# Looking for the Best Spectral Resolution in Automatic Speaker Recognition.

H. Sayoud*, S. Ouamour**

USTHB, Electronics Institute, BP 32 Bab-Ezzouar, Alger, Algeria
Email: * sayoud@hotpop.com ** ouamour@hotpop.com

*Abstract* — **In this research work we look for the optimal spectral resolution for speaker authentication in quiet and noisy environment, using the speech signal (microphonic and telephonic bandwidths). This problem is investigated according to several conditions. For this purpose, we investigated the effect of the spectral resolution in speaker identification performance. During this research work, we implemented a statistical approach based on second order statistical measures and using the normalised Mel-spectral energies (MFSC).**

**In order to find the optimal spectral resolution, in microphonic and telephonic bandwidth, we tested several dimensions for the MFSC vector (*Normalised Mel energies*) ranging from 12 to 60 and several types of additive noise (*white noise, car noise and racket noise*) at several SNR ratios.**

**Results show that the optimal spectral dimension depends on the experimental conditions. So, we noticed the importance of the high spectral resolution of 60 coefficients / 8 kHz for the [0-8 kHz] bandwidth and the resolution of 48 coefficients / 8 kHz for the [0.3-3.4 kHz] bandwidth (*especially in noisy environment*), whereas the actual works have always favoured resolutions less than 24 coefficients in such tasks. For example, we note an improvement of about 11% in the recognition score, since we increase the resolution from 24 to 48 MFSC for the telephonic bandwidth.**

*Index Terms* — **Speech processing, Speaker recognition**

## I. INTRODUCTION

The vocal expression is a particular characterization for the speaker: thus it is possible, in normal conditions, to recognise his corresponding talker during a telephonic conversation. *Speaker characterization* is a generic term for discriminating between several persons thanks to their voices.

Our task, in this research domain, is to recognise not what it is said, but the identity of the speaker who is talking, only by his vocal characterizations. It is true that we can use several features for this purpose; but the real problem, we are interested in, would be the optimal resolution to use (features dimension) and the effect of noises / bandwidth on this resolution.

In order to make a judicious choice of the dimension of the speaker's spectral characterization in those conditions, we tested several resolutions ranging between 12 and 60 spectral coefficients (MFSC) during the speech signal analysis.

Concerning the task of speaker recognition, we choose a statistical approach based on 2$^{nd}$ order statistical measures [1].

Moreover, for the evaluation of the robustness of the system implemented, we expanded our investigation to noisy environments by testing 3 types of noise and with different SNR between 0 and 24 dB.

Finally, the great quantity of results obtained during these experiments leads us to several discussions which are presented at the end of this paper.

## II. SPEECH DATABASE

The speech database is extracted from TIMIT [2] and FTIMIT [3] corresponding to 37 different speakers. The approximate duration of an utterance is 9 s for the statistical training and 7s for the test.

A second investigation is made in noisy environment [5] and with three types of noise [4]: - the Gaussian white noise (GWN), - the car noise, and - the racket noise.

These noises are added during the training and the test [5] at the following rates: 0 dB, 6 dB, 12 dB, 18 dB, 24 dB.

Each database is processed with 5 different spectral resolutions (size of the filter bank) [6]: 12, 24, 36, 48 and 60 filters.

Furthermore, two types of statistical distances are tested: $\mu_G$ measure and $\mu_{Gc}$ measure [7]. This difference also implies 2 tests and 2 different results. Finally, we must recall that the computation of these MFSC coefficients is done for each segment of 32 ms by applying energy normalisation.

## III. COMPROMISE BETWEEN "LOW SPECTRAL RESOLUTION" AND "HIGH SPECTRAL RESOLUTION".

An important question may be asked during the choice of the optimal dimension of the speaker's acoustic features. This question is: *Is there a relationship between the spectral dimension (dimension of the MFSC coefficients) and the modelization reliability of the speaker's features ?*

For this purpose, firstly, we tried to investigate several 3D spectrums obtained with variable spectral resolutions in order to find a link between these resolutions and the accuracy of the speaker's spectral characterization (figure 1 and 2).

According to these results, we noticed the dependence between the MFSC dimension and both the formantic envelope and the high-frequency spectrum. This dependence allows us to adjust the MFSC dimension in order to get a balanced spectral characterization for the speaker [8].

We can easily observe the good formantic envelope in the 12 or 24-MFSC curve (fig. 1). But in the other hand, we really find more details in the high-frequency part of the curve in the case of 48 or 60-MFSC (fig. 2).

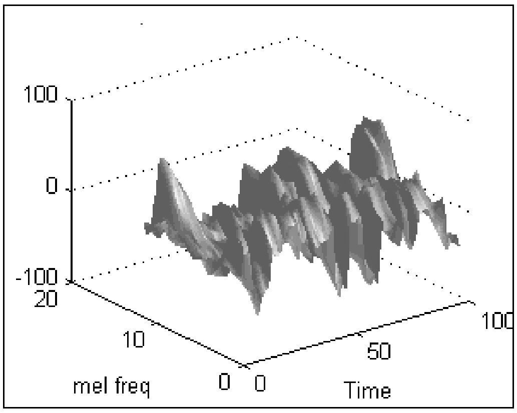

Fig.1. 3D-Mel spectrum Representation with12 coefficients

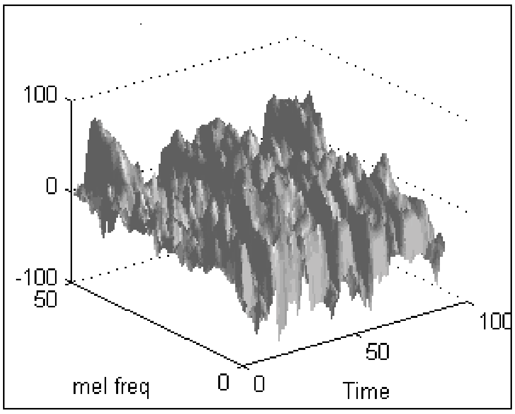

Fig. 2. 3D-Mel spectrum Representation with 48 coefficients

Given the fact that the information contained in the formantic envelope are important, we have a great interest in reducing the number of the Mel filters, but as described in previous investigations the high-frequency information can bring an appreciable improvement in speaker identification (especially for corrupted speech): then, it will be beneficial to introduce the effect of those information in the MFSC parameterisation if we want to obtain a higher identification quality. This means that the filter-bank size should be great.

However, we must recall that when we use very high dimensions we shall deal with two problems: the complexity of calculation and the bad modelization of the covariance matrix. Thus, theoretically, the speaker characterization needs a balanced compromise between low and high spectral resolutions.

A second question may occur then: *What is the optimal number of the Mel-spectral coefficients for this task ?*

To answer this question, we performed several practical experiments using different filter-bank sizes and with a thorough comparison.

## IV. Description of the Statistical Method of Speaker Identification: *Second Order Statistical Measures*

Our speaker identification method [5], based on mono-Gaussian model [9] [7] [1], uses some measures of similarity, which are called Second Order Statistical Measures (SOSM). These measures are used in order to recognise the speaker at each segment of the speech signal.

We recall bellow the most important properties of the approach [9] [7] [1].

Let $\{x_t\}_{1\le t\le M}$ be a sequence of M vectors resulting from the *P*-dimensional acoustic analysis of a speech signal uttered by speaker ***x***. These vectors are summarised by the mean vector $\bar{x}$ and the covariance matrix X:

Similarly, for a speech signal uttered by speaker ***y***, a sequence of N vectors $\{y_t\}_{1\le t\le M}$ can be extracted.

By assuming that all acoustic vectors extracted from the speech signal uttered by speaker ***x*** are distributed like a Gaussian function, the likelihood of a single vector $y_t$ uttered by speaker ***y*** is

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \quad (1)$$

If we assume that all vectors $y_t$ are independent observations, the average log-likelihood of $\{y_t\}_{1\le t\le M}$ can be written as

$$\bar{L}x(y_1^N) = \frac{1}{N}\log G(y_1...y_N|\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t|\boldsymbol{x}) \quad (2)$$

We also define the minus-log-likelihood $\mu(\boldsymbol{x},y_t)$ which is equivalent to similarity measure between vector $y_t$ (uttered by ***y***) and the model of speaker ***x***, so that

$$\underset{x}{Arg\max}\ G(y_t/\boldsymbol{x}) = \underset{x}{Arg\min}\ \mu(\boldsymbol{x},y_t) \quad (3)$$

We have then:

$$\mu(\boldsymbol{x},y_t) = -\log\ G(y_t/\boldsymbol{x}) \quad (4)$$

The similarity measure between test utterance $\{y_t\}_{1\le t\le M}$ of speaker ***y*** and the model of speaker ***x*** is then

$$\mu(\boldsymbol{x},\boldsymbol{y}) = \mu(\boldsymbol{x},y_1^N) = \tfrac{1}{N}\sum_{t=1}^{N}\mu(\boldsymbol{x},y_t) \quad (5)$$

$$= -\overline{L}x(y_1^N) \qquad (6)$$

After simplifications, we obtain

$$\mu(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1}) + (\overline{y}-\overline{x})^T X^{-1}(\overline{y}-\overline{x})\right]-1 \qquad (7)$$

This measure is equivalent to the standard Gaussian likelihood measure (asymmetric $\mu_G$) defined in [7] [5]. A variant of this measure called $\mu_{Gc}$ is deduced from the previous one by assuming that $\overline{y} = \overline{x}$ (i.e. the inter-speaker variability of the mean vector is negligible). Thus, the new formula becomes:

$$\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right]-1 \qquad (8)$$

In our research work we used the symmetric form of this statistical measure: namely the average between μ(***x,y***) and μ(***y,x***).

## V. Recognition experiments in quiet environment (without noise)

The first experiment consists in the recognition of 37 speakers of TIMIT [2] by the SOSM method [7] in a quiet environment (without noise). Two cases are investigated: the identification in the microphonic bandwidth [0-8 kHz] and the identification in the telephonic bandwidth: [300-3400 Hz]. The MFSC size varies from 12 to 60 coefficients. See figure 3.

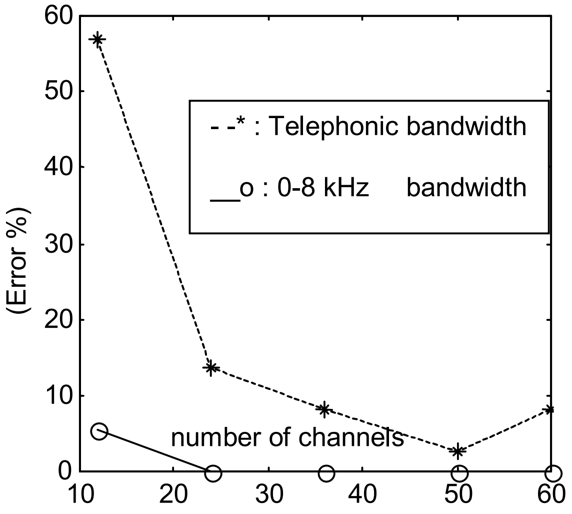

Fig. 3**:** False identification score for the case of quiet environment.

- **For the 0-8000 Hz bandwidth,** the false identification score is 5.4% with 12 MFSC and it is 0% for all other cases (24, 36, 48 et and 60 MFSC) which involves a good identification from 24 channels.

- **For the telephonic bandwidth,** the false identification score is 56.7% with 12 MFSC, it is 13.5% with 24 MFSC, it is 8.1% with 36 MFSC, it is 2.7% with 48 MFSC and it is 8.1% with 60 MFSC. Therefore, the size of 48 channels seems to be the best one on telephonic bandwidth.

## VI. Recognition experiments in noisy environment

The second investigation consists in identifying all the previous speakers, in noisy environment: SNR ranging from 0 dB to 18 dB. We used three types of noises [4], which are: the Gaussian white noise, the racket noise and the car noise;

### *A. Observation of the results*

Figures 4 to 6 concern the experiments of speaker identification on the microphonic bandwidth [0-8 kHz] and figures 7 to 9 concern the experiments of speaker identification on the telephonic bandwidth: [300-3400 Hz].

The following points are briefly noticed:
In the audible bandwidth 0-8 kHz, the best scores are got with 60 coefficients, which means that the high spectral resolution provides more protection against noises.
In the telephonic bandwidth 300-3400 Hz, the best scores are got once with 48 coefficients and once with 60 coefficients.
In 0-8 kHz, the most disruptive noise is the racket noise followed by the white noise and finally the car noise.
In the telephonic bandwidth the most disruptive noise is the racket noise followed by the two other noises.
In the 0-8 kHz bandwidth, a very strong noise at 0 dB causes a devaluation of the identification score for over 20%, except for the case of the car noise where the score remains high (97.3%) even at 0 dB and with 60 channels.
In the telephonic bandwidth, the white noise and car noise at 0 dB provoke a score devaluation of over 20%. Concerning the racket noise, the score devaluation is more then 40%, which involves a failure of the identification system in this case.

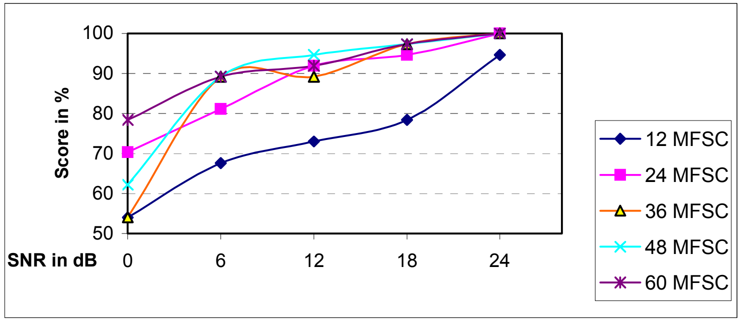

Fig. 4. Recognition scores in noisy environment (white noise), in the 0-8 kHz bandwidth

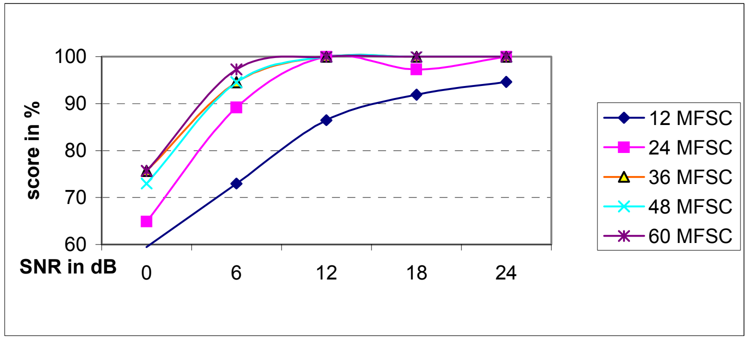

Fig. 5. Recognition scores in noisy environment (racket noise), in the 0-8 kHz bandwidth

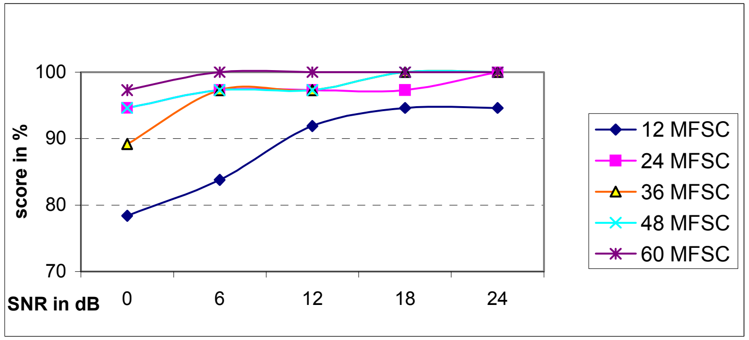

Fig. 6. Recognition scores in noisy environment (car noise), in the 0-8 kHz bandwidth

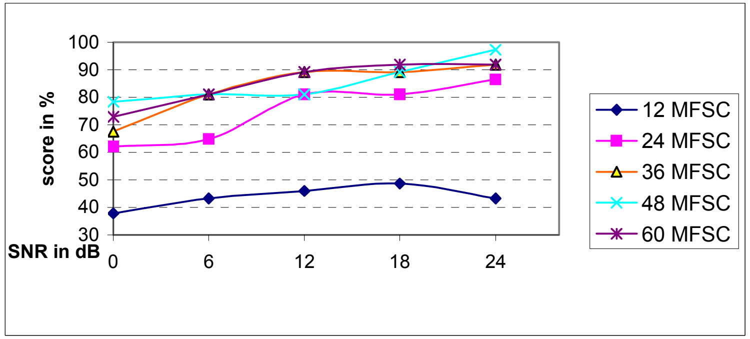

Fig. 7. Recognition scores in noisy environment (white noise), in the telephonic bandwidth

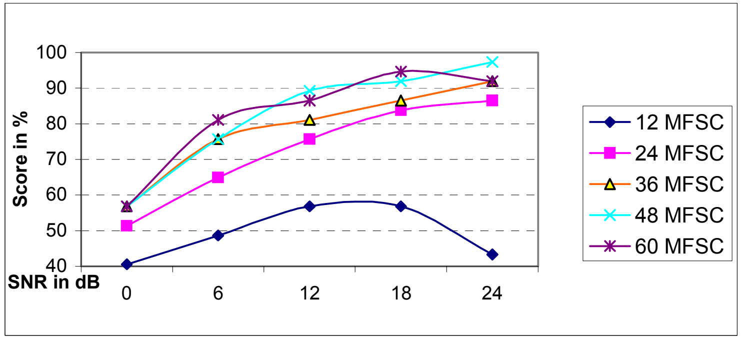

Fig. 8. Recognition scores in noisy environment (racket noise), in the telephonic bandwidth

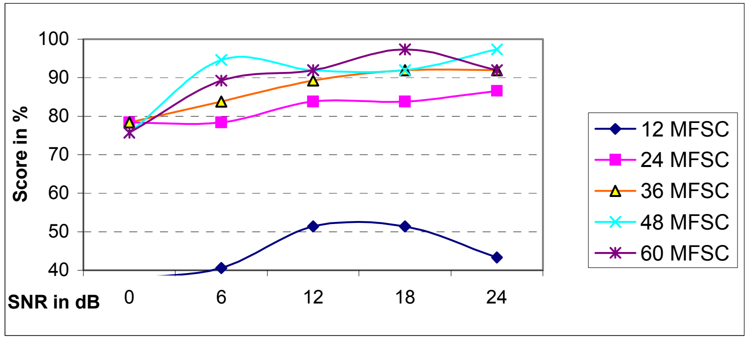

Fig. 9. Recognition scores in noisy environment (car noise), in the telephonic bandwidth

### *B. Discussion and conclusion*

The first conclusion we can deduce is that the high spectral resolution provides a lot of information characterizing the speaker, which improve the speaker recognition especially in noisy environment. Thus the 60 channels resolution should be an interesting resolution for the systems of speaker recognition implemented in noisy environment. However, on telephonic bandwidth, the optimal resolution should be between 48 and 60 channels.

The second conclusion we can deduce is that the car noise is not disruptive in speaker recognition; unlike we could think, because of the auditory disturbing provoked by this noise. In the other hand, the racket noise, which is well filtered by our brain, seems to be extremely disruptive in speaker recognition, even more disruptive then the white noise. The most probable cause of this failure is that the racket noise and the speech signal have the same type of features.

## VII. General conclusion

Usually, in statistical approaches using spectral features, one prefers to use low-resolution dimensions (*from 12 to 24 coefficients*) in the Mel-spectral modelization. This low spectral resolution has two advantages: a simplification of calculations and a good representation of the formantic envelope in the spectrum.

In another hand, when we use high dimensions, we shall deal with two problems: the complexity of calculation and the bad modelization of the covariance matrix. However, given the fact that the high-frequency information can enhance the discrimination between the speakers, it will be interesting to introduce that by rising the spectral resolution. This issue needs to find an optimal compromise in speaker characterization.

During this research work, we have proved the importance of high spectral resolutions in speaker authentication and we have found that the optimal spectral resolution depends on several parameters, especially the spectral bandwidth, the SNR and the type of noise mixed with the speech signal (if any).

The experiments done in both noisy and quiet environment (white noise, racket and car noise) showed that the high spectral resolution provides very important information for the speaker and helps recognise him more accurately, particularly in noisy environment.

Again, in the case of noisy environment, the resolution of 60 coefficients / 8 kHz seems to be very interesting for speaker recognition in the microphonic bandwidth. In the other hand, in the [300-3400 Hz] bandwidth, the optimal would be 48 coefficients / 8 kHz in non-noisy environment and would be a compromise between the two dimensions (48 and 60) when the speech is corrupted. We think that this little reduction is caused by the limitation of the telephonic filtering which rejects the entire part of the spectrum over 3400 Hz and which also means a significant loss of the high-frequency information.

Finally, we should recall that the results of this research work are specific to only one approach of speaker recognition; namely the second order statistical approach associated with the Mel-spectral modelization. And we should not expand these results for other approaches used in the same task, without redoing the experiments described in this paper.

## References

[1] I. MAGRIN-CHAGNOLLEAU, J. F. BONASTRE and F. BIMBOT 1995, "Effect of Utterance Duration and phonetic Content on Speaker identification Using Second-Order Statistical Methods", ESCA EUROSPEECH'95, vol. 1, pp 337-340, sep. 95, Madrid.

[2] W. FISHER, V. ZUE , J. BERNSTEIN and D. PALLET 1986, "An acoustic-phontic database", JASA, suppl. A, Vol. 81(S92) 1986.

[3] I. MAGRIN-CHAGNOLLEAU, J. WILKE and F. BIMBOT 1996, "A Further Investigation on AR-Vector Models for Text-Independent Speaker Identification", ICASSP, pp 401-404, May 7-10 1996, Atlanta, GA.

[4] H.S. LEE, A.C. TSOI 1995, " Application of multi-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment ». Speech Communication, Volume. 17, Number, 1-2, August 1995, pp. 59-76.

[5] H. SAYOUD et al 2003, 'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation. Publication dans la revue Acta Acustica, No4, Vol 89, 2003. pp 702-710.

[6] B.A. DAUTRICH, L.R. RABINER and T.B. MARTIN 1983, "The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer", Bell System Technical Journal, 1983.

[7] F. BIMBOT, I. MAGRIN-CHAGNOLLEAU, and L. MATHAN 1995, "Second-Order Statistical Measures for text-independent Broadcaster Identification". Speech Communication, Vol. 17, No 1-2, Aug. 95, pp. 177-192.

[8] F. BONASTRE, L. BESACIER 1997, "Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur". Actes du 4e Congrès Français d'Acoust., pp 357-360, Marseille 14-18 Apr. 97.

[9] H. Gish 1990, Robust discrimination in automatic speaker identification. IEEE Intern. Conference on Acoustics Speech and Signal Processing. April 90, New Mexico, 289-292.

[10] G. R. DODDINGTON 1998, " Speaker Recognition Evaluation Methodology. An Overview and Perspectives", RLA2C Avignon, 20-23 Apr 98, pp 60-66.

[11] D.A. REYNOLDS 1994, "Speaker identification and verification using Gaussian Mixture speaker models", Workshop on Automatic Speaker Recognition, identification and verification", April 1994, Martigny, Switzerland, pp. 27-30.