

# AN INSTANTIABLE SPEECH BIOMETRICS MODULE WITH NATURAL LANGUAGE INTERFACE: IMPLEMENTATION IN THE TELEPHONY ENVIRONMENT

Jiří Navrátil, Jan Kleindienst, Stéphane H. Maes

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA  
e-mail: {jiri,jankle,smaes}@watson.ibm.com

## ABSTRACT

This paper describes an implementation of the concept of conversational speech biometrics approach to personal authentication in telephony environment. An application-independent module including a NL-enabled part for verbal verification and identification and an acoustic speaker recognition engine for voice-print analysis are combined using a special verification/identification protocol which allows the application to adapt the session according to the dialog development and the query/command security. The results validate the feasibility and advantages of the concept of integrating speaker and speech recognition technology. Users familiar with the system can log into the system 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. This makes speaker recognition for the first time deployable for high security applications even with today's technology - a claim that can't be made with other speaker recognition technology. The system has a client-server architecture and is suitable for various applications and platforms.

## 1. INTRODUCTION

Personal authentication plays an essential role in all applications that are to be accessed by specific sets of users. Ranging from retina scans to finger prints, many modalities and techniques have been applied to achieve this task. Voice, as one of the modalities, is especially important in applications such as telephony dialog systems where it is the natural communication means and, besides the dialing keyboard, the only one available. Speaker recognition technology analyzing and modeling the speaker's voice prints has been a major research effort for the past decades and is reaching maturity. Extending the voice-print-based approach to a combined verification based on voice and knowledge has been pursuit in recent work [1][2][3][4][7][8]. In [2] a concept of conversational biometrics is presented which allows for enrolling, identifying and verifying a person by exploiting two different personal features by the system, namely what the person *is* (represented by the voice) as well as what he or she *knows* (knowledge of passwords or other person- and application-related information). The concept of conversational speech biometrics (CSB) brings several advantages to the application: 1) increased system robustness against impostors, 2) system flexibility in carrying out the authentication, e.g. adaptive length of a verification session dialog dependent on a voice-print confidence, extensibility to identification and speaking style adaptation/recognition, 3) possibility of continuous voice-print enrollment [1], 4) continuous verbal information collection (enrollment), 5) transparency and non-obtrusivity with respect to business-logic

dialog. Unlike the purely voice-print-based systems, the CSB is able to handle new types of channels without a-priori voice enrollment by first backing off to verbal verification with subsequent voice-print creation. The CSB appears especially natural in the framework of speech-enabled applications where the transition from the verification dialog to the application dialog can be designed seamless.

This paper describes an application-independent module that implements the CSB concept in the telephony environment incorporating a NL-enabled part for the verbal information verification, so-called Verbal Identification and Verification Agent (VIVA) and a speaker recognition engine for voice-print analysis. A special verification protocol is introduced which allows the application to adapt the session according to the dialog development and the query/command security. The system has a client-server architecture and is suitable for various applications and platforms.

## 2. VIVA ARCHITECTURE

The system overview of the Verbal Identification and Verification Agent (VIVA) is shown in Fig. 1. This client-server architecture consists of the following functional parts: 1) the VIVA server which handles requests for verbal verification sessions and maintains the database records, 2) the VIVA client interface, 3) the speech biometrics module which interacts with both the user and the application via a speech and a proprietary interface respectively. Further on this module requests sessions from the VIVA server, triggers the voice-print analysis, processes and combines both results. The following subsections detail on the individual components. In order to have additional application-oriented information about the users (such as nicknames, speaking style, type of queries and other preferences) the VIVA also accesses a "personality server" which, however, is not a principal component of the CSB. It can be used to improve the recognition by adding other conversational features to model the speaker (speech rate, accent, way to express requests, type of requests), as well as to customize or increase the personalization of the application or the conversational engines and to help disambiguate attribute-value pairs in the dialog.

### 2.1. Verification Sessions, Interviews

The VIVA server is the core of the verbal knowledge-based part of the CSB system as it maintains and accesses the database with users' personal information and it generates the CSB questions and evaluates user's answers. The server accepts NL text strings as input and carries out NL parsing tasks internally (see below). This makes the server independent of application, platform and speech domain (telephony, desktop). For the interaction between the server and its clients a special protocol has been developed involving the notions "session" and "interview" which provides the necessary flexibility of the verification sessions on one hand,

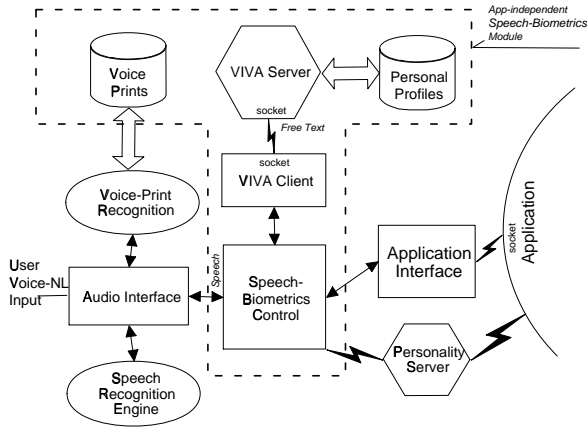


Figure 1. VIVA overview

and helps maintain security of the personal database on the other hand. All clients must first create a session for a given claimant. Within a single session a client may request one or several successive interviews for the session's claimant. The notion of an interview stands for an instance of a verbal verification trial for which certain parameters may be changed. In our case such parameters are the security policy determining the length of the interview and the minimum number of questions that must be answered correctly for a positive verification. Also, optional information about the topics of the questions to be generated may be passed along within an interview. The VIVA server alone decides about the outcome of an interview. The client, even though it is an integral part of the system, does not obtain information about whether or not a specific answer to a question was correct nor how many questions were finally answered correctly. Since the server is remote and the personal information is not passed through the interface, there is an additional security level against data capture and fraud. An important property of the session-interview protocol is the fact that the server keeps a complete record from all interviews within the session and designs the new interviews appropriately, e.g. it does not repeat the questions from previous turns. Note, that the questions are generated randomly from a larger question ensemble to prevent recording or eavesdropping and fraud.

Using multiple interviews the client can adapt the total length of the initial verification according to other criteria such as the voice-print-based confidence by opening several successive interviews with increasing security policy. Further on, after a successful initial verification and after passing the control to the actual application, the client may still collect the audio data from the user-application communication and, in case of a low voice-print confidence, may initiate an additional short interview for re-verification thus handling speaker changes.

As mentioned above, the client-server interaction is a text-based communication consisting of protocol commands and question/answer strings. To process the user's answers which are generally natural language a simplified NL processing was implemented within VIVA server to canonicalize the text input and extract the values for a given attribute (attribute-value-pair method [9]). The server then checks whether the values match user's database entries whereby it is capable of handling multiple correct answers for one question (e.g. favourite color might be "blue" or "white" or

both) and also detects "cheating", i.e. more than a certain number of values in the answer (e.g. "my favourite color is blue, red, white, green...").

## 2.2. Speech Biometrics Control

Via the VIVA client block the Speech Biometrics Control (see Fig 1) initiates verification sessions and individual interviews with the server. This control module is also an intermediary between the application and the user in that it handles all CSB dialogs with the user using speech recognition and informs the application about the final outcome of the verification evaluating both the verbal session and the voice-print.

The typical initial procedure looks like the following: The application creates an instance of the CSB module when the user tries to log on (e.g. an incoming call is detected). The CSB module then takes over the control and first tries to obtain the user's ID claim. This is achieved through a direct prompt or through an identification procedure based on voice and verbal information. In our implementation the claim ID is a digit string (extension number). If the user does not explicitly specify this number and starts talking to the application directly, the CSB module suspends the speech command and starts determining the claim in a short dialog using an open-set acoustic speaker identification. In case of an unsuccessful identification the user is prompted for the claim number explicitly. A similar identification procedure for large populations using dialogs for narrowing the target group with subsequent acoustic ID is described in [2]. Using the claim ID the speech biometrics control creates a verification session and an initial interview with the VIVA server. Questions generated by the server are synthesized for the user and user's response is decoded using a telephone-speech recognition engine. Since the VIVA server also passes additional information about the *topic* associated to each question (e.g. "color", "digit", "year", "hobby" etc.), appropriate vocabularies and grammars can be switched by the control module on a question basis. Decoded answer is returned back to the server for evaluation. In parallel, the control module collects the audio data in a buffer. Once the security policy for the open interview is satisfied the server returns a positive result and closes the verbal interview. Subsequently the control module triggers the voice-print verification with the collected audio.

The interview is closed with a negative result in case there were too many ambiguous or wrong answers (given an interview policy) which results in the control module terminating the complete session and rejecting the user.

After this initial interview the control is given back to the application. The instance of the CSB module may be terminated if the application does not require any further authentication or may remain persistent. In the latter case the speech biometrics control creates a listener associated with the audio stream and collects the speech from the regular user-application communication. Then, a voice-print re-verification can be requested from the application at any time (especially before committing crucial operations) thus achieving continuous speaker tracking and detecting potential speaker changes.

## 2.3. Speaker and Speech Recognition

For voice-print analysis our VIVA implementation uses a text-independent speaker recognition engine as described in [5][6]. Via the SVAPI interface [11] the engine can be used for speaker identification, verification and enrollment. Each speaker is modeled by means of Gaussian mixture models estimated from Mel-Frequency Cepstral Coefficients and their derivatives, transformed using a Maximum Likelihood Linear Transform. The performance of the acoustic speaker verification is given in Sec. 4.

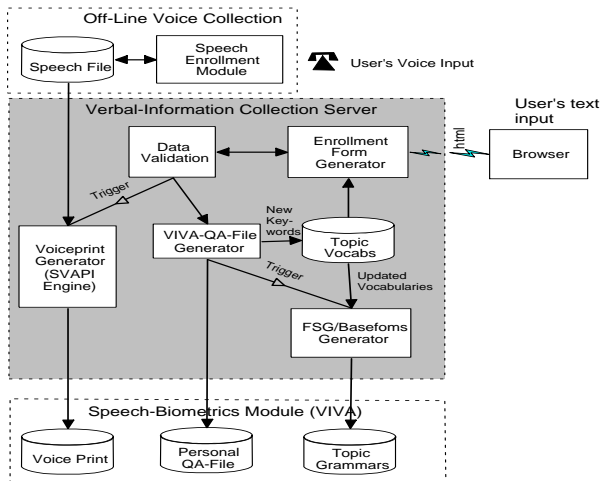


Figure 2. On-Line Enrollment

For decoding the user's responses the IBM ViaVoice Telephony engine [10] was used. In each question-answer turn a special topic-dependent finite-state grammar was activated aiming at detecting all possible target values of this topic-attribute as well as allowing the user to formulate the answers in a natural way. Additional control actions, such as requests for repeating the questions, were also added. Statistical NL processing can also be used as described in [9][10].

### 3. ENROLLMENT

Since the CSB concept involves a twofold information acquisition: verbal knowledge and the voice-print, a combined enrollment procedure was developed in our implementation including a telephone voice enrollment and an on-line verbal enrollment using an html form. The structure of the enrollment is shown in 2.

As a first step the new user has to record his/her speech by calling a telephone collection script. Once there is an existent recording the user is allowed to invoke an enrollment form and specify personal data, such as passwords and answers to questions on various topics, as suggested by the server. The answers can be selections from predetermined value lists, e.g. selected cities or colors, or user's own new keywords. It is also possible to add new questions within the existent topics or dynamically generate these based on contexts or history of previous transactions or other events. After validation the form data is used for generating user's VIVA profile (the question-answer (QA) file) which will be used by the VIVA server during runtime. Further on, all new keywords are included in the system attribute-value registry (NL files) and their pronunciation baseforms with the corresponding FSG are generated. In parallel, the voiceprint enrollment starts based on the pre-recorded speech. If the generation process was successful the user can log into the system immediately.

### 4. RESULTS

In order to assess the implemented system we first consider the acoustic and the verbal verification separately and

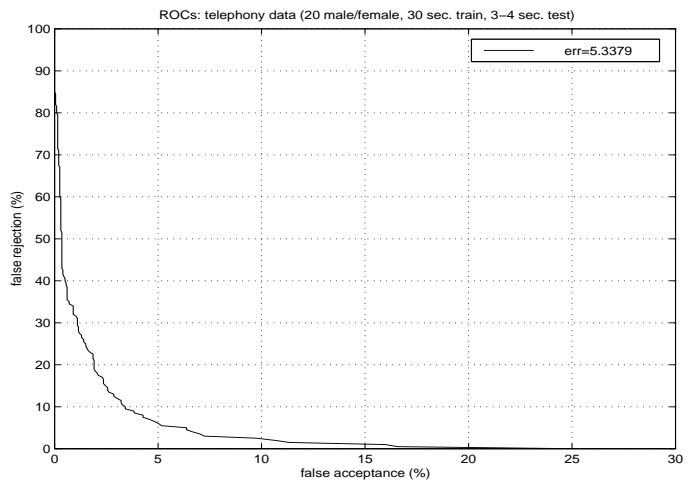


Figure 3. The ROC curve of the speaker recognition engine

then evaluate the overall performance of the CSB module.

A detailed description of the speaker recognition engine can be found in [5][6]. Fig. 3 shows the ROC curves measured on the Lincoln-Lab-Handset-Database evaluation data consisting of 20 target speakers and telephone-quality speech. The speakers enrolled using the "rainbow" sentences (ca. 30 sec). The test utterances were ca. 3-4 sec long resulting in an acoustic EER=5.3%. The acoustic EER will significantly decrease on longer test utterances which is what happens when collecting utterances during a dialog. Thus, the final error rates calculated below may be lower in reality.

It is clear that accuracy of the speech recognition has an impact on the false rejection rate of verbal part of the CSB. Inaccurately decoded answers containing wrong or no values will be considered as incorrect by the VIVA server. The overall false rejection (FR) rate of the CSB resulting from the acoustic FR  $p_{FR}(X)$  and the erroneous-dialog probability  $p_{err}(\mathcal{D})$  (assuming correctly formulated answers by the user) within an interview can be estimated as [2]:

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

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

$$p(FA(X)) \propto p_{FA}(X) \left(\frac{1}{M_q}\right)^k \quad (2)$$

where  $\left(\frac{1}{M_q}\right)^k$  stands for the expected perplexity of  $k$  questions with  $M_q$  possible answers. 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

$$p_{err}(\mathcal{D}) = \sum_{l=N-k+1}^N \binom{N}{l} p_{err}(q)^l (1 - p_{err}(q))^{N-l} \quad (3)$$

whereby  $p_{err}(q)$  is the probability of a question answered incorrectly (assuming topic independence). However, in the reported experiment an interview policy was implemented such that the interview is closed already when the  $l = N - k + 1$  incorrect answers are detected. Thus the dialog error calculation (3) must take the variable interview length into account as a probability of observing the  $l$ -th

Policy	FA %	FR %	Avg. Interview length in sec.
3-2	$5 \cdot 10^{-6}$	8.4	20
5-4	$1.3 \cdot 10^{-11}$	14.6	35
5-3	$2.5 \cdot 10^{-7}$	6.2	30
6-5	$2.5 \cdot 10^{-15}$	18.4	45
6-4	$1.3 \cdot 10^{-11}$	7.1	40
3-2 <i>OP</i>	$1 \cdot 10^{-5}$	3.2	20
6-4 <i>OP</i>	$2.6 \cdot 10^{-11}$	2.7	40

**Table 1. False acceptance and rejection rates for various interview policies. Calculated for an acoustic EER=5% (*OP*timistic rows calculated for system-acquainted users with  $P_{err}(q) = 0.05$  and an ac. FR=2.5%)**

error and ending the interview:

$$p_{err}(D) = \sum_{n=l}^N \binom{n-1}{l-1} p_{err}(q)^l (1 - p_{err}(q))^{n-l} \quad (4)$$

where  $l = N - k + 1$ .

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. Disregarding the grammar part modeling the spontaneous non-value speech,  $M_q$  in (2) was dependent on the question topic: ranging from 20 (colors) over ca. 50 (years) to  $> 10^6$  for digit strings. It has to be noted that no exact value of the real attribute perplexity can be determined because 1) the NL-part of the FSG adds a certain perplexity and can catch some out-of-vocabulary values by decoding them as non-value words, i.e. the decoding is open-set (the same applies to statistical NL modeling), 2) the perplexity of certain attributes, e.g. years, is reduced by their meaning and predictability in real context. 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) were considered as incorrect (representing ca. 5-8% of the answers) which might be potentially recovered by an appropriate dialog, thus reducing the  $p_{err}(D)$  in (2). It has to be noted that the first-time users spoke highly spontaneously and inserted longer passages containing no relevant attribute values. For the experienced users the answer error rate was less than 5%.

By loosening the interview policy strictness it is possible to compensate for dialog errors arising from speech recognition errors. The Table 1 shows FA and FR rates for various policies in which the first parameter stands for the maximum number of questions to be asked and the second for minimum number of correct answers. The average interview length takes into account that in cases of satisfying the minimum policy requirement the interview is closed with success before the maximum number of questions is reached.

For the calculations in Table 1 an equal-error-rate (EER) of the acoustic speaker recognition 5.0% and the realistic question perplexities  $1/2 \cdot 10^4$  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.

Smaller FR can be obtained by increasing the acoustic FA to 10% entailing acoustic FR of 2.5% and by assuming that users familiar with the system achieve  $p_{err}(q) = 0.05$  (last two "optimistic" rows). Higher acoustic FA rates, however, can make the system vulnerable to data fraud. As mentioned above, with longer utterances collected over several dialog turns the total amount of speech will often be greater than 3-4 sec resulting in corresponding decrease of the EER.

#### 4.1. Conclusion

The results obtained using our conversational speech biometrics system in telephony environment prove the feasibility and robustness of the concept of closely integrating speaker and speech recognition technology. Users familiar with the system can log into the system 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. Using additional recovery mechanisms the dialog errors may be further reduced. With today's technology the CSB allows to design dialog for wide range of preset operating points. The protocols introduced in the VIVA system make this design flexible for various verification procedures while maintaining security and robustness against fraud. The CSB makes speaker recognition for the first time deployable for high security applications as a primary security system even with today's technology - a claim that can't be made with other speaker recognition technology.

#### REFERENCES

- [1] S. Maes, D. Kanevsky, "Apparatus and methods for speaker verification/identification/classification employing non-acoustic and/or acoustic models and databases," Pat. US 5897616, June 1997.
- [2] S.H. Maes, "Conversational Biometrics," In Proc. of Eurospeech, Budapest, Hungary, September, 1999.
- [3] S.H. Maes, "Conversational Biometrics: integration of conversational dialog systems and text-independent speaker recognition," Submitted to IEEE Computer, Special Issue on Biometrics, 1999.
- [4] S.H. Maes, H.S.M. Beigi, "Open Sesame! Speech Password or Key to Secure Your Door," In Proc. ACCV, January, 1998. Invited paper.
- [5] U.V. Chaudhari, H.S.M. Beigi, S.H. Maes, J.S. Sorensen, "Multi-Environment Speaker Verification," in Proc. AUTOID'99, New Jersey, 1999.
- [6] U.V. Chaudhari, S.H. Maes, "Pattern-Specific Maximum Likelihood Transformations and Speaker Recognition With Sparse Training Data," Preprint - 1999.
- [7] Q. Li, B.-H. Juang, Q. Zhou, C.-H. Lee, "Verbal Information Verification," In Proc. of Eurospeech, Vol. 2, pp. 839-842, 1997.
- [8] Q. Li, B.-H. Juang, "Speaker Verification Using Verbal Information Verification for Automatic Enrollment," In Proc. of ICASSP-98, Seattle, WA, May, 1998, pp. 133-136.
- [9] K.A. Papineni, S. Roukos, R.T. Ward, "Free-flow dialog management using forms," In Proc. of Eurospeech, Budapest, Hungary, September, 1999.
- [10] K. Davies et al., "The IBM conversational telephony system for financial applications," In Proc. of Eurospeech, Budapest, Hungary, September, 1999.
- [11] Speaker Verification API, <http://www.srapl.com/svapi>