A&QT-R 2002 (THETA 13)
International Conference on Automation, Quality, and Testing, Robotics
May 23-25, 2002, Cluj-Napoca, Romania

ON THE INFLUENCE OF THE FIRST FEATURES COEFFCIENTS OVER SPEAKER RECOGNITION

Pop G. Petre, Toderean Gavril, Lupu Eugen

Technical University, Cluj-Napoca
Gh. Baritiu str. 25-27,fax: 0040-64-191689, Romania, E-mail:Petre.Pop@com.utcluj.ro

Abstract : Speaker recognition requires speaker characteristic features, independent of the particular spoken word if possible. Speaker identity is correlated with the physiological and behavioral characteristics of the speaker, both encoded in the spectral envelope and in the supra-segmental features (voice source characteristics and dynamic feature spanning over several segments). Features based on short-term spectral estimate have a strong dependece on individual speakers and consequently are used in speaker recognition. However, these features also contain information about the lexical content of the utterance. That's why some coefficients need to be removed and some need to be emphasize while other need to be deemphasized.
We study the influence of eliminating some coefficients from features (LPC, LPC cepstrum, MFCC) on speaker recognition using DTW.

Key words : *LPC, LPC cepstrum, MFCC, speaker recognition, DTW.*

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. The speech waveform is sampled at a rate between 8kHz and 22kHz and processed to produce a new reprezentation as a sequence of vectors containing values of what are generally called parameters. The vectors typically comprise between 10 and 20 parameters, and are usually computed every 15 to 30 msec. Representations used in speaker recognition concentrate primarly on properties of the speech signal atributtable to the shape of the vocal tract. Representations are almost always derived from the short time power spectrum, ignoring the phase structure, primarly because human ears are very insensitive to phase effects.
Usually the speech signal is preproccesed before extracting parameters by [4] :

- preemphazis : the speech samples are filtered by a pass-high digital filter in order to offset the natural slope (attenuation of 20dB/dec) due to physiological characteristics of the speech production system, thereby improving the efficiency of analysis;
- frame blocking : the speech signal is blocked into frames of N samples with an overlap factor of frames M (M<= N);
- windowing : favor the samples towards the center of the window; this fact coupled with the overlapping analysis performs an important function in obtaining smoothly varying parametric estimates; the Hamming window is the typical window in speech processing.

However, these spectral parameters also contain information about the lexical content of the utterance. That's why some coefficients need to be removed and some need to be emphasize while other need to be deemphasized.
We study the influence of eliminating some of the first coefficients from features (LPC, LPC cepstrum, MFCC) on speech derived spectrum and on speaker recognition using DTW.

2. THE INFLUENCE ON THE DERIVED SPECTRUM

2.1 LPC analysis

LPC (Linear Predictive Coding) analysis provide a good model for speech signal and works well in recognition systems [3]. This technique fits the parameters of an all-pole model to the speech spectrum, though the spectrum itself is not computed explicitly. First, the autocorrelation is evaluate and the results are converted into a set of LPC coefficients folowing the Levinson-Durbin recursive method [2].

The LPC derived spectrum is more smoothed than FFT spectrum (fig. 1) :

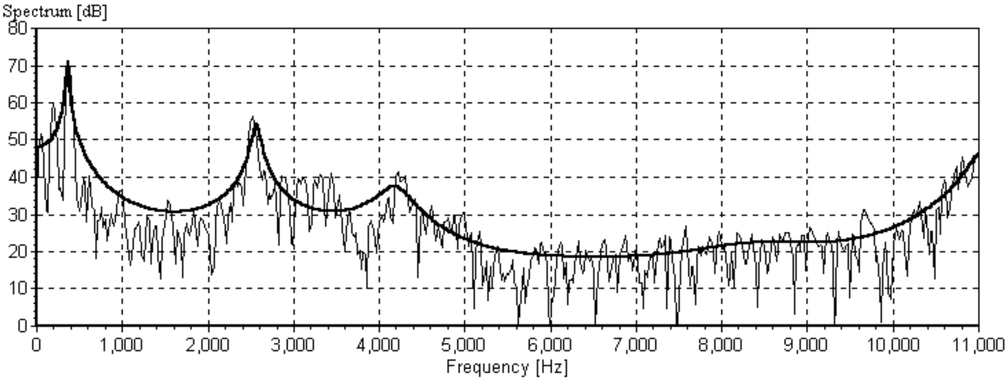

Figure 1. Power spectrum and LPC derived spectrum.

The effect of elimating first k_0 coefficients over derived spectrum is shown in fig. 2.

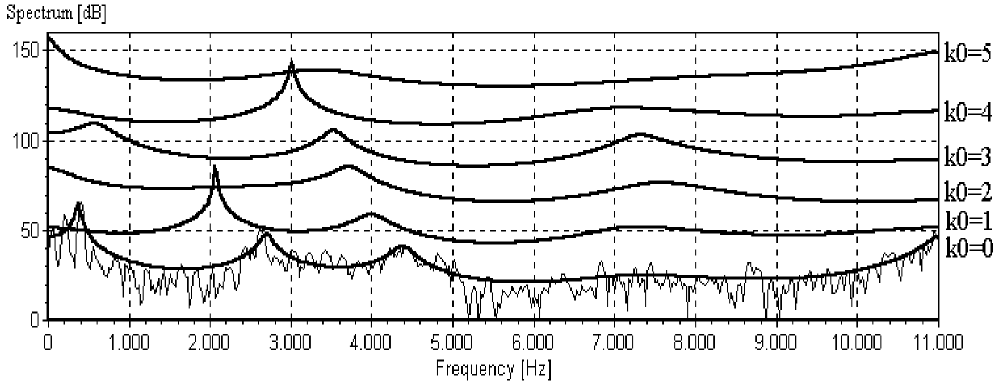

Figure 2. Power spectrum and LPC derived spectrum for k_0=0…5.

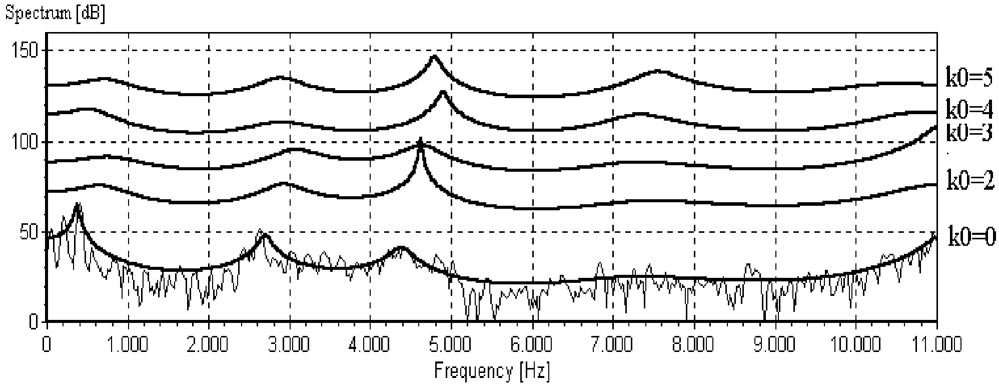

Figure 3. Power spectrum and LPC derived spectrum for k_0=0, 2…5.

It is clear that the spectrum is distorted. For k_0=3, the number of maxima (corresponding to speaker formants) are preserved but the maxima's positions are altered. The first coefficient (c_0) correspond to the frame energy and is usually removed from the parameter set and replaced with the energy evaluated before any other processing (preemphasize and windowing). So, we maintain first coefficient and remove the subsequent coefficients as in fig. 3. In this case, the spectrum is not so much affected and the formants positions are very closely to that of the original spectrum. Additionaly, maximas from higher frequency domain are accentuated.

2.2 Cepstral analysis

Cepstrum is defined as the inverse FFT of the logarithm of the spectrum. This allow to separate the contribution of vocal tract from that of the source in the speech signal. Usually LPC coefficients are converted directly to cepstrum coefficients using some recursive relations [5].

The LPC derived cepstrum give a much smoother spectrum than that obtained from LPC (fig. 4).

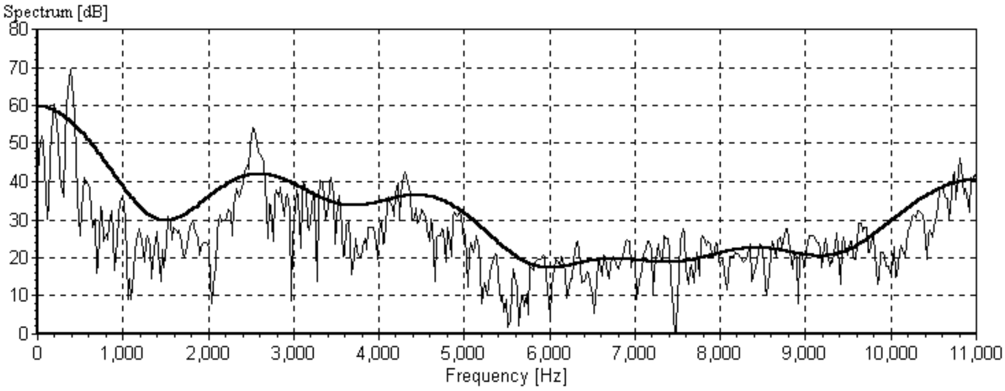

Figure 4. Power spectrum and LPC cepstrum derived spectrum.

Therefore LPC cepstrum provides a stabler representation from one repetition to another of a speaker's uttrenaces. Aditionally, LPCC coefficients are uncorrelated which is a premise for good results in speaker recognition.

The effect of elimating first k_0 coefficients over derived spectrum is shown in fig. 5. In this case, the specra are not so much affected than in LPC case. For k_0=1, the new spectrum is very closed to the original, especially in first half of frequency domain. By maintaing first coefficient and removing subsequents there is no major changes in spectrum (fig. 6).

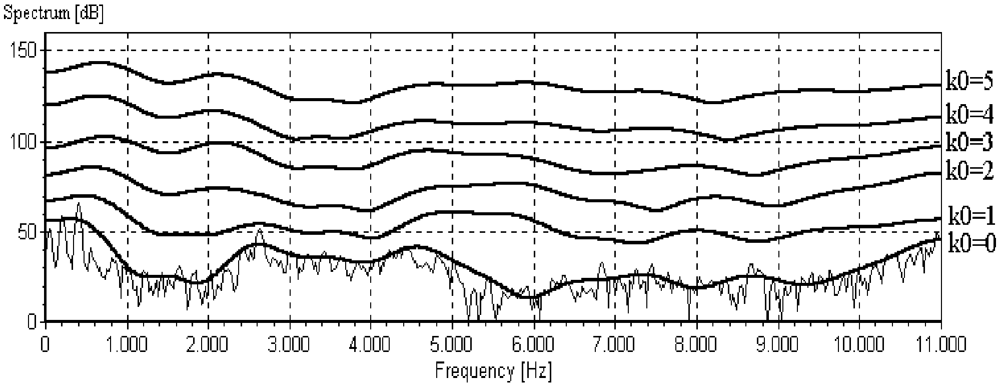

Figure 5. Power spectrum and LPC cepstrum derived spectrum for k_0=0...5.

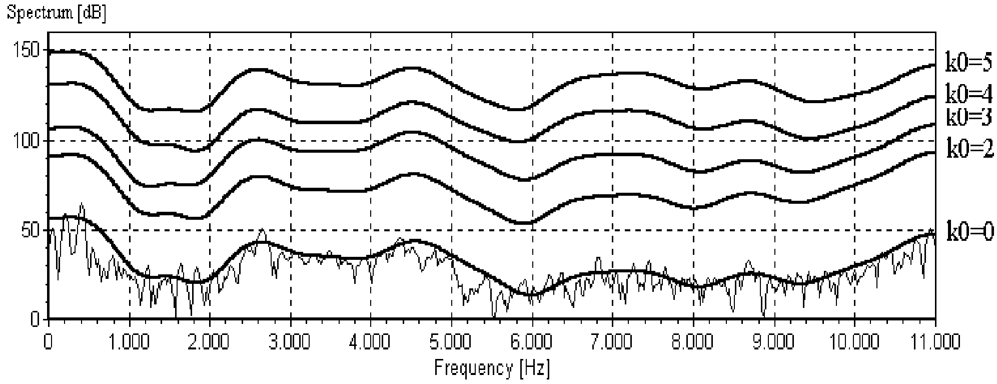

Figure 6. Power spectrum and LPC cepstrum derived spectrum for k_0=0, 2...5.

2.3 Mel cepstral analysis

Perceptual analysis emulate human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands. Mel cepstral analysis use the Mel scale and a cepstral smoothing in order to get the final smoothed spectrum. First the short-term spectrum of the vocal segment is evaluated. This spectrum is integrated over gradually widening frequency intervals on the Mel scale. The resulting Mel-warped spectrum is projected on a cosine basis and the Mel frequency cepstral coefficients (MFCC) are obtained.
The first cepstral coefficient (c_0) describe the shape of the log spectrum independent of its overall level. Next coefficient (c_1) measures the balance between the upper and lower halves of the spectrum, and higher order coefficients are concerned with increasingly finer features in the spectrum [1].

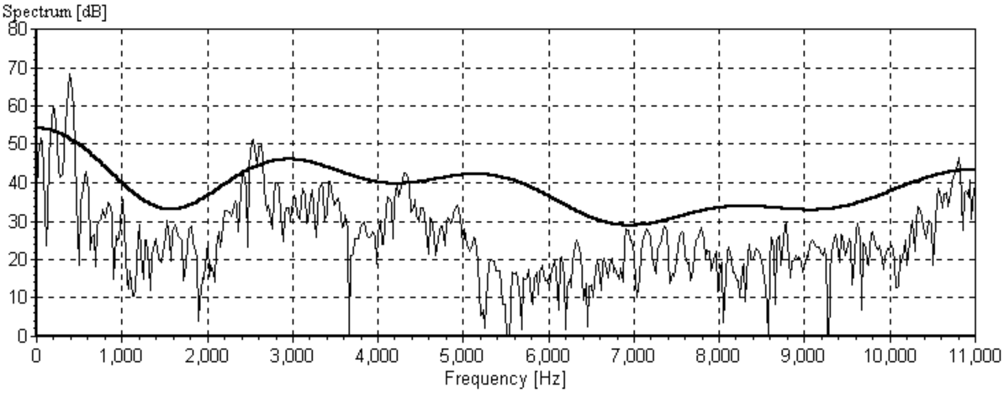

Figure 7. Power spectrum and MFCC derived spectrum.

The MFCC derived spectrum is even more smoothed than LPC or LPC cepstrum derived spectrum as in fig. 7. MFCC are the parametrisation of choice for many speech/speaker recognition applications because they give good discrimination.
The effect of elimating first k_0 coefficients over derived spectrum is shown in fig. 8.

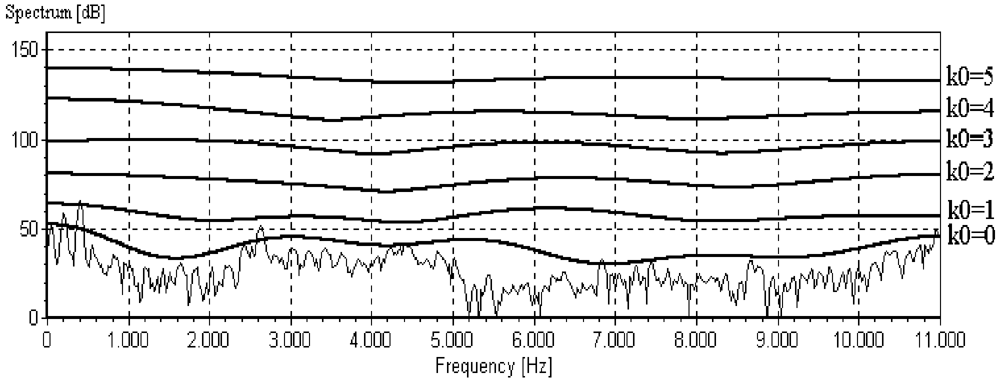

Figure 8. Power spectrum and MFCC derived spectrum for $k_0=0 \ldots 5$.

The new spectra are seriously damaged, even for $k_0=1$. By maintaing first coefficient and removing subsequents there is no major changes in spectrum (fig. 9).

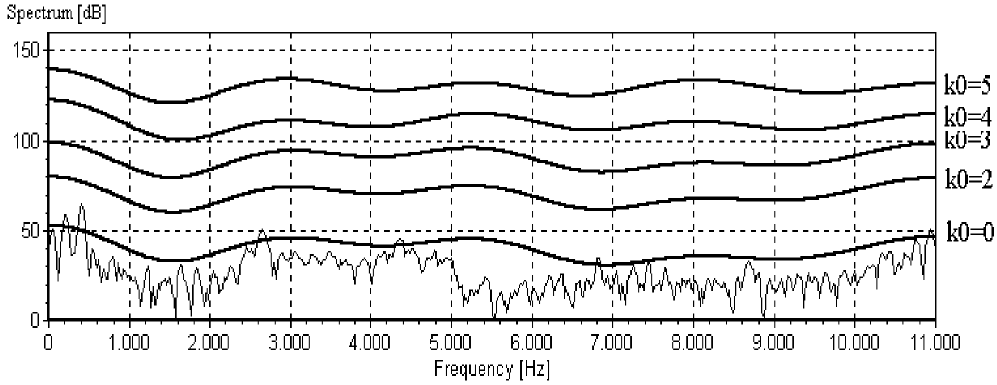

Figure 9. Power spectrum and MFCC derived spectrum for $k_0=0,\ 2 \ldots 5$.

3. THE INFLUENCE ON THE SPEAKER RECOGNITION

We realize a set of experiments on speaker recognition using DTW. Dynamic time warping (DTW) is a well known method used in isolated word recognition. In DTW method the unknown test pattern is compared with each sound reference pattern and a measure of similarity (distance) between the test pattern and each reference pattern is computed. DTW imply both a local distance measure between two spectral vectors and a global time alignment procedure which compensates the different rates of speaking of the two patterns.

In our experiments we used the Euclidian distance as local distance.

The data set consist of digit utterances from 20 speakers, 16 males and 4 females. For each type of paramers (LPC, LPC cepstrum, MFCC), the number of first k_0 coefficients eliminated was varied between 0 and 5.

The following results was obtained :

- feature type : LPC
  - the sequence $c_3,\ldots,c_n$ gave the best results;

- feature type : LPC cepstrum
  - the sequence $c_2,c_3,\ldots,c_n$ gave the best results;

- feature type : MFCC
  - the sequence $c_1,\ldots,c_n$ gave the best results.

In case of LPC and LPC cepstrum parameters, for $k_0=3$ and $k_0=2$ respectevelly, the interspeaker distance is reduced for "distant" speakers without afecting the recognition decision and is increased for "closely" speakers and consequently improve the recognition decision.

4. CONCLUSIONS

This paper try to figure out the influence of the first coefficients from the parameter set (LPC, LPC cepstrum, MFCC) used in speaker recognition over the shape of the derived spectrum and finally over the process of recognition.

In LPC case, the first coefficient has a great influence over the spectrum shape. Removing the subsequent coefficients has no major effects over the spectrum shape.

In LPC cepstrum case, the first coefficient has a smaller influence over the spectrum shape than in LPC case. Removing the subsequent coefficients still preserve the spectrum shape.

In MFFC case, all first coefficients affect the spectrum shape.

A set of experiments on speaker recognition was realized using DTW in which we try to figure out the influence of the first coefficients on the recognition results.

In LPC case, we found that first coefficients (c_1, c_2) are harmfull for speaker recognition.

In LPC cepstrum case, we found that only one coefficient (c_1) has negative influence on recognition.

In MFCC case, all first coefficients must be used in recognition.

References

[1] Cole, A. R. (editor) [1995] "Survey of the State of the Art in Human Language Technology", ftp://cslu.cse.ogi.edu/hlt/

[2] Rabiner, L., Juang, B.H. [1993] "Funfamentals of Speech Recognition", Prentice Hall.

[3] Huang, X., Acero, A., Hon, H. [2001] "Spoken Language Processing", Prentice Hall .

[4] H. Hermansky [1999] "Mel cepstrum, deltas,double-deltas,.. - What else is new?", in *Proceedings of Robust Methods for Speech Recognition in Adverse Conditions, 1999, Tampere, Finland.*

[5] Picone, J.,W. [1993] "Signal modeling techniques in speech recognition", *Proc.IEEE* vol. 81 sept. 1993, pp.1215-1246.