INTEGRATING TIME-ALIGNMENT INFORMATION INTO THE DECISION MAKING FOR
TEXT-DEPENDENT HMM-BASED SPEAKER VERIFICATION

D. Charlet
France Telecom, BD/CNET
DIH/DIPS, 2 av. Pierre Marzin
22307 Lannion Cedex - FRANCE
delphine.charlet@cnet.francetelecom.fr

ABSTRACT

This paper proposes an integration of the time-alignment information in the decision making for HMM-based text-dependent speaker verification. The principle is to consider acoustical score and time-alignment as joint observations for which a log-likelihood ratio is computed and compared to a threshold. It is shown that such integration has two distincts aspects, one being a kind of adaptation of the acoustical score threshold to the observed alignment, the other being the integration of the speaker-specificity information of the alignment in the decision making. Experiments on a large-scale and realistic database are reported. They show the interest of the proposed method and encourage further investigation in such an approach.

1. INTRODUCTION

For HMM-based text-dependent speaker verification, previous works [1] [2] have shown that some speaker-characteristic information was contained in the alignment, namely the number of frames assigned to each state. [2] estimates a probability density function of the duration in each state and joins this pdf to the HMM model to rescore the utterance after Viterbi decoding. In [1], we have proposed a discrete duration model which appears to contain important speaker-specific information. In this work, we study the way to integrate speaker-specific alignment information in the decision making. Usually, decision is made by comparing the acoustical score (classically a time-normalized log-likelihood ratio) of the speech utterance \( X \) to a threshold. Here, we propose to consider the resulting acoustical score and alignment as joint observations for which a log-likelihood ratio (llk-ratio) is computed. This resulting llk-ratio is used as the decision score, i.e., when compared to a threshold, it leads to the final decision.

In the next section, the HMM-based text-dependent system used in this work is presented. Then in section 3 the principle of the integration of alignment information is detailed. Eventually, experiments are reported in section 4.

2. VERIFICATION SYSTEM

The system used is the one proposed in [1]: a HMM-based system in which the claimed-speaker model score and the world model (i.e., speaker-independent model) score are computed for a common alignment made on the speaker-independent model of the password. The claimed-speaker model is built by retraining in a bayesian-like way the speaker-independent model (variances are not retrained hence speaker-model and speaker-independent model share the same state-dependent variances). It was shown that such a system preserves some speaker-specific information contained in the alignment and makes the acoustical score more consistent in emphasizing remarkable parts of the claimed-speaker model. The acoustical score for a speech utterance \( X \) of length \( T \) is:

\[
S(X) = \frac{1}{T} \log \frac{P(X|\lambda)}{P(X|I)}
\]  

where \( \lambda \) is the claimed-speaker model and \( I \) the speaker-independent model of the password. It can be rewritten as:

\[
S(X) = \frac{1}{T} \sum_t \left( \frac{\mu_{q[t]} - \mu_{\lambda[t]}}{\sigma_{q[t]}} \right) \left( X_t - \frac{1}{2}(\mu_{q[t]} + \mu_{\lambda[t]}) \right)
\]  

where \( q[t] \) is the index of the state the frame \( X_t \) is aligned on, \( \mu_{q[t]} \) the mean of the speaker model in state \( q[t] \), \( \mu_{\lambda[t]} \) the mean of the speaker-independent model in state \( q[t] \) and \( \sigma_{q[t]} \) the standard-deviation in state \( q[t] \) (for sake of simplicity we consider that we are in a one-dimensional space).

The principle of the integration of the alignment information in the decision making does not depend on the special structure of our HMM-based system. It could be applied to any HMM-based speaker verification system. Nevertheless, computation is drastically simplified in case of unique alignment, and the formulae given below are only valid for the unique alignment system.
3. INTEGRATION OF THE ALIGNMENT IN THE DECISION MAKING

3.1. Principle

A speech utterance $X$ is processed again the speaker-independent HMM (world-model) and the HMM of the claimed-speaker according to the alignment obtained on the world model. The resulting acoustical score $S(X)$ (given by equation 2) and alignment $Alg(X)$ are now considered as observations.

In the framework of hypothesis testing, $H_0$ is the null hypothesis that these observations come from the true-speaker whereas $H_1$ is the alternative hypothesis that these observations come from an impostor. Hence the decision is made with the llk-ratio:

$$L(X) = \log \left( \frac{P(S(X), Alg(X) | H_0)}{P(S(X), Alg(X) | H_1)} \right)$$  

(3)

It can be rewritten as:

$$L(X) = \log \left( \frac{P(S(X) | Alg(X), H_0) P(Alg(X) | H_0)}{P(S(X) | Alg(X), H_1) P(Alg(X) | H_1)} \right)$$

$$L(X) = \log \left( \frac{P(S(X) | Alg(X), H_0)}{P(S(X) | Alg(X), H_1)} \right) + \log \left( \frac{P(Alg(X) | H_0)}{P(Alg(X) | H_1)} \right)$$

$$L(X) = L_{S|Alg}(X) + L_{Alg}(X)$$  

(4)

Hence, the new decision score $L(X)$ is the sum of 2 terms, $L_{S|Alg}(X)$ being the llk-ratio of the acoustical score given the alignment, and $L_{Alg}(X)$ being the llk-ratio of the alignment. Details to compute both terms are given below.

3.2. Likelihood ratio for the acoustical score given the alignment

To compute $L_{S|Alg}(X)$, we assign a probability density function to $S$ in both hypothesis $H_0$ and $H_1$. Assumption is made that these pdf are gaussian, which is justified if $X$ is long enough (see [3]). Hence, the mean and standard deviation of $S$ must be estimated in both hypothesis. The mean is the expectation of the acoustical score for the given alignment and the given hypothesis. In the case of $H_0$ (true-speaker hypothesis) it is then estimated in the following way:

$$E(S(X) | Alg(X), H_0) = E \left( \left( \frac{1}{T} \sum_t S(X_t) \right) | Alg(X), H_0 \right)$$  

(5)

Where $X_t$ is the speech frame at time $t$. Because alignment is given ($X_t$ is assigned to state $i$):

$$E(S(X) | Alg(X), H_0) = \frac{1}{T} \sum_t E(S(X_t) | Alg(X), H_0)$$  

(6)

and

$$E(S(X_t) | Alg(X), H_0) = m_{\text{aut}}^i$$  

(7)

where $m_{\text{aut}}^i$ is the mean of true-speaker score in state $i$. Consequently:

$$E(S(X) | Alg(X), H_0) = \frac{1}{T} \sum_i T_i m_{\text{aut}}^i$$  

(8)

with $T_i$ being the number of frames assigned to state $i$. [3] originally presents this principle, but in that work the fact the alignment is given, which is, to our mind, a necessary condition for the validity of the formula, was not required.

Hence, the mean of the pdf of the acoustical score in the hypothesis $H_0$ depends on the observed alignment. The mean is adapted to the observed alignment. It can be viewed as a kind of duration normalization.

The same idea is also used for the hypothesis $H_1$. In that case:

$$E(S(X) | Alg(X), H_1) = \frac{1}{T} \sum_i T_i m_{\text{imp}}^i$$  

(9)

where $m_{\text{imp}}^i$ is the mean of impostor score in state $i$.

The same principle could also be applied to estimate standard-deviation. Nevertheless, for sake of simplicity, we prefer not to adapt the standard-deviation with the observed alignment and consider a global standard deviation of the acoustical score at the word-level in the true-speaker hypothesis $s_{\text{aut}}$ and in the impostor hypothesis $s_{\text{imp}}$.

It is then necessary to estimate the state-level means $m_{\text{aut}}^i$ and $m_{\text{imp}}^i$ and the whole-word-level standard-deviation $s_{\text{aut}}$ and $s_{\text{imp}}$. [3] analytically derives the state-level parameters from the HMM-parameter, based on the following idea:

$$E(X_t | Alg(X), H_0) = \mu^i$$  

(10)

where $\mu^i$ is the mean of state $i$ ($i = q[t]$ in equation 2) in the HMM of speaker $\lambda$. Hence, the state-level statistic estimated analytically are model-dependent, therefore speaker-dependent and leads to a speaker-dependent decision making.

Unfortunately, because of the very small amount of training data in practical application, speaker-model parameters are undertrained and the formula above is far from being satisfied. There is an important bias between the analytically-derived mean and the empirically observed mean. Therefore, we choose to estimate empirically the score parameters on a development database. Hence, it was impossible to estimate speaker-dependent means (speakers from the development database are distinct from speaker of the evaluation database). From the development database, speaker-independent state-level true-speaker score means ($m_{\text{aut}}^i$) and impostor score
means \( m_{imp}^i \) were estimated as well as whole-word-level true-speaker score standard-deviation \( s_{out} \) and impostor score standard-deviation \( s_{imp} \).

Hence, the llk-ratio of the acoustical score given the alignment can be expressed as:

\[
L_{S/Alg}(X) = \log \left( \frac{P(S(X)/Alg(X), H_0)}{P(S(X)/Alg(X), H_1)} \right) \tag{11}
\]

\[
= \log \frac{s_{imp}}{s_{out}} - \frac{(S(X) - \frac{1}{N} \sum_{i=1}^{N} T_i m_{aut}^i)^2}{2 s_{aut}^2} + \frac{(S(X) - \frac{1}{N} \sum_{i=1}^{N} T_i m_{imp}^i)^2}{2 s_{imp}^2}
\]

At that point, it is worth noticing that if the llk-ratio of the acoustical score was estimated without the hypothesis of given alignment, \( \hat{L}_S(X) \) could be expressed as:

\[
\hat{L}_S(X) = \log \left( \frac{P(S(X)/H_0)}{P(S(X)/H_1)} \right) \tag{12}
\]

\[
= \log \frac{s_{imp}}{s_{out}} - \frac{(S(X) - m_{aut})^2}{2 s_{aut}^2} + \frac{(S(X) - m_{imp})^2}{2 s_{imp}^2}
\]

where \( m_{aut} \) and \( m_{imp} \) are the word-level score means. As these parameters are estimated in a speaker-independent way, the function that transform \( S(X) \) into \( \hat{L}_S(X) \) is unique and monotonous thus would not change anything in the discrimination. Therefore, the differences between performances obtained with \( S(X) \) and performances obtained with \( L_{S/Alg}(X) \) will be only due to the adaptation of the mean with the observed alignment.

### 3.3. Likelihood ratio for the alignment

In [1], we have developed a discrete duration model for alignment that appears to be quite effective though simple. We briefly recall it here. Namely, for each speech utterance \( X \), we define a vector \( Track(X) \):

\[
Track(X) = \{ tr_i(X) \}_{i=1, \ldots, N} \tag{13}
\]

where \( N \) is the number of states of the model, and \( tr_i(X) = \frac{T_i(X)}{T(X)} \) is the number of frames associated to state \( i \) during the alignment, normalized by the total number of frames of the utterance. For each true speaker \( \lambda \), a temporal model, \( Track_\lambda \), is built by averaging the tracks obtained on the utterances used for training. For each test utterance \( X \), a distortion \( D_\lambda(X) \) is measured between the duration model of the claimed speaker \( Track_\lambda \) and the observed \( Track(X) \):

\[
D_\lambda(X) = \sum_{i=1}^{N} [ tr_i(X) - tr_i^\lambda ] \tag{14}
\]

As there is some speaker characteristic information contained in the alignment, the distortion \( D_\lambda(X) \) is lower for true-speaker utterance than for impostor utterance, and discrimination between true speakers and impostors is done by comparing \( D_\lambda(X) \) to a threshold. Hence the pdf of \( Alg \) in hypothesis \( H_0 \) is replaced with \( D_\lambda(X) \). The distribution of \( Alg \) in hypothesis \( H_1 \) is supposed to be uniform, therefore negligible in the decision making. Hence, the llk-ratio of the alignment is replaced with:

\[
L_{Alg}(X) = k.D_\lambda(X) \tag{15}
\]

where \( k \) is an empirically optimized numerical factor to compensate for the approximation of the llk-ratio (0.2 in our experiments).

### 4. EXPERIMENTS

#### 4.1. Database Description and Experimental Setup

For experimental evaluation, we use a telephone speech database collected over long distance telephone lines, that contains a set of 43 true speakers (male and female) and a distinct set of 600 impostors. The speech data consists of 5 short sentences (average duration of each sentence: 1.5s) recorded in various and realistic conditions (including background noise, such as TV). The sentences are referred to as sen 1 to sen 5. For each true speaker, training is performed with 3 repetitions of the password uttered during a single call. For testing, a verification test is performed on each utterance of the right password collected during different calls. The number of true testing calls varies from one speaker to another, between 1 and 25, with an average of 13 per speaker. An equal number of impostors attempts (same-sex and cross-sex trials) is performed. For each true speaker, there is on average a period of 2 months between the call used for training and those used for testing. The development database that serves for estimating the statistics of the acoustical score consists of a distinct part of the same database.

An unique threshold is set a posteriori for all passwords and all speakers (based on 3114 true speaker attempts and 3114 impostors attempts), so as to obtain the so-called global Equal Error Rate (i.e. when false rejection rate equals false acceptence rate). A threshold per sentence is also set, so as to evaluate the EER for each password separately.
<table>
<thead>
<tr>
<th>score</th>
<th>S(X)</th>
<th>L_{S/Alg}(X)</th>
<th>L_{Alg}(X)</th>
<th>L_{L}(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sen 1</td>
<td>6.20</td>
<td>5.88</td>
<td>18.1</td>
<td>5.72</td>
</tr>
<tr>
<td>sen 2</td>
<td>3.09</td>
<td>2.94</td>
<td>14.4</td>
<td>2.50</td>
</tr>
<tr>
<td>sen 3</td>
<td>4.26</td>
<td>3.92</td>
<td>18.9</td>
<td>3.75</td>
</tr>
<tr>
<td>sen 4</td>
<td>1.74</td>
<td>1.48</td>
<td>12.0</td>
<td>1.22</td>
</tr>
<tr>
<td>sen 5</td>
<td>4.19</td>
<td>3.57</td>
<td>17.4</td>
<td>3.57</td>
</tr>
<tr>
<td>global</td>
<td>3.90</td>
<td>3.56</td>
<td>16.4</td>
<td>3.35</td>
</tr>
</tbody>
</table>

Table 1. EER (in %) for different decision scores

The speaker-independent model is trained with approximately 150 repetitions of the password uttered by 50 speakers, that are distinct from those used in the impostor set.

4.2. Results and Discussion

In table 1, results are presented in terms of EER. The first line indicates the type of decision score (i.e. the score compared to the threshold): S(X) means acoustical score, \( L_{S/Alg}(X) \) means llk-ratio of the acoustical score given the alignment, \( L_{Alg}(X) \) means alignment score and \( L_{L}(X) \) means \( L_{S/Alg}(X) + L_{Alg}(X) \).

First of all, when comparing the first two columns, \( L_{S/Alg}(X) \) leads to better results than \( S(X) \) for each sentence. The global EER is decreased of about 9%. Hence, the adaptation of the llk-ratio to the observed alignment is valid.

Performances of \( L_{Alg} \) confirm that there is actually a speaker-characteristic information contained in the alignment. This effect combined with acoustical llk enables improvement. From our point of view, the improvement obtained when using both llk ratio for acoustical score and alignment score is remarkable though small (15% global EER reduction), because it combines 2 effects that are a priori antagonist. Indeed, the llk-ratio for acoustical score given the alignment is a kind of compensation for alignment variations, whereas the alignment score emphasizes the speaker-specificity contained in the alignment.

5. CONCLUSION

In this work, we have proposed an original way to integrate time-alignment information in the decision making. This integration has 2 aspects, one being the adaptation of the llk-ratio of acoustical score (a kind of alignment-dependent threshold when decision is made only with \( L_{S/Alg}(X) \)), the other emphasizing the speaker-specific information contained in time-alignment. This was evaluated on a particular HMM-based system in which alignment is made only on the speaker-independent model. As we have seen that alignment is even more speaker-specific when obtained on claimed-speaker model [1], further work would naturally be to adapt this integration in the context of model-dependent alignment.