A Multilingual Speaker Verification System: Architecture and Performance Evaluation

Daniel Tapias[1], Javier Caminero[2], Joaquín González-Rodríguez[3], Javier Ortega-García[3], Luis Hernández[4], Mercedes Solá[5]

(1) Telefónica Móviles, (2) Telefónica Investigación y Desarrollo, (3) ATVS-DIAC-Univ.Politécnica de Madrid,
(4)GAPS-SSR- Univ.Politécnica de Madrid, (5) Informática El Corte Inglés
tapias_d@tsm.es; fjcg@tid.es; jgonzalez@diac.upm.es; jortega@diac.upm.es;
luis@gaps.ssr.upm.es; 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 or the IP network (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 develop a product prototype that will increase the services access security with no increase of the service complexity. The system will initially work in six European Languages (Spanish, English, French, Catalan, Galician and Basque) even though the system architecture will easily allow the addition of new languages. Our system is being developed through a Trial EC project called SAFE[*].

1. Introduction

One of the most important utilities of a Multilingual Speaker Verification System is to act as a Secure Access Front End (SAFE) to any service to the user in any network (phone banking, phone purchasing, private groups information, etc). Our system is able to serve multiple simultaneous lines in different European languages. The system can be used in real applications to either allow or deny users the access to a particular service. The confirmation or rejection of the user's identity is carried out by combining speech recognition, to recognise the personal identification number (PIN), and speaker verification, to analyse the user's voice (Who says that PIN?), over the mobile and landline telephone network and the IP network (G.729 and G.723). Additionally, the system will also allow the user to modify his/her PIN.

This system increases the services access security with no increase of the services complexity. In this way 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 his/her PIN. The system works in six European Languages, three of them widely spoken in the European Union (Spanish, English and French) and the three regional languages from Spain (Catalan, Galician and Basque) even though the system architecture will very easily allow the addition of other languages.

This system has been developed with the following goals:

✓ To increase the usage of the mobile telephone network and the IP network in secure environments 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, e-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: The speaker verification technology has been adapted and tested in the fixed telephone network in previous projects and this project adapts this technology to the mobile and the IP network, so that allows the same service to be accessible from any network as far as service access security is concerned.

✓ To increase user satisfaction by introducing new and more sophisticated services based on these new technologies.

✓ To keep user loyalty by providing safer and easy-to-use services.

2. System Description

The technological areas that are involved in SAFE are:

✓ Multilingual Speech Recognition, that allow users to pronounce their PIN and gain access to a service using the language of their choice.

✓ 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 analysing the voice that the user produces while he/she is pronouncing his/her PIN.

The main characteristics of the system are:

✓ Multilingual input and output

✓ Widely spoken languages: English, French and Spanish

✓ Regional languages: Catalan, Galician and Basque

✓ Adapted to the mobile (GSM) and IP networks (G.723, G.729)

✓ High level of service access security with no increase of the complexity of the service (i.e.: it is based on PIN pronunciation).

✓ Combination of speech recognition and speaker verification techniques to increase the reliability of the system since it:
  - o Authenticates the user's identity using speaker verification techniques
  - o Checks that the pronounced PIN is correct using speech recognition

✓ Easy addition of new languages.

[*] http://www.atvs.diac.upm.es/safe/

Then, the block diagram of the SAFE system is the following:

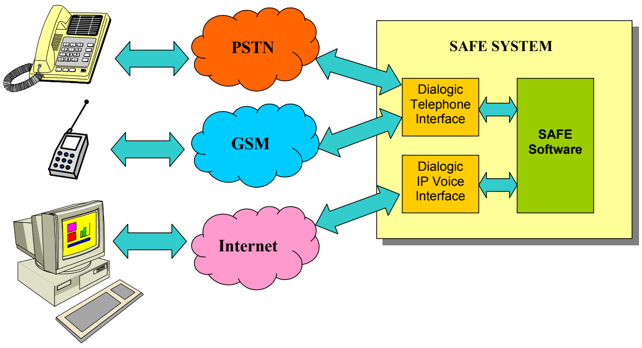

Figure 1.- SAFE block diagram

The SAFE final system is composed of two main different components:

- ✓ Speech recognizer API: this API was developed previously by TI+D and it is actually a commercial product. In the SAFE system, it is responsible of the recognition of the string digit uttered in the PIN.
- ✓ Speaker verification API: while the speaker verification technology was available previously to the SAFE project from previous joint work from ATVS-UPM and Telefónica, the API has been developed in this project.

The SAFE prototype is accessible from three different networks, closely linked:

- ✓ PSTN: landline conventional phones.
- ✓ GSM: from digital mobile phones.
- ✓ IP Network (G. 729 and G.723 Voice Coding Standards): from personal computers with IP access.

In order to take into account the different incoming languages to the system, the user will dial a different telephone number per language. However, the processing boards and industrial PCs are the same, activating in each one of the cases the different language-dependent modules (e.g., a single PC can attend different calls in different languages simultaneously).

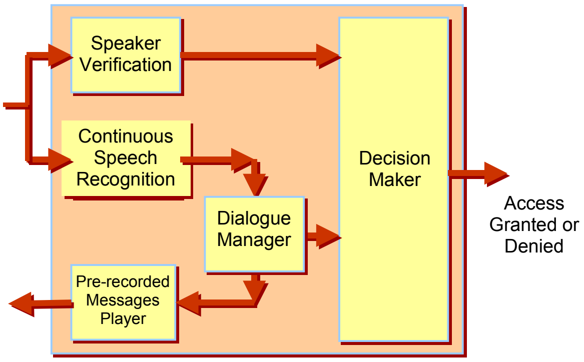

Figure 2.- SAFE modules

The system will have three main working modes:

1. *Enrollment*: the first time the user calls the system, he will be requested to provide his/her name and family name (just the first family name in Spanish, we usually use two), his/her PIN through keyboard, and 3 repetitions (digit by digit) of the uttered PIN.
2. *User verification*: 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 detailed later.
3. *Password change*: the user can optionally change or update his/her password, after a first validation (through mode 2) of the user identity.

The final SAFE system will allow mixed initiative dialogues. However, in order to focus in the authentication process, the dialogue will be directed in the SAFE prototype and governed through a decision tree, as we can see in figure 3.

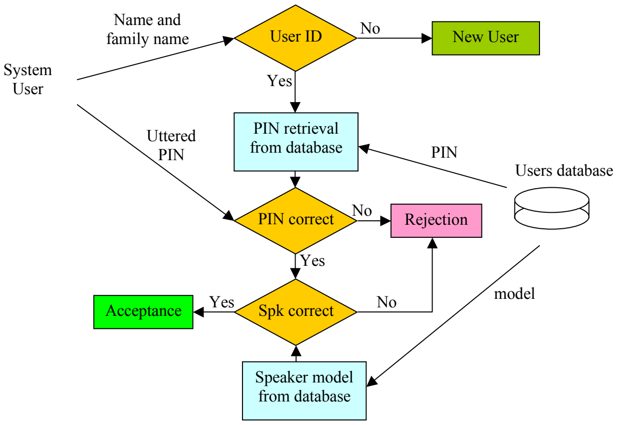

Figure 3.- SAFE User Verification mode

In order to improve user acceptability, the speaker identity is only checked with the speaker verifier if the PIN has been correctly recognized. In this way, the possible case of a true speaker misspeaking his/her correct PIN is not accepted. This policy decision has been assumed in order to improve the user confidence in the system.

Different spoken messages, in each one of the 6 accepted languages of the system, have been recorded for the system prototype, namely:

- ✓ System welcome and user prompt
- ✓ New user welcome and user prompt
- ✓ PIN utterance prompt (first one and repetitions)
- ✓ "End of user enrollment" information
- ✓ Acceptance or rejection information
- ✓ Error messages
- ✓ End of session message

3. Technology Description

Main body text.

3.1. Speech Recognition Technology

Main body text.

3.1.1. Heading 3

Main body text.

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:

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

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:

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

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 from the point where the false acceptances equal the false rejections, known as the equal error rate (EER), which will be used in the following to describe and compare the performance of the system in different conditions.

One of the biggest problem 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 this 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):

$$\log \mathrm{L} = \log \mathrm{p}\left(\mathrm{X} \mid \lambda = \lambda_{\mathrm{q}}\right) - \log \mathrm{p}\left(\mathrm{X} \mid \lambda = \lambda_{\mathrm{UBM}}\right)$$

3.2.1. Heading 3

Main body text.

4. Applications

The goal of the SAFE project was to develop, in a multilingual environment, a prototype for secure remote authetication based on the analysis of the user's voice. This prototype is adapted to work both in the IP and the mobile networks and 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.

Now, we focus on the kind of voice enabled applications in which our prototype is suitable to be used. We have defined a classification of all speech enabled services as a combination of three distinct aspects summarised by the keywords Information, Communication and Transaction, which is given below:

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 example, a non exhaustive list of possible services in this class may be:

- ✓ Personal name dialer
- ✓ Unified messaging
- ✓ Scheduling
- ✓ Phone banking (balance information)
- ✓ Worked hours assignment
- ✓ Voice voting (for TV and radio programs)

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 autheticating the caller. These services are typically call center applications.

For example, a non exhaustive list of possible services in this class may be:

- ✓ Flight information service
- ✓ Hotel information
- ✓ Fares information
- ✓ Car rental information (fares, model, location)
- ✓ Traffic information
- ✓ Train timetable information
- ✓ Travel agencies offers
- ✓ Wheather forecast information
- ✓ Stock quotes information
- ✓ Movie and Theater information, etc.

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 require some kind of caller authentication procedure.

For example, a non exhaustive list of possible services in this class may be:

- ✓ Flight reservation
- ✓ Hotel reservation
- ✓ Car rental
- ✓ Travel reservation
- ✓ Phone shopping
- ✓ Phone banking.
- ✓ Brockerage, etc.

Both Information and Communication based services are based on a cost saving business model while Transaction based services are based on a revenue generating business model. There are some straightforward applications which are easy to position on one particular class, though most complex ones will be a combination of the different classes.

4.1. Heading 2

Main body text.

4.1.1. Heading 3

Main body text.

5. System Evaluation

The evaluation of the system has been performed in two differents stages. Firstly, the different involved technologies have been tested in operational conditions closely similar to the actual ones with speech databases. Once the technology was evaluated, and the system prototype available to be used on-line, several objective and subjective tests have been performed with real users.

5.1. Technology Objective Evaluation

The two involved technologies, namely multilingual digit string recognition and speaker verification, both adapted to the GSM and IP environments, have been tested initially in separate tests, in order to validate its abilities in operational conditions.

5.1.1. Speech Recognizer Evaluation

Main body text.

5.1.2. Speaker Verifier Evaluation

This evaluation have included different aspects, all of them important to take into account for the final application.

Adaptation to mobile and IP networks

Several tests have been 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, adaptation to the mobile and IP networks is necessary in order to achieve optimal results.

The speech files used in these 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 users 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 normalised 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 results 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 test are showed 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

The results from the pin test are the following:

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 has been done. In this test a comparison between a language-dependent UBM normalization and a multilingual UBM normalization has been done.

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 test are showed in the following table:

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

The results from the pin test are showed in the following table:

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 sections 2.2.3 and 4.2.3 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 and the little 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 ATVS speaker recognizer is suitable to be used in a multilingual environment over the mobile, IP and landline telephone networks as the overall system secutity rely on the combination of the speaker and speech recognizers. The use of a single UBM is a very good compromise between the system complexity and the system performance and the time course influence can be decreased by using inter-session training strategies for the speaker models.

5.2. Perceptual Evaluation from Users

Main body text.

6. Conclusions

Main body text.

6.1. Heading 2

Main body text.

6.1.1. Heading 3

Main body text.

7. References

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

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.

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.

Nowadays, it is right to believe that some of the most interesting commercial opportunities in the market are on the union between the Internet and the PSTN (including mobile). Hence network operators move towards the IP world and there are several initiatives to ensure effective interoperability between networks.