

Conversational Speech Biometrics

Stéphane H. Maes, Jiří Navrátil, and Upendra V. Chaudhari

IBM T.J. Watson Research Center Rt. 134, Yorktown Heights, NY, USA
{smaes,jiri,uvc}@us.ibm.com

Abstract. This paper discusses a new modality for speaker recognition - conversational biometrics - as a high security voice-based authentication method for E-commerce applications. By combining diverse simultaneous conversational technologies, high accuracy transparent speaker recognition becomes possible even in channel or environment mismatches. For speaker identification over very large populations, we combine dialogs to reduce the set of confusable speakers and text-independent speaker identification to pin-point the actual speaker. Similarly, dialogs with personal random or predefined questions are used to perform simultaneously knowledge-based and acoustic-based verifications of the user. Adequate design of the dialog allows to tailor the ROC curves to the needs of most applications. We demonstrate the conceptual advantages using our telephony prototype. Users familiar with the system can log into the system with 0.8% or 1.3% false rejection and ca. $5 \cdot 10^{-12}\%$ or $2 \cdot 10^{-6}\%$ false acceptance rates in about 40 sec or 20 sec respectively which is an impressive result as compared to purely voice-print based authentication.

1 Introduction

With the rapid development of automated applications for finances and E-commerce and in the context of the evolving internet and wireless communication technology, the importance of reliable, high-security, and non-intrusive methods for personal authentication has been growing significantly. Today, the quality of these methods plays an essential role for the acceptability and ease of use of the target applications. Many modalities and techniques have been applied to achieve the task of authentication, ranging from retina scans to finger prints. In this paper, a particular modality is of interest - the voice. This modality has a unique advantage over other biometrics by relying on speech, the primary vector of communication and is especially important in applications such as telephony dialog systems where it is a natural and, besides the touchtone keypad, also the only communication means. By extracting appropriate features from a person's voice the uniqueness of the physiology of the vocal tract and the articulatory properties can be captured to a high degree and can serve the purpose of authentication. Speaker recognition technology analyzing and modeling the voice prints has been a major research effort for the past decades, today gradually reaching maturity. Despite impressive results multiple unknown factors in the

acoustic speaker recognition still exist: unicity, uncooperative speakers, robustness etc. Multiple commercial systems are already available, in most cases, however, as field prototypes or secondary systems. Text-constrained methods which are technically more simple and achieve higher accuracies, are prone to fraud by recorded speech or using future-generation speech synthesizers which can mimic the target person. Text-constrained methods are also intrusive and therefore can not be closely integrated within an application dialog flow: either the speaker recognition is performed as a separate process or it is performed once in the business logic, but it is not an underlying process or always invocable option. Text-independent systems, on the other hand, are technically more challenging, and the accuracy rates may be somewhat lower compared to text-dependent systems, but they open new perspectives and application possibilities. In particular these methods are non-obstrusive. As a result they can be closely integrated within a dialog, run in parallel, contribute to the dialog flow or application business logic and be invoked at any time. In general, the voice-print recognition accuracy tends to deteriorate in adverse acoustic conditions, such in the telephony environment which introduces highly variable and unknown transducer properties to the source speech. Hitherto, the problems of robustness and accuracy have been major obstacles for deploying voice-print based speaker recognition for remote authentication applications. In this paper a concept of combining two authentication modalities, the voice-print and the speaker's knowledge is presented that allows for flexible identification and verification with a high degree of security: a concept called Conversational Speech Biometrics [16, 17]. The following sections detail on the principles of this concept and its both functional components: the speaker and the speech recognition technologies. Corresponding application scenarios together with a description of a prototype implementation in the telephony environment and experimental results are presented. We show that the speech biometrics is a powerful framework for remote authentication which enables speech, as a single communication modality, to serve as a primary security key for a wide range of applications.

2 Conversational Biometrics

Classical authentication relies on one of these three items: what you own, what you are and what you know. Key or card-based systems characterize what you own. PIN and password based systems rely on what you know. Voice passwords have been proposed: utterance verification for access control and password compliance [22, 15, 13]. Biometrics and in particular speaker recognition rely on what you are. The new approach of speech biometrics or conversational biometrics [17] employs text-independent speaker recognition to acoustically identify or verify answers from the user in dialog with the system. The questions addressed to the user can be randomly selected, follow a pre-defined sequence or follow a business logic. With this approach, user verification and identification rely on acoustic recognition and on the content of the answers to the questions. Be-

yond eliminating the problem of prerecorded speech and increased security, this combination has many application-related advantages as will be discussed later.

2.1 Acoustic Speaker Recognition

The speaker recognition problem can be divided into four different functional modi:

- *Speaker identification*, aiming at determining the identity of a speaker based on his or her voice. The speakers are already enrolled in the system. No identity claim is provided. If the set of speakers to be identified is restricted to be the enrolled speakers, we speak of closed-set identification. The ability of the system to also detect unknown speakers extends the task to so-called open-set speaker identification.

In terms of biometrics, speaker identification is a “many-to-many” recognition task. The decision alternatives are equal to the size of the enrolled speakers (+ 1 in open-set case). Therefore, the accuracy of speaker identification degrades as the size of the speaker population increases.

Besides classical speaker identification, some extensions exist with added functionality of providing N -best lists or confidence scores. In the former case, a speaker identification system returns a sorted list of N identities that match the best the current speaker. The latter case rather implies that the identifier will produce a confidence level for each enrolled speaker that he or she matches the current speaker. Open set speaker identification requires rejection features that can usually be directly used for verification purposes. The recognition rates for closed-identification range from ca. 95% for small populations (100 speakers) to 70-90% for large populations (few thousands of speakers) based on 3-5 sec telephony-quality speech [6].

- *Speaker verification*, a task of verifying the identity claim of a speaker based on his or her voice.

In terms of biometrics, speaker verification is a “one-to-many” recognition task. In contrast to speaker identification, the accuracy of speaker verification does not directly depend on the population size. However, as it is typical in biometrics, the estimate of this accuracy depends on the representation of the population samples used to evaluate the accuracy. In contrast to other biometrics, these estimators also strongly depend on the channel effects and noise corruption of the signal. As mentioned above, speaker recognition performances vary dramatically from matched conditions (same type of microphone, channel characteristics and background noise) to mismatched conditions.

Besides classical speaker verification, we must also mention extensions where instead of hard accept or reject decisions, confidence levels are returned. Typical performances, represented as equal-error-rates, lie between 2 and 5% for 2-4 sec telephone-quality speech. [6]

- *speaker classification*, performing the speaker recognition over an unknown number of unknown (unenrolled) speakers. Usually, it means to be able to

detect speaker changes, also called speaker separation, and index the resulting segments according to the identity.

This function is specifically speech related. Only portions of the concept are met in other biometrics. However, the capabilities that it offers to distinguish between different undeclared successive users of a system may also be implemented with other biometrics.

- *speaker enrollment*. In order to recognize the user based on his or her voice, samples of the user's voice need to be acquired and the speaker model (voice-print) created. Often, the models used for speaker identification differ from those used for speaker verification. By analogy to fingerprints, voice-prints refer to the minimum set of characteristics of a speaker required to create the speaker models used for identification and verification. Voice-prints are algorithm-dependent.

Similarly to speech recognition, the principle here is that there is no better enrollment data than more data! The more data available for a speaker the more accurate the resulting voice-prints. Especially if this data can be collected over multiple mismatched conditions representative of the actual mismatches experienced during recognition. [5]

Further on, the task complexity can be distinguished w.r.t. the type of vocabulary presented in the enrollment and during the recognition. Text-dependent and text-constrained recognition restricts the words to be spoken to a certain small set, e.g. a password (global or user-selected), or a digit string. Similarly, text-prompted systems restrict the input utterance whereby the words to be spoken are generated by the system itself, which reduces the chances of using prerecorded speech. Finally, the text-independent speaker recognition offers most freedom as for the use of vocabulary and belongs to the technically most demanding tasks. As for the conversational speech biometrics the text-independency is an essential feature as it allows for analysis of all the user-application conversation regardless of whether related or unrelated to the act of authentication, e.g. as a continuous background listener.

The literature on voice-print modeling and recognition comprises areas of template-matching, statistical modeling and artificial neural networks [1, 8, 19, 9, 12, 10, 11, 4, 16]. Our speaker recognition engine is based on structured speaker modeling using Gaussian mixture models in all four functional modi listed above [3, 5, 2, 6]. Depending on the text-modus, the voice-prints are created in the enrollment stage from the user's speech transformed to a sequence of feature vectors and collected over several different channels. Typically, the amount of enrollment speech ranges between 30 and 120 sec. The incoming speech is internally segmented into specific phonetic units on multiple levels of granularity (e.g. phone level, phone-class level, global level) using speaker-independent Hidden Markov Models. The Gaussian mixtures are with diagonal covariances and are initialized with estimates from a global, speaker-independent model (seed) in order to alleviate the problem of data sparseness, which strongly applies with the enrollment amounts mentioned above. For each model grain unit a linear feature transform is estimated so as to minimize the loss of likelihood mass due to the diagonal

covariance assumption. During the test, a likelihood measure between the test utterance and the voice-print is calculated as the accumulated maximum observation probability of the feature vectors over all granularity levels and their associated units [6]. Whereas the identification consists of calculating the test likelihoods based on all models of enrolled speakers, the verification is posed a hypothesis test with a discriminant function calculated as a likelihood-ratio test. In the hypothesis test the target speaker's likelihood is obtained from the target voice-print and the non-target speaker hypothesis is represented by the likelihood of a certain number of competing models (cohorts). The cohorts are determined either in the enrollment or on-the-fly during the test. Combining both methods the identification and the verification, also the open-set identification task can be performed in order to detect (and to reject) unknown speakers.

Typical performance rates measured for telephone-quality speech for the described speaker recognition engine, using 30 sec speech for enrollment and ca. 3-5 sec utterances for testing. On a population of 100 speakers the identification error is 4.8% and increases to 10.0% with a larger population size consisting of 1000 speakers [6]. The text-independent verification performance measured in the operating point of equal level of false acceptances and false rejections (equal error rate) for 3 sec tests is ca. 2.0%. Fusion with additional decisions based on algorithms to be disclosed elsewhere can further reduce this number to 1.2% These values will also be used for estimating the performance of the overall speech biometrics system in connection with the experiments described in section 3.

2.2 Speech Biometrics: Integration of Speaker Recognition and Speech Recognition

Consider the system described in figure 1, which simultaneously performs speech recognition and speaker recognition on the input utterances. The audio stream is provided to the acoustic front-end as isolated utterances (command and control mode or answers to a directed dialog) or as a continuous stream. The front-end captures the audio and extracts the acoustic features (e.g. MFCC). The features can be further compressed using the algorithm described in [21]. Note that the acoustic features can be shared by the speech recognition engine and any post processing system (e.g. natural language understanding module [20]). In networked applications, where the acoustic front-end can be on the client side while the other conversational functions are performed on the server side, sharable features allow one data stream at data rates as low as 4 to 5 kb/s, quite suitable for wireless modem connections. On embedded systems, only one signal processing task is performed. This reduces the CPU, memory and power requirements. The feature stream is then split up between the speech recognition engine and the text-independent speaker recognition engine.

Besides numerous advantages for speaker adaptation in the speech recognition engine and command disambiguation, the integrated framework offers the basis for implementing the conversational speech biometrics (CSB) concept. Simultaneous speech recognition and text-independent speaker verification can be

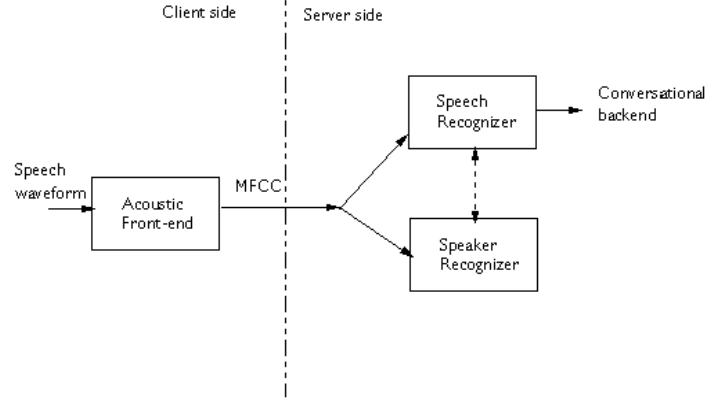


Fig. 1. Integration of speech and speaker recognition engines.

used for continuous access control. For example, in a command and control application or directed dialog, each command or transaction request can be executed only upon verification of the speaker. The verification can be performed on a continuous streaming input, on a command by command basis or on a set of utterances. This also provides continuous background monitoring capabilities to certify that no speaker change took place during a transaction, or after the authentication.

Obviously, such integration of the speaker recognition capability allows transparent recognition in a non-obtrusive manner to the user and the transaction. Also, since it is well known that with more data a more robust recognition can be achieved, it is particularly advantageous to postpone recognition decisions later in the transaction when a final decision is required.

Speech biometrics, as we originally called it [16] requires a close integration of the text-independent engine with the entire conversational system. As illustrated in figure 2 conversational systems consist of speech recognition, speech synthesis, natural language understanding, natural language generation and dialog management [20, 7]. Indeed, the dialog management now carries a conversation with the speaker aimed at automatically identifying a cooperative user or verifying a claimant.

Conversational identification consists of a dialog that reduces the set of confusable speakers handled by the speaker identification engine, assuming cooperative users. For example, an IVR (Interactive Voice Response) system interrogates the speaker as follows:

- *"What is your name"*
- I am John Doe
- *"What city are you calling from?"*

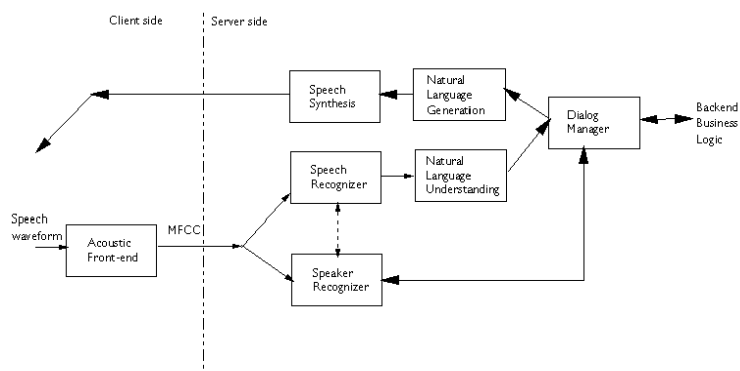


Fig. 2. Conversational biometrics architecture.

- I am in Manhattan
- ...

By now, out of the pool of millions of users enrolled in the system, the dialog has reduced the set of candidates to a subset for example smaller than say twenty to fifty speakers. If the sub-set is still bigger the dialog can continue. Text-independent speaker identification can now benefit from the reduced number of candidates and can perform identification or generate a N -best candidate list [17].

Conversational verification consists of a dialog to perform a knowledge-based verification of the person in parallel with the acoustic-based verification. It is a powerful mechanism to combat the limitations still inherent to speaker verification systems. Consider an automated phone banking application driven by an IVR. The following dialog takes place:

- *“What is your mother’s maiden name?”*
- My mother’s name was Doe1
- *“What is your favorite color?”*
- I like red
- ...

The questions can be randomly generated out of information collected during enrollment or they can be dynamically generated based on past transactions history. With an appropriate recovery dialog the false rejections can be reduced to an arbitrary level.

It has been shown empirically [23] that this method allows authentication based on information that is easy to remember for the enrolled user, while at the same time being hard to guess. These “cognitive” passwords are easy to remember because they are based on factual events regarding the user and on his or her opinions. While some of these data are known to others, the amount

of unrelated information is large. It was thus observed that only a small fraction of this information could be guessed, even by persons of close relationship to the enrolled user. In comparison, standard passwords, which are conventionally alphanumeric strings that are either user generated or randomly constructed, were difficult to remember and insecure by virtue of the steps users took in order to be able to remember them. Speech biometrics provides an additional benefit by leveraging information in the acoustic signal to make the overall system even more robust.

The described verification scenaria can be exercised and are applicable to virtually all E-commerce transactions with users acting in networks with voice or mixed voice-data connections. One example is a CSB verification that follows a transaction request previously completed in a purely data-based network connection. The verification is achieved by creating an automated voice connection from the network to the user terminal, typically a mobile phone, in order to carry out a CSB session. CSB and the steadily growing number of mobile-phone terminals, their convergence with the internet involving a variety of applications such as banking and shopping, represent a particularly attractive basis to establish a universal voice-supported user authentication modality.

The following items summarize the advantages of the concept:

- improved system robustness against impostors
- 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
- possibility of continuous voice-print enrollment. 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
- continuous verbal information collection (enrollment)
- transparency and non-obtrusivity with respect to business-logic dialog

3 Implementing CSB: The Voice Identification and Verification Agent

The practicability and performance of the CSB concept has been studied by implementing the principles described above in the telephony environment and measuring the performance on an authentication task with real speech. In this framework, an application-independent module was developed [18] that incorporates a natural-language-enabled part for the verbal information verification, so-called Verbal Identification and Verification Agent (VIVA), and the speaker recognition engine (see Section 2.1 engine for the voice-print analysis).

The VIVA system (see Fig. 3) has a client-server architecture and is suitable for various applications and platforms because of its independence in terms of maintaining own databases and handling dialogs. The architecture consists of the following functional parts (Fig. 3) 1) the VIVA server which handles requests

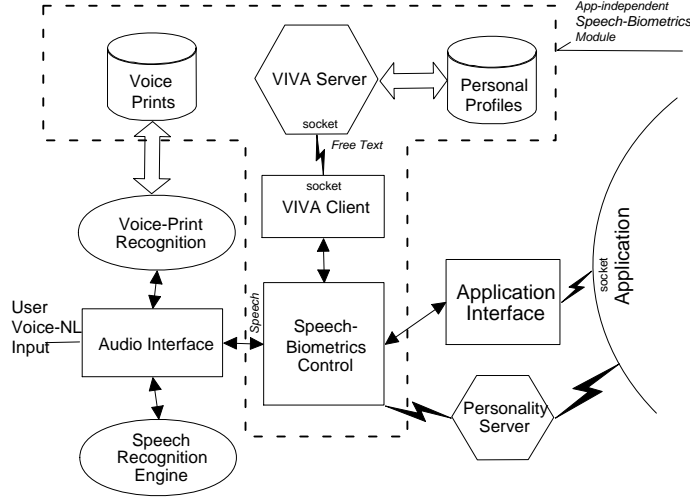


Fig. 3. VIVA overview

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. For the communication between a VIVA client and the server a proprietary protocol was introduced involving the notion of a verification “session” and “interview.” An interview represents an elementary dialog consisting of a small number of questions given a security policy. The policy can be defined as the ratio of the maximum number of questions and the minimum number of correct answers per interview. Within one session, multiple interviews with varying policies can be opened which allows for adapting the session length to the current voice-print match confidence. The questions asked within one session are generated so as to prevent repetitions across interviews and also to guarantee for sufficient security by an appropriate topic coverage, e.g. there will be at least one password question among questions about family, hobbies, or favourite colors.

The typical procedure looks like the following: The application creates an instance of the CSB VIVA module when the user tries to log on (e.g. an incoming call is detected). The VIVA then takes over the control and first tries to obtain the user’s ID claim. This is achieved through a direct prompt or using 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 issues commands to the application

directly, the VIVA 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. 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 the IBM ViaVoice Telephony engine [7]. Appropriate vocabularies and grammars are switched by the biometrics control module on a question basis according to the current topic. 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 (session remains still open). The control module triggers the voice-print verification based on the the collected audio and subsequently decides whether to accept the speaker, or to continue the verification session by creating an additional interview, or possibly to reject the speaker due to too many unsuccessful interviews (incorrect or ambiguous answers) or a too poor acoustic match.

After this initial verification, if closed positive, 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 instantiated. 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.

The VIVA system supports an automated user enrollment via HTML for the knowledge database and a telephone voice collection for the acoustic information.

3.1 Experimental Results

It is obvious that accuracy of both the speaker and the speech recognition has an impact on the false rejection rate of the overall CSB system. 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 system resulting from the acoustic FR $P_{FR}(X)$ and the FR due to erroneous-dialog with probability $P_{FR}(\mathcal{D})$ (both conditioned on the acoustic and dialog analysis and assuming correctly formulated answers by the user in the dialog) within an interview can be estimated as [17]:

$$P(FR(X)) = P_{FR}(X) + P_{FR}(\mathcal{D}) - P_{FR}(X, \mathcal{D}) \quad (1)$$

$$= P_{FR}(X) + P_{FR}(\mathcal{D}) - P_{FR}(X)P_{FR}(\mathcal{D}) \quad (2)$$

in which the the acoustic voice-print performance and the rejection rate due to dialog errors are seen statically independent. The corresponding false acceptance (FA) rate of the CSB is written as

$$P(FA(X)) \propto P_{FA}(X) \frac{1}{M_q^k} \quad (3)$$

where $\frac{1}{M_q^k}$ stands for the expected perplexity of k questions with M_q possible answers which approximates the probability of false acceptance in a dialog due to a correct random guess of an impostor as well as due to a speech recognition error. In general, it is possible to apply various policies to combine the outcomes of both the acoustic-based and the dialog-based match. For example, using the information about the quality of the current acoustic channel the decision might rely more on the verbal part of the verification without causing a too high false rejection due to poor acoustic channel. In fact, the equations (2) and (3) represent an upper bound for the FR and a lower bound for the FA respectively. 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

$$P_{FR}(\mathcal{D}) = \sum_{l=N-k+1}^N \binom{N}{l} p_{err}(q)^l (1 - p_{err}(q))^{N-l} \quad (4)$$

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 (4) must take the variable interview length into account as a probability of observing the l -th error and ending the interview:

$$P_{FR}(\mathcal{D}) = \sum_{n=l}^N \binom{n-1}{l-1} p_{err}(q)^l (1 - p_{err}(q))^{n-l} \quad (5)$$

where $l = N - k + 1$.

To estimate the dialog recognition error a data collection was carried out [18] 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 (3) 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 natural-language-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 $P_{FR}(\mathcal{D})$ in (3). For the experienced users the answer error rate was less than 5%. Further experimental details can be found in [18].

For the calculations in Table 1 an equal-error-rate (EER) of the acoustic speaker recognition 2.0% 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.

The Table 1 shows FA and FR rates for various security 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. Allowing a small number of the answers to be incorrect prevents too high false acceptance due to speech recognition errors.

Policy	FA %	FR %	Avg. Interview length in sec.
3-2	$2 \cdot 10^{-6}$	5.1	20
5-4	$5.0 \cdot 10^{-12}$	11.9	35
5-3	$1.0 \cdot 10^{-7}$	3.2	30
6-5	$1.0 \cdot 10^{-13}$	15.8	45
6-4	$5.0 \cdot 10^{-12}$	4.2	40
3-2 <i>OP</i>	$4.0 \cdot 10^{-6}$	1.7	20
6-4 <i>OP</i>	$1.0 \cdot 10^{-11}$	1.2	40

Table 1. False acceptance and rejection rates for various interview policies. Calculated for an acoustic EER=2.0% (*OP*timistic rows calculated for system-acquainted users with $p_{err}(q) = 0.05$ and an ac. FR=1.0%)

Smaller FR can be obtained by decreasing the acoustic FR (1.0%) entailing a higher FA rate (4.0%) and by assuming that users familiar with the system achieve $p_{err}(q) = 0.05$ (last two “optimistic” rows in Table 1). The acoustic EER might also be lower in reality, even though 2.0% was assumed in this calculation, especially for longer interviews where the amount of collected speech exceeds 3 sec the EER will be roughly halved, reducing the overall error rates correspondingly. Further improvements of the acoustic EER to 1.2% (or roughly 0.6 and 2.4% for the optimistic operating point), as mentioned earlier, change the last two rows to $FA/FR = 2.2 \cdot 10^{-6}\%/1.3\%$ and $FA/FR = 5.5 \cdot 10^{-12}\%/0.8\%$ for the policies 3-2 and 6-4 respectively.

4 Conclusion

We have demonstrated the advantages of integrating speaker recognition and conversational systems to implement conversational 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 results obtained using our telephony prototype prove the feasibility and robustness of the CSB concept. Users familiar with the system can log into the system with 0.8% or 1.3% false rejection and ca. $5 \cdot 10^{-12}\%$ or $2 \cdot 10^{-6}\%$ 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 system even with today's technology - a claim that can't be made with other speaker recognition technology.

References

1. B. S. Atal. Automatic recognition of speakers from their voices. *Proc. IEEE*, 64:pp. 460–475, 1976.
2. H. S. Beigi, S. H. Maes, J. S. Sorensen, and U. V. Chaudhari. A hierarchical approach to large-scale speaker recognition. In *Proc. Eurospeech*, 1999.
3. H. S. M. Beigi, S. Maes, and J. Sorensen. A frame-based statistical method for speaker recognition. In *Proc. RLA2C*, Avignon, France, April 1998.
4. J. Campbell. Automatic speech and speaker recognition, advanced topics. In Lee et al. [14].
5. U. V. Chaudhari, H. S. Beigi, and S. H. Maes. Multi-environment speaker verification. In *Proc. AutoID*, 1999.
6. U.V. Chaudhari, J. Navrátil, and S.H. Maes. Multi-grained data modeling for speaker recognition with sparse training and test data. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, October 2000. submitted.
7. K. Davies and al. The IBM conversational telephony system for financial applications. In *Proc. Eurospeech*, 1999.
8. G. R. Doddington. Speaker recognition - identifying people by their voices. *Proc. IEEE*, 76(11):pp. 1651–1664, 1985.
9. K.R. Farrell, R.J. Mammone, and K.T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 2(1):194–205, January 1994.
10. S. Furui. Automatic speech and speaker recognition, advanced topics. In Lee et al. [14].
11. S. Furui. Recent advances in speaker recognition. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Proc. Audio- and Video-based biometric person authentication*, pages 237–252. Springer-Verlag, 1997.
12. S. Furui and M. Sondhi, editors. *Advances in speech signal processing*. Marcel Dekker, New York, NY, 1991.
13. O. Kimball, M. Schmidt, H. Gish, and J. Waterman. Speaker verification with limited enrollment data. In *Proc. Eurospeech*, volume 2, pages 967–970, 1997.
14. C.-H. Lee, F. K. Soong, and K. K. Paliwal, editors. *Automatic speech and speaker recognition, advanced topics*. Kluwer Academic Publishers, Norwell, MA, 1996.
15. Q. Li, B.-H. Juang, Q. Zhou, and C.-H. Lee. Verbal information verification. In *Proc. Eurospeech*, volume 2, pages 839–842, 1997.
16. S. H. Maes and H. S. Beigi. Open Sesame! Speech password or key to secure your door. In *Proc. ACCV*, 1998. invited paper.

17. S.H. Maes. Conversational biometrics. In *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Budapest, Hungary, 1999.
18. J. Navrátil, J. Kleindienst, and S.H. Maes. An instantiable speech biometrics module with natural language interface: Implementation in the telephony environment. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000. IEEE.
19. D. O'Shaughnessy. Speaker recognition. *IEEE ASSP Magazine*, 3(4):pp. 4–17, October 1986.
20. K. A. Papineni, S. Roukos, and R. T. Ward. Free-flow dialog management using forms. In *Proc. Eurospeech*, 1999.
21. G. Ramaswamy and P. Gopalakrishnan. Compression of acoustic features for speech recognition in network environments. In *Proc. ICASSP*, volume 2, pages 977–980, 1998.
22. A. E. Rosenberg and S. Parthasarathy. Speaker identification with user-selected password phrases. In *Proc. Eurospeech*, volume 3, pages 1371–1374, 1997.
23. M. Zviran and W.J. Haga. User authentication by cognitive passwords: An empirical assessment. *IEEE*, 1990.