A SPEECH BIOMETRICS SYSTEM WITH MULTI-GRAINED SPEAKER MODELING

J. Navrtil, U. V. Chaudhuri, S. H. Maes
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
e-mail: {jiri, uvc, smaes}@us.ibm.com

ABSTRACT
The paper describes a system for voice-based personal authentication in a conceptual framework of conversational speech biometrics relying on two sources of authentication - the acoustic voice-print and the user knowledge. Several technologies are closely integrated: speaker recognition and speech recognition as well as natural language understanding and dialog management. The first part of this paper details the algorithms for the acoustic voice-print modeling and recognition and presents experimental results on two databases with telephony speech. The second part describes the integration of acoustic voice-print recognition and speech recognition in a prototype that allows for natural language input and dynamic authentication scenario in the telephony environment, together with an experimental evaluation carried out on a speech-biometrics database.

1. ACOUSTIC SPEAKER RECOGNITION
Voice plays an important role in remote access applications such as telephony dialog systems where it is the sole means of communication. Besides conveying content, speech also carries certain physical biometrics of the speaker reflecting the properties of the vocal tract and the speaker’s way of articulating. Based on this, acoustic speaker recognition technology analyzing and modeling the speaker’s voice prints has been a major research effort for the past decades [1, 4, 10, 5, 6, 11, 7].

Speaker recognition comprises the tasks of verification, defined by an existence of a claim that is to be accepted or rejected, identification, i.e. determining an identity of a previously enrolled speaker, and classification which involves segmenting and processing speech utterances containing an unknown number of speakers. As for the degree of difficulty and with respect to the vocabulary, the tasks can be further divided into: text-dependent (same text for all users), user-selected (user-specific text), text-prompted (e.g. a digit string generated by the system), and text-independent speaker recognition. Algorithms for the latter task - being the most challenging - enable the speaker verification technology to be used in a variety of applications without restrictions to particular types of utterances and thus allow for a flexible design of authentication scenarios such as in conversational biometrics [8]. In particular, it is unobtrusive to the user and the transactions as it can run in parallel and independent of any dialog carried by the user with another human being or a machine. The acoustic algorithms described in this paper all follow the strategy of text-independent speaker recognition.

1.1. Multi-Grained Speaker Modeling
An overview of the training part of our system is shown in Fig. 1. The core model structure is Gaussian Mixture Models (GMM), which have been successfully applied in other systems as well [11]. In order to achieve a refined modeling with respect to the underlying phonetic structure of the speech, a HMM-based phonemic decoder is used in the training to label each speech frame (feature vector) with a corresponding phoneme. Based on this information the frames are further structured on multiple levels of phonetic “coarseness” (or granularity), starting from phoneme level, over broad phone-classes (e.g. “vowels”, “fricatives”, “plosives” etc.), to a global level which includes all frames regardless of their phonetic association. Vectors in each such “bin” then serve for estimating a GMM in that particular bin. In order to alleviate the problem of lacking robustness, a speaker-independent (SI) model with the same structure trained on sufficient amounts of data from many speakers provides the initial values in the estimation of the speaker model. The resulting GMM’s in each grain (bin) and on each level of granularity constitute the voice-print in the original feature space. In [3] the use of a maximum likelihood linear transform (MLLT) method in speaker recognition is presented that rotates the feature space so as to minimize the loss of the probability mass due to the diagonal assumption of covariance matrices in Gaussian mixture models. In our system, the MLLT is estimated and applied to the original feature space delimited by the respective grains in the original voice-print. Thus, a new voice-print in a multiply transformed feature space, optimized for diagonal GMM’s is obtained, which leads to significant improvements in the performance as will be shown in Sec. 1.5.

1.2. Scoring
The use of multi-grained models described above involves calculation of scores on multiple granularity levels and thus multiple GMM’s. In order to obtain a single likelihood value for a particular test utterance the individual scores need to be combined properly. In the context of phonetic labeling, one possible method is to calculate the interpolated probability of a labeled vector (e.g. phone “a”) given the corresponding grains on several granularity levels (e.g. “a”-“vowels”-
“global”), as follows:

$$S(X|M) = \sum_i S(\mathbf{x}_i, M) = \sum_i \log \sum_{y_i} a_i \cdot p(\mathbf{x}_i|M|y_i)$$  \hspace{1cm} (1)$$

with $X = \mathbf{x}_1, ..., \mathbf{x}_T$ denoting the feature vector sequence (test utterance), $Y = y_1, ..., y_T$ denoting the label sequence, $M[y(t, y_t)]$ being the model grain on level $l$ determined by the label $y_t$, and $a_i$ denoting the interpolation constants corresponding to each level. Besides tuning the interpolation constants this scoring method needs explicit label information and thus entails HMM decoding.

In our system, a simplified scoring mechanism that follows the principle of selecting a single grain across all levels and units giving the maximum likelihood for a particular feature frame. Thus (1) becomes:

$$S(X|M) = \sum_i S(\mathbf{x}_i, M) = \sum_i \log \max_{(l,i)} p(\mathbf{x}_i|M|y_i|l,i))$$  \hspace{1cm} (2)$$

In verification, the maximum operator has a justification in the observation that model competitiveness plays an important role for the performance. Further on, since the likelihood is calculated for all grains regardless of the phonetic correspondence the labeling information is not needed in (2) thus reducing the computation costs during the test.

### 1.3. Verification Task

The verification is viewed as a binary hypothesis test in which an identity claim, supplied by the user, is to be verified, i.e. accepted or rejected. A particular form of hypothesis testing widely used in speaker verification is the log-likelihood ratio test that involves a model of the target (client) class $M$ and a “world” model representing acoustic observations not belonging to the target speaker $\overline{M}$ (background):

$$\theta \leq \frac{S(X|M) - S(X|\overline{M})}{S(X|\overline{M})}$$  \hspace{1cm} (3)$$

whereby $\theta$ is an acceptance threshold determining the level of security and the relative number of false rejections vs. false acceptances (operating point). The world model $\overline{M}$ is often approximated by a set of ($N$) speaker models selected for a particular target model based on their competitiveness (or similarity) w.r.t. the target, so-called cohorts:

$$\theta \leq \frac{S(X|M) - \frac{1}{N} \sum_{i=1}^{N} S(X|C_i)}{S(X|\overline{M})}$$  \hspace{1cm} (4)$$

Unlike in the standard procedure of determining the cohort set in the training stage and using this set in (4) for the tests, we introduce a weighting into the calculation of the background part of the ratio, as follows

$$\theta \leq \frac{S(X|M) - \sum_{i=1}^{N} w_i \cdot S(X|C_i)}{\sum_{i=1}^{N} w_i = 1}$$  \hspace{1cm} (5)$$

and propose several functions for deriving the weights $\{w_i\}$ so as to improve the model competitiveness: 1) linear weighting function emphasizing the significance of cohort models that are closest to the target model given a test utterance (Fig. 2-a) 2) non-linear weighting function deriving the weights from the utterance likelihood (Fig. 2-b), and 3) a function as in 2) with a pruning threshold eliminating cohort models whose weight is too small. All three functions adapt an adaptation of the cohort set to the particular test utterance with the objective to increase cohort competitiveness and prove to bring performance gains and incur negligible additional computation.

### 1.4. Identification task

The identification is divided into two modes: closed-set and open-set. In our system, the latter is achieved by carrying out a closed-set identification and a subsequent verification of the result allowing for rejecting tests that are inconclusive or appear to be spoken by an unknown speaker.

Given a test utterance $X$ the maximum-likelihood classifier is used for this task calculating the scores over the target speaker population

$$I^* = \arg \max_i S(X|M_i)$$  \hspace{1cm} (6)$$
Table 1. Equal-error rates of the acoustic verification on the Lincoln Lab (LL) and the continuous digit strings databases

<table>
<thead>
<tr>
<th>Config</th>
<th>LL</th>
<th>MLL/T</th>
<th>LL</th>
<th>MLL/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lev 0 only</td>
<td>6.9%</td>
<td>2.4%</td>
<td>10.9%</td>
<td>9.2%</td>
</tr>
<tr>
<td>Multi-grain</td>
<td>3.9%</td>
<td>2.0%</td>
<td>5.5%</td>
<td>7.7%</td>
</tr>
</tbody>
</table>

Table 2. Performance of the individual background weighting methods (LL database)

<table>
<thead>
<tr>
<th>Weighting</th>
<th>% EER</th>
<th>Lev 0</th>
<th>MLL/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>3.7%</td>
<td>2.5%</td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>3.2%</td>
<td>2.3%</td>
<td></td>
</tr>
<tr>
<td>Likelihood</td>
<td>3.0</td>
<td>2.4%</td>
<td></td>
</tr>
<tr>
<td>Pruning</td>
<td>2.4%</td>
<td>2.0%</td>
<td></td>
</tr>
</tbody>
</table>

Clearly, the task complexity in terms of classification as well as computation is directly proportional to the target population size. Issues in large population speaker identification have been addressed in previous work [2].

1.5. Experimental Results

The acoustic speaker recognition system was evaluated on two independent databases with telephony-quality speech. The Lincoln Lab handset database was used to train and test the system in text-independent mode. For verification, 20 speakers from this database were enrolled as target models and 20 other speakers served as impostors, gaining a total number of 200 target and 3980 impostor utterances. The test length varied between 3 and 5 seconds. The second database consisted of 40 target speakers and approx. 450 impostor speakers and the utterances contained continuously spoken digit strings with the test length between 1-2 sec (200 target and about 14000 impostor tests). The amount of speech for enrollment in both databases was roughly 30 sec. The background population was created using internal large telephony database involving approximately 500 speakers and used in both experiments.

The features extracted from the speech signal were 12 mel-frequency cepstral coefficients (without C0) and their 1st and 2nd derivatives resulting in a 30-dimensional feature vector. Cepstral mean subtraction was carried out on an utterance basis to compensate for channel variations.

Table 1 shows the system performance in the verification task measured on the two databases for several configurations. The multi-grained models using the scoring method in Eq. (2) consistently outperform configurations without phonetic structuring, i.e. models with one (global) level (" Lev 0"). Further on, the MLL/T in both the global-level models and the multi-level models reduces the error on most of the cases as compared to the baseline ("BSL") system without the transform. Despite the fact that the digit task, in terms of the vocabulary, is a simpler problem, the small test length of 1-2 sec results in higher equal-error rates compared to the 3-5 sec tests in the LL database.

The results in Tab. 1 were obtained using the likelihood-derived background model weighting with pruning. Details on the contribution of the individual weighting methods are shown in Tab. 2. Clearly, the alternative non-uniform weighting functions, in particular combined with the model pruning, have significant impact on the final performance of the log-likelihood ratio discriminant without incurring considerable additional computation.

2. CONVERSATIONAL SPEECH BIOMETRICS FRAMEWORK

The concept of conversational speech biometrics (CSB) closely integrates the speaker recognition and speech recognition technology in order to exploit advantages of both the voice-print and verbal-content-based personal authentication. Speech biometrics, as introduced in [7], drastically improves the overall system robustness against impostors and thus enables voice to serve as a primary security key for a range of applications, particularly in the telephony environment. Besides the increased reliability, the CSB framework brings further advantages such as better flexibility in carrying out the authentication, e.g. adaptive length of a verification session, dialog dependent on a voice-print confidence, and extensibility to individuation and speaking style adaptation/recognition. Moreover, a continuous (on-the-fly) verbal or voice-print enrollment is possible by backing off to the respective alternative component thus maintaining a basic level of security while the enrollment data is being collected.

2.1. VIVA Prototype

To show the effectivity of this concept a prototype was developed that implements the essential CSB features in the telephony environment - called Voice Identification and Verification Agent (VIVA). The VIVA is an application and platform independent module capable of identification and verification based on a biometrics dialog conducted with the user.

The VIVA client-server architecture is depicted in Fig. 3 with two core parts: the VIVA server that maintains the database with user profiles and handles multiple client connection requests, and the speech biometrics control acting as an intermediate piece between the user and the application, that creates sessions for a particular user and controls both the speech recognition (IBM ViaVoice Telephony) and speaker recognition engine. Since the input is allowed to be natural language the speech recognition is supported by special topic-dependent finite state grammars. Within a session the control opens one or several interviews containing a variable number of verification questions out of a question pool for the user. The length and number of the interviews (and thus questions) depends on the security policy set by the application as well as on the history of the answers and the current confidence of the acoustic voice-print match. The security level for an interview is defined as the maximum number of questions asked and the minimum number of answers required to be correct.

The VIVA also provides an on-line user enrollment via an html-form and a telephony server to create the user profile.
and the voice-print. Further architectural details can be found in [9].

2.2. Experiments

Besides other design-oriented advantages, the main strength of the CSB concept lies in that it relies on two independent authentication sources and can thus achieve a higher accuracy and reliability. The overall false rejection (FR) rate of the CSB system resulting from the acoustic FR $p_{FR}(X)$ and the erroneous-dialog probability $p_{err}(D)$ (assuming correctly formulated answers by the user and also assuming the statistical independence of the two probabilities) within an interview can be estimated as [8]:

$$p(FR(X)) = p_{FR}(X) + (1 - p_{FR}(X))p_{err}(D)$$  

(7)

with the corresponding false acceptance (FA) rate of the CSB

$$p(FA(X)) \propto p_{FA}(X)\left(\frac{1}{M_q}\right)^k$$  

(8)

where $\left(\frac{1}{M_q}\right)^k$ stands for the expected perplexity of $k$ questions with $M_q$ possible answers. (8) neglects the probability of a correct answer from an incorrect input due to recognition error. In the case of $N$ questions in an interview and a minimum required number of $k$ correct answers the dialog error obeys the binomial distribution and, given the fact that the VIVA closes the interview already when the first $t = N - k + 1$ incorrect answers are detected, can be written as [9]

$$p_{err}(D) = \sum_{n=t}^{N} \binom{n-1}{t-1} p_{err}(q)^t (1 - p_{err}(q))^{n-t}$$  

(9)

whereby $p_{err}(q)$ is the probability of a question decoded incorrectly (assuming topic independence) and $t = N - k + 1$.

### Table 3

False acceptance and rejection rates for various interview policies. Calculated for an acoustic EER=5% (OFtimistic rows calculated for system-acquainted users with $P_{err}(q) = 0.05$ and an ac. FR=2.5%)

<table>
<thead>
<tr>
<th>Policy</th>
<th>FA %</th>
<th>FR %</th>
<th>Avg. Interview length in sec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2</td>
<td>$5 \times 10^{-9}$</td>
<td>8.4</td>
<td>20</td>
</tr>
<tr>
<td>6-4</td>
<td>$1.3 \times 10^{-11}$</td>
<td>7.1</td>
<td>40</td>
</tr>
<tr>
<td>3-2 OP</td>
<td>$1 \times 10^{-9}$</td>
<td>3.2</td>
<td>20</td>
</tr>
<tr>
<td>6-4 OP</td>
<td>$2.6 \times 10^{-11}$</td>
<td>2.7</td>
<td>40</td>
</tr>
</tbody>
</table>

To estimate the dialog recognition error a data collection was carried out on speakers who were asked to answer 25 speech biometrics questions to several topics: digits, years, cities, states, colors, hobbies and favourite food. some of the speakers were acquainted with the system, the rest were first-time users. The attribute perplexity varied from topic to topic. $M_q$ in (8) was dependent on the question topic: ranging from 20 (colors) over ca. 50 (years) to $> 10^6$ for digit strings. The overall answer correctness, defined as containing the correct answer without insertions of multiple incorrect values due to erroneous recognition, was 11.3% ranging from 0% for hobbies to 15% for cities and states. Note that empty answers (i.e. containing no relevant attribute values) representing ca. 5-8% of the answers were handled by an appropriate recovery dialog, thus reducing the $p_{err}(D)$ in (8). For the experienced users the answer error rate was less than 5%. Further experimental details can be found in [9].

For the calculations in Table 3 an equal-error-rate (EER) of the acoustic speaker recognition 5.0% and the realistic question perplexities $1/2 \times 10^5$ for digits, and $1/50$ or $1/20$ as representative values for other topics were used, assuming that in an interview with $k = 3$ there is one question for each of these perplexities, for $k = 4$ additional $1/50$ and for $k = 5$ additional $1/20$ factors were taken.

The Table 3 shows FA and FR rates for two security policies 3-2 and 6-4 according to the definition above. Allowing a small number of the answers to be incorrect prevents too high false acceptance due to speech recognition errors.

Smaller FR can be obtained by decreasing the acoustic FR (2.5%) entailing a higher FA rate (10%) and by assuming that users familiar with the system achieve $p_{err}(q) = 0.05$ (last two “optimistic” rows in Table 3). The acoustic EER might also be lower in reality, even though 5% was assumed in this calculation, especially for longer interviews where the amount of collected speech will increase, reducing the overall error rates correspondingly.

### 3. CONCLUSIONS

We have presented an acoustic speaker recognition system with refined speaker modeling using multi-grained models and a corresponding scoring technique and have demonstrated the advantages of integrating speaker recognition and conversational systems to implement voice-based biometrics. Appropriate design of the application allows to perform simultaneous content/knowledge-based recognition
with high accuracy even in challenging conditions or over very large populations. The acoustic voice-print recognition achieves equal-error rates of 2% and 5.5% for 3-5 sec and 1-2 sec of speech respectively. When combined with the knowledge-based authentication with the conversational speech biometrics framework, users familiar with the system can log in with 2.7% or 3.2% false rejection and ca. $3 \cdot 10^{-11} \%$ or $10^{-5} \%$ false acceptance rates in about 40 sec or 20 sec respectively which is an impressive result as compared to purely voice-print based authentication.

The concept of Conversational Speech Biometrics makes speaker recognition for the first time deployable for high security applications as a primary security key even with today's technology.

REFERENCES


