Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 51(65), Fascicola 2, 2006

Features Extraction from Romanian Vowels Using Matlab

Alina Nica[1], Alexandru Căruntu[1], Gavril Toderean[1], Ovidiu Buza[1]

Abstract –In this paper we developed in MATLAB a software environment in order to extract the main features from the Romanian vowels, which we intend to use in the synthesis step. We estimated speech parameters such as: energy, zero crossing rate (ZCR), fundamental frequency, formants. The used methods for obtaining the parameters are time domain analysis, cepstral analysis and Linear Predictive Coding (LPC). The analyzed vowels were uttered by several speakers and some experimental results are presented.

Keywords: speech analysis, vowels, features extraction, cepstrum, LPC

I. INTRODUCTION

In most applications of speech processing (e.g., synthesis, recognition), the speech analysis is the first step, and it involves the features extraction from the speech signal. By analyzing the speech signal, we want to obtain a useful representation of the speech waveform, in terms of parameters, that contain relevant information for speech synthesis. We used methods of speech analysis both in time domain (in this case the analysis is performed directly on the speech waveform), and in frequency domain (after a spectral transformation of the speech).

The representation of the speech waveform in terms of time-domain measurements include average zero-crossing rate, energy. These representations are very simple to implement and on their basis it is possible to estimate important features of the speech signal [1].

A very important speech analysis technique is the method of linear predictive analysis. This method provides an accurate representation of the basic speech parameters and is relatively efficient for computation [2].

The cepstral analysis involves the process of separating two convolutionally related proprieties by transforming the relationship into a summation. The importance of the cepstrum stems from the fact that it allows for the separate representation of the spectral envelope and fine structure.

We used the analysis methods mentioned above in order to extract the romanian vowels parameters, and the software environment, to perform it, was developed in MATLAB.

The Romanian language has seven vowels: *a ,e ,i, o, u, ă, î* and accordingly to the phonetically researches they appear with a frequency of 45,16%[3]. Generally romanian vowels are oral sounds, excepting the situations when they are neighboring with a nasal consonant; in this case they become also nasal sounds[4].

Vowels are voiced sounds and they are produced with the vibration of the vocal folds. The frequency of the vocal-folds oscillation is the fundamental frequency of the speech signal [5], [6]. The mean value for the fundamental frequency for male voices is 125 Hz and for female voices is 250 Hz [2].

The principal resonant structure, particularly for the vowels, is known as the vocal tract, and the resonance frequencies are called formants and by convention they are numbered from the low-frequency end and are usually referred to as F_1, F_2, F_3, etc. The most significant formants in determining the phonetic properties of speech sounds are generally F_1 and F_2 (ranged between 250 Hz and 3 kHz); but for certain phonemes some higher-frequency formants can also be important.

The the paper is organized as follows: in Section II we review briefly the analysis methods used to extract the speech parameters and some examples on different vowels are presented, too; in Section III are given a few of the experimental results; finally, in Section IV are the conclusions.

II. FEATURES EXTRACTION

Many analysis techniques have been developed in time. In our experiments we have chosen the most used of them. This section presents shortly a description and some details of their implementation.

A. *Sound spectrogram analysis*

An important tool for speech analysis is the *spectrogram*, which converts a two-dimensional speech waveform (amplitude-time) into a three-dimensional pattern (time-frequency-amplitude). Thus, on a spectrogram, time and frequency are displayed in its horizontal and vertical axes, respectively, and amplitude is noted by the darkness of the display [7], [8]. Peaks in the spectrum, corresponding to the formants, appear as dark

[1] Technical University of Cluj-Napoca,
e-mail: Alina.Nica@com.utcluj.ro

BUPT

horizontal bands and the vertical stripes correspond to the fundamental frequency. To compute a spectrogram the short-time Fourier analysis is used. The speech signal is time-varying, but speech analysis assumes that the signal properties change relatively slowly on short periods of time (10 to 30 ms), so that the signal characteristics can be considered uniform in that regions. Consequently, the speech signal is decomposed into a sequence of short segments, referred to as *analysis frames* and each one is analyzed independently. This technique is called *short-time analysis*.

For a given signal *s[m]*, the short-time signal $s_n[m]$ of frame *n* is defined as:

$$s_n[m] = s[m] w_n[m]. \qquad (1)$$

the product of *s[m]* by a *window function* $w_n[m]$, which is zero everywhere except in a small region.

Prior to frequency analysis, the frames are multiplied by a tapered window, in order to reduce any discontinuities at the edges of the selected region; otherwise it could appear some spurios high-frequency components into the spectrum. The most used windows are Hamming and Hanning windows. The length of the analysis window must give an adequate time and frequency resolution. A common compromise is to use a 20-30 ms window applied at 10 ms intervals.

There are two types of spectrograms: *wide-band spectrograms*, which use short windows (<10 ms) and *narrow-band spectrograms*, which use long windows (>20 ms). Wide-band spectrograms are useful in viewing vocal tract parameters (formant frequencies), while narrow-band spectrograms are good for fundamental frequency estimation.

B. *Time domain analysis*

By analyzing the speech signal in time domain, some important features can be estimated: maximum and medium amplitude, energy, zero-crossing rate, fundamental frequency.

Short-time energy of the speech wave is defined with the equation:

$$E_n = \frac{1}{N} \sum_m [s(m) w(n-m)]^2 . \qquad (2)$$

where *N* is the number of samples, *w* is a window used for analysis and *s* is the speech signal. It provides a convenient represantation of the amplitude variation over time. Energy emphasizes high amplitudes (the signal is squared in calculating the energy). Voiced segments have high energy and unvoiced segments have much lower energy. In order to reflect accurately the variations of the signal amplitude, the choise of window duration is very important: if it is too short, the energy will depend exactly of the waveform and if it is too large, the variations of the signal amplitude will not be reflected very corectly.

Zero-crossing rate (ZCR) is a simple measurement and provides adequate spectral information at a low cost. The short-time average zero-crossing rate is defined as:

$$ZCR_n = \frac{1}{2} \sum_m \left| \mathrm{sgn}[s(m)] - \mathrm{sgn}[s(m-1)] \right| w(n-m). \quad (3)$$

where *s* is the signal, sgn is the *signum* function and *w* is a window. This parameter is a simple measure of the dominant frequency of a signal. It is useful in differentiating between voiced and unvoiced signals, because unvoiced speech have much higher ZCR values than voiced speech. A suggested boundary is 2500 crossing/s, since unvoiced and voiced speech average about 4900 and 1400 crossing/s, respectively [7]. The zero-crossing rate can be used in the phone segmentation when preparing a database for concatenative Text-to-Speech synthesis [9].

C. *LPC analysis*

Linear Predictive Coding is a very important tool in speech analysis. A parametric model is computed based on least mean squared error theory, this technique being known as linear prediction (LP). LPC has been used to estimate fundamental frequency, vocal tract area functions, but it primarily provides a small set of speech parameters (called LPC coefficients) that represent the configuration of the vocal tract (the formants) [7], [10], [11].

LPC estimates the speech signal based on a linear combination of its *p* previous samples [1]. A larger *predictor order p* enables a more accurate model. In Figure 1 is an example of LPC spectra using different values of *p*, for vowel *a* (a1.wav).

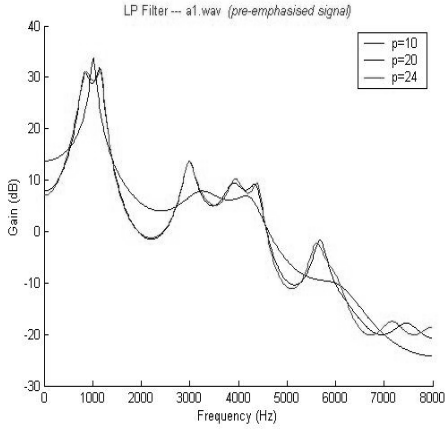


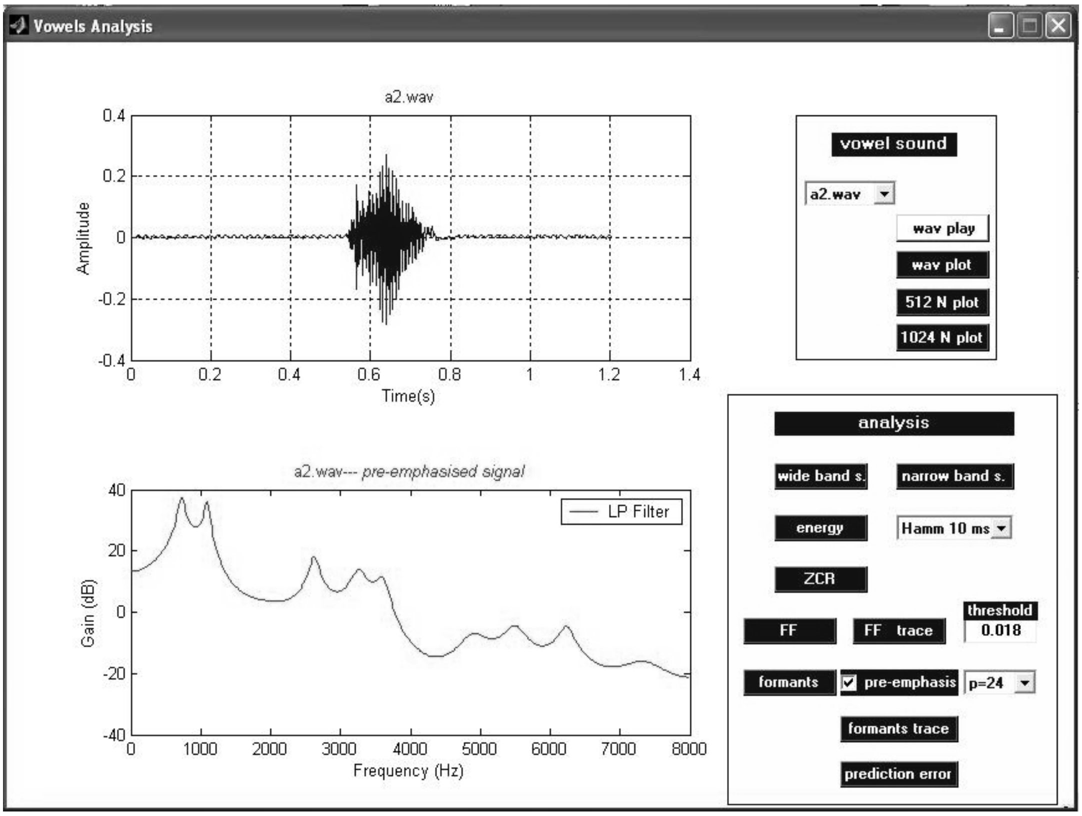
Fig. 1. LPC spectrum for vowel *a* (a1.wav) for different values of predictor order *p*

By removing the formant effects from the speech signal, the residual signal is obtained. Prior to speech analysis is recommended to pre-emphasis the signal, in order to emphasize the low frequencies in the speech spectrum.

D. *Cepstral analysis*

Cepstral analysis provides a method for separating

BUPT



a2.wav

Hamm 10 ms

threshold 0.018

☑ pre-emphasis

p=24

Fig. 2. The Vowels Analysis interface

the vocal tract information from excitation.

The process of passing an excitation signal through a vocal-tract filter to generate a speech signal can be represented as a process of convolution in time domain, which is equivalent to multiplying the spectral magnitudes of the source and filter components. If the spectrum is represented logarithmically, these components are additive and it is much easier to separate them using filtering techniques. *Cepstrum* is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum [12]. The cepstrum can be used to determine the fundamental frequency of voiced speech, because the part of the cepstrum corresponding to the source is often manifested as a single pike.

In Figure 3 is presented an example of fundamental frequency estimation for vowel *e* (e1.wav).

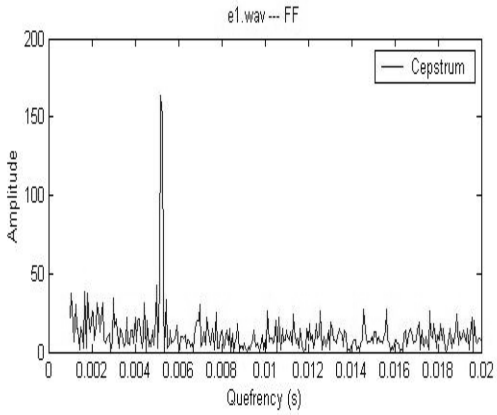


Fig. 3. Fundamental frequency estimation for vowel *e* (e1.wav) using cepstral method

The location of this peak gives the measurement for the frequency of the source signal.

III. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Our application was developed in MATLAB [13], [14]. We named it *Vowels Analysis* and its interface is presented in Figure 2. It has two panels: *vowel sound* and *analysis*, and on them, there are some buttons, which perform different tasks. Thus, it can be displayed the entire waveform of the speech signal (or only a few samples of it) and the corresponding narrow-band and wide-band spectrograms.

Also, the user can visualize some of the speech signal features: energy, zero-crossing rate, fundamental frequency, formants, formants trace, prediction error.

The sounds used in our experiments were recorded in wave format, at a sampling rate of 16 kHz, and they were uttered by several speakers (men and women).

We obtained one of the most important feature, which characterizes the vowels, the formants. In order to obtain the formants, we have done experiments using different values for the prediction order *p*, and varying the degree of pre-emphasis.

In Figure 2 it is illustrated the estimation of the formants for vowel *a* (a2.wav) using pre-emphasized signal and a value of the prediction order *p* of 24.

For example, according to [15], the romanian vowel *a* has the average value for the first three formants as follows: F_1=700 Hz, F_2=1300 Hz and F_3=2600 Hz.

BUPT

The mean values obtained by us, for the first three formants of vowel *a* are: F_1=744 Hz, F_2=1172 Hz and F_3=2744 Hz.

Also, we extracted the fundamental frequency of the speech waves and the fundamental frequency trace can be visualized. In Figure 4 is presented an example of fundamental frequency trace for vowel *e* (e1.wav)

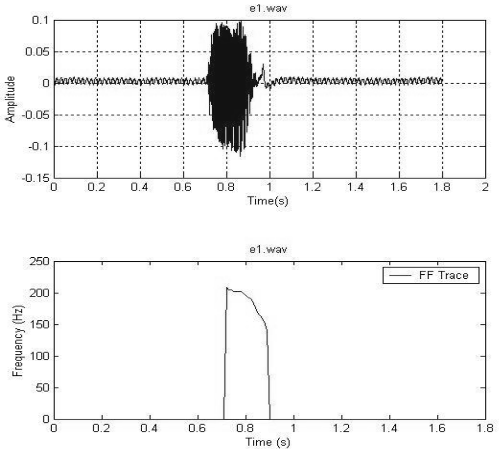


Fig. 4. Waveform and fundamental frequency trace for vowel *e* (e1.wav)

As well, we have measured the energy (we used Hamming window of different lengths) and the zero-crossing rate.

IV. CONCLUSIONS

In this paper, a few analysis techniques, which we have used to extract the main features of the speech signal, were briefly described. Next, we presented the interface of the software tool which we implemented in MATLAB, in order to extract the speech parameters. This application provided the means of some of the aspects of speech processing theory in a graphical manner. As well, we obtained the most important features for the vowels, which will be necessary in the synthesis stage.

We intend to develop this software environment, in order to obtain a database with the speech signal parameters. The goal of our future work is to experiment the concatenative synthesis method and to perform some prosody modifications on the synthesized speech signal.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, 1978.

[2] S. Furui, *Digital Speech Processing, Synthesis, and Recognition,* Second Edition, Revised and Expanded, Marcel Dekker, Inc., 2001.

[3] S. Puşcariu, *Limba română-Rostire*, vol. I, Editura Minerva, Bucureşti, 1976.

[4] V. Şerban, *Fonetica*, Editura Augusta, Timişoara, 1997.

[5] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, Second Edition, Taylor&Francis, 2001.

[6] B. Yegnanarayana and N.J. Velduis, "Extraction of Vocal-Tract System Characteristics from Speech Signal", *IEEE Transaction on Speech and Audio Processing,* Vol. 6, No. 4, pp. 313-327, July 1998.

[7] D. O'Shaghnessy, *Speech Communications, Human and Machine*, Second Edition, IEEE Press, Inc., New York, 2000.

[8] T. Dutoit, *Introduction au traitement automatique de la parole*, *Notes de cours*, Faculte Polytechnique de Mons, Belgium, 2000.

[9] T. Zang and C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification",", *IEEE Transaction on Speech and Audio Processing,* Vol. 9, No. 4, pp. 441-457, May 2001.

[10] X. Huang, A. Acero and H.W. Hon, *Spoken Language Processing: A Guide to Theory, Algoritm and System Development*, First Edition, Prentice-Hall, 2001.

[11] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels", *Speech Communication,* Vol. 38, pp. 141-160, 2002.

[12] http://www.utdallas.edu/~loizou/ee6362/lec6.pdf

[13] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[14] http://svr-ww.eng.cam.ac.uk/~ajr/SA95/SpeechAnalysis.html

[15] I.T. Stan, *Fonetica*, Editura Presa Universitară Clujeană, Cluj-Napoca, 1996.