Unsupervised intra-speaker variability compensation based on Gestalt and model adaptation in speaker verification with telephone speech

Nestor Becerra Yoma, Claudio Garretón, Carlos Molina, Fernando Huenupán

Speech Processing and Transmission Laboratory

Department of Electrical Engineering

Universidad de Chile, Santiago, Chile

nbecerra@ing.uchile.cl

Telephone: +56-2-678 4205   Fax: +56-2-695 3881

Abstract

In this paper an unsupervised compensation method based on Gestalt, ISVC, is proposed to address the problem of limited enrolling data and noise robustness in text-dependent speaker verification (SV). Reductions in EER and in the integral below the ROC curve as high as 20% or 40% and 30% or 60%, respectively, can be achieved by ISVC independently of the number of enrolling utterances. In contrast to model adaptation methods, ISVC is memoryless with respect to previous verification attempts. As shown here, unsupervised model adaptation can lead to substantial improvements in EER but is highly dependent on the sequence of client/impostor verification events. In adverse scenarios, such as massive impostor attacks and verification from alternated telephone line, unsupervised model adaptation might even provide reductions in verification accuracy when compared with the baseline system. In those cases, ISVC can even outperform adaptation schemes. It is worth emphasizing that ISVC and unsupervised model adaptation are compatible and the combination of both methods always improves the performance of model adaptation. The combination of both schemes can lead to improvements in EER as high as 33.7%. Due to the restrictions of commercially available databases for text-dependent SV research, the results presented here are based on local databases in Spanish. By doing so, the visibility of research in Iberian Languages is highlighted.

I. Introduction

From the usability point of view, the enrolling procedure in a speaker verification (SV) system over the telephone network should be fast and efficient. However, limited enrolling data leads to poorly trained models, which in turn seriously degrades the accuracy of SV engines. Moreover, additive and convolutional noise is usually one of the most important problems faced by speech and speaker recognition systems in real applications. Several noise canceling technique have been proposed to handle additive and convolutional noise (Hardt and Fellbaum, 1997; Ortega-Garcia and Gonzalez-Rodriguez, 1996; Yiu et. al., 2007). These noise cancellation techniques can substantially reduce the mismatch between training and testing conditions as far as additive and convolution distortion is concerned. However, they do not improve the generalization ability of trained models from the intra-speaker variability point of view.

The limited enrolling data problem in SV has been addressed by several authors using HMM adaptation methods. Those techniques adapt HMM parameters employing speech data that is input by the user in verification events after enrolling. The HMM parameters are usually adapted by mean of applying techniques based on Bayesian Maximum a Posteriori, MAP, (Barras et. al., 2004; Gauvain and Lee, 1994; Yu and Mason, 1996) adaptation and Maximum Likelihood Linear Regression, MLLR, (Ahn and Ko, 2000; Leggetter and Woodland, 1997). Those methods are classified as supervised or unsupervised depending on the requirement of human assistance to transcribe and label the adaptation data. Supervised adaptation techniques, although more effective than unsupervised approaches, are impractical on largescale SV based services. On the other hand, the unsupervised classification of adaptation data can introduce an error in the HHM parameter re-estimation procedure, which in turn will be propagated into further verification events. In that sense, the time order of client and impostors has a direct effect on the performance of unsupervised adaptation schemes.

In this paper intra-speaker variability compensation (ISVC) inspired on the Gestalt principle of shape perception (Wertheomer, M., 1950) is proposed to reduce the distortion between verification signals and the client HMM. Instead of adapting the client HMM, the approach described here modifies the verification signals using MAP estimation. Gestalt principle is one of the first studies proposed by a group philosophers and psychologists to model shape and pattern perception in the 19th century. According to Gestalt, the whole is prioritized and the parts are understood within the systematic whole (Wertheomer, M., 1950). Gestalt principle has recently been employed in the field of image processing (Gao et. al., 2005; Jiang, 2006; Kim and Kweon, 2006.) but, surprisingly, it has not been exhaustively applied to speech or speaker recognition method. In contrast to Gestalt theory, current pattern recognition algorithms in speech tend to process every signal segment with the same priority. The proposed intra-speaker variability compensation (ISVC) based on Gestalt model attempts to modify the input signal by reducing not relevant differences between testing signal and reference pattern or model (see Fig. 1), if those differences are low and comparable to intra-speaker variability. By doing so, it is expected to force the pattern recognition algorithm to focus on the most relevant features of the signal input. Due to the fact that the client HMM is not modified, the error caused by misclassification of adaptation data is avoided. Moreover, the proposed compensation scheme also leads to a noise removal effect. In order to contrast the results obtained with the compensation scheme proposed in this paper, a conventional unsupervised model adaptation strategy based on MAP is evaluated, compared and combined with compensation scheme presented here. The contributions of this paper concern: a) an unsupervised intra-speaker variability compensation (ISVC) method for SV; b) a comparison of ISVC with unsupervised HMM adaptation; and, c) combination of ISVC with model adaptation. Experiments with telephone speech in matched and unmatched conditions suggest that: ISVC alone can lead to reductions in EER and in the area of FA/FR ROC curve as high as 20% or 40% and 30% or 60%, respectively; in contrast to model adaptation, ISVC does not depend on the sequence client-impostor signals that are verified and may give similar or superior reductions in EER in some cases; and, in combination with unsupervised HMM adaptation, ISVC provided reductions of 34% and 44% in EER and in the area of FA/FR ROC curve, respectively. Observe that the strategy followed in this paper corresponds to compare ISVC with unsupervised model adaptation philosophy instead of comparing ISVC with a specific unsupervised adaptation scheme. It is worth emphasizing that, due to the restrictions of commercially available databases for text-dependent SV research, local databases in Spanish had to be recorded to obtain the results presented here. Consequently, the visibility of research in Iberian Languages is highlighted by this research. Finally, the approach and analysis presented in this paper have not been found in the specialized literature.

II. Intra-speaker variability modeling

In the text-dependent SV task considered here, each utterance is processed with the forced-Viterbi algorithm in order to estimate the normalized log likelihood, $\log L(O)$ (Furui, 1997):

$$\log L(O) = \log P(O/\lambda_{SD}) - \log P(O/\lambda_{SI}) \qquad (1)$$

where O is the observation sequence; and, $P(O/\lambda_{SD})$ and $P(O/\lambda_{SI})$ represent the likelihood related to the speaker dependent (λ_{SD}) and independent (λ_{SI}) models, respectively. Both models, λ_{SD} and λ_{SI}, correspond to the sequence of triphone HMM's that compose the testing sequence O. In order to estimate the false-rejection and false-acceptance error curves, the normalized log likelihood $\log L(O)$ is divided by the number of frames (*T*) in the verification utterance: $\log L(O)' = \frac{\log L(O)}{T}$. It is worth highlighting that

λ_{SD} is computed with the enrolling data pronounced by the client, and λ_{SI} is estimated with a set of impostors. In this paper one multivariate Gaussian density per state was employed in λ_{SD}.

Given a state *s* in λ_{SD} and the enrolling data, the intra-speaker variability is modeled in this paper as the vector $d(t) = \left[d(t,0), d(t,1), ..., d(t,n), ..., d(t,N-1)\right]$ where:

$$d(t,n) = \left|D(t,n)\right| \tag{2}$$

and $D(t,n) = \mu_{s(t),n} - O(t,n)$, $\mu_{s(t),n}$ is the n^{th} component in the mean vector of the observation probability function in state *s* that was allocated to frame $O(t) = \left[O(t,0), O(t,1), ..., O(t,n), ..., O(t,N-1)\right]$ as a result of the forced Viterbi alignment; and *N* is the number of parameters. This alignment associates a state within the HMM sequence to every frame. As a consequence, the state allocated to frame *O(t)* is denoted by *s(t).*

In order to estimate the p.d.f. of the intra-speaker variability, the histogram of *d(t,n)* was obtained by using enrolling utterances from an evaluation database composed of 13 speakers after training the speaker dependent HMM's. In this paper the intra-speaker variability is considered state and speaker independent. Examples of the resulted histograms are shown in Fig. 1. As can be seen in Fig.1, the p.d.f. of *d(t,n),* $f\left[d(n)\right]$, can be modeled with a gamma distribution (Rao, 1965):

$$f\left[d(n)\right] = A \cdot \exp\left(-\alpha(n) \cdot d(n)\right) \cdot d(n)^{p(n)-1} \tag{3}$$

where $\alpha(n) = \frac{E[d(n)]}{Var[d(n)]}$ *;* $p(n) = \frac{E[d(n)]^2}{Var[d(n)]}$; *A* is a normalizing term; and, $E[d(n)]$ and $Var[d(n)]$ are the mean and variance of the histogram of *d(n),* respectively. To simplify the notation, the argument *t* was withdrawn from *d(t,n)* in (3).

III. Intra-speaker Variability Compensation (ISVC)

ISVC aims to modify the input observation by reducing not relevant differences between test utterance and client HMM if those differences are low and comparable to intra-speaker variability. ISVC is inspired in Gestalt principle and is graphically illustrated in Fig.1. As can be seen in Fig.1, a noisy input signal that corresponds to a distorted triangle is compared with two reference templates: a triangle and a square. When compared with the triangle, the distance (black regions) between the input signal and the triangle is reduced because this difference is low. In contrast, the distance between the input signal and the square keeps unchanged because this difference is comparatively high. By adopting this procedure, the matching algorithm could focus its decision by analyzing the most relevant parts of the input signal (e.g. its vertices).

If $\tilde{O}(t,n)$ and $O(t,n)$ denote the n^{th} feature in the compensated and observed frames, respectively, the compensation is expressed with:

$$\tilde{O}(t,n) \ = \ O(t,n) + \left[\Delta O(t,n)\right]^{optimal} \tag{4}$$

where $\left[\Delta O(t,n)\right]^{optimal}$ is the correction component at instant *t*.

$\left[\Delta O(t,n)\right]^{optimal}$ is modeled here as a fraction of the multivariate vector difference between $O(t)$ and $\mu_{s(t)}$:

$$\left[\Delta O(t,n)\right]^{optimal} = D(t,n) \cdot \left[K(t,n)\right]^{optimal} \tag{5}$$

where $\left[K(t,n)\right]^{optimal}$ represents the optimal fraction of difference $D(t,n)$. A graphical comparison of ISVC with a model adaptation approach can be seen in Fig. 2. The compensation component $\left[\Delta O(t,n)\right]^{optimal}$ is estimated by maximizing the a posteriori p.d.f. $\Pr\left[\mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n) \middle| O(t,n), s(t)\right]$, where the difference $\mu_{s(t),n} - \tilde{O}(t,n)$ is equivalent to $\mu_{s(t),n} - \left(\Delta O(t,n) + O(t,n)\right) = D(t,n) - \Delta O(t,n)$. This term defines the optimal distance between the adapted observation vector $\tilde{O}(t,n)$ and mean vector $\mu_{s(t),n}$, given a state $s(t)$. By using the Bayes theorem, the maximization can be expressed as:

$$\begin{aligned}\left[\Delta O(t,n)\right]^{optimal} &= \arg\max_{\Delta O(t,n)} \left\{ \Pr\left[\mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n) \middle| O(t,n), s(t)\right]\right\} \\ &= \arg\max_{\Delta O(t,n)} \left\{ \frac{\Pr\left[O(t,n) \mid \mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n), s(t)\right] \cdot \Pr\left[\mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n) \middle| s(t)\right]}{\Pr\left[O(t,n) \middle| s(t)\right]} \right\}\end{aligned} \tag{6}$$

In (6), $\Pr\left[O(t,n) \middle| s(t)\right]$, the observation probability, does not depend on $\mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n)$. By defining $\tilde{\mu}_{s(t),n} = \mu_{s(t),n} - \Delta O(t,n)$, it is possible to say that $\tilde{\mu}_{s(t),n} - O(t,n) = \mu_{s(t),n} - \Delta O(t,n) - O(t,n) = \mu_{s(t),n} - \tilde{O}(t,n)$. If diagonal covariance matrix is considered, the single Gaussian observation probability corresponds to:

$$\Pr\left[O(t,n) \mid \mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n), s(t)\right] = \prod_{n=1}^{N} \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2_{s(t),n}}} \cdot e^{-\frac{1}{2} \cdot \frac{\left(\mu_{s(t),n} - O(t,n) - \Delta O(t,n)\right)^2}{\sigma^2_{s(t),n}}} \tag{7}$$

where N is the number of feature parameters; and $\sigma^2_{s(t),n}$ is the n^{th} component in the variance vector of the observation probability in state *s(t)*. Notice that $\tilde{\mu}_{s(t),n}$ denotes the mean associated to the n^{th} feature in the mean of the adapted observation probability in state *s(t)* if model adaptation took place. As can be seen in (7), replacing $\mu_{s(t),n} - \Delta O(t,n)$ with $\tilde{\mu}_{s(t),n}$ is equivalent to evaluate the current observation probability with the observation vector modified by $\Delta O(t,n)$. Then, $\Pr\left[O(t,n) \mid \mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n), s(t)\right] = \Pr\left[O(t,n) + \Delta O(t,n) \mid \mu_{s(t),n}, s(t)\right]$.

Consequently, $\Pr\left[O(t,n) \middle| \mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n), s(t)\right]$ is also equivalent to $\Pr\left[O(t,n) \middle| \tilde{\mu}_{s(t),n} = \mu_{s(t),n} - \Delta O(t,n), s(t)\right]$. Then, $\Pr\left[O(t,n) \middle| \mu_{s(t),n} - \tilde{O}(t,n) = D(t,n) - \Delta O(t,n), s(t)\right]$ can be written as $\Pr\left[O(t,n) + \Delta O(t,n) \middle| s(t)\right]$. Moreover, $D(t,n) - \Delta O(t,n)$ is modeled with $\left|D(t,n) - \Delta O(t,n)\right|$ in

$\Pr\left[D(t,n)-\Delta O(t,n)\middle|s(t)\right]$, which in turn is supposed independent of *s(t)* and is replaced with $f\left[\left|D(t,n)-\Delta O(t,n)\right|\right]$ that is described with a gamma p.d.f. as indicated in (3). Then, the optimization in (6) is reduced to:

$$\left[\Delta O(t,n)\right]^{optimal}=\underset{\Delta O(t,n)}{\arg\max}\left\{f\left[\left|D(t,n)-\Delta O(t,n)\right|\right]\cdot\Pr\left[\tilde{O}(t,n)\middle|s(t)\right]\right\} \tag{8}$$

Replacing $\Delta O(t,n)$ with $K(t,n)\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]$ as shown in (5), the maximization expression (8) is equivalent to:

$$\left[K(t,n)\right]^{optimal}=\underset{K(t,n)}{\arg\max}\left\{f\left[\left|\left(1-K(t,n)\right)\cdot\left(\mu_{s(t),n}\text{ - }O(t,n)\right)\right|\right]\cdot\Pr\left[\tilde{O}(t,n)\middle|s(t)\right]\right\} \tag{9}$$

As mentioned above, the speaker-dependent observation probability $\Pr\left[\tilde{O}(t,n)\middle|s(t)\right]$ is modeled with a single Gaussian with diagonal covariance matrices, then (9) can be rewritten as:

$$\left[K(t,n)\right]^{optimal}=\underset{K(t,n)}{\arg\max}\left\{\begin{array}{l}A(n)\cdot\left(\left|\left[1-K(t,n)\right]\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]\right|\right)^{p(n)-1}\cdot\exp\left(-\alpha(n)\cdot\left[\left|\left[1-K(t,n)\right]\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]\right|\right]\right)\\ \cdot\exp\left[-\dfrac{\left(O(t,n)+K(t,n)\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]-\mu_{s(t),n}\right)^{2}}{2\cdot\sigma_{s(t),n}^{2}}\right]\end{array}\right\} \tag{10}$$

In the Log domain, equation (10) can be expressed as:

$$\left[K(t,n)\right]^{optimal}=\underset{K(t,n)}{\arg\max}\left\{\begin{array}{l}\log\left[A(n)\right]+\left(p(n)-1\right)\cdot\log\left(\left|\left[1-K(t,n)\right]\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]\right|\right)\\ -\alpha(n)\cdot\left(\left|\left[1-K(t,n)\right]\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]\right|\right)-\dfrac{\left(O(t,n)+K(t,n)\cdot\left[\mu_{s(t),n}\text{ - }O(t,n)\right]-\mu_{s(t),n}\right)^{2}}{2\cdot\sigma_{s(t),n}^{2}}\end{array}\right\} \tag{11}$$

Computing the partial derivate with respect to *K(t,n)* and setting it to zero:

$$\left[1-K(t,n)\right]\cdot\frac{\left(\mu_{s(t),n}-O(t,n)\right)^{2}}{\sigma_{s(t),n}^{2}}+\alpha(n)\cdot\left[\left|\mu_{s(t),n}-O(t,n)\right|\right]-\frac{p(n)-1}{1-K(t,n)}=0 \tag{12}$$

This quadratic equation provides two solutions:

$$\left[K(t,n)\right]^{optimal} = 1 - \frac{1}{2} \cdot \left(\frac{-\alpha(n) \cdot \left(\left| \mu_{s(t),n} - O(t,n) \right| \right)}{\Omega(t,n)} \pm \sqrt{ \left[\frac{\alpha(n) \cdot \left(\left| \mu_{s(t),n} - O(t,n) \right| \right)}{\Omega(t,n)} \right]^2 + \frac{4 \cdot \left(p(n) - 1 \right)}{\Omega(t,n)} } \right) \quad (13)$$

where $\Omega(t,n) = \frac{\left[\mu_{s(t),n} - O(t,n) \right]^2}{\sigma^2_{s(t),n}}$ and the solution $\left| K(t,n) \right| \geq 1$ was discarded. Finally, the compensation scheme is applied as follows:

$$\left[\Delta O(t,n)\right]^{optimal} = \begin{cases} \left[K(t,n)\right]^{optimal} \cdot \left[\mu_{s(t),n} - O(t,n) \right], & \text{if } \left| \mu_{s(t)} \text{ - } O(t) \right| \leq R \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where *R* is a threshold that defines a compensation region.

IV. Feature Compensation vs. Model Adaptation

Model adaptation approaches have successfully been applied on speaker and environment adaptation in speech and speaker recognition. However, in speech recognition conventional adaptation techniques (e.g. ML, MAP and MLLR) dramatically degrade when just a few adapting utterances are available (Cui and Alwan, 2005; Leggetter and Woodland, 1995; Myrvoll et. al., 2000). For example, in (Cui and Alwan, 2005) MLLR does not always lead to improvements in WER with five adaptation utterances at moderate or high SNR (>15dB). Moreover, the effectiveness of unsupervised adaptation is also significantly degraded when compared with supervised schemes (Afify et. al., 1998; Myrvoll et. al., 2000; Uebel L. F. and Woodland, 2001). In SV "unsupervised" mainly means that the user identity is not known, which in turn is the most common situation. If the selection of adaptation data is adequate (e.g. an accurate discrimination between clients and impostors in SV), the system can improve its robustness by the proper use of adaptation methods. On the other hand, if the classification of adaptation data is not precise, errors can be introduced in the re-estimated model parameters, which are propagated into further verification attempts. These adaptation errors can also result from time variability of mismatch conditions between enrolling and testing when the telephone line or handset is not the same from one verification attempt to the other. This mismatch certainly reduces the accuracy of client/impostor discrimination in data classification and should degrade the effectiveness of any model adaptation scheme.

In contrast to model adaptation, ISVC has no temporal memory between consecutive verification events. ISVC does not modify users´ models and client/impostor discrimination errors of signals or frames do not propagated into further verification events. As a result, there will be no sustained degradation or improvement of the system performance from one verification event to the other. As shown in this paper, if the a priori parameters used by ISVC are well selected, system accuracy can be improved despite the fact that the enrolling data is increased.

It is worth emphasizing that ISVC and model adaptation schemes are not incompatible. Actually, as suggested by results presented here, the combination of ISVC with a standard unsupervised model adaptation technique can lead to higher reductions in EER than both approaches isolated. Observe that the strategy followed in this paper corresponds to compare ISVC with unsupervised model adaptation philosophy instead of comparing ISVC with a specific unsupervised adaptation scheme.

V. Experiments

Limited enrolling data and robustness to mismatch conditions are certainly the main problems faced by SV in real applications. Surprisingly, commercially available databases for text-dependent SV[1] do not include speech recorded from different telephone lines. Such sort of database is essential to evaluate the effectiveness and robustness of noise canceling, feature compensation and model adaptation schemes for channel mismatch. Moreover, no commercial data for text-dependent SV in Spanish has been found for purchasing. Consequently, given this situation, recording a local database was an essential step to carry on this research. The databases employed in this paper are similar to those mentioned elsewhere (Asami et. al., 2005; Yegnanarayana et. al., 2005). It is worth emphasizing that several authors show results with databases elaborated by artificially introducing additive and convolutional noise to clean speech. However, the authors believe that the procedure adopted here is more realistic and emulates more effectively operating conditions. Moreover, it is a way to improve the visibility of research in Iberian Languages.

Most of text-dependent SV tasks employ digit or number sequences. However, the usability of SV-based systems could be improved in some cultures if users are prompted to pronounce their first and family names. This sort of database is not available either, and local speech data was also recorded in Spanish for further experiments employing first and family names in Spanish.

5.1. Database 1

The first database is composed of 40 speakers (20 males and 20 females). The vocabulary corresponds to Spanish digits. Each speaker pronounced the 10-digit sequence “0-1-2-3-4-5-6-7-8-9” six times for enrolling. For verification, every speaker uttered the four-digit sequences “1-8-6-4”, “4-5-2-0” and “9-5-7-3” three times each. The enrolling speech signals were recorded on the same telephone line at Speech Processing and Transmission Laboratory (LPTV). The verification signals were recorded in two conditions: matched, from LPTV telephone line; and unmatched, where every user phoned from his house. Consequently, two channel mismatch conditions between enrolling and testing are available: a) same telephone channel in enrolling and verification; and, b) telephone channel depending on the user. The term telephone channel considers telephone handset, twisted-pair copper wire and processing in the CO´s, which in turn varies from call to call.

At each channel mismatch condition, the baseline FA and FR error rates are computed as follows: FR curve is estimated with 40 speakers x 9 verification signals per client = 360 signals; and, FA curve is computed by avoiding cross-gender impostor trials with 19 impostors x 9 verification signals per impostor x 40 users = 6840 experiments. The speaker-independent HMM used in the likelihood normalization in (1) was trained with 60 speakers (30 male + 30 female). Each speaker uttered three times the digit sequence “0-1-2-3-4-5-6-7-8-9” on LPTV telephone line.

5.2. Database 2

This database is composed of 31 speakers (11 males and 20 females). All the speech signals were recorded on LPTV telephone line. Each speaker pronounced his/her first and family names in Spanish eight times (3 for enrolling and 5 for verification) and the corresponding impostors repeated the client’s first and family names one time each. As a consequence, every speaker-dependent HMM is trained with only three utterances. The impostor universe for a given client is defined by the speakers with the same genre. FR curves are estimated with (11 male-speakers + 20 female-speakers) x 5 verification utterances = 155 signals. FA curves were obtained with (11 male-speakers) x 10 impostor signals plus (20 female-speakers) x 19 impostor signals = 490 signals. The speaker-independent HMM used in the likelihood normalization (1), is trained with 150 speakers, mixed gender. Utterances used for training the speaker-independent model were recorded on multiple telephone lines.

5.3. Experimental setup

Enrolling and verification utterances are decomposed as a sequence of triphones. Thirty-three cepstral coefficients are computed per frame: the frame energy plus ten static coefficients and their first and second time derivatives. The HMM’s were trained with the Viterbi algorithm. Each triphone was modeled with a three-state left-to-right HMM topology without skip-state transition, with one multivariate Gaussian density per state in speaker-dependent models, and eight multivariate Gaussian densities per state in the speaker- independent model. Both models employed diagonal covariance matrices.

5.4. ISVC

ISVC is tested with database 1 and 2 with no telephone channel mismatch. Distributions *d(n)* defined in (3) and shown in Fig. 3 was estimated with an evaluation database composed of 13 speakers that were different from the testing database. The evaluation database was also recorded on LPTV telephone line. The baseline system gave an EER equal to 6.29% and 11.09% in databases 1 and 2, respectively. Results are presented in Tables 1-2 and Figs. 4-8.

5.5. Comparing and combining ISVC with unsupervised model adaptation

ISVC is compared and combined with an unsupervised incremental adaptation approach based on MAP re-estimation of mean vectors (Barras et. al., 2004; Gauvain and Lee, 1994; Yu and Mason, 1996). Unsupervised online MAP adaptation was applied with a fixed adaptation weight τ and a posteriori probability of the target client given the score, $\Pr\left(\text{client}\,\middle|\log L(O)\right)$, where $\log L(O)$ is defined in (1), as proposed in (Barras et. al., 2004), $\Pr\left(\text{client}\,\middle|\log L(O)\right)$ is estimated using the a priori distribution of true client score $\log L(O)$. The update equation for HMM vector mean is:

$$\hat{\mu}_s = \frac{\mu_s + \tau \cdot \Pr\left(\text{client}\,\middle|\log L(O)\right) \cdot \bar{O}}{1 + \tau \cdot \Pr\left(\text{client}\,\middle|\log L(O)\right)} \tag{15}$$

where τ is an adaptation weight; $\bar{O}$ is the average of frames that are allocated to state *s* as a result of the forced Viterbi alignment; and, μ_s and $\hat{\mu}_s$ are the original and compensated vector means at state *s*, respectively. In unsupervised adaptation schemes, it is possible to assume that an error on adaptation data selection (false acceptance) could cause speaker model degradation. Therefore, experiments must represent the behavior of the adaptation scheme on different scenarios of client and impostor verification events. The same strategy adopted elsewhere (Fredouille et. al., 2000) was also followed here and more than one scenario was employed to evaluate the unsupervised adaptation strategy:

- **Scenario 1**. The purpose of this experiment is to simulate a set of massive client verification events, followed by a set of massive impostor attempts. Nine client verification utterances are followed by 171 impostor verification utterances per user.
- **Scenario 2**: This experiment aims to simulate a balanced sequence of client and impostor verification events. In this scenario, one client verification attempt followed by two impostor verification attempts are constantly alternating. The same utterances from scenario 1 were used.

The model adaptation algorithm is applied with an adaptation window. The size of the adaptation window represents how many utterances from previous verification attempts will be considered for adaptation. Weight τ and the size of the adaptation window are adjusted using 20 users from database 1 (ten male plus ten female), channel matched version, in the both scenarios explained above. Finally, unsupervised adaptation is properly compared and combined with ISVC by making use of optimal weight τ and adaptation window. Both scenarios are tested with telephone channel matching condition between enrolling and verification. Results are presented in Tables 3 and 4, and Figs. 9-12.

5.6. Telephone Channel mismatch

Matched and unmatched versions of database 1 are used in this set of experiments. Matched and unmatched utterances are alternated in order to simulate verification attempts from different telephone lines. FR curve is estimated with 40 speakers x (9 matched + 9 unmatched) verification signals per client = 720 signals. FA curve are obtained with 19 impostors x (9 matched + 9 unmatched) verification signals per impostor x 40 users = 13680 experiments. Testing utterances are processed with an unsupervised maximum likelihood signal bias removal algorithm (Afify et. al., 1998; Rahim and Huang, 1996) to compensate for channel mismatch between train and test conditions. Scenarios 1 and 2 are also employed in these tests to generate different sequences of client/impostor. Results are presented in Figs. 13-14.

VI. Discussion

According to Figs. 4-7, ISVC can lead to reductions as high as 20% or 40% and 30% or 60%, in EER on databases 1 (matched condition) and 2, respectively, when compared with the baseline systems. Although the reduction in EER is dependent on R in (14), Figures 4-5 shows that there is a wide range of values for R where ISVC provides significant improvements in speaker verification accuracy. Surprisingly, although both tasks are very different, the optimal range of values for R is the same for databases 1 and 2. This should be due to the fact that R in (14) contributes to discriminate client/impostor frames. Observe that intra-speaker variability as defined in (2), which should be task independent, also attempts to discriminate client/impostor frames. As can be seen in Figs. 6-7 and Tables 1-2, the integral below the ROC curves is 40% and 62% lower when ISVC is applied on databases 1 and 2, respectively, with R in (14) equal to 35 and 40. As can be seen in Fig. 8, database 1, the difference in EER provided by six and three enrolling signals is reduced by 35.3% when ISVC was incorporated to the baseline system. According to Fig 8, ISCV always improves the system accuracy independently of the number of enrolling utterances.

Figures 9 and 10 present the results with unsupervised HMM adaptation, as in (15), with a subset of 20 speakers of database 1, matched version. Figures 9 and 10 show EER versus adaptation window width in two scenarios for the sequence of client/impostor verification attempts, as explained in 5.5. Adaptation window width is tuned for every adaptation weight τ that is evaluated. Then an optimal pair (adaptation window, adaptation weight τ) is chosen. The idea is to assess the effect on unsupervised adaptation of discrimination ability between clients and impostors. As can be seen in Fig. 9 (scenario 1), the implemented adaptation scheme leads to reductions in EER as high as 54.5%. When combined with ISVC the reduction in EER is higher and equal to 63.7%. Nevertheless, according to Fig. 10, unsupervised model adaptation shows a non-consistent behavior in scenario 2 where a low improvement in accuracy is observed (the highest reduction in EER is equal to 18%). Actually, in some cases EER increases. But, when combined with ISVC, the adaptation scheme leads to a reduction in EER equal to

45.5%. This result suggests that an unsupervised adaptation method might even degrade the accuracy of the system in an adverse scenario of persistent impostor verification attempts. In contrast, ISVC may be outperformed by model adaptation schemes in some situations, but it is robust to massive or persistent impostor attack due to the fact that it does not have temporal memory. These results can also be observed in Figs. 11 and 12 that present DET curves provided by baseline system, ISVC, unsupervised model adaptation (UnsAdap) and the combination of ISVC with model adaptation in scenario 1 and 2, respectively, with the whole database 1, matched version. The combination of both methods gives a reduction in the area below the ROC curve equal to 44% and 19% in scenario 1 and scenario 2, respectively.

Results with matched and unmatched verification signals from database 1, as explained in section 5.6, in scenarios 1 and 2 are shown in Figs. 13 and 14, respectively. As can be seen in Fig. 13, ISVC also outperforms the baseline system. In contrast to Fig. 11, the unsupervised adaptation scheme is more significantly superior to ISVC. However, the highest reduction in EER and in the area below the ROC curve, 27.5% and 35.4% compared with the baseline system, respectively, takes place when unsupervised adaptation is combined with ISVC. However, when compared with adaptation only, the combination of both schemes gives a lower improvement than in Fig.11. This must be due to the fact that the relative improvement due to ISVC increases in telephone channel matched conditions. In scenario 2 (Fig. 14) the combination with ISVC also improves the accuracy of model adaptation scheme when compared with baseline results. However, the difference between ISVC and model adaptation is less significant than in scenario 1 (Fig. 13). This is certainly due to the temporal memory of the unsupervised adaptation method that is misled by not reliable adaptation data.

VII. Conclusions

The proposed unsupervised compensation method based on Gestalt, ISVC, can lead to reductions in EER and in the integral below the ROC curve as high as 20% or 40% and 30% or 60%, respectively, independently of the number of enrolling utterances. ISVC is memoryless with respect to previous verification attempts. Unsupervised model adaptation can lead to substantial improvements in EER depending on the sequence of client/impostor verification events. For instance, an initial set of client signals certainly make a user model more robust. However, in adverse scenarios, such as massive or persistent impostor, unsupervised model adaptation might even provide reductions in verification accuracy when compared with the baseline system. In those cases, ISVC can even outperform adaptation schemes due to the fact that ISVC lacks temporal memory. It is worth emphasizing that ISVC and unsupervised model adaptation are compatible and the combination of both methods always improves the performance of model adaptation. To improve the accuracy of ISVC by including the dependence of intra-speaker variability on speaker and phonetic class, to model the combination of ISVC with methods for removal of telephone line mismatch and to model the effect of ISVC on the threshold of EER can be proposed as future research.

References

Afify, M., Gong, Y., Haton, J., 1998. A general joint additive and convolutive bias compensation approach applied to noise Lombard speech recognition, IEEE Trans. on Speech and Audio Process. 6(6), pp. 524-538.

Ahn, S., Ko, H., 2000. Speaker adaptations in sparse training data for improved speaker verification, IEE Electronics Letters. 36, pp. 371– 376.

Asami, T., Iwano, K., Furui, S., 2005. Stream-weight optimization by LDA and adaboost for multi-stream speaker verification". In: Proc. of ICSLP, Lisbon, Portugal, pp. 2185-2188.

Barras, C., Meignier, S., Gauvain, J.L., 2004. Unsupervised online adaptation for speaker verification over the telephone, In: Proc. of Odyssey, Toledo, Spain.

Cui, X., Alwan, A., 2005. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR". In IEEE Trans. on Speech and Audio Process. 13(6), pp.1161-1172.

Fredouille, C., Mariethoz, J., Jaboulet, C., Hennebert, J., Bonnastre, J.F., Mokbel, C., Bimbot, F., 2000. Behaviour of a Bayesian adaptation method for incremental enrollment in speaker verification, In: Proc. ICASSP, Istambul, Turkey, pp.

Furui, S., 1997. Recent advances in speaker recognition, Pattern Recognition Letters. 18, pp. 859-872.

Gao, Q., Zhang, Y., Parslow, A., 2005. The influence of perceptual grouping on motion detection, Computer Vision and Image Understanding. 100(3), pp. 442-457.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains, IEEE Trans. on Speech Audio Process. 2, 291-298.

Hardt, D., Fellbaum, K., 1997. Spectral subtraction and RASTA filtering in text-dependent HMM-based speaker verification, In: Proc. ICASSP, Munich, Germany, pp. 867-870.

Jiang, H., Ngo, C.W., Tan, H.K., 2006. Gestalt-based feature similarity measure in trademark database, Pattern Recognition. 39(5), pp 988-1001.

Kim, S., Kweon, I.S., 2006. 3D target recognition using cooperative feature map binding under Markov Chain Monte Carlo, Pattern Recognition Letters. 27(7), pp. 811-821

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs, Computer Speech and Language. 9, pp. 171- 185.

Myrvoll, T., Siohan, O., Lee, C.H., Chou, W., 2000. Structural maximum a posteriori linear regression for unsupervised speaker adaptation. In: Proc. ICSLP, Beijing, China, pp. 540-543.

Ortega-Garcia, J., Gonzalez-Rodriguez, J., 1996. Overview of speech enhancement techniques for automatic speaker recognition, In: Proc. ICSLP, Philadelphia, USA, pp. 929-932.

Rahim, M.G., Huang, B.H., 1996. Signal bias removal by maximum likelihood for robust telephone speech recognition, IEEE Trans. on Speech and Audio Process. 4(1), pp. 19-30.

Rao, C.R., 1965. Linear statistical inference and its applications, John Wiley and Sons.

Uebel, L.F., Woodland, P.C., 2001. Speaker adaptation using lattice-based MLLR. ITRW on Adaptation Methods for Speech Recognition, Sophia Antipolis, France.

Wertheomer, M., 1950. Laws of organization in perceptual forms, Humanities Press.

Yegnanarayana, B., Mahadeva Prasanna, S.R., Zachariah, J. M., Gupta, C. S., 2005. Combining evidence from source suprasegmental and spectral features for a fixed-text speaker verification system, IEEE Trans. on Audio Speech and Language Process. 13(4), pp. 575-582.

Yiu, K.K., Mak, M.W., Kung, S.Y., 2007. Environment adaptation for robust speaker verification by cascading maximum likelihood linear regression and reinforced learning, Computer Speech and Language. 21, pp. 231-246.

Yu, K., Mason, J.S., 1996. On-line incremental adaptation for speaker verification using maximum likelihood estimates of CDHMM parameters, In: Proc. ICSLP, Philadelphia, USA, pp. 1752-1755.

R	ROC Area	Reduction compared with the baseline system
0 (Baseline)	133.08	0.00%
28	101.43	23.80%
35	80.29	39.70%
40	84.60	36.40%
45	92.09	30.80%

Table 1: Integral below ROC curve vs. *R* as defined in (14) with ISVC in experiments with database 1, matched version. The results are compared with the baseline system. Three utterances were employed for enrolling.

R	ROC Area	Reduction compared with the baseline system
0 (Baseline)	487.16	0.00%
27	358.06	26.50%
35	205.80	57.80%
40	185.65	61.90%
45	240.01	50.70%

Table 2: Integral below ROC curve vs. *R* with ISVC in experiments with database 2. The results are compared with the baseline system. Three utterances were employed for enrolling.

	Baseline	ISVC	UnsAdap	ISVC+ UnsAdap
Scenario 1, matched	6.29	5.00	5.97	4.17
Scenario 2, matched	6.07	4.46	6.45	5.53
Scenario 1, matched + unmatched	13.60	12.84	10.72	9.86
Scenario 2, matched + unmatched	16.04	12.13	13.04	10.91

Table 3: EER (%) in experiments with database 1 (matched and matched + unmatched versions) in scenarios 1 and 2. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is employed with R equal to 35. Three utterances were employed for enrolling.

	Baseline	ISVC	UnsAdap	ISVC+UnsAdap
Scenario 1, matched	133.10	83.90	159.60	74.03
Scenario 2, matched	113.43	61.06	145.00	92.23
Scenario 1, matched + unmatched	630.67	565.64	443.55	407.17
Scenario 2, matched + unmatched	958.58	660.63	602.51	501.03

Table 4: Integral below ROC curve in experiments with database 1 (matched and matched + unmatched versions) in scenarios 1 and 2. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is employed with R equal to 35. Three utterances were employed for enrolling.

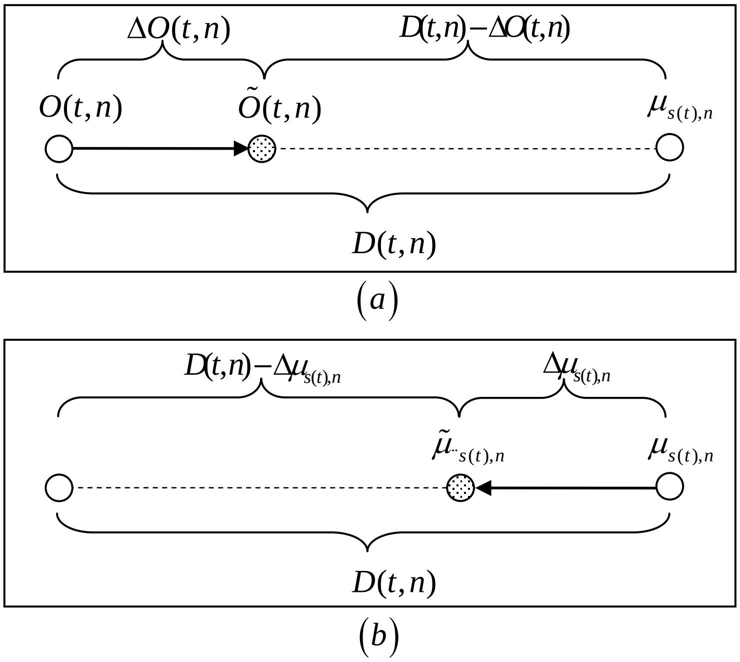

Figure 1: Graphical comparison of (*a*) ISVC with (*b*) model adaptation approach, where $\Delta\mu_{s(t),n}$ represents the mean adaptation component associated to unsupervised model adaptation.

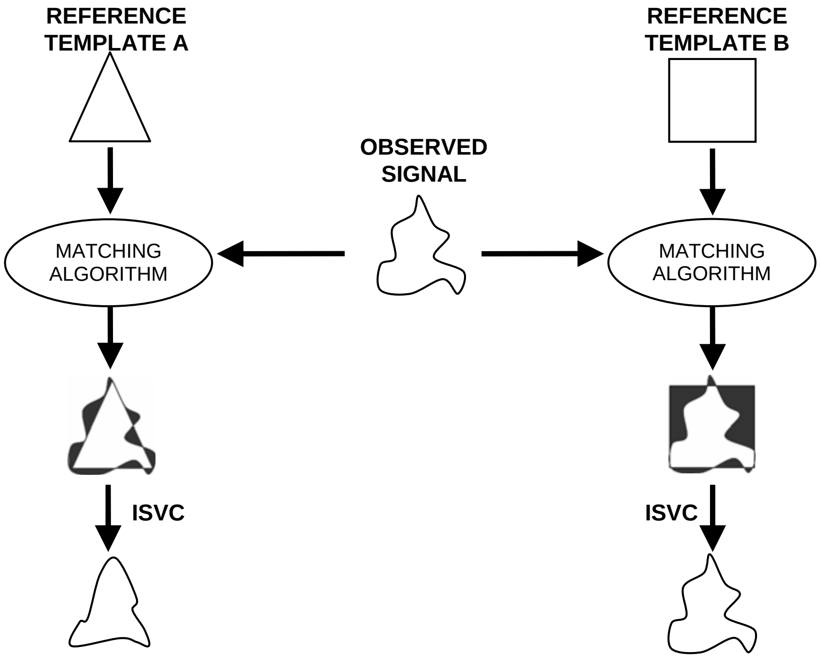

Figure 2: Graphical representation of the proposed feature compensation scheme based on Gestalt principle, ISVC.

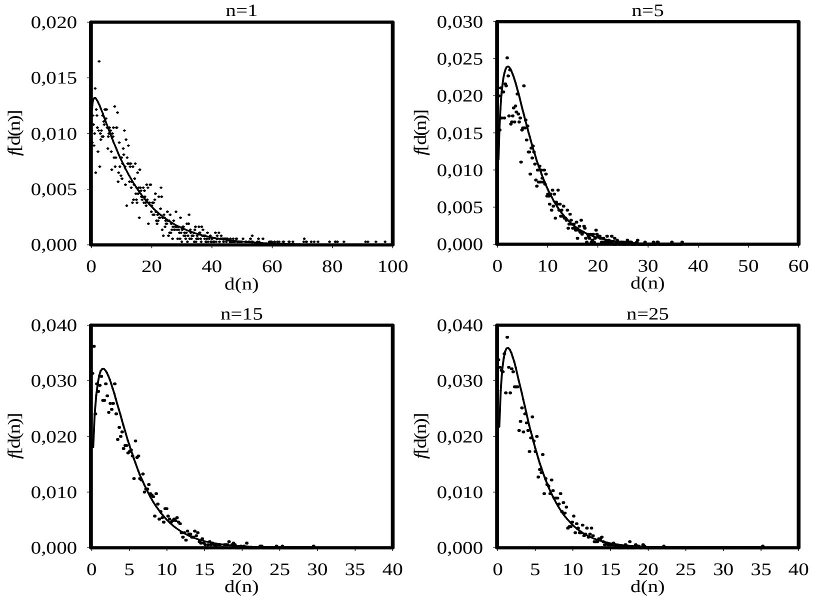

Figure 3: Distribution of *d(n)* as defined in (2): observed histogram and approximated gamma function *f[d(n)]* as represented in (3). The parameters employed in the figure corresponds to static (1 and 5), delta (15) and delta-delta (25) cepstral coefficients.

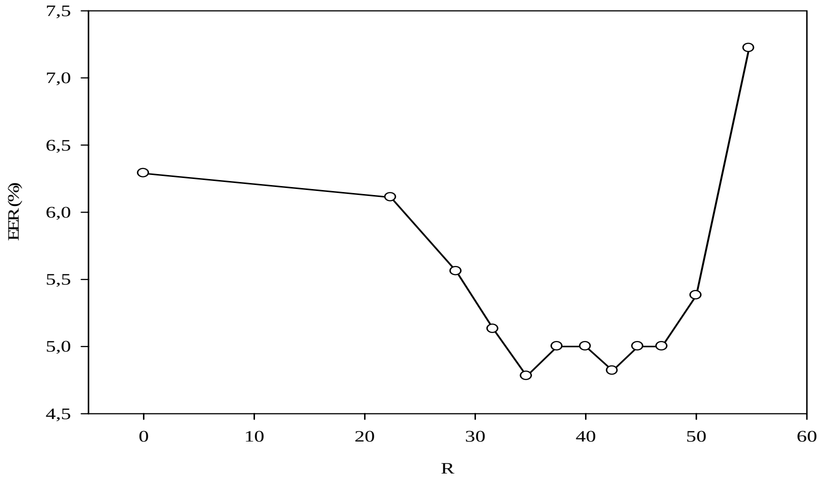

Figure 4: EER (%) vs. *R* as defined in (14) with ISVC in database 1, matched version. Three utterances were employed for enrolling.

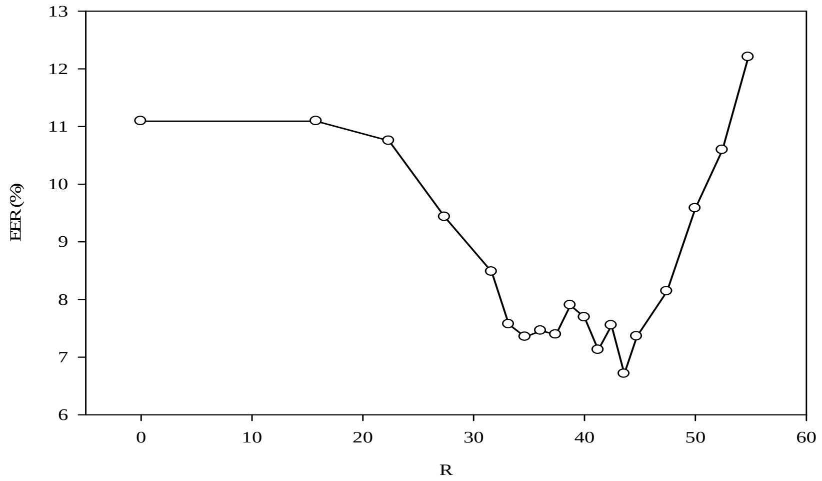

Figure 5: EER (%) vs. *R* as defined in (14) with ISVC in database 2. Three utterances were employed for enrolling.

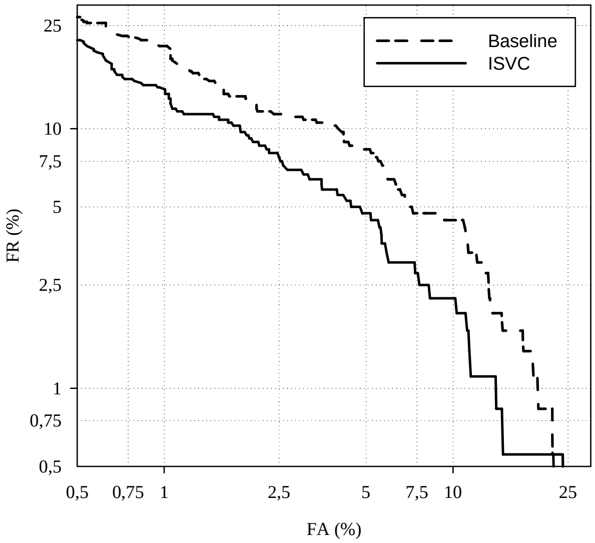

Figure 6: DET curve given by the baseline system and by ISVC with *R* equal to 35 in database 1, matched version. Three utterances were employed for enrolling.

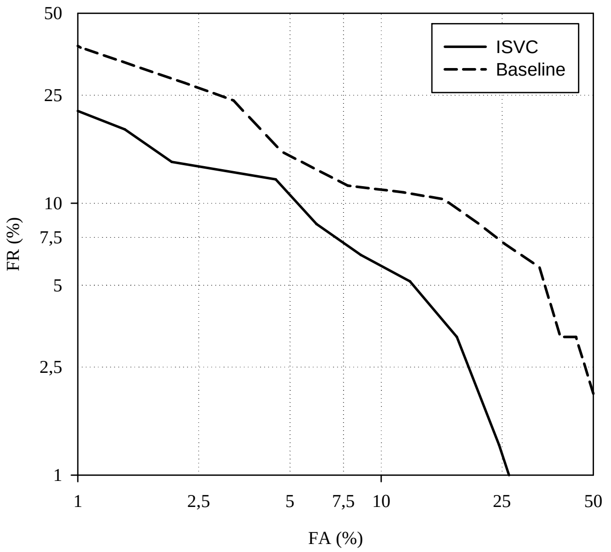

Figure 7: DET curve given by the baseline system and by ISVC with *R* equal to 35 in database 2. Three utterances were employed for enrolling.

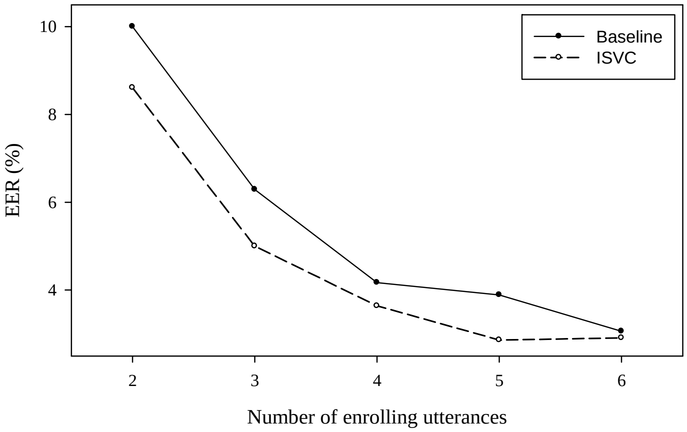

Figure 8: EER (%) vs. number of enrolling utterances given by the baseline system and by ISVC with *R* equal to 35 in database 1, matched version.

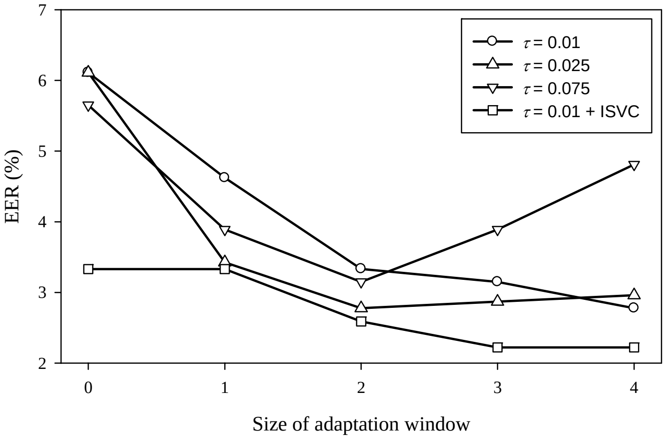

Figure 9: EER (%) vs. number of utterances in adaptation window in database 1, matched version, scenario 1, with several values of τ using the unsupervised model adaptation according to (15). The optimal value for τ is chosen in order to combine unsupervised model adaptation with ISVC (*R* = 35). A reduced set of database 1 is employed. In this set of data EER is equal to 6.11% and 3.33% with the baseline system and with ISVC, respectively.

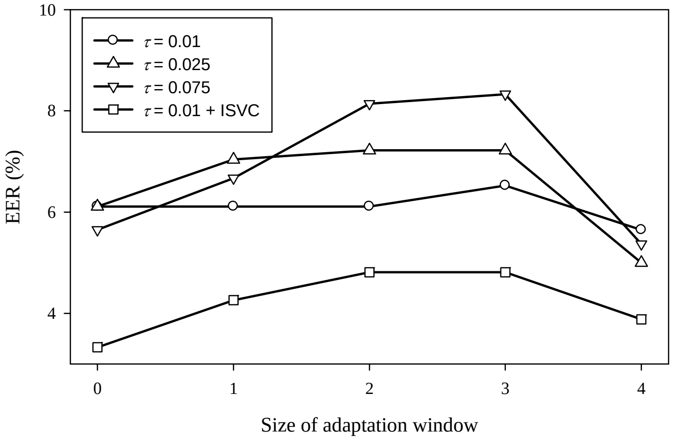

Figure 10: EER (%) vs. number of utterances in adaptation window in database 1, matched version, scenario 2, with several values of τ using the unsupervised model adaptation according to (15). The optimal value for τ is chosen in order to combine unsupervised model adaptation with ISVC (*R* = 35). A reduced set of database 1 is employed. In this set of data EER is equal to 6.11% and 3.33% with the baseline system and with ISVC, respectively.

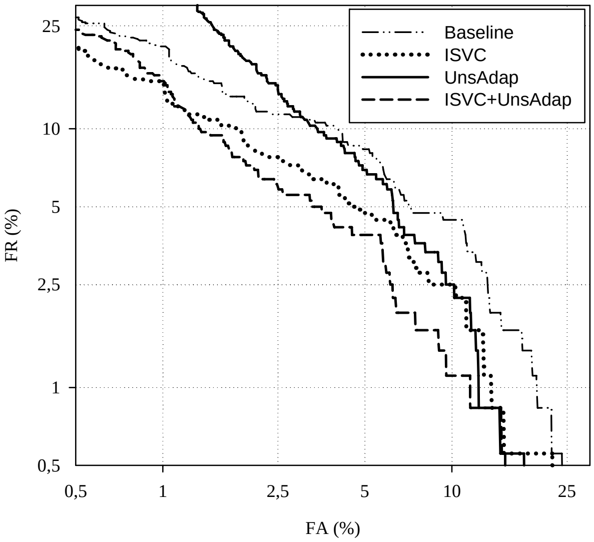

Figure 11: DET curves with database 1, matched version, scenario 1. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is used with R equal to 35. Three utterances were employed for enrolling.

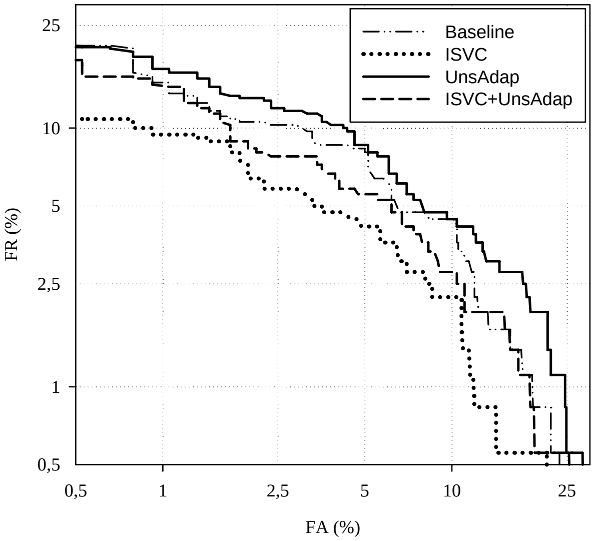

Figure 12: DET curves with database 1, matched version, scenario 2. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is used with R equal to 35. Three utterances were employed for enrolling.

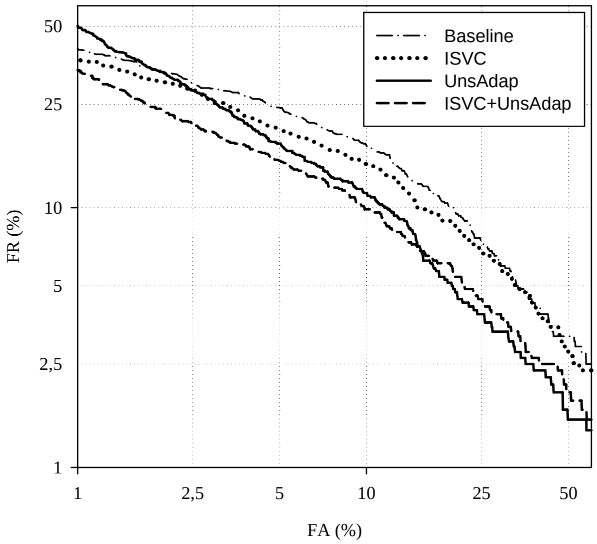

Figure 13: DET curves with database 1, matched and unmatched versions, scenario 1. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is used with *R* equal to 35. Three utterances were employed for enrolling.

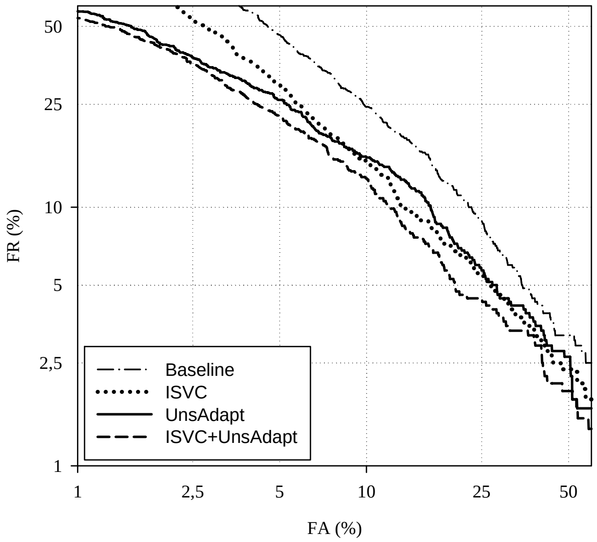

Figure 14: DET curves with database 1, matched and unmatched versions, scenario 2. Unsupervised model adaptation as in (15) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is used with *R* equal to 35. Three utterances were employed for enrolling.