A Multilingual Speaker Verification System: Architecture and Performance Evaluation

Javier Caminero[(1)], Joaquín González-Rodríguez[(2)], Javier Ortega-García[(2)], Daniel Tapias[(3)], Pedro M. Ruz[(3)], Mercedes Solá[(4)]

(1) Telefónica Investigación y Desarrollo; Emilio Vargas 6, E-28043 Madrid, Spain; fjcg@tid.es
(2) ATVS-DIAC-Univ.Politécnica de Madrid; Campus Sur, E-28031 Madrid, Spain; {jgonzalez, jortega}@diac.upm.es
(3) Telefónica Móviles España; Serrano Galvache 56, E-28033 Madrid, Spain; {tapias_d, ruth_pm}@tsm.es
(4) Informática El Corte Inglés; Travesía Costa Brava 4, E-28034 Madrid, Spain; m_sola@ieci.es

Abstract

In this contribution we present a multilingual secure access front-end that checks the identity of the user of a service through the mobile, PSTN and the IP networks (G.723, G.729). Our system prototype is based on speech recognition and speaker verification technologies and it uses a decision mechanism to combine the outputs of both modules. The main objective of the system is to increase the services access security with no increase of the service complexity. The system initially works in six European Languages (Spanish, English, French, Catalan, Galician and Basque) even though the system architecture easily allows the addition of new languages. The system has been developed through a EC funded project called SAFE[(*)] (Secure-Access Front End, IST-1999-20959).

1. Introduction

One of the main obstacles to develop the information society is the lack of confidence of end users on the security of the transactions and accesses to confidential, private or group-restricted information. More than 25% of the potential e-commerce customers are worried about security related issues, so that security is a key factor in the development of this kind of services. For this reason, Telefónica Investigación y Desarrollo (TID) and ATVS-DIAC have made an important research and development effort in the speaker verification area in the last years. This effort has ended up with the implementation of a first version of the SAFE system, that we describe in this paper together with the evaluation results.

A key advantage of the SAFE system is that it increases the service access security with no increase of the service complexity. This is due to the fact that there is no need of modifying the dialogue of the service with no speaker verification. The only exception is the enrollment phase, in which the system creates a user voice model through a training process. With this approach and once the training phase has been completed, users will not feel any difference neither in access time nor in the dialogue with already existing services but will be more confident in the service access control. This goal is achieved by doing the speaker verification process on the speech produced by the user while he/she is pronouncing the PIN. The system works in six European Languages: Spanish, English, French, Catalan, Galician and Basque even though the system architecture easily allows the addition of other languages.

The main goals of the SAFE project were the following:

- To increase the usage of the mobile, PSTN, and IP networks by improving the service access security and therefore allowing the creation of new services that require the use of this kind of technology (phone banking, v-commerce, etc.).
- To go one step further in the integration of the different communication networks, so that all services in the distinct networks can be accessed from any network.
- To increase user satisfaction by introducing new and more sophisticated services based on these new technologies.
- To increase user loyalty and, therefore, reduce the customer churn by providing safe and easy-to-use services.

2. System Description

The SAFE system is based on the combination of two different technologies:

- Multilingual Speech Recognition, that recognises the user's PIN in the language previously selected, and checks if the recognised PIN is correct.
- Multilingual Speaker Verification, that will be used to check whether the voice of the user matches with the voice of the real owner of the PIN. This process is carried out by analyzing the voice that the user produces while he/she is pronouncing his/her PIN.

Both technologies, developed at TID and ATVS-DIAC respectively, are integrated into the speech technology platform developed at TID. The block diagram of the SAFE system is the one showed in Figures 1 and 2:

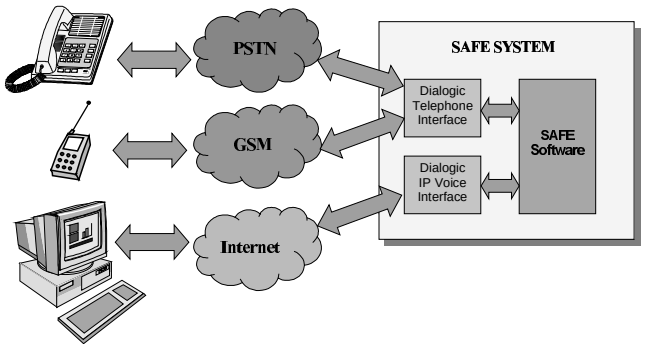

Figure 1- SAFE system block diagram

(*) http://www.atvs.diac.upm.es/safe/

As can be observed, the spoken input goes directly both to the speech recogniser and the speaker verifier. The output of both modules goes to the decision-maker. The dialogue manager is in charge of interacting with the user in three different scenarios or working modes and is able to recover from speech recognition errors. The three working modes are described next:

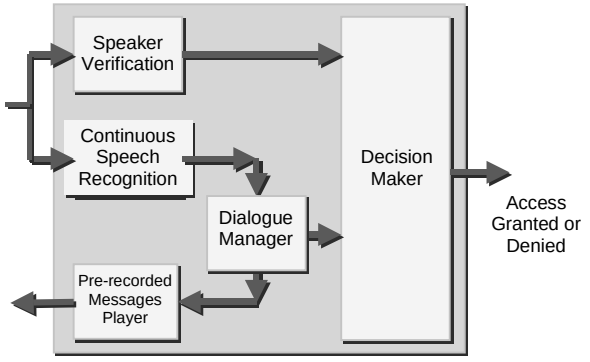

Figure 2.- SAFE modules

1. *Enrollment mode*: the first time the user calls the system, he will be requested to provide his/her first name and family name, his/her PIN through keyboard, and three utterances of the PIN.
2. *User verification mode*: this is the normal operational mode of the system. The SAFE system will give or deny access to services in this mode. This mode is described in more detail later.
3. *Password change mode*: the user can optionally change or update his/her PIN, after a first validation (through mode 2) of the user identity.

The SAFE system uses directed dialogues that are governed by a decision tree, which is shown in Figure 3.

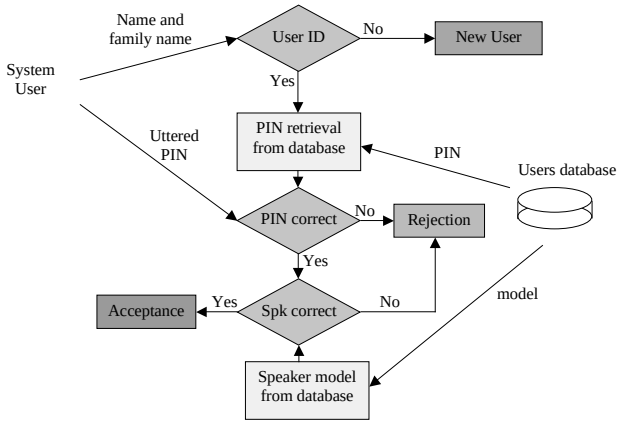

Figure 3.- SAFE User Verification mode

In the current implementation of the SAFE system, the user can be rejected in two different situations: the recognised PIN is not the right one or the input voice does not match the "real" user's voice.

3. Technology Description

In this section, we describe both the Speech Recognition and Speaker verification Technology used in the SAFE system.

3.1. Speech Recognition Technology

The Speech Recogniser uses the front-end developed at TID and described in [Villarrubia, 2001]. Basically, the speech signal is digitalized at 8KHz, and pre-emphasized by a factor alpha=0.97. The speech is then blocked into frames of 24ms every 12ms. A total of 27 Mel cepstral parameters are extracted: 8 mel-cepstra, 8 delta-mel, energy, delta-energy and 9 acceleration parameters. The triphones are modeled using CDHMM (Continuous Density Hidden Markov Models).

The evaluation results for the different communication networks (PSTN, GSM and IP) and conditions are described in section 5.

3.2. Speaker Verification Technology

The speaker characteristics are obtained from MFCC (Mel-Frequency Cepstral Coefficients) vectors, including temporal information through delta and delta-delta coefficients and CMN (Cepstral Mean Normalization). From these vectors, text-independent speaker verification is performed, where the speaker model is obtained through state-of-the-art Gaussian Mixture Models (GMMs) [Reynolds, 1995; Ortega-Garcia, 1998], trained with Maximum Likelihood.

In GMM systems, each speaker model λ is given by:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \qquad i = 1, \ldots, M$$

with mean vector μ_i and covariance matrix Σ_i; a gaussian mixture density is given by a weighted sum of component densities:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

where x is our L-dimensional cepstral vector, with mixture weights p_i and component densities $b_i(x)$ given by the equation:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{L/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \, \Sigma_i^{-1} \, (\vec{x}-\vec{\mu}_i)\right\}$$

The speaker recognition system models the speaker characteristics with a one-state model per speaker with a discrete set of gaussian mixtures corresponding to the probabilistic distribution of the MFCC vectors obtained from the speaker data.

The task of automatic verification of a identity from a speaker voice is performed by the system from two different inputs: the test utterance and a claimed identity. With these inputs, the system computes the likelihood of the test utterance against the claimed model, and compares it with the claimed-speaker threshold, accepting the speaker as the correct user, or rejecting him as an impostor.

In order to describe the performance of speaker verification systems, ROC (Receiver Operating Characteristic) and DET (Detection Error Tradeoff) curves are usually used [Doddington, 1998]. However, we can obtain a simple estimate of the performance of the system using the point where the false acceptance rate equals the false rejection rate. This point is known as the equal error rate (EER), which will be used in section 5 to describe and compare the performance of the system in different conditions.

One of the biggest problems in speaker verification is the intra-speaker variability in the likelihood scores for different repetitions of the same utterance. Then, instead of using fixed thresholds for every speaker, we can make use of likelihood-ratios, obtaining in this way an utterance-dependent threshold. Several possibilities for these likelihood ratios exist, from cohort-speakers to universal background models. In this work, we have used the following equation for the likelihood ratio [Higgins, 1991]:

$$\log L = \log p(X \mid \lambda = \lambda_q) - \log p(X \mid \lambda = \lambda_{UBM})$$

4. Applications

The goal of the SAFE project was to develop, in a multilingual environment, a prototype for secure remote authentication based on the analysis of the user's voice. This prototype will be used in speech enabled applications where the verification of the user's identity is required to avoid fraud and increase the user's confidence in the security of any transaction or private information access.

Next, we show a classification of speech-enabled services as a combination of three distinct aspects that can be summarized by the keywords Information, Communication and Transaction:

- Communication based services, which are all the voice-enabled applications where the caller manages his personal information. In these applications there is a need for authentication, and then, the caller can either ask for personal information, or submit information. For instance, personal voice dialing, unified messaging or calendar may be examples of this kind of applications.
- Information based services, which are all the voice-enabled applications where any caller from anywhere calls to obtain public or corporate information. In these applications there is no need of authenticating the caller. For instance, flight information, hotel information or stock quotes information are examples of this kind of services.
- Transaction based services, which are voice-enabled applications where the dialogue with the caller will end up with a trade, a booking or a good delivery. Therefore, this kind of applications requires some kind of caller authentication procedure. For instance, flight or hotel reservation, brokerage, etc. are examples of this kind of services.

5. System Evaluation

The evaluation of the system has been performed in two different stages. First, the different technologies involved were tested in simulated operational conditions. Once the technology was evaluated, several objective and subjective tests were performed with real users.

5.1. Technology Objective Evaluation

The speech recogniser and the speaker verifier, were tested initially in separate tests that are reported next.

5.1.1. Speech Recognizer Evaluation

The main goal of the Speech Recognizer Evaluation is to study the influence of the codification in the performance of the recognizer and to decide whether it would be appropriate to have different sets of HMM models for each codification.

The speech files used in these tests come from the VESLIM database, which is a Castilian Spanish database that contains recordings of company names, cinemas and theatres for evaluation of speech recognizers and phonetically balanced sentences for acoustic model training. It has approximately 400 different speakers from the central region of Spain and 600 speakers from the rest of the regions. The regions, which the files belong to, are West Andalusia, East Andalusia, Canary Islands, Catalonia, Central Region and Galicia. This database was recorded at TID in 1995 and the recordings were done using a high quality microphone in the absence of background noise.

Stationary noise was artificially added to these files, obtaining test files at 0dB, 5dB, 10dB, 15dB and 20dB of SNR. After the noise addition both the training and the testing files were codified with the different coders at the different rates.

Using a set of 1291 speech files, recognition tests were performed for the GSM (full and half rate), AMR and IP coders at the different SNRs. The results are shown in terms of average word error rate in the following table. The HMM acoustic models were retrained with a set of 2644 GSM-encoded phonetically balanced sentences.

Coder	%ER (0dB)	%ER (5dB)	%ER(10dB)	%ER (15dB)	%ER (20dB)
AMR (12.2 Kb/s)	19.5%	6.51%	3.72%	1.86%	1.47%
GSM	29.6%	8.13%	2.71%	1.86%	1.63%
GSM HR	25.17 %	9.99%	5.27%	3.25%	2.71%
G723.1 (6.3Kb/s)	26.1%	8.60%	3.72%	2.94%	1.70%
G723.1 (5.3 Kb/s)	26.7%	9.84%	4.18%	2.63%	2.09%
G729	19.3%	6.97%	3.64%	2.32%	1.63%

In and second test, the acoustic models were trained using speech coded with four different coders: AMR, GSM-HR, G723.1 and G729, so that the final models were estimated with information from the four codification standards.

As before, the testing data corresponding to the different coders was tested using these new models and the results are showed in the following table:

Coder	%ER (0dB)	%ER (5dB)	%ER (10dB)	%ER (15dB)	%ER (20dB)
AMR (12.2 Kb/s)	13.5%	4.73%	2.79%	2.01%	1.24%
GSM	24.5%	7.28%	3.33%	1.55%	1.08%
GSM HR	17.5%	6.43%	4.18%	2.63%	1.86%
G723.1 (6.3Kb/s)	20.4%	6.74%	3.41%	2.01%	1.63%
G723.1 (5.3 Kb/s)	22.6%	6.82%	3.49%	2.56%	1.55%
G729	14.7%	5.34%	2.94%	1.39%	1.24%

The comparison of both tables shows that there is a significant improvement in terms of average word error rate at low and high SNRs, when we use generic models for all the codifications. For example, the word error rate reduction at 20 dB of SNR ranges from 4% for G723.1 (6.3 Kb/s) to 33.7% for GSM.

Finally, it is important to remark that although this study of the influence of the codification has just been carried out in Castilian Spanish, can be extended to the other languages.

5.1.2. Speaker Verifier Evaluation

This evaluation has included the following different and important aspects: adaptation to mobile and IP networks, time course influence, language influence, and universal background model.

Adaptation to mobile and IP networks

Several tests were done to evaluate the channel influence on the overall performance of the speaker recognizer. Due to the fact that the original system (previous to the SAFE project) was designed to operate over landline telephone network, it was adapted to the mobile and IP networks to achieve optimal results.

The speech files used in this test come from the database SAFEDAT, which is a subset of 70 speakers from the Castilian Spanish database named GAUDI. The GAUDI files were collected over a microphone and a fixed telephone line. Besides these two formats, the SAFEDAT database contains coded versions of these original files codified using GSM at Full Rate (FR), Enhanced Full Rate (EFR) and IP (G.723.1 and G.729).

The tests were performed over the following communication channels: GSM-FR, GSM-EFR, G.723.1, G.729 and fixed telephone line. For each of these channels, 70 speaker models were trained with five utterances of number strings per speaker. To compute the miss and false alarm probabilities a set of detection output scores was obtained using three Personal Identity Number (PIN) utterances from each speaker. In this way, each speaker model was tested with three user attempts to access the system and the remaining files as impostor attempts. All the tests were made under matched conditions, which means that training and testing files correspond to the same channel. The scores were normalized using a channel-dependent Universal Background Model (UBM).

For the test conditions detailed in the previous section, recognition tests results are showed in the following table in terms of average equal error rate over the 70 speaker models:

EER (%)	PSTN	GSM-FR	GSM-EFR	G.723.1	G.729
Matched conditions	5.51	8.58	6.27	8.76	7.35

From the results shown in the previous table we can see that there is a slight increase in the average EER over the GSM an IP. Despite that fact, we conclude that the speaker recognition system also obtains good performance over the other channels, as the system security will rely on the combination of the speaker and the speech recognizers.

Time course influence

One of the major problems speaker recognizers have to deal with is the variability of the speaker features due to time course influence. Several tests will be done to measure the influence of this factor on the system performance, and different training strategies will be tested in order to obtain the optimal system settings.

To perform this test, the distribution for release 1.1 of the Speaker Recognition Corpus from the Oregon Graduate Institute has been used. This corpus consists of speech recorded from approximately 90 people over a two-year period. Each person recorded speech in twelve sessions spread out over those two years. The speech files are in English and were collected using the landline telephone network. Although the corpus consists of seven different tasks, only two of them will be used for this test. The first one is the numbers task, where each participant repeated six different number strings four times during each recording session (for a total of 24 utterances). And the second one is the password task where each participant was prompted to create a password, which on subsequent recording sessions would be asked to repeat four times (for a total of 24 utterances).

Each of the two mentioned tasks has been used to perform a different test. For both tests, 50 participants were used as system users and the remaining speakers as system impostors. The speaker models for both tests were trained using three different strategies:

- S1: Four utterances of the pin or password (depending on the test) from the first session were used to train each user model.
- S2: The same files used in S1 plus four utterances from session 7, which corresponds with one year later from the first session.
- S3: The same files used in S1 and S2 plus four utterances from session 12, which corresponds with two years later from the first session.

In order to compute the miss and false alarm probabilities for the passwords test, a set of detection output scores was obtained using the password utterances of each system user from the nine remaining sessions and one password utterance from each impostor. None of the impostor passwords matched any of the user passwords.

As for the pin test, the detection output scores were obtained in a similar way to the password test but each user was assigned one of the six possible pins and the impostors that tried to access the system knew the user password.

All the scores from both tests were normalized with a language-dependent UBM.

The following results were obtained using the test conditions detailed above. For both password and pin tests, the results are shown in terms of average equal error rate. The fourth row of the table corresponds to different result analysis:

- Short-term results: computed using files from sessions: 2,3 and 4.
- Mid-term results: computed using files from sessions: 5,6 and 8.
- Long-term results: computed using files from sessions: 9,10 and 11.

- All results: computed using files from all the sessions mentioned above.

The results from the password and PIN tests are shown in the following table:

PASSWORD EER (%)	S1	S2	S3
Short- term	3.59	1.97	1.68
Middle- term	5.53	1.95	1.10
Long- term	5.35	1.7	0.74
ALL	6.01	2.43	1.6

PIN EER (%)	S1	S2	S3
Short- term	3.31	2.18	1.76
Middle- term	6.39	3.15	2.65
Long- term	5.83	2.38	1.59
ALL	6.7	3.52	2.78

From the analysis of both tests results, several conclusions are shown. First of all, making a comparison between the results from the password test and the pin test we can see that the first one shows results slightly better than the second one. The main reason for this is that in the password test the impostors didn't know the real password of the users while in the pin test they did. This shows that although GMM technology is text independent, there is some phonetic dependence within the model when it is not trained with a large amount of data.

Regarding to the training strategies, we can conclude that S3 always shows better results than S2 and the same for S2 with respect to S1. One of the reasons is that the amount of data used to build the models is not the same, so for those who were trained with more data, the results are better. Another conclusion is that the inter-session strategy decreases the time course influence on the performance of the system.

Language influence measure

In order to be able to use the speaker recognizer in multilingual environments the language influence on the system performance has been tested. Given that the speaker recognizer has shown good results for Castilian Spanish language, similar tests have been performed done in the other languages included in this project to evaluate the language dependence of the system.

Four different languages have been used to perform the tests. The databases has been provided by TID and the files used for the tests are the following:

- *French:* 50 speakers with five number strings files each. All 250 number strings are different for each speaker.
- *Galician:* 50 speakers with six number strings files each. All 300 number strings are different for each speaker.
- *Basque:* 50 speakers with nine number strings files each. All 450 number strings are different for each speaker.
- *Catalan:* 50 speakers with six number strings files each. All 300 number strings are different for each speaker.

Given that none of the speakers repeated the same number string, all the tests were text independent. For each language, the files used to compute the miss and false alarm probabilities were the following:

- *French:* 3 number strings for each model.
- *Galician:* 4 number strings for each model.
- *Basque:* 7 number strings for each model.
- *Catalan:* 4 number strings for each model.

To compute the miss detection probabilities each speaker model was tested with the two remaining number strings of the speaker. The false alarm probabilities were computed using those two number strings from the remaining set of speakers. All the output detection scores were normalized using a language dependent UBM.

The following table shows the results obtained by performing the tests detailed in the previous section:

EER (%)	FRENCH	GALICIAN	BASQUE	CATALAN
Matched conditions	15.3	13.5	6.90	7.05

Due to the fact that each language database was composed of a different amount of data, the training conditions were different for each language so we can appreciate some increases in the average EER for those whose speaker models were trained with less data.

Universal Background Model

The scores computed in all the tests were normalized with a UBM in order to discriminate the speaker identities by their own voice features and not by the shared features among all the speakers. Due to the importance of the UBM normalization a test to measure the language dependence influence on the UBM was done. We also did a comparison between a language-dependent UBM normalization and a multilingual UBM normalization.

All the test details are the same as those described previously but in this test the UBM was trained with multilingual data from: Spanish Castilian, Basque, French, English, Galician and Catalan databases.

The results from the password and PIN tests are shown in the following tables:

PASSWORD EER (%)	S1	S2	S3
Short- term	3.73	2.05	1.76
Middle- term	5.67	2.03	1.06
Long- term	5.40	1.79	0.79
ALL	6.14	2.5	1.61

PIN EER (%)	S1	S2	S3
Short-term	3.45	2.48	2.07
Middle- term	6.87	3.16	2.87
Long- term	5.92	2.55	1.89
ALL	7.01	3.66	2.99

Comparing the results showed in these tables with the ones shown in the time course influence section, we can see that there is a slight increase in the average EER when the multilingual UBM normalization is used. Considering that using a multilingual UBM will allow the system to use the same UBM for all the languages at the expense of a small increase in the EER we can conclude that using a single language-independent UBM is a good compromise between system complexity and system performance.

As final conclusions for the objective evaluation of the speaker recognizer, we can state that the use of a single multilingual UBM is a very good compromise between the system complexity and the system performance and the time course influence can be reduced by means of inter-session training strategies.

5.2. Perceptual Evaluation from Users

In order to test the security improvements and the subjective quality of the system, a subjective evaluation was performed. In this test, a limited set of Spanish users (10 male and 10 female) tried the system and then, they answered a small questionnaire:

1. Perceived security of the system:
   very high / high / normal / low / very low
2. Operational complexity:
   very high / high / normal / low / very low
3. If this service was available through your phone, and its use was optional, select the applications where you would use it:
   3.1. Domotic control.
   3.2. Access to bank account information.
   3.3. Access to bank account operations.
   3.4. Transactions services over the phone.

The results of the evaluation are the following:

1. Perceived security of the system:
   - Very high: 15%
   - High: 80%
   - Normal: 5%
   - Low: 0%
   - Very low: 0%
2. Operational complexity:
   - Very high: 0%
   - High: 10%
   - Normal: 5%
   - Low: 75%
   - Very low: 10%
3. Percentage of potential users:
   3.1 Domotic control: 90%
   3.2 Access to bank account information: 85%
   3.3 Access to bank account operations: 60%
   3.4 Transaction services over the phone: 80%

From the results, we see that users show a high confidence in the system, as can be observed from the result about perceived security, and an easy-to-use system, from the system complexity result. However, there is still some lack of confidence in the technology, since there is still some people reluctant to access services by voice despite the good system performance perceived.

6. Conclusions

In this paper we have described the SAFE speaker verification system, that is based on the combination of speech recognition and speaker verification techniques. In this way, an impostor can be rejected in two different situations: the PIN or password is not the right one or the incoming speech does not match the "real" user's voice.

We have also presented the results of the objective and subjective evaluations. The objective evaluations were carried out on the speech recogniser and the speaker verifier separately in order to be able to simulate different working conditions and make system design decisions. We have shown that the speech recogniser performance is better if we use generic acoustic models that have been estimated using several training databases (one for each speech coder). This is probably due to the fact that in this way, the amount of training data is sufficient to obtain precise estimates of the model parameters. On the other hand, the effects of codification on the word accuracy are not crucial as we have observed in other research works and the use of a generic model for all the codifications is a good trade-off between complexity and performance of the system. As for the objective evaluation of the speaker verifier, we can conclude that the use of a single multilingual UBM is also a good trade-off between complexity and performance of the system. Additionally, the time course influence can be reduced by means of inter-session training strategies.

Finally, the subjective measures show that despite the high level of security and the low complexity perceived by users, there are still some people reluctant to use this technology in applications like bank account operations.

At this moment, we have implemented a new version of the SAFE system, which takes into account the results of the evaluation we report in this paper. This new version is going to be evaluated with real users in a couple of real services. This evaluation will also combine objective and subjective measures and will focus on the effect of different dialogue strategies in the quality of the system.

7. References

Doddington, G.R. (1998) "Speaker Recognition Evaluation Methodology – An Overview and Perspective", Proc. of ESCA-IEEE Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 60-66, Avignon (France).

Higgins A. et al. (1991), "Speaker Verification Using Randomized Phrase Prompting", *Digital Signal Processing* (Academic Press), vol. 1, pp. 89-106.

Ortega-Garcia, J., Gonzalez-Rodriguez, J. et al. (2000) "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification", *Speech Communication* (Elsevier), vol. 31, pp. 255-264.

Reynolds, D. A. and Rose, R.C. (1995) "Robust Text-independent Speaker Identification using Gaussian Mixture Speakers Models", *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1.

Villarrubia, L. et al (2001). "Speech Recognition in the new UMTS Communication Networks and Internet Environments", Comunicaciones de Telefónica I+D, No. 23, pp. 99-112.