KEYSTONES OF A SPEAKER VERIFICATION R&D GROUP

Roxana Saint-Nom[*]

Cochairman of the Department of Electrical Engineer - BUENOS AIRES INSTITUTE of TECHNOLOGY (ITBA) Argentina - saintnom@itba.edu.ar

ABSTRACT

Starting a new R&D group is always a difficult task. But without funding available, few full time faculty and no tools, its conception becomes a challenge.
This paper describes how to overcome these problems. It also deals with technical aspects of Speaker Verification methodology to justify why it is a proper topic for the new-born group.
Preliminary results obtained from self developed tools are presented, as well as the strategy designed to get funding and political support.

Index Terms— Speaker Recognition, Education

1. INTRODUCTION

ITBA is a small University devoted to Engineering. The undergraduate courses excel in many aspects. A five-year degree builds outstanding professionals for our market. They are all well prepared to face an always changing environment.
However industry is not highly developed. Technical know-how is usually imported. Our engineers can manage it well, but they are rarely their creators.
During the last five years, our economy improved. Being a welcomed fact, it brought winds of change. As the global world demanded manufactured products, our people needed modern technologies. Relative lower costs opened new opportunities to build our own way to progress.
Government got aware of the situation and began to promote higher education in Science and Technology. Some money was available to research and development in engineering fields. This is perhaps something very usual in developed countries, but it is certainly not the case in Argentina.
This opportunity fitted well with ITBA plans to promote a graduate school of engineering. Some political steps have been already taken, but the areas have to be created with care.
The purpose of this paper is to describe with some detail both aspects of a R&D: its conception and the technical interests that keep it working.

2. HOW A R&D GROUP WAS BORN

I've been working as an undergraduate SP professor for many years. I had nice experiences teaching signal processing [1] and I could realize that it is a very important subject to promote further studies at graduate level. I began introducing research into my courses some years ago, with very encouraging results [2]. My new position as a co chair in the electrical engineering department got me in contact with strategies that broadened my vision. Those undergraduate students, who performed so well in R&D, should be encouraged to keep working beyond course objectives.
During the semester, students had a clear motivation to work. It was a regular assignment, mandatory for evaluation purposes. But after the course was approved, they find new responsibilities and new challenges to pursue.
How could we attract a young researcher into a team which does not exist, into a subject that has no follow-up at graduate school level and whose applications are still unknown? Moreover, how can we provide the basic requirements our researching group would need?
These tough questions have relative simple answers.
Inertia: The idea of encouraging students to keep working after the R&D assignment is through gives results quite easy. They are proud of their accomplishment, and eager to hear that they can do more. They are acquainted with the material, and had some training with required tools. This inertia is present there to be profited.
Opportunity: Sometimes, the best ideas stay undeveloped because of a bad sense of opportunity. However, a good understanding of the university policy, experience about student maturity and knowledge of their available time were key issues for the group's creation.
Determination: there are many factors required to begin a R&D group. Among them, I am certain that determination, my own determination, is the engine. I could overcome the many difficulties because I never gave up.
Application field: When resources are scarce, selection of the field is quite complicated. It should interest both students and possible sponsors. Application tools should be inexpensive. And indeed, a mentor should be available.
That is why and how my Speaker Verification R&D group was born.

3. SV TEACHING CORE

This item describes the technical aspects of a SV system in order to be able to point out later how it can fit into my purposes.

3.1. SV Overview

Speaker verification is a process performed by a speech processing system that accepts or rejects a person's claimed identity. The biometric signal used is voice, a natural approach for access control to communication systems. It is a system with a wide range of applications, under constant research.

[*] IEEE Senior Member

The idea is to match a voice sample acquired in the recognition phase with a speaker model built in the training phase [3]. The decision yields to a true claimer or an impostor, but it also generates two types of errors: the false acceptance of an invalid user (FA) and the false rejection of a valid costumer (FR).
The general speaker verification problem may involve a closed set of speakers or an open one. In the latter, no previous model of the speaker exists.
There is also another important classification between text-dependent (TD) and text-independent (TI) speaker verification. TD only models the speaker for a limited set of words in a known context. When the sequence of spoken words is unknown, the problem becomes more difficult and recognition rates decrease.
Verification process consists in five steps [4]:

- Data acquisition: Test speech from a microphone is transformed into digital speech.
- Feature extraction: a short interval of speech is mapped into a multidimensional feature space.
- Matching Method: the sequence of feature vectors is compared to speaker models by a matching method, producing a matching score.
- Decision making: depending on normalization and threshold level, a decision is accomplished in an hypothesis-testing problem.
- Training: the acquired speech of a customer is processed to obtain a speaker model. It can be a template, a codebook or a statistical model, depending on the matching method used.

3.2. SV issues that may be addressed in a project

Feature extraction is the estimation of variables obtained from parameters of a speech signal. To reduce the dimension of the feature vector, a selection is made, meaning that a transformation that preserves important speaker information is performed. This new feature space enables simple measures of similarity.
One transformation widely used in SV systems is Mel warping. It changes the frequency scale to place less emphasis on high frequencies, based on the nonlinear human perception of the spectrum.
Stochastic Models: Pattern matching is performed measuring the conditional probability (likelihood) of the feature vector of the input sequence, given the model. The focus is made in the concept of modeling a speaker with a conditional pdf, and to compare it against that of the claimed costumer. The most widely used stochastic approach for speaker verification is the **GMM**: Gaussian Mixture Model [5].
This approach provides a probabilistic model of the underlying sounds of a person's voice, without imposing markovian constraints among sound classes.
It has many advantages, like noise and channel robustness, efficient computation time, insensitivity to model initialization and higher identification performance that preceding methods.
As Fig. 1 shows, a GMM is a weighted (p_i) sum of M component probability density functions b_i of the random vector $\vec{X}$. Each Gaussian pdf has a mean μ_i and a covariance σ_i^2. The model is parameterized by the notation:

$$\lambda = \{p_i, \vec{\mu}_i, \sigma_i^2\} \quad i = 1, \ldots, M$$

Speech corpora are data bases specifically created for development and evaluation in ASR (Automatic Speaker Recognition) and ASV research. Four factors have to be available in a corpus to evaluate its applicability to a speaker verification system: [6]

- Number and diversity of speakers, classified as customers and impostors.
- Number and time separation of sessions per speaker
- Type of speech (phrase, digit, read sentence, conversational speech)
- Channel, microphone and recording environment description.

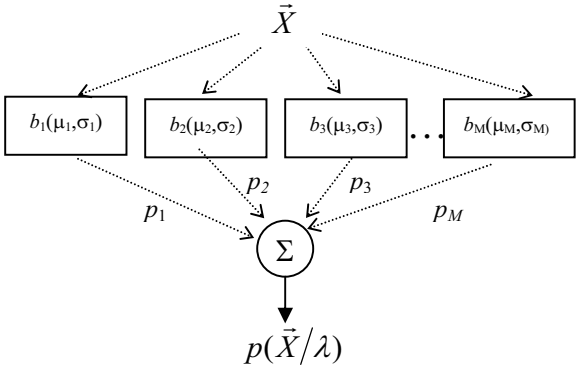

Figure 1. GMM Model

Decision making is being studied in depth because it depends on applications, training data sessions, speech variability and channel conditions.
Taking into account the way corpora have been collected, actual speaker verification is contaminated by factors of speech variability. In forensic context, for example, results may change considerably. [7]
There are two main categories:

- Peculiar intra-speaker variability:
  Characteristics to be taken into account are: manner of speaking, age, gender, inter-session variability, dialectal variations, emotional conditions, health condition.
- Forced intra-speaker variability
  These effects may change speaker-dependent features: Lombard effect, external-influenced stress, cocktail-party effects.

To perform verification tasks in variable conditions, matching methods must be carefully chosen to include some type of score normalization.
Condition between recording session of training data and actual test speech signal may vary in many uncontrolled situations.
Differing microphones, transmission link effects and acoustic environment are possible causes of changes in test signal features. They pose a hard challenge to matching methods.
One way to deal with a part of the problem is to "clean" the input data during training. Compensation techniques widely used include cepstral mean subtraction and RASTA filtering. Newer methods are speaker model synthesis and feature mapping [8]. On the output side, score domain compensation tries to act on score scales and shifts due to channel variations

3.3. Actual SV Applications

Examples of applications for ASV systems [9] [10] are mostly in the low security area. Voice activated door locks, computer and website access controls, telephone banking systems for transaction

authentication that does not require a PIN, law enforcement in home-parole and prison call monitoring, as well as voice samples [11] for forensic analysis.

4. PROBLEM SPECIFICATIONS

From the technical point of view, my R&D Project must deal with some problems.

4.1. Tools Development

In order to perform the feature extraction the researcher team has programmed a MATLAB GUI. [12] This parameterization system [13] is shown in Figure 2.

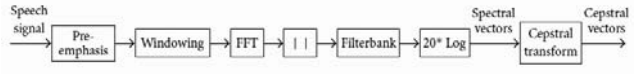

Figure 2. Ceptral parameterization of the voice signal

Having designed our own preprocessing system allows us to perform experiments as well as changing some block parameters to investigate results. As an example of the Mel Scale filterbank design, Figure 3 displays a space-time response obtained from the platform.

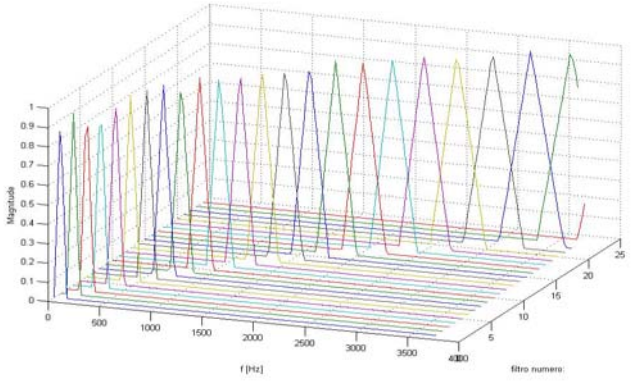

Figure 3. Mel Scale Filterbank Plot

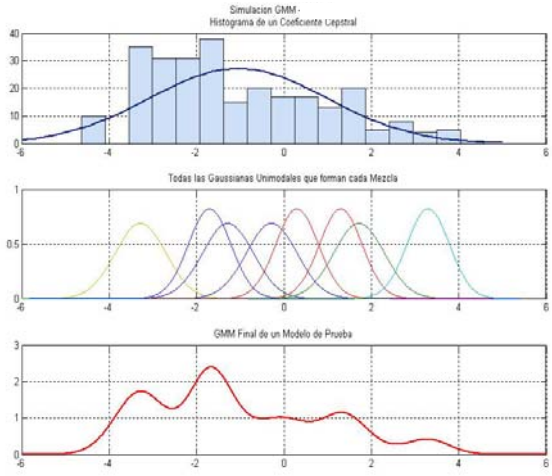

Figure 4. GMM's Model Platform

The second tool to develop is the GMM model creator. Up to this point, preliminary experiments are carried out in another MATLAB platform, as shown in Figure 4. But the next step is to adapt free code downloaded from the Internet based on the HTK language [14]. It was originally written for speaker recognition purposes but the author [15] encourages further work to implement speaker verification tasks.

4.2. Speech Corpus Acquisition

It is important that our results could be compared to other SV system performances in order to assess our work. Although many variables can be considered, a reliable corpus is the most difficult to match. In other contexts it would be the easiest, given that the system makes use of standard corpora. But in our case, funding to buy it is not yet available.

For that reason I decomposed the problem in two steps. Firstly, we developed a prototype system. Its purpose is to attract investors to get the material support we lack. After having succeeded step one, we will acquire a proper corpus. Meanwhile our databases are from two sources: internet and "homemade" speaker sessions.

4.3. Decision making

This is the central topic for research purposes. It is not only the development of an evaluation tool but also it requires that the system be oriented to a certain application.

In our case we found that Speaker Verification for forensic purposes may be easier to exploit.

In conventional speaker verification systems, scores obtained from claimant's utterances are averaged and the resulting mean score is used for decision making. In the forensic framework, the Bayesian Likelihood-Ratio approach is firmly established as a theoretical method.

Given this condition, research focuses in the performance of a forensic system in presence of speaker or channel variability, assessed according to the bayesian approach.

5. RESULTS

In order to test the experimental tools created, the team has performed a set of experiments.

5.1. Set-up sketch

Referring to Figure 2, the parameterization sequence is characterized as follows:

- Pre-emphasis: 0.98
- Windowing / FFT: Hamming window of 30ms with 10ms delay
- Mel filterbank: 24 filters

Experiments were carried out in a closed group of 10 voices.

GMM's are generated from 10 training sessions, in a number varying from 8 to 32, depending on the case.

5.2. Outcome

In this initial phase, results are shown as a percentage of true identifications. There were 100 tests performed.

Conclusions from table 1 are oriented to system performance. Processor's capability is seriously challenged when the number of Gaussians is 32. It also poses a problem to long training sessions.

It is not conclusive from the last column of table 1 that recognition is better when we increase the number of different training samples. We used a closed set of speakers with no background

model. But the point was to check that the tendency was right, indicating that our algorithms were working.

Number of voice samples	Number of GMM´s	% True ID
1 wav file	8	56
	16	64
	32	55
3 wav files	8	94
	16	93
	32	96
5 wav files	8	98
	16	100
	32	100

Table 1. Tool testing results

The exercise has given us many positive responses, like experience managing voice samples and Gaussian Mixture Models. It also showed us how difficult is to present a result that could be appraised by others.

6. FUNDING AND SUPPORT

Forensic Speaker Verification is a subject that is suitable of funding because of two reasons:

- It is easy to explain to non technical investors, with applications that are of common interest.
- It involves technology complicated enough to justify our proposed budgets

Our strategy of fund raising has overcome phase one. Political support to start the group and maintain student scholarships have been granted. The team has performed a state of the art research in the matter. We are already testing our initial tools.

Phase two involves some money to acquire a proper corpus to test research advances in decision making. We hope we will be presenting these results after the summer.

7. CONCLUSIONS

As the beginning of a new line of investigation in a University, R&D project's topic election should pursue some criteria that allow its survival.

However there are some other issues: sense of opportunity to get proper funding and political support, mentors available, motivated students and attracting applications that can be easily shown. Under these concepts we have successfully started a new R&D group that may be the ground basis of a graduate curriculum. FSV fulfills the aforementioned requirements.

If you have similar restrictions around you, but you count with your determination, look for a subject that meet your needs and start your engines.

8. ACKNOWLEDGEMENTS

I would like to show my appreciation to our first research team of students in speaker verification:

- Daniel Bennasar
- Juan Filips
- Bernardo Herrero

9. REFERENCES

[1] Jacoby D., Saint-Nom R., "Nice Experiences teaching SP in Argentina", *ICASSP 2001 Proceedings*, Vol. V

[2] Saint-Nom R., Jacoby D., "Building the first steps into SP Research", *ICASSP 2005 Proceedings*. V – 545-548

[3] S. Furui, "An Overview of Speaker Recognition Technology," *Workshop on ASR, ASI and ASV - ESCA '94 Proceedings, pages 1-9.*

[4] J. P. Campbell, ``Speaker recognition: A tutorial," *Proceedings of the IEEE, vol. 85, pp. 1437-1462, September 1997.*

[5] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.*

[6] J. P. Campbell, and D. Reynolds, "Corpora for the evaluation of Speaker Recognition Systems," *Proceedings ICASSP 99* http://www.apl.jhu.edu/Classes/Notes/Campbell/SpkrRec/icassp99.htm

[7] J. Ortega-Garcia et al., "Speech Variability in Automatic Speaker Recognition Systems for Commercial and Forensic Purposes", *IEEE AES Systems Magazine, November 2000.*

[8] D. Reynolds, "CHANNEL robust Speaker Verification via Feature Mapping". *ICASSP Proceedings II page 53, 2003.*

[9] D. Reynolds, "An Overview of Automatic Speaker Recognition Technology". *ICASSP Proceedings IV page 4072, 2002.*

[10] W. Roberts and J. Willmore, "Automatic Speaker Recognition using Gaussian Mixture Models".- *Crown Copyright of the Commonwealth of Australia, page 465, 1999.*

[11] J. Gonzalez Rodriguez et al., "Forensic Identification Reporting using Automatic Speaker Recognition Systems". *ICASSP Proceedings II, page 93, 2003.*

[12] **MATLAB** ®: http://www.mathworks.com/products/

[13] Frédéric Bimbot et al., "A Tutorial on Text-Independent Speaker Verification".- *EURASIP Journal on Applied Signal Processing 2004, 430–451 - Hindawi Publishing Corporation*

[14] Cambridge Hidden Markov modelling toolkit (HTK) http://htk.eng.cam.ac.uk/

[15] Ulrich T¨urk, Florian Schiel, "Speaker Verification Based on the German VeriDat Database" - *EUROSPEECH 2003 – GENEVA Pages 3025-3028* - http://www.bas.uni-muenchen.de/Bas/SV/