Speaker identification system using LPC -Application on Berber language

Article · December 2015

DOI: 10.5281/zenodo.45765.

1 author:

Fatma Zohra Chelali
University of Science and Technology Houari Boumediene
35 PUBLICATIONS **135** CITATIONS

SEE PROFILE

Speaker identification system using LPC -Application on Berber language

F. Z. Chelali, K. Sadeddine, A. Djeradi

LCPTS Laboratory
University Houari-Boumediene of Sciences and Technology

***Abstract*—** The aim of our article is to design a speaker identification system applied to the Berber language, based on linear predictive coding parameterization and maximum likelihood classification.
Our objective is to establish a system for speaker identification in dependent and independent mode using continuous speech with duration of 2 or 3 seconds. To ensure that task, we have decided to employ some features based on the Linear Predictive Coding (LPC). The used classifier is based on the maximum likelihood principle. Furthermore, a speech dataset from Berber language and using emphatic syllables was recorded in laboratory and analysed off-line.
The experimental evaluations were performed using the first and second derivatives of the descriptors, and the fundamental frequency (i.e. pitch). Results show that the proposed features and approach are interesting in speaker identification.

Keywords—Speech Processing; Automatic Speaker Identification; Berber Language.

I. Introduction

Automatic speech and speaker recognition have got, these recent years, a great interest from many researchers in the design of systems that recognize sounds or to identify speakers. In this article, we are interested primarily in speaker recognition in dependent and independent mode of the text applied to the Berber language. We distinguish between dependent and independent speaker recognition while presenting our results.

In the dependent mode, the text uttered by the speaker is the same used in the learning phase but in the independent mode, the speaker uttered a different text from the existing in the learning phase.

For this purpose, an acoustic corpus for sentences in continuous speech of the Berber language has been recorded at speech communication and signal processing laboratory. Like many studies in different languages; the Berber language, widely used by a lot of people in North Africa, presents an important area of research interest for many scientists. A dataset of continuous speech containing Berber emphatic for 6 speakers was used. Several studies have been presented in many languages; and the intra-personal and extra-personal variability present some problems to the designers of automatic speech/speaker recognition systems.

(Saeed and Kheir Nammoud, 2003) in their speaker identification system in dependent mode for the digit (zeros at ten) in the Arabic language permits a recognition rate of 98.8 per cent using the Burg estimation model.

(Tanprasert & al, 2000) in their work presents a system for speaker identification of dependent mode applied to the Thai-language using a neural network, whose characteristic parameters are the LPC coefficients, which permited a recognition rate of 95 %.

(Toutios and Margaritis, 2002) for the development of a system called TOOLKIT GIO (identification in speaker dependent mode) have extracted the MCCF coefficients and the PLP parameters and used a neural network , they obtained a recognition rate of 93.10 per cent for the first dataset in the English language.

A speaker recognition system typically consists of three stages: feature extraction, speaker modeling, and decision making using pattern classification methods.

Therefore, our work is to establish a system of speaker identification in dependent and independent mode using continuous speech with a duration of 2 or 3 seconds. For this reason, our choice focused on the coefficients extracted from the linear predictive analysis (LPC) and the fundamental frequency. The used classifier is based on the maximum likelihood.

The second section will give a brief description of the phonetic and phonological characteristics of the Berber language. The third section discusses the presentation of the corpus, the characterization phase using the LPC vectors descriptors. To use dynamic information contained in speech, our approach consists in extracting first and second order derivatives of the LPC features, this section presents results obtained during the identification process. Furthermore, problems in speaker variability are discussed using the fundamental frequency. A conclusion summarizes the results obtained and future work.

II. Berber language

The phonetic classification is distinguished into classes and sub classes; we distinguish between the consonants and vowels. We have a vowel when the mode of production is characterized by the free passage of air in cavities located above the glottis. Most of the vowels used in the languages are audible; they are uttered with a vibration of the vocal cords to the contrary of the deaf vowels that are pronounced without vibration of the vocal cords. We present in the following section a few properties of the Berber language.

The Amazigh language, known as Berber or Tamazight, is a branch of Hamito-Semitic (Afro-Asiatic) languages. It is, spoken in a vast geographical area of North Africa. Amazigh covers the northern part of Africa which extends from the Red Sea to the Canary Islands and from the Niger and Mali (Touareg) in the Sahara to the Mediterranean Sea (Satori &al, 2014).

In Algeria, the principal Berber-speaking region is Kabylia. In a relatively limited but densely populated surface area, Kabylia alone has two-thirds of Algeria's Berber speakers. The other significant Berber-speaking groups are: the Chaouias of the Aures region, having in all likelihood a million people, and the people of the Mzab (in Ghardaia and other Ibadhite cities), having a population of between 150,000 and 200,000. There are in fact other Berber-speaking groups in Algeria, but these are modest linguistic islands of only several thousands to tens of thousands of speakers (Chaker,2004).

Kabyle has three phonemic vowels: open /a/, and close /i/, /u/, similarly to Classical Arabic. <e> is used to write the epenthetic schwa vowel [ə] which occurs frequently in Kabyle. Historically it is thought to be the result of a pan-Berber reduction or merger of three other vowels. The phonetic realization of the vowels, especially /a/, is influenced by the character of the surrounding consonants; emphatic consonants invite a more open realization of the vowel, e.g. aẓru = [azˤru] 'stone' vs. amud = [æmud] 'seed'. Often /a, i, u/ are realized as [æ, ɪ, ʊ].

III. Simulation results

Identification and automatic speaker verification has grouped several researchers in this area: more recently, the application needs have given rise to new tasks such as speaker indexing of audio stream, or the monitoring of speakers or new variants such as detection of speaker in a conversation.

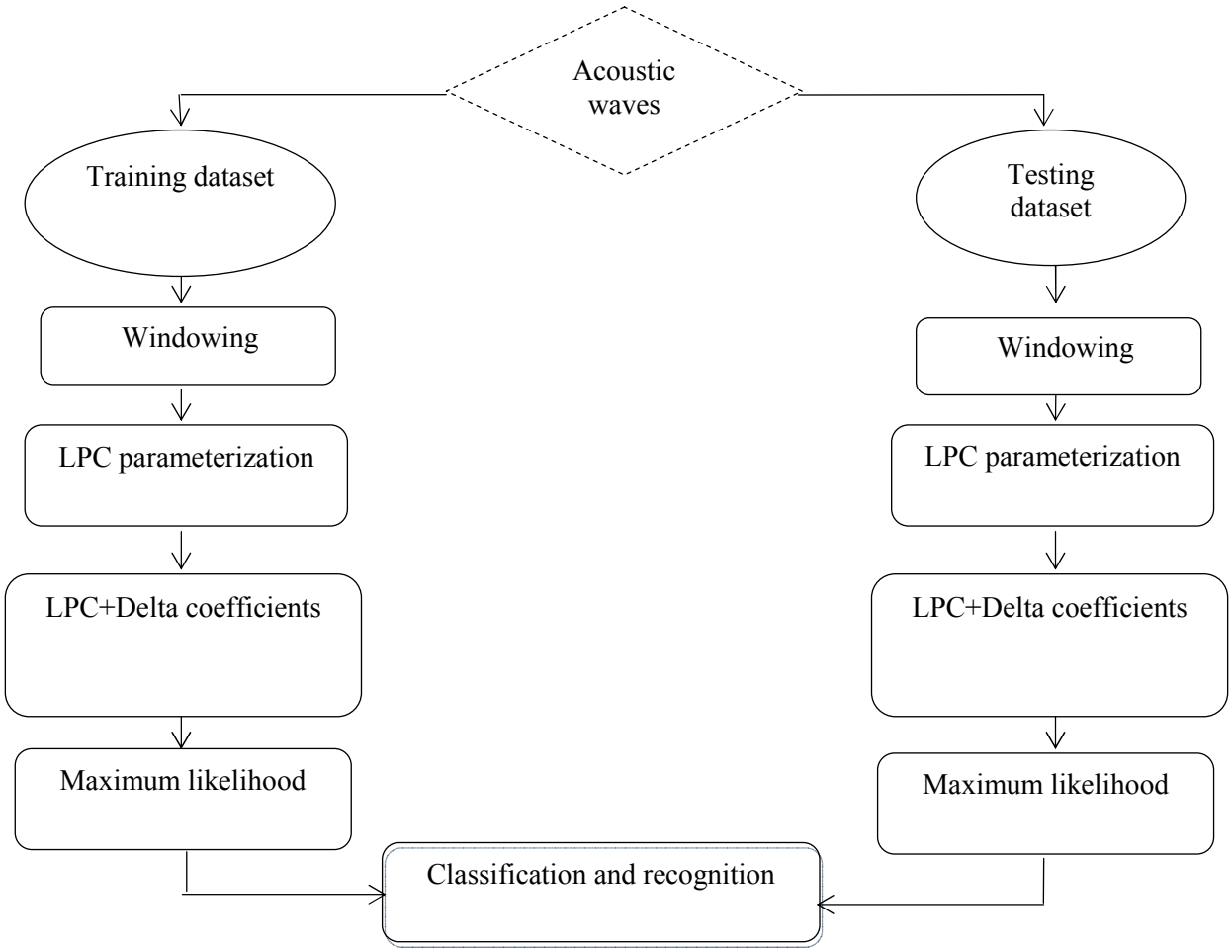

Figure 1. Architecture used in the speaker identification system

III.1 Speech dataset

We chose six speakers for the Berber language (kabyle) which are: Nacera, Zohra,Gassi, Kaci, and Smail, all native to the region of Kabylia. These speakers repeated with an average speed and an average energy six sentences in continuous speech. The sound card we used is called MobilePre USB. MobilePre USB is a preamplifier mobile integrating an audio interface efficient for recordings on computer (laptop or mobile).

The dataset includes speech signals from six (6) different subjects. The speech signals are acquired during 2 s or 3s with different sessions to consider all variations at a sampling rate of 16 kHz. After, a manual segmentation with a specialized software is done to consider only the

part of the voice. During the recording, each repetition has been analyzed to ensure that the entire sequences have been properly stored and avoiding any external interference. The chosen sequences are as follows:

- *"assagui yelha lhal";*
- *"azeka anrouh arthemourth";*
- *"anzoum remthane g nevdhou";*
- *"yetchak izeme yellozen";*
- *"yeghine wagroud aghour yemasse" and*
- *"thelha thazalith gakhem rebi".*

The first four sentences have duration of two seconds, while the last sequences have equal duration of 3 seconds. The segmentation has been performed manually by using the specialized recording software in order to take only the essential part that has been produced. Furthermore, a standardisation of the sound files has been performed.

III.2 Speech parameterization

There are several methods to extract the characteristics of a voice signal, such as the spectral coefficients, cepstral and others. Among the most used descriptors, and simple to implement are those derived from the linear predictive coding (LPC). This method allows converting information into a vector containing the most important characteristics.

Linear Predictive Coding (LPC) is used to compress the speech signal without losing its audibility. The speech is divided into segments and then transmits the voiced or unvoiced information, the pitch period and the coefficients for the filter. It is one of the most powerful methods used in audio and speech signal processing which extracts speech parameters like pitch formants and spectra. The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech and estimated speech signal over a finite duration (Ambika & al, 2012).

It allows encoding of good quality speech at a low bit rate and it also provides extremely accurate estimation of speech parameters. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples. The predictor coefficients are represented by a_k and it is normally estimated in every frame with size of 20 ms long. Another important parameter is the gain (G). The transfer function of the time varying is an all-pole filter with a sufficient number of poles. It is a good approximation for speech signals. Thus, we can model this filter H(z) as follows (Ambika&al, 2012):

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \quad (1)$$

The summation is computed starting at k=1 up to p (where p is the order of the LPC) analysis. This means that only the first p coefficient are transmitted to the LPC synthesizer. Hence guaranteeing the stability of system H (z), Levinson – Durbin recursion will be utilized to compute the required parameters for the auto-correlation method (Ambika&al, 2012).

Our goal is to calculate the characteristic vectors of LPC coefficients, characterizing the speech signals, for the whole of the repetitions and the set of speakers. We present the important steps to the calculation of the LPC coefficients. We start by standardizing the segmented signals of our dataset in order to achieve the same size of the sound file for each phrase taken alone.

Each voice signal is decomposed into a number of frames. For each frame (or window) we shall determine p coefficients LPC, which will provide a vector of p x m coefficients, where m is the number of frames for each sound file.

The LPC coefficients are calculated for the new standardized signals and stored in a matrix of j lines and i columns (i= 5) for each speaker for the text dependent mode, and j lines and i columns (i= 5) for each sentence for the text independent mode. We recall that 5 (i.e. i=5) represents the five repetitions of each matrix.

The decision module will allow us to designate the recognized speaker. We will use in our work the correlation method for speaker recognition.

To estimate the quality of the recognition, it is necessary first, to calculate the recognition rate (RR %). Note that the threshold of the correlation coefficient (R) that we have chosen is 0.75 and which is considered suitable for our data.

$$RR(\%) = \frac{Number\ of\ files\ whose\ coefficients\ R > 0.75}{Total\ number\ of\ files} * 100 \quad (2)$$

The following histograms show the recognition results obtained for the Berber corpus in a dependent mode, where the order of prediction varies from 10 to 18. The orders of prediction that we have chosen, are respectively: 10, 12, 14, 16 and 18 in order to analyze its influence on the recognition rate.

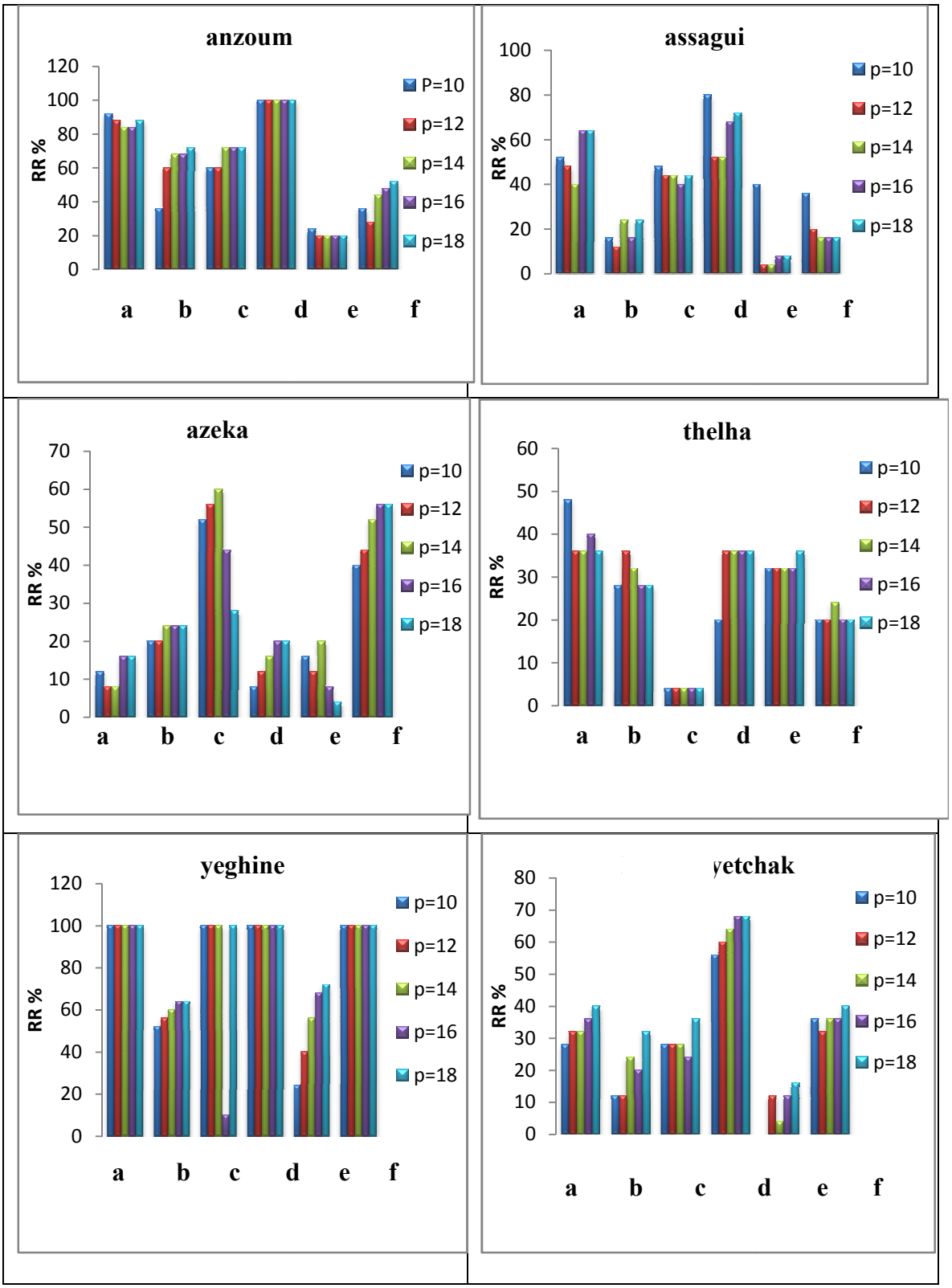

Figure 2. Recognition rate (RR %) in dependent mode by varying the order of prediction;
(a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f) speaker"Gassi".

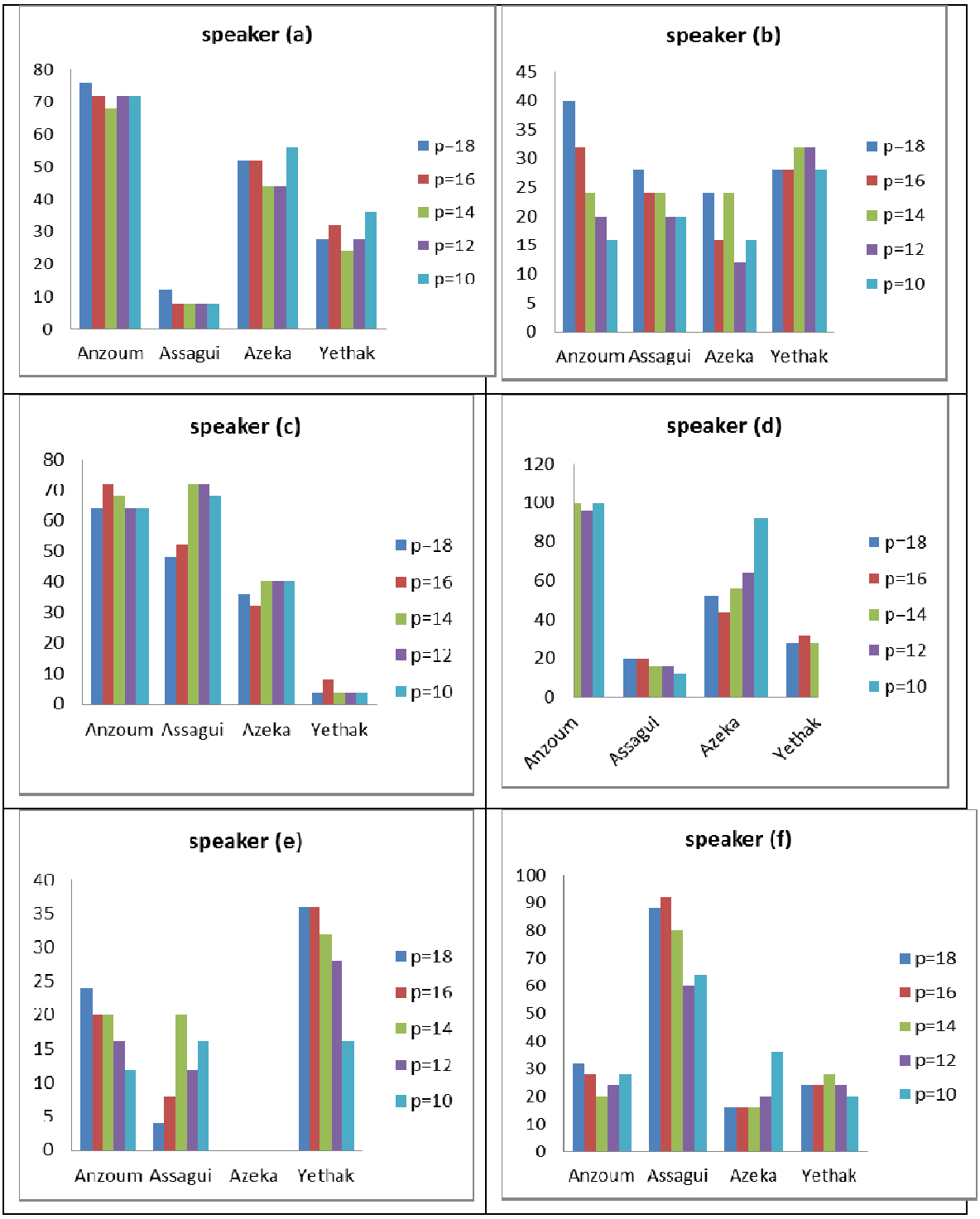

Figure 3. Recognition rate in independent mode by varying the order of prediction (sequences of 3 seconds) ; (a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f)speaker"Gassi".

According to the obtained results, the recognition score is higher in dependent mode of text comparably to the independent mode. In addition, we note that the score of correct classification is higher for an order of prediction of 16 and 18 whereas it decreases for a lower order to 12.

We will subsequently be interested in first and second derivatives of LPC coefficients, widely used in speech/speaker recognition, we will analyse their influence on the recognition rate and discuss the results obtained for each mode and for the total dataset.

The derivatives of first and second order are often used in speech or speaker recognition. This allows us to add information concerning the dynamics of the signal (Holmes, 2003).

Dynamic features are a measure of the change in the static features. These dynamic features are often referred to time derivatives or deltas. One way of computing the delta features is by simply differencing between the feature values for two frames (Holmes, 2003):

$$\Delta y_t = y_{t+D} - y_{t-D} \quad (3)$$

Where *D* represents the number of frames to offset either side of the current frame and thus controls the width of the window over which the differencing operation is carried out. Typically, *D* is set to a value of 1 or 2. Although time-difference features have been used successfully in many systems, they are sensitive to random fluctuations in the original static features and therefore tend to be 'noisy'. A more robust measure of local change is obtained by applying linear regression over a sequence of frames (Holmes, 2003).

The delta coefficients (derived from the first order) are often estimated with the following equation:

$$\Delta y_t = \frac{\sum_{\tau=1}^{D}(y_{t+\tau} - y_{t-\tau})}{2\sum_{\tau=1}^{D}\tau^2} \quad (4)$$

where D = 1 is the habitual choice for an analyzed window.

where Δy_t is the delta coefficient calculated at frame t for LPC features vector y_t and τ is the time advance and delay for the delta computation. In most cases, all Delta LPC coefficients were concatenated to the original feature vector to form a new delta enhanced feature vector.

The coefficients derived from the second order delta-deltas are estimated in the same way from coefficients of the first order. The concatenation of these vectors can improve the efficiency of the recognition system.

The following histograms allow better visualization of the recognition results for the first and second derivatives vectors for the total sentences.

We therefore conclude that the first and second derivatives improve the recognition rates. In fact, the addition of the second derivatives (which capture the changes in the dynamics of first order) allows an additional improvement of the identification rate but remain small in front of the first derivative.

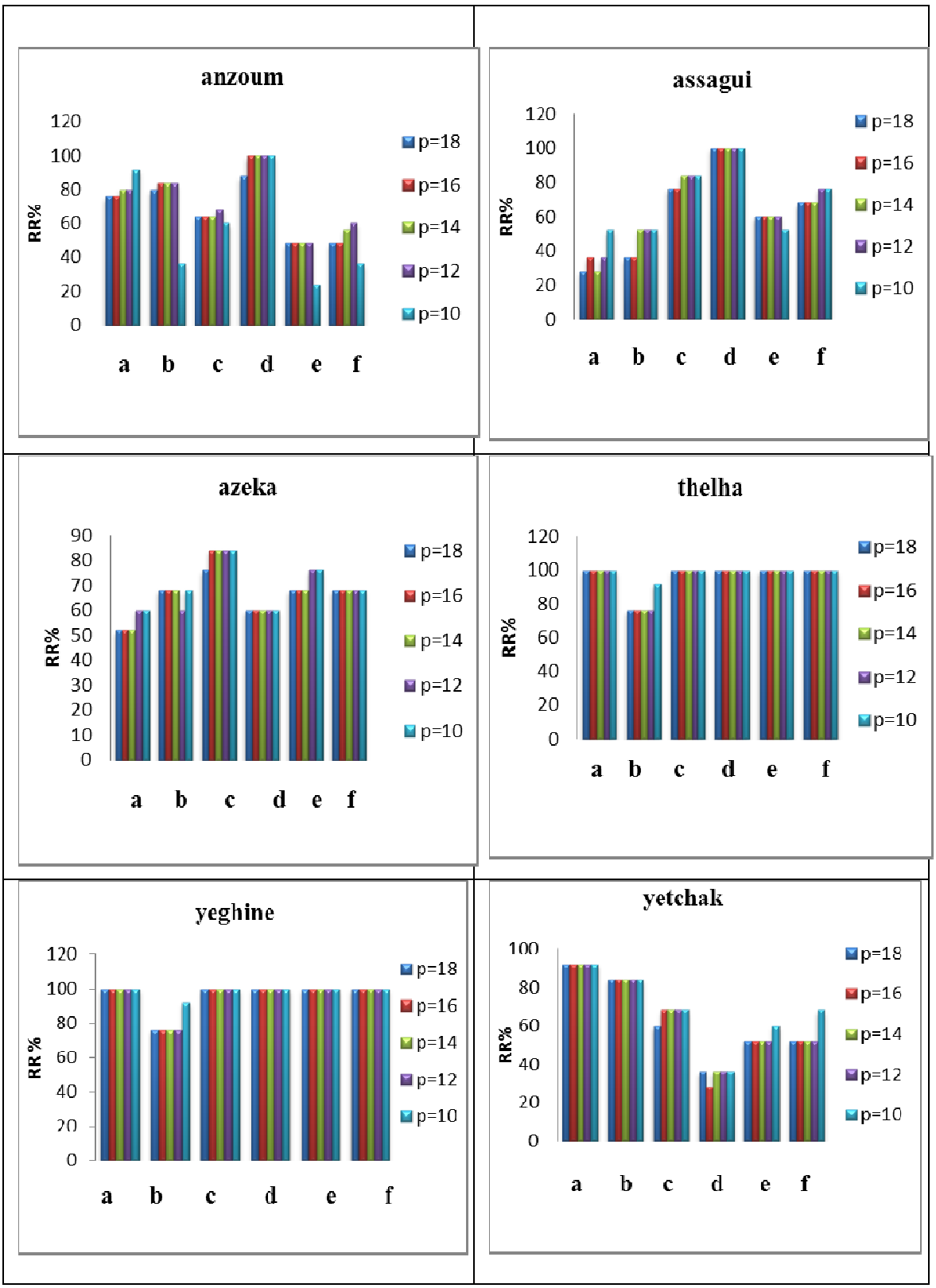

Figure 4. Recognition rate using the first derivative in dependent mode; (a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f)speaker"Gassi".

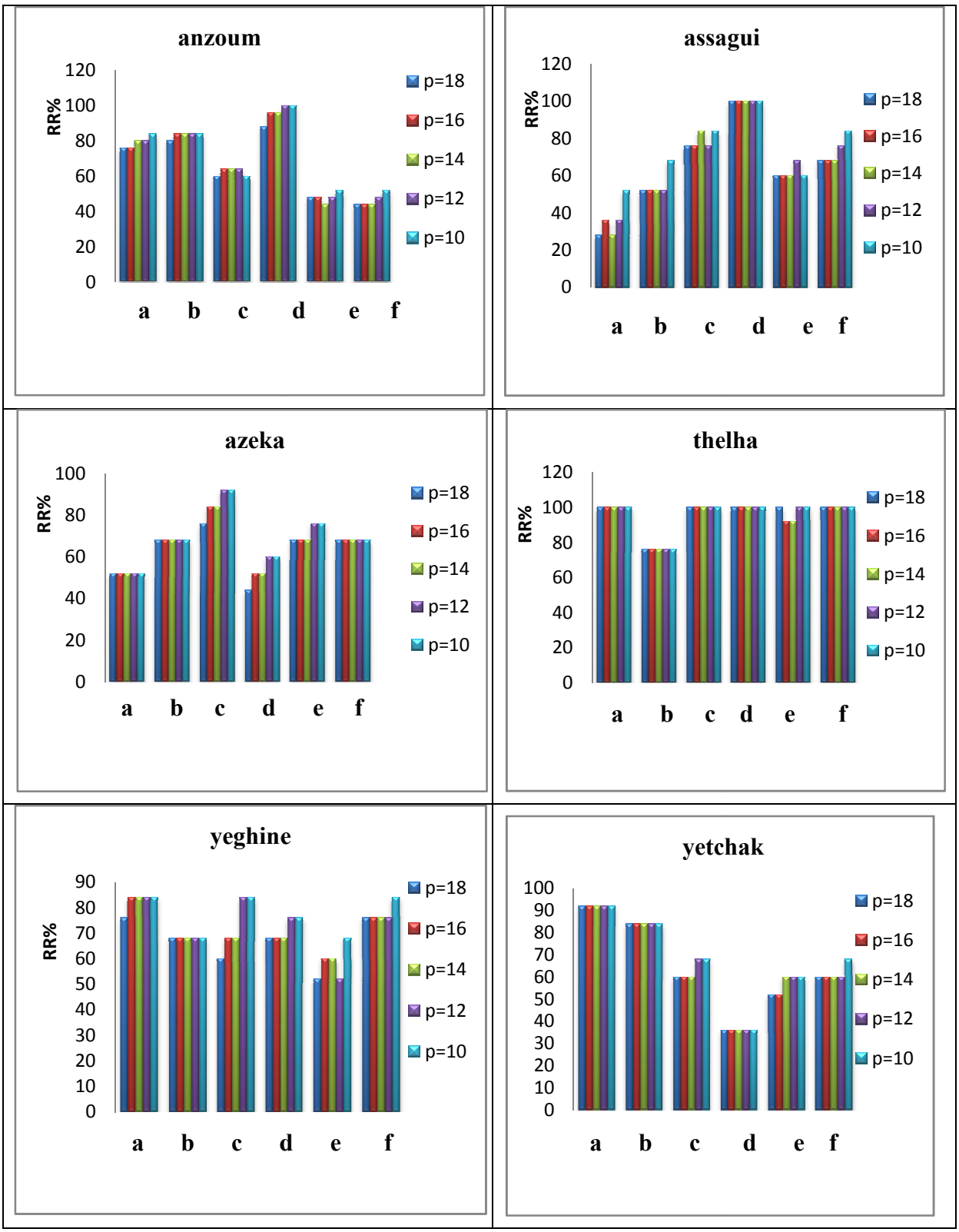

Figure 5. Recognition rate using the second derivative in dependent mode; (a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f)speaker"Gassi".

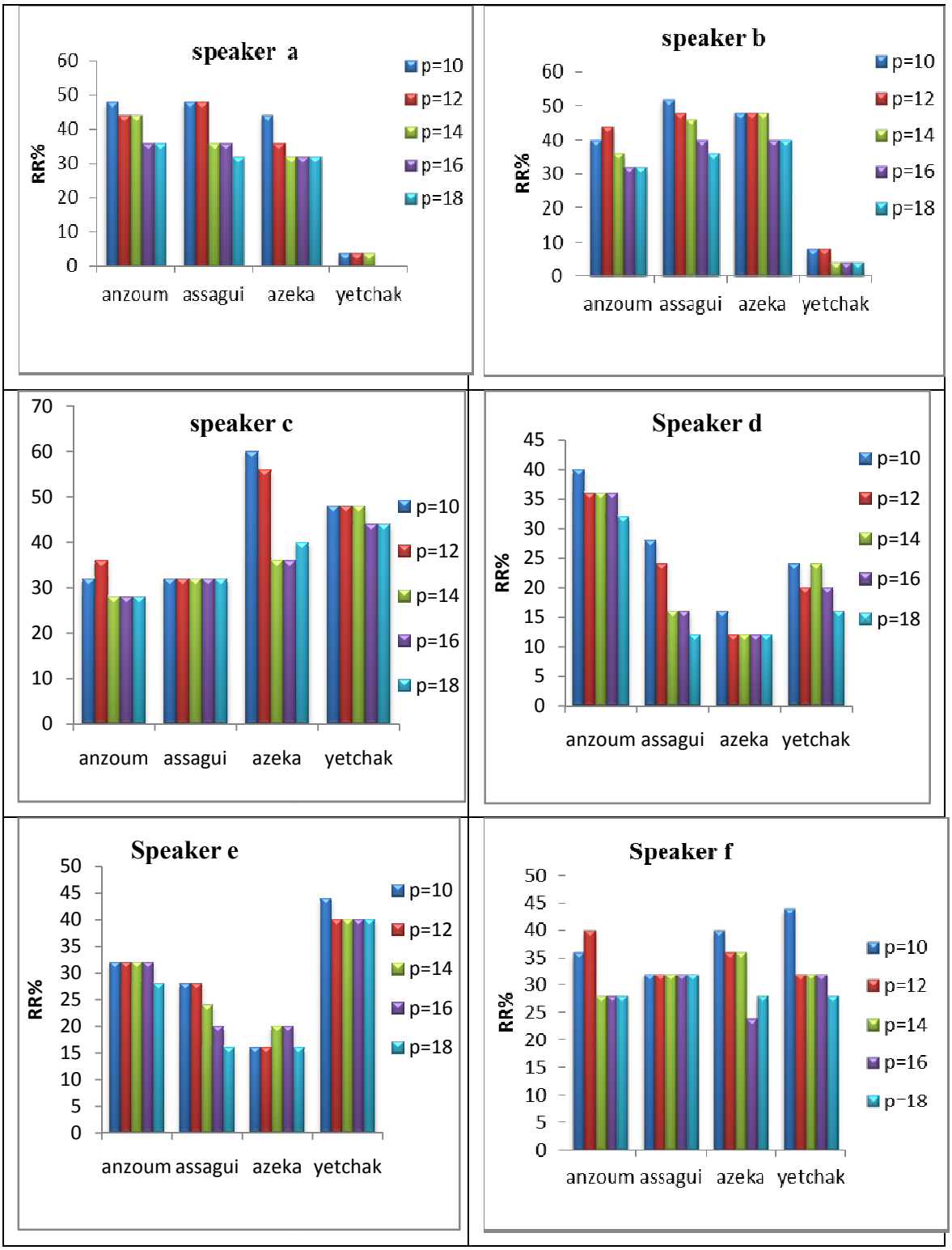

Figure 6. Recognition rate using the first derivative in independent mode (duration of 02 seconds) ; (a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f)speaker"Gassi".

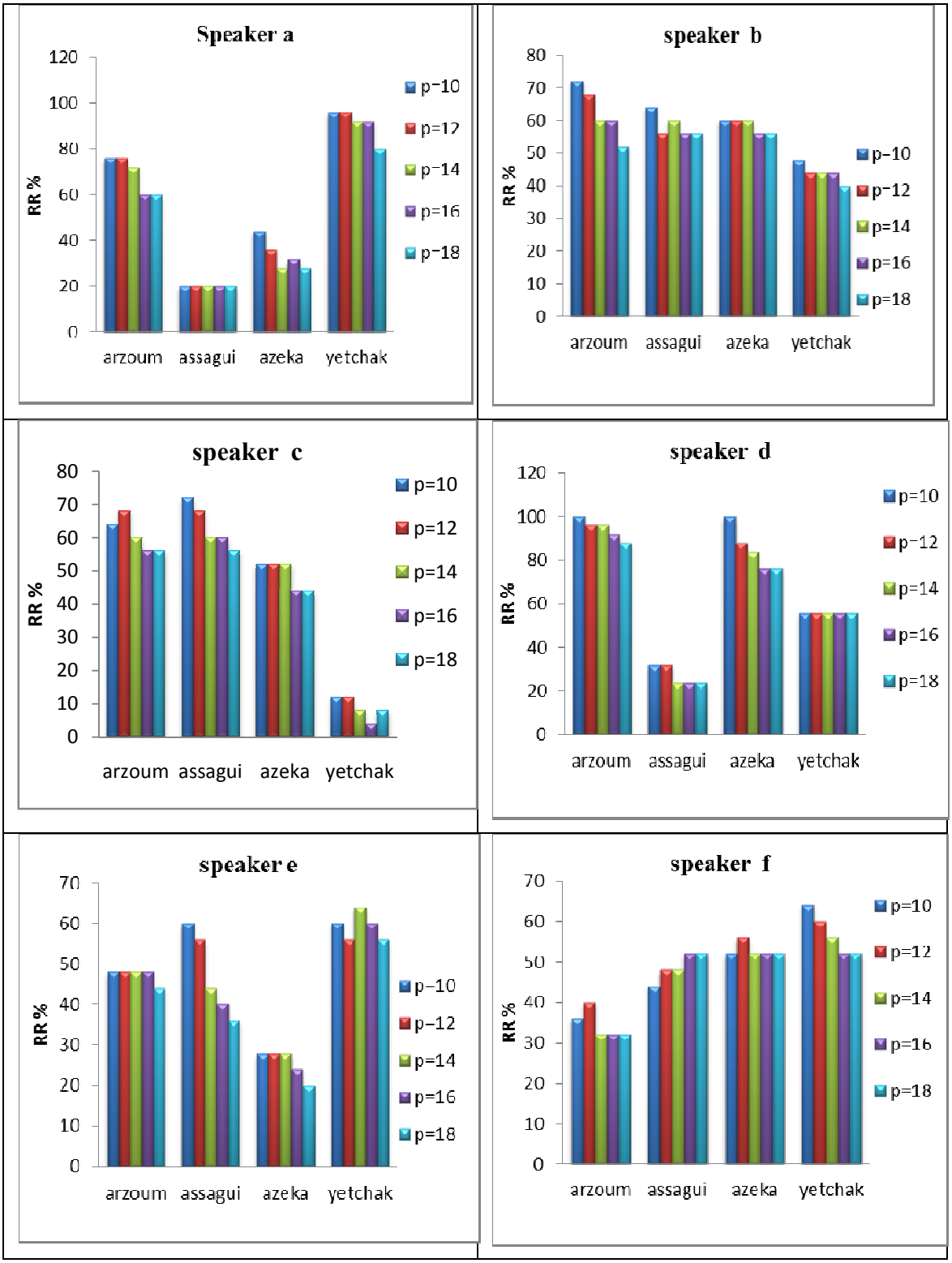

Figure 7. Recognition rate using the second derivative in independent mode (duration of 02 seconds) ; (a) speaker"Kaci"; (b) speaker"Nacera";(c) speaker"Ghania";(d)speaker"Smail"; (e)speaker"Zohra"; (f)speaker"Gassi".

The advantages of deriving the features are mainly due to their ability to capture the dynamic information. Most of the current systems include features derived from the first order to what it adds a function of energy, pitch and many include the second derivatives.

III.3 Speaker recognition using the pitch information

There are several methods of determining the fundamental frequency (pitch). We present in this paragraph the important steps in the calculation of the pitch using the LPC model. The calculation of the fundamental frequency is summarized as follows:

Begin

For each speech signal;

Apply the Butterworth filter;

Fix the maximal frequency of research to 500 Hz and the minimal to 80Hz;

Calculate the LPC coefficients for the total frames of each signal;

Find the excitation signal such as $H(z) = (S(z))/(E(z)) = 1/(A(z))$ and $E(z) = A(z).S(z)$

Then $e(n) = a(n) * s(n)$. So, we search the autocorrelation function of e(n);

Find the position of a maximum and search the first pick between Tmin et Tmax, where Tmin=1 /Fmin and Tmax=1/Fmax;

Affect the position of the first pick to the value of the pitch F_0;

End.

The following figure shows the variation of the pitch for the same sentence spoken by the same speaker " Kaci" at two different moments.

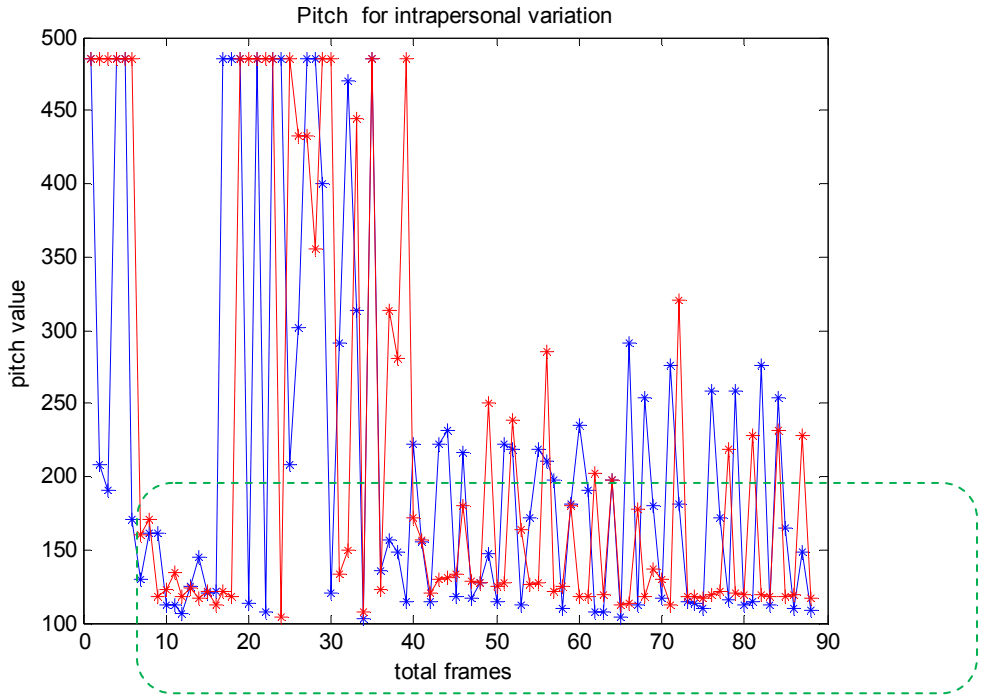

Figure 8. Pitch variation in intrapersonal variability

We note clearly a slight variation of the pitch for the same speaker. This shows the intra-personal variation. The same phenomenon is noted when two different speakers utter the same sentence (extra-personal variation.).

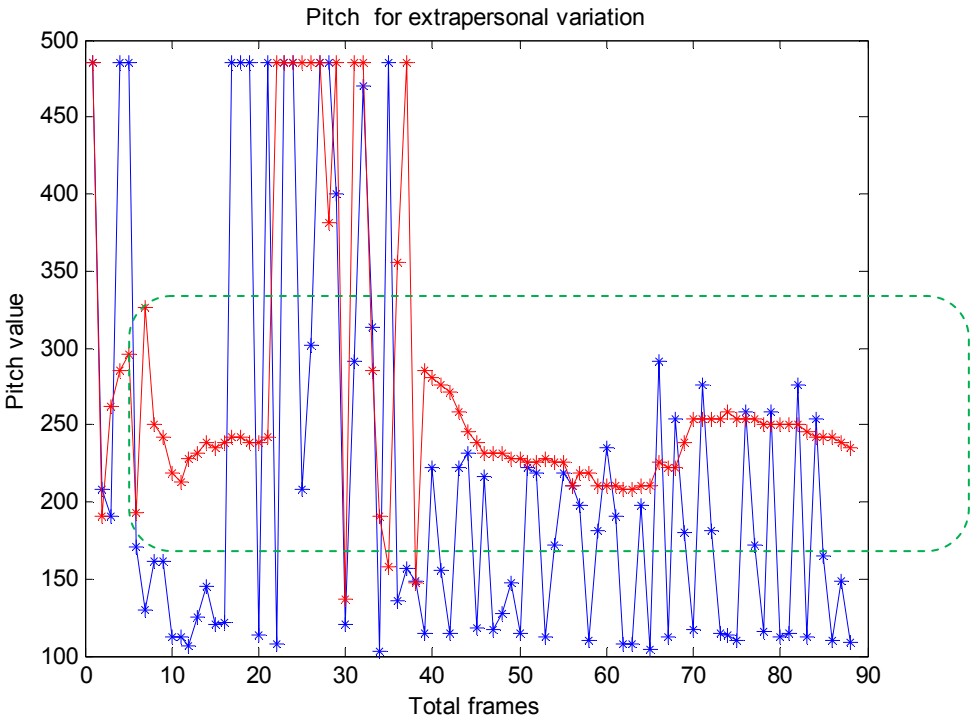

Figure 9. Pitch variation in extra personal variability

The main cause of inter-speakers differences is due to physiological nature. For the speaker recognition, it is important to extract the characteristics of the speech signal that present a high inter-speaker variability to be able to differentiate the speakers between them, and a low intra-speaker variability to ensure the robustness of the system (Josse, 2003).

Speech is mainly produced through the vocal cords, which generate a sound to a low frequency called the fundamental. This frequency is different from one individual to another and more generally between men and women.

The inter-speaker variability also has its origin in the differences of pronunciation, which exist within the same language and who constitute the regional accents. This property is valid for the standard Arabic or Berber language for instance.

That is, we have presented the results obtained for the development of an identification system based on the LPC analysis and the maximum likelihood. We have also improved the scores obtained by adding the derivatives of the first and second order. The results obtained by calculating the derived LPC coefficients are satisfactory.

In addition, the system based on the computation of the fundamental frequency shows the degradation of the identification system and this is probably due to the large variability existing within the same speaker.

IV. Conclusion

In this article we were interested in designing a speaker identification system; first by using the derived coefficients from the LPC analysis (i.e. their first and second derivatives); and then by using the fundamental frequency. The used classifier is based on the maximum likelihood scheme.

We have used a dataset of acoustic data recorded in the LCPTS Laboratory; it concerns phrases recorded in Berber. We have extracted, for both training and testing, the LPC coefficients, their derivatives of first and second order as well as the fundamental frequency. Subsequently, the classification process with a specific threshold of 0.75 allowed us to provide a good score of speaker identification (for each speaker or in the overall).

The obtained recognition scores were interesting: ranging from 80% to 96 %, for the combination LPC-derivatives (first and second), which appear to be better than the scores of the LPC coefficients used alone.

On the other hand, the performances provided by the fundamental frequency are relatively low, which is due to the large intra-variability of this parameter during the articulation of the acoustic signals.

Future work should investigate other types of features and/or classifiers to try improve the global recognition scores, hopefully.

References

Ambika D., Radha V., (2012), "A Comparative Study between Discrete Wavelet Transform and Linear Predictive Coding",,2012, ISBN 978-1-4673-4805-8/12/$31.00 , IEEE.

Hassan Satori & Fatima El Haoussi," Investigation Amazigh speech recognition using CMU tools", International Journal of Speech Technology, ISSN 1381-2416, DOI 10.1007/s10772-014-9223-y.2014.

Holmes J. and Holmes W. (2003). Chapter 10: Introduction to Front-end Analysis for Automatic Speech Recognition .2ème edition, Speech Synthesis and Recognition. Taylor and Francis e-Library.

Josse V. (2003). Identification nommée du locuteur: Exploitation conjointe du signal sonore et de sa transcription. Thèse de doctorat, Ecole doctorale, Académie de Nantes. Université du Maine. France.

Saeed K .et Kheir Nammous M. (2003). A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image. IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, VOL. (54), NO. 2.887-897 APRIL 2007.

Tanprasert C., Wutiwiwatchai C. et Sae-tang S.(2000). Text-dependent Speaker Identification Using Neural Network On Distinctive Thai Tone Marks. Technical Journal.vol(1), NO.6,249-253.

Toutios A. et K. G. Margaritis.(2002). Development of a Text-Depe ndent Speaker Identification System with the OGI Toolkit. Second Hellenic Conference on Al, SETN-2002,Thessaloniki, Greece, Proceeding, Companion Volume,525-530.

Salem CHAKER. (2004). “Berber, A “Long forgotten” Language of france”, Professor of Berber Language, INALCO (Paris). Report. Translated by Laurie and Amar Chaker).