On comparing and combining intra-speaker variability compensation and unsupervised model adaptation in speaker verification

Claudio Garreton, Nestor Becerra Yoma, Fernando Huenupan and Carlos Molina

Speech Processing and Transmission Laboratory
Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
nbecerra@ing.uchile.cl

Abstract

In this paper an unsupervised intra-speaker variability compensation method, ISVC, and unsupervised model adaptation are tested to address the problem of limited enrolling data in text-dependent speaker verification. 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, unsupervised model adaptation might even provide reductions in verification accuracy when compared with the baseline system. In those cases, ISVC may 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 34%.

Index Terms: speaker verification, limited enrolling data, intra-speaker variability, model adaptation, telephone speech.

1. Introduction

The amount and duration of enrollment utterances in a speaker verification (SV) system are key factors to guarantee low error rates during verification. However, the enrolling procedure in a SV system over the telephone network should be fast and efficient from the usability point of view. As a consequence, limited enrolling data is a common situation that leads to poorly trained models, which in turn seriously degrades the accuracy of SV engines.

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 Maximum Likelihood (ML) [1], Bayesian Maximum a Posteriori (MAP) [2] and Maximum Likelihood Linear Regression (MLLR) [3-4]. Supervised adaptation techniques, although more effective than unsupervised approaches, are impractical on large-scale SV based services. On the other hand, the unsupervised classification of adaptation data can introduce an error in the HMM 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) [5] is used to reduce the distortion between verification signals and the client HMM. Instead of adapting the client HMM, ISVC attempts to modify the input signal by reducing not relevant differences between testing signal and reference pattern or model with MAP estimation, 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.

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) a comparison of ISVC with unsupervised HMM adaptation; and, b) a combination of ISVC with model adaptation. Experiments with telephone speech suggest that: ISVC alone can lead to reductions in EER and in the area below FA/FR ROC curve as high as 24% and 40%, respectively; and, in combination with unsupervised HMM adaptation, ISVC provided reductions of 34% and 44% in EER and in the area below 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. Finally, the approach and analysis presented in this paper have not been found in the specialized literature.

2. Intra-speaker variability compensation

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, \( \log L(O) \) [6]:

\[
\log L(O) = \log Pr(O|\lambda_{SD}) - \log Pr(O|\lambda_{SI})
\]  

where \( O \) is the observation sequence; and, \( Pr(O|\lambda_{SD}) \) and \( Pr(O|\lambda_{SI}) \) represent the likelihood related to the speaker dependent (\( \lambda_{SD} \)) and independent (\( \lambda_{SI} \)) models, respectively. Both models, \( \lambda_{SD} \) and \( \lambda_{SI} \), 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: \( \frac{\log L(O)}{T} \). It is worth highlighting that \( \lambda_{SD} \) is computed with the enrolling data pronounced by the client, and \( \lambda_{SI} \) is estimated with a set of impostors. In this paper one multivariate Gaussian density per state is employed in \( \lambda_{SD} \).

The approach described in [5] models the intra-speaker variability given a state \( s \) in \( \lambda_{SD} \) and the enrolling data, as the vector \( d(t) = [d(t,0),...,d(t,n),...,d(t,N-1)] \) where
\[ d(t,n) = D(t,n) \] and \( D(t,n) = \mu_{t(n)} - O(t,n) \), \( \mu_{t(n)} \) is the \( n^\text{th} \) component in the mean vector of the observation probability function in state \( s \) that was allocated to frame \( O(t) = \{ O(t,0), ..., O(t,n), ..., O(t,N-1) \} \) 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 intra-speaker variability p.d.f., the histogram of \( d(t,n) \) was obtained with data different from training utterances. The intra-speaker variability is considered state and speaker independent. The p.d.f. of \( d(t,n) \), \( f(d(n)) \), is modeled with a gamma distribution:

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

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

ISVC aims to modify the input observation by reducing not relevant differences between test utterance and client HMM. If \( O(t,n) \) and \( O(t,n) \) denote the \( n^\text{th} \) feature in the compensated and observed frames, respectively, the compensation is expressed with:

\[
 \hat{O}(t,n) = O(t,n) + \left[ \Delta O(t,n) \right]^{\text{optimal}}
\]  

where \( \left[ \Delta O(t,n) \right]^{\text{optimal}} \) is the correction component at instant \( t \) and is modeled as a fraction of the multivariate vector difference between \( O(t,n) \) and \( \mu_{t(n)} \):

\[
 \left[ \Delta O(t,n) \right]^{\text{optimal}} = D(t,n) \left[ K(t,n) \right]^{\text{optimal}}
\]  

where \( \left[ K(t,n) \right]^{\text{optimal}} \) represents the optimal fraction of difference \( D(t,n) \). The compensation component \( \left[ \Delta O(t,n) \right]^{\text{optimal}} \) is estimated by maximizing the a posteriori p.d.f. \( P[\mu_{t(n)} - \hat{O}(t,n) | O(t,n), s(t)] \), where the difference \( \mu_{t(n)} - \hat{O}(t,n) \) defines the optimal distance between the adapted observation vector \( \hat{O}(t,n) \) and mean vector \( \mu_{t(n)} \), given a state \( s(t) \). As shown in [5], the maximization can be expressed as:

\[
 \left[ K(t,n) \right]^{\text{optimal}} = \arg \max_{K(t,n)} \left\{ \frac{\left[ 1 - K(t,n) \right] \left[ \mu_{t(n)} - \hat{O}(t,n) \right]}{\Pr[\hat{O}(t,n) | O(t,n)]} \right\}
\]  

then the solution provided by (5) is:

\[
 \left[ K(t,n) \right]^{\text{optimal}} = 
\left\{ \begin{array}{ll}
-\frac{\alpha(n)}{\Omega(t,n)} \left[ \mu_{t(n)} - \hat{O}(t,n) \right] + \\
1 - \frac{1}{2} \left[ \frac{\alpha(n)}{\Omega(t,n)} \left[ \mu_{t(n)} - \hat{O}(t,n) \right]^2 + 4 \left( p(n) - 1 \right) \right]^{\frac{1}{2}}
\end{array} \right.
\]  

where \( \Omega(t,n) = \left[ \mu_{t(n)} - \hat{O}(t,n) \right]^2 \).

The compensation scheme is applied as follows:

\[
 \left[ \Delta O(t,n) \right]^{\text{optimal}} = 
\left\{ \begin{array}{ll}
\left[ K(t,n) \right]^{\text{optimal}} \left[ \mu_{t(n)} - \hat{O}(t,n) \right] & \text{if } |\mu_{t(n)} - O(t,n)| \leq R \\
0, & \text{otherwise}
\end{array} \right.
\]  

where \( R \) is a threshold that defines a compensation region [5].

3. 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 [7-8]. Moreover, the effectiveness of unsupervised adaptation is also significantly degraded when compared with supervised schemes [8-9]. 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 inadequate 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.

4. Experiments

The 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” three 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 and verification speech signals were recorded on the same telephone channel. 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.
are the original and compensated

Figure 1: EER (%) vs. number of utterances in adaptation window in scenarios 1 (a) and 2 (b), with several values for $W$ using UnsAdap according to (8).

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”.

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.

ISVC is compared and combined with an unsupervised incremental adaptation approach (UnsAdap) based on MAP re-estimation of mean vectors as proposed in [10]. Unsupervised online MAP adaptation was applied with a fixed adaptation weight $\tau$ and a posteriori probability of the target client given the score, $Pr_{\text{client}}(\log L(O))$, where $\log L(O)$ is defined in (1), $Pr_{\text{client}}(\log L(O))$ is estimated using the a priori distribution of true client score $\log L(O)$. The update equation for HMM vector mean is:

$$\hat{\mu}_s = \mu_s + \tau \cdot Pr_{\text{client}}(\log L(O)) \cdot \hat{O} \over 1 + \tau \cdot Pr_{\text{client}}(\log L(O))$$

where $\tau$ is an adaptation weight; $\hat{O}$ is the average of frames that are allocated to state $s$ as a result of the forced Viterbi alignment; and, $\mu_s$ 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 [11] 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 significant client verification events, followed by a set of massive impostor attempts. Nine client verification utterances are followed by 171 impostor verification utterances per user. Results are obtained using 40 speakers x 180 verification signals per client = 7200 signals.

- Scenario 2: This experiment aims to simulate a balanced sequence of client and impostor verification events. In this scenario, one client and two impostor verification attempts are constantly alternating. Results are estimated with 40 speakers x 28 verification signals per client = 1120 signals.

UnsAdapt 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 $\tau$ and the size of the adaptation window are adjusted using 20 users from the database (ten male plus ten female), in both scenarios explained above. Results can be seen in Fig. 1. Finally, unsupervised adaptation is properly compared and combined with ISVC by making use of optimal weight $\tau$ and adaptation window.

5. Discussion and Conclusions

According to Table 1, ISVC can lead to reductions as high as 25% and 40%, in EER and in the integral below the ROC curve, when compared with the baseline system. Although the reduction in EER is dependent on $R$ in (7), Table 1 shows that there is a wide range of values for $R$ where ISVC provides significant improvements in speaker verification accuracy.

<table>
<thead>
<tr>
<th>$R^2$</th>
<th>EER (%)</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (baseline)</td>
<td>6.29</td>
<td>133.1</td>
</tr>
<tr>
<td>1000</td>
<td>5.13</td>
<td>84.6</td>
</tr>
<tr>
<td>1400</td>
<td>5.00</td>
<td>83.9</td>
</tr>
<tr>
<td>1800</td>
<td>4.82</td>
<td>85.75</td>
</tr>
<tr>
<td>2200</td>
<td>5.00</td>
<td>101.78</td>
</tr>
</tbody>
</table>

Table 1: EER (%) and integral below FA/FR ROC vs. $R^2$ given by the baseline system and by ISVC as in (7).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Baseline</th>
<th>ISVC</th>
<th>UnsAdap</th>
<th>ISVC+UnsAdap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.29</td>
<td>5.00</td>
<td>5.97</td>
<td>4.17</td>
</tr>
<tr>
<td>2</td>
<td>6.07</td>
<td>4.46</td>
<td>6.45</td>
<td>5.53</td>
</tr>
</tbody>
</table>

Table 2: EER (%) in scenarios 1 and 2. UnsAdap as in (8) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is employed with $R = 35$.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Baseline</th>
<th>ISVC</th>
<th>UnsAdap</th>
<th>ISVC+UnsAdap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>133.1</td>
<td>83.9</td>
<td>159.6</td>
<td>74.0</td>
</tr>
<tr>
<td>2</td>
<td>113.4</td>
<td>61.1</td>
<td>145.0</td>
<td>92.2</td>
</tr>
</tbody>
</table>

Table 3: Integral below FA/FR ROC curve in scenarios 1 and 2. UnsAdap as in (8) is applied with an adaptation window of 4 utterances and $\tau = 0.01$. ISVC is employed with $R = 35$. 
As can be seen in Fig. 1, the optimal values for adaptation weight and adaptation window are 0.01 and 4, respectively. In scenario 1, UnsAdapt leads to significant improvements in EER. Nevertheless, in scenario 2 UnsAdap shows a non-consistent behavior where a low improvement in accuracy is observed. Actually, in some cases EER increases. In both scenarios, UnsAdap leads to higher reductions when combined with ISVC.

Tables 2 and 3 present experiment provided by the baseline system, ISVC, UnsAdap and both approaches combined in scenarios 1 and 2, respectively. The combination of both methods gives an average reduction in scenarios 1 and 2 equal to 21% and 32% in EER and the area below the ROC curve, respectively.

As shown in Figs. 2 and 3, ISVC can lead to significant reductions in EER and in the integral below the ROC curve. ISVC is memoryless with respect to previous verification attempts. On the other hand, UnsAdap 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 attacks, UnsAdap might even provide reductions in verification accuracy when compared with the baseline system. The observed dependence of UnsAdap on the client/impostor sequence is consistent with the results presented in [11]. 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 UnsAdap are compatible and the combination of both methods always outperforms the results obtained with UnsAdap. 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.

6. Acknowledgement
This work was funded by Conicyt - Chile under grants Fondef D021-1089, D051-10243 and Fondecyt 1070382.

7. References