker Identification with User-Selected Password Phrases", *In Proceedings EUROSPEECH'97*, pp. 1371-1374, 1997.

[9] H. Bourlard, Y. Kamp, H. Ney and C. J Wellekens, "Speaker-Dependent Connected Speech Recognition via Dynamic Programing and Stat istical Methods", *In Speech and Speaker Recognition*, Vol. 2, pp. 115-148, 1985, Karger,Basel .

[10] L. Rodriguez-Linres and C. Garcia-Mateo and J. L. Alba-Castro, "On Combining Classifiers for Speaker Authentication" *Pattern Recognition*, Vol. 36(2) pp. 347-359, 2003.

[11] M. Hebert and S. D. Peters, "Improved Normalization Without Recourse To An Impostor Database For Speaker Ve rification", *In Proceedings EUROSPEECH'01*, pp. 2557-2560, 2001.

[12] O. Siohan and C. -H. Lee and A. C. Surendran and Q. Li, "Background Model Design For Flexible and Portable Speaker Verification", *Proc. of Inter. Conf. on Speech and Signal Processing*,pp. 825-828, 1999.

[13] A. Viterbi, "Error Bounds For Convolutional Codes and Asymptotically Optimum Decoding Algor ithm", *IEEE trans. on Information Theory*,pp. 260-269, 1967

[14] 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.

[15] 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.

[16] L. R. Rabiner, J. G. Wilpon and B.-H. Juang, "A Segmental K-means Training Procedure for Connected Word Recognition, *AT&T Bell Labs Tech. J*, pp. 21-31, 1986

[17] R. Collobert, S. Bengio and J. Mariethoz, "Torch: A Modular Machine Learning Software Library" *Technical Report* IDIAP-RR 02-46, 2002.

[18] M. F. BenZeghiba and H. Bourlard, "Confidence Measures In Multiple Pronunciations Modeling for Speaker Verificati on" *Proc. of Inter. Conf. on Speech and Signal Processing*, pp 103-106, 2004, M ontreal, Canada

[19] D. A. Reynolds and T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, pp. 19-41, 2000.