# NON DIRECTLY ACOUSTIC PROCESS FOR COSTLESS SPEAKER RECOGNITION AND INDEXATION

Teva Merlin, Jean-François Bonastre, Corinne Fredouille

# NON DIRECTLY ACOUSTIC PROCESS FOR COSTLESS SPEAKER RECOGNITION AND INDEXATION

*Teva MERLIN, Jean-François BONASTRE, Corinne FREDOUILLE*

LIA - University of Avignon
Agroparc - 339, chemin des Meinajaries BP1228
84911 Avignon Cedex 9 (France)
e-mail : (teva.merlin,jean-francois.bonastre,corinne.fredouille)@lia.univ-avignon.fr

**ABSTRACT**

This paper presents a new approach to speaker recognition and indexation systems, based on non-directly-acoustic processing. This new method is specifically designed to lower the complexity of the modeling phase, compared to classical techniques, as well as to decrease the required amount of learning data, making it particularly well-suited to on-line learning (needed for speaker indexation) and use on embedded systems.

## 1. INTRODUCTION

Classical speaker recognition systems usually require a complex acoustic parameterization and modeling phase. This complexity has increased over the last few years with widespread use of multiclass modeling (Gaussian Mixture Models, HMM based models, LVCSR methods, data driven models …).

The computational heaviness implied, as well as the requirement for large amounts of acoustic data to build and/or adapt complex speaker models, make it particularly difficult to implement on-line speaker recognition systems.

In the framework of speaker indexation and/or segmentation, this point is of high importance: in this case, speaker models are not available beforehand and have to be built from a small amount of data.

This paper presents a new approach to speaker recognition and indexation systems, developed to provide an answer to these two problems: it allows lower complexity, with the ability to compute speaker models from very few data. The proposed method is well-suited to on-line processing and embedded systems.

## 2. OVERVIEW OF THE STRUCTURE OF A CLASSICAL SPEAKER RECOGNITION SYSTEM

In speaker recognition systems, the learning phase classically consists of two steps: first, parameterization of the acoustic data, and then modeling (figure 1).

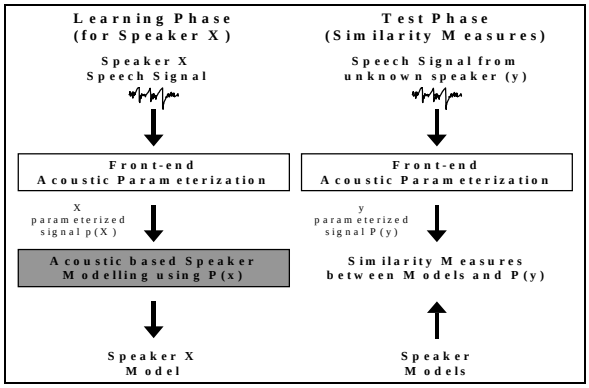


**Figure 1:** Graphic description of classical methods (acoustic driven) learning and test phases

Previous work shows the importance of parameterization and front-end signal processing phases for speaker recognition process [2][3]. For embedded systems a compromise between performance, parameterization complexity and size has to be made.

The modeling methods most frequently used are statistics-based ones: Second Order Statistical Methods based on Mono Gaussian Models [1] or Gaussian Mixture Models [6], HMM modeling [4] and Auto Regressive Vector methods [5]. In fact, the majority of the text independent speaker recognition systems developed for the 1999 NIST evaluation

campaign[1] were based on Gaussian Mixture Models, using generally 128 gaussians.

Once this modeling phase has been achieved, the test phase consists, given a speech signal coming from an unknown speaker, in computing a similarity measure between this signal and a known speaker model (as shown in figure 1). For speaker identification tasks, this step is repeated for each known speaker.

The main drawback of statistics-based modeling methods is that they require a minimum amount of data to be accurate. Moreover, due to the acoustic nature of the data, learning has to be done using several sets of data reflecting the acoustic variability resulting from the phonetic content, the channel and the recording conditions, in order for the speaker models to take this variability into account.

Besides the availability of such data for each speaker for which a model is to be built - which involves a long duration learning signal - this implies an increase of the model size and complexity as well as of the amount of computing resources needed.

This is what leads speaker recognition systems based on such methods to perform poorly on short learning duration, while also being difficult to implement on embedded systems.

## 3. DESCRIPTION OF THE PROPOSED METHOD

We propose a method which allows to define the modeling phase independently from the acoustic constraints, thus leading to decrease the need for large amounts of data.

This is done by introducing a preliminary step into the process, which yields a new representation of the speech signal. This representation being defined specifically for the speaker recognition problem, it is designed to be clearly speaker-oriented.

The underlying idea is to consider the speech signal through a set of speaker-distinguishing characteristics. Given a speech sample, its representation is obtained by computing a valuation of each of the characteristics. This process can be seen as a projection of the speech sample from the acoustic space into a new space. A referential for this space is defined by the considered characteristics, each corresponding to an axis.

The entire recognition process (modeling and test phases) is then designed to be carried out on data projected into this characteristic-based space, no matter how the projection is computed from corresponding acoustic data. From this point of view, the projection step may be seen as a "black box", for which only the input and output have to be known. We will refer to it using this term from now on.

The projection is performed using classical recognition techniques. In fact, there can be as many different types of techniques as characteristics to be valued. Each of these recognizers needs its own learning phase, which may be complex and require a large amount of data (see figure 2). However, this heavy learning phase has to be carried out only once, to set up the projection system. Afterwards, the recognizers only have to be run in test phase, to output a value corresponding to the acoustic data to project. This is usually a rather simple task.

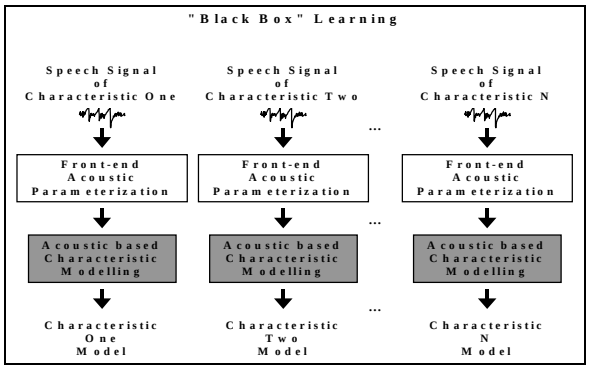


**Figure 2:** Description of the "BlackBox" Projection System learning

The whole problem of acoustic variability is taken into account by the recognizers composing the projection system. The projection result is then theoretically independent regarding this variability.

This leads the post-projection process to be rather easy, as it is no longer necessary to deal with this problem.

The modeling phase is here far simpler than for acoustic models, and doesn't require to handle as much data. It can be performed using "standard", non speaker recognition-specific classification techniques.

[1] Since 1996, the National Institute of Standards and Technologies (NIST/NSA) organizes some benchmark evaluations in text independent speaker recognition over the telephone. See http://www.itl.nist.gov/div894/894.01/spkrec.htm for more details.

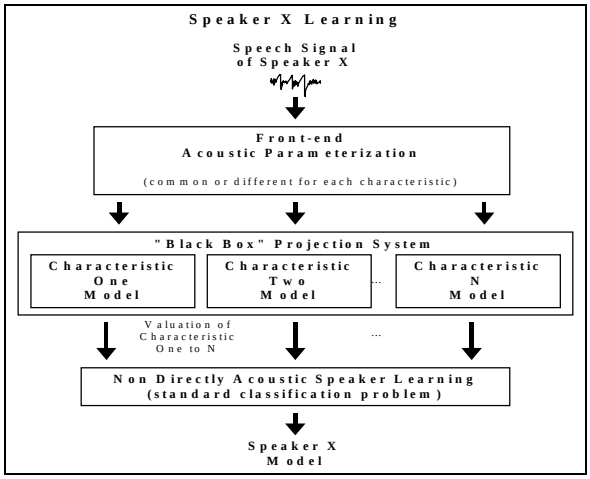


**Figure 3** : Learning in the non-directly acoustic approach

This simplicity, allowing to build a model from a very small data set with very low computing resources requirements, is particularly useful when it comes to automatic indexation. In this case, where speaker models have to be constantly reestimated, the gain in terms of learning complexity compared to classical systems is even more obvious: the projection process has to be done only once for a given set of data; the learning process can then be iterated numerous times on the projected data without seeing major raise in complexity.

The test phase also benefits from working on projected data, as the decision scheme no longer has to take acoustic variability into account. It is based on simple distance computation within the projection space between the projection of the test sample and speaker models built during the previous phase. This step is also made easy due to the rather small size of the speaker models in the projection space.

## 4. A FIRST IMPLEMENTATION OF THE PROPOSED METHOD

A simple implementation has been achieved as an example of the proposed method.

The projection system (the “black box”) retained for this implementation is not based on an explicit specification of the considered characteristics.

Instead, each characteristic is evaluated according to similarity between the test signal and a chosen voice. This likeness is evaluated using a classical speaker recognition system (see section 4.1). One acoustic model is computed for each characteristic, using a record pronounced by the corresponding speaker. A classical similarity measure between a model and the test signal gives the valuation of a specific characteristic.

### 4.1 Parameterization and basic speaker recognition technique

The classical recognition technique used here is based on second order statistical modeling and maximum likelihood computation.

The signal is characterized each 10 ms by a 24-coefficient spectral vector using a linear scale. No other front-end processing is performed.

A speaker (or a characteristic) is modeled by a global mean vector and a covariance matrix (mono-gaussian modeling).

The similarity measure between a signal and a model corresponds to maximum-likelihood estimation.

### 4.2 Description of the projection system

As seen above, the projection process consists here in computing a similarity measure between the data to project and a set of voices, using the technique described in section 4.1. This set is made of 40 randomly-picked male speakers.

None of these speakers is used during the test phase.

### 4.3 Speaker Training

A model for a given speaker is built by classifying the points resulting from the projection of his learning data. The classification technique is a variant of the k-means algorithm. Each of the resulting classes is then represented by two vectors: its center of gravity and its standard deviation. 15 class models are used for this paper.

### 4.4 Similarity measure

The distance between a frame of speech signal (represented by a vector in the projection space) and a speaker model is given by the minimum non-oriented angle measure between this frame vector and the center of gravity of each class of the model.

## 5. SPEAKER IDENTIFICATION EXPERIMENTS

In order to test the system described above, an experiment has been carried out within the framework of close-set text independent speaker identification over telephone lines.

The data set used for this experimentation is a subset of the SWITCHBOARD database used for the 1997 NIST Speaker Verification System Evaluation.

This set is composed of some parts of real telephone conversations (natural speaking). It includes various recording conditions and noisy segments.

35 speakers have been used for the test phase. This set is distinct from the one used for the projection "black box". For each of these, 60 seconds of speech have been used as learning data, and around 60 other seconds have been dedicated to the test phase.

Decisions were taken for speech segments of 3 seconds (corresponding to 300 frames), based on the computation of the mean distance between these frames and each of the models of the 35 “client” speakers. A good identification is recorded if the corresponding speaker obtains a distance strictly inferior to the other speakers.

The tests are conducted for different learning durations: from 60s down to 0.5s. For a short duration, the test are repeated several times (10 times for 0.5s to 3.75s duration and 5 for 7.5s duration), randomly selecting the learning segment.

In order to evaluate the possibilities of the proposed method, we carried out the same tests using the basic speaker recognition system shown in section 4.1 (it is exactly the system used within the projection "black box").

Table 1 shows results obtained for different learning durations. The percentage of good identification is given for the new "projection" method and for the classical one, according to the learning duration. For short duration, the maximum and the mean identification rate (obtained on the 10 or 5 randomly-picked learning segments) are specified.

| | | Identification rate | | | |
|---|---|---|---|---|---|
| | | MGM | | **Projection** | |
| | | Mean | Max | **Mean** | **Max** |
| Learning duration (s) | 0.5 | 6.12 | 8.55 | **28.59** | **36.23** |
| | 1 | 11.35 | 16.37 | **38.20** | **45.65** |
| | 2 | 23.67 | 30.43 | **44.88** | **51.59** |
| | 3.75 | 47.82 | 53.33 | **56.26** | **64.06** |
| | 7.5 | 70.43 | 73.91 | **63.79** | **70.72** |
| | 15 | 87.53 | | **72.03** | |
| | 30 | 95.36 | | **70.72** | |
| | 60 | 97.53 | | **74.06** | |

**Table1**: Speaker identification results (35 speakers -- 690 tests). The table shows % of good identification for classical MGM method and the proposed method (projection) for different learning durations. For short durations, different learning segments are chosen and the mean/max of identification rates are given.

Comments :

- On long duration learning (60s to 15s), the results obtained by the classical, mono-gaussian based method seem to be less than state-of-the-art ones. This loss should not appear on short duration, due to the impossibility to learn or adapt 128 (or higher) based mixture models on 50 occurrences (0.5s duration).

- The new "projection" method obtains encouraging results on long learning duration (around 74% of good identification) and the results remain at a good level for 0.5s duration (between 28 to 36% of identification).

- For long duration learning, these results actually show a loss when compared to the identification rate obtained on the same data set with the basic monogaussian system. However, this loss must be understood more as a consequence of the oversimple nature of the implementation, than as the result of the intrinsic limits of the method. Notably, the definition of the projection "black box" is extremely simplified and does not allow to exploit all the speaker-specific information found in the speech signal.

- For short training duration (less than 7.5s), the very simple implementation of our method obtains better results than the classical method. For example, for 2s duration the respective scores are 30.43% for the classical method and 51.59% for the projection method.

- It is to be noticed that - in terms of complexity - the projection phase demands less than one iteration of the EM learning algorithm of classical GMM systems. The learning phase on the projected data is of extremely low cost, allowing our system to require few computing resources compared to classical systems.

- As shown by Table 2, the size of the speaker models is reduced significantly by the projection method.

| | Size of the models (number of components) |
|---|---|
| Projection method (40 char. and 1 class) 24 and 72 coef. | 40*2*1 = 80 |
| Projection method (40 char. and 15 classes) 24 and 72 coef. | 40*2*15 = 1200 |
| Mono Gaussian (complete cov. Matrix) 24 coef. acoustic vectors. | (24*25)/2 + 24 = 324 |
| Mono Gaussian (complete cov. Matrix) 72 coef. acoustic vectors | (72*73)/2 + 72 = 2700 |
| GMM 128 (diagonal cov. matrix) 24 coef. Acoustic vectors | 24*2*128+128 = 6272 |
| GMM 128 (diagonal cov. matrix) 72 coef. acoustic vectors | 72*2*128+128 = 18560 |

**Table 2**: Size of the speaker models for 3 different methods, in case of 24 and 72 coefficient acoustic vectors.

## 6. CONCLUSION

The method presented here answers the needs for lighter speaker recognition and indexation systems.

The main advantages of this method are :

- The non acoustic nature of the learning and recognition phase allow a low cost processing needing few learning data. It's particularly interesting for on-line learning (needed for speaker indexation) and/or embedded systems.
- The speaker models are built on non-acoustic data. Usual classification techniques can be used for this task.
- The size of the speaker models is really small compared to classical methods.
- The proposed projection method allows to split the whole process into two distinct phases: first, expressing in the new referential the data for a given speaker; and then, the actual recognition task.

The obtained results are promising, given the triviality of the implementation. This method shows some potential, which has yet to be exploited. To do so, two things have to be improved: the methods used to exploit the projected data (using better classification techniques and similarity measure), and the definition of the projection “black box”. This second point remains far less straightforward than the first one.

## 7. ACKNOWLEDGEMENTS

This work was supported by the LIARMA Project (LIA Avignon and RMA/Bruxelles).

## 8. REFERENCES

[1] F. Bimbot et al., *Second-order statistical measures for text-independent speaker identification,* Speech Communication, 17:pp. 177-192, 1995

[2] D. Charlet and D. Jouvet, *Optimizing Feature Set for Speaker Verification,* AVBPA 97, LNCS n° 1207, pp. 203-210

[3] G. Gravier et al., *Model dependent spectral representations for speaker recognition,* In Eurospeech, 1997

[4] L.F. Lamel, J.L. Gauvain, *Speaker verification over the telephone*, In RLA2C, Avignon, April 1998, pp. 76-79

[5] I. Magrin-Chagnolleau et al, *A further investigation on AR-Vector models for text-independent speaker verification*, Proc. ICASSP-96, Atlanta, pp. 401-404

[6] D. Reynolds and R. Rose, *Robust text-independent speaker identification using Gaussian mixture models,* IEEE Trans. On Speech Audio Processing, 1995, pp. 72-83