Marzena MIĘSIKOWSKA
KIELCE UNIVERSITY OF TECHNOLOGY, FACULTY OF ELECTRICAL ENGINEERING, AUTOMATICS AND COMPUTER SCIENCE

Speech signal processing and analysis tool

M.Sc. Marzena MIĘSIKOWSKA

Assistant in the Department of Computer Science, Faculty of Electrical Engineering, Automatics and Computer Science, Kielce University of Technology. The researcher is interested in digital signal processing, designing and managing database systems.

e-mail: marzena@tu.kielce.pl

Abstract

The project's objective is to create a tool intended for processing, analysis, and parameterizing human speech signal. The main aim is to obtain a speech signal image with some selected parameterization methods. The methods include use of 2D parameterization grid [1, 2] as well as cepstral coefficients CC [3]. Obtaining signal image as well as its further analysis without signal preprocessing is extremely difficult and the process doesn't guarantee desirable results. For this reason the tool is based on two main modules. The first one is intended for signal preprocessing, preparing it for further analysis. The other one provides signal parameterization methods. The tool was implemented in Java language.

Keywords: speech signal processing, cepstral coefficients.

Narzędzie do przetwarzania i analizy sygnału mowy

Streszczenie

W pracy podjęto próbę stworzenia narzędzia umożliwiającego przetwarzanie, analizę i parametryzację sygnału mowy. Głównym celem jest pozyskanie obrazu sygnału mowy za pomocą wybranych metod parametryzacji. Wybrane metody parametryzacji sygnału mowy to parametryzacja za pomocą siatki dwuwymiarowej [1, 2] oraz współczynniki cepstralne [3]. Zobrazowanie sygnału oraz jego dalsza analiza bez operacji wstępnego przetworzenia sygnału jest procesem trudnym i nie zawsze przynosi pożądane rezultaty. Wobec tego narzędzie wyposażono w dwa zasadnicze moduły. Pierwszy moduł odpowiedzialny jest za wstępne przetworzenie sygnału, przygotowujące sygnał do dalszej analizy. Drugi moduł dostarcza metod parametryzacji sygnału mowy. Narzędzie zaimplementowano w języku Java.

Słowa kluczowe: przetwarzanie sygnału mowy, współczynniki cepstralne.

1. Introduction

The paper attempts to create a tool for processing, analysis, and parameterization of speech signal. The main objective is to compare parameterization methods such as 2D parameterization grid [1, 2] and cepstral coefficients CC [3].

Currently, Matlab is the most popular computing language and interactive environment for algorithm development. In case of selecting a different environment, the user must implement digital signal processing algorithms. Such an implementation involves optimization of the applied algorithms.

Implementation of digital processing of speech in Java environment may prove very useful. Java language is widely used especially in developing database interfaces. Some database systems use Java as access point to various elements of the database. Another advantage of the application of the Java language is the existence of the implemented voice recognition systems such as Sphinx-4 [4], which are known to work correctly. Java allows its API *javax.speech* to be use in synthesis and speech recognition of English phonemes. Java speech recognition requiring preprocessing methods may become a multipurpose mechanism that is widely used on a number of platforms and operating systems.

The Java Sound API provides the lowest level of sound support on the Java platform. It provides application programs with a great amount of control over sound operations, and it is extensible. For example, the Java Sound API supplies mechanisms for installing, accessing, and manipulating system resources such as audio mixers, MIDI synthesizers, other audio or MIDI devices. The Java Sound API does not include sophisticated sound editors or graphical tools, but it provides capabilities upon which such programs can be built. It emphasizes low-level control beyond that commonly expected by the end user.

Analysis of a speech signal is usually carried out on the basis of the graphical image of the signal. The signal imagery as well as a further analysis require signal preprocessing. The tool makes it possible to process the signal in the frequency and time domains. Application of the digital signal processing involves pre-emphasis, 2N Point Real FFT, DIF FFT radix-2, IDFT, and windowing of the speech signal.

The sections of the article are as follows: section II reviews implemented methods of signal processing and parameterization. Section III demonstrates the efficiency of the applied methods of signal parameterization. Section IV presents the results of speech signal imagery. Conclusions are specified in section V, while section VI provides a summary of the article.

2. Tool and the speech signal processing and parameterization methods

The tool consist of two modules. The first module is responsible for digital processing of a signal. It delivers data concerning time and frequency of the signal. The second module provides signal parameterization methods. The structure of the tool is represented in Figure 1.

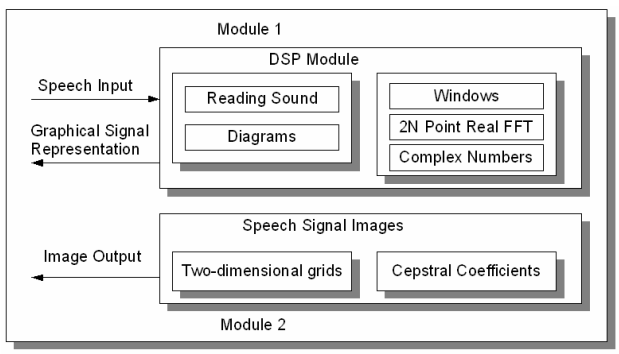

Fig. 1. The structure of the tool
Rys. 1. Struktura narzędzia

The first module, *Digital Signal Processing (DSP) Module*, enables the following:

- to read wave files
- graphical signal representation in the time and frequency domains
- digital signal processing – windowing, fast Fourier transform, pre-emphasis.

The function of the DSP module is to transmit a processed speech signal to the second module so that the signal image could be obtained.

Time domain signal constitutes an input to the digital signal processing. Analysis of the signal in the time domain may refer to

the signal's amplitude and changes. The speed of signal changes in the time domain may be additionally increased in the process of pre-emphasis. Upon the application of pre-emphasis the numerous elements of the signal that were invisible become readable and subject to analysis. Pre-emphasis is crucial in the process of determining cepstral coefficients, using linear prediction method.

For various purposes, the spectrum analysis can be carried out with a number of methods. The method used for determining the frequency content of the speech signal is the algorithm 2N Point Real FFT. The 2N Point Real FFT algorithm utilizes a mechanism for determining the fast Fourier transform DIF FFT radix-2. The FFT suffers from negative results of a spectral leak. The time data is multiplied by window functions. If addition of zero's is necessary in the extension of time data sequence, zero value samples are added after multiplication of original time data sequence by the window function. The windowing will decrease the leak, but it will not eliminate it completely. Values of the spectral strength of the FFT is calculated on the basis of the following formula:

$$X_{PS}(m) = \left|X(m)\right|^2 = X_{real}(m)^2 + X_{imag}(m)^2 \qquad (1)$$

which will allow to determine the power spectrum in dB as well as normalized power spectrum in dB.

In order to reduce the number of calculations necessary in windowing the FFT input data leading to the reduction of spectral leak, a frequency domain windowing was used. The method entails calculation of FFT for non-windowed data in the initial stage, subsequently frequency domain windowing is carried out. Since the formulas for Hanning and Hamming windows have a cosine form $w(n) = \alpha - \beta\cos(2\pi n / N)$ for n=0,1,2,…..,N-1, therefore, in the frequency domain the value of an m spectral line for non-windowed $X(m)$ is equal to:

$$X_o(m) = \alpha X(m) - \frac{\beta}{2}X(m-1) - \frac{\beta}{2}X(m+1) \qquad (2)$$

In order to calculate the N-point FFT, a non-windowed FTT can make use of the formula (2), avoiding windowing in the time domain that require time-consuming multiplication. The frequency windowing is done for selected FFT samples.

The second module, *Speech Signal Images*, possesses the signal parameterization methods utilizing 2D grid [1, 2] as well as cepstral coefficients CC [3].

Two-dimensional Grids module is responsible for parameterization of the speech signal in recognition of Polish phonemes [2]. The input to this module is a signal processed in the time domain by the first module. The signal processing involves the reduction of noise at the beginning and at the end of a recording, so that the signal contains as much information as necessary. The function of the module in the initial stage is to determine the location of the local minimums, which will be used as a basis for calculating the location and width of the basic pattern of laryngeal tones. The second stage involves application of 2D grid that has a specific resolution into one or more basic patterns in such a way so that the height of the grid is automatically adjusted to the amplitude of the signal , and the grid width to the duration of the basic pattern. The 5x7 grid was used. One's are written in the cells in which the signal occurs whereas zero's are added to the signal-free cells.

In order to reduce the influence of noise into the combination of bits in the grid, it is recommended to apply the gird several times on the basic patterns and to calculate a single matrix average [2].

The imagery obtained through cepstral coefficients can be obtained using two methods [3]:

- logarithmization of the spectrum module and recalculation of simple or reverse discrete Fourier transform,
- linear prediction method.

Cepstral Coefficients module is responsible for parameterization of the signal. The module utilizes the methods of digital signal processing of the DSP module. The standard cepstral coefficients are calculated for a section of a speech signal in the time domain. In this case the signal section consists of 240 samples and the signal shift is equal to 180 samples. Twelve cepstral coefficients are calculated for the purpose of speech recognition.

In case of the first method concerning the calculation of cepstral coefficients, a speech signal section is multiplied by the Hamming window, which is followed by fast Fourier transform. The next step involves logarithmization of the spectral module with subsequent simple or reverse Fourier transform, which is expressed by the following formula:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \ln\left|\sum_{m=0}^{N-1} w(m)x(m)e^{-j2\pi km/N}\right| e^{\pm j2\pi kn/N} \qquad (3)$$

The second method involves linear prediction. A prediction filter p=10 identical to LPC 10 as well as determined cepstral coefficients q=12 are used in speech recognition. Prediction modeling is not applied on the original signal section x(n), but on the signal version that underwent pre-emphasis, which was filtered by non-recursive FIR filter :

$$x'(n) = x(n) - 0{,}9375x(n-1) \qquad (4)$$

The filter increases the high frequencies of the human speech signal. The speech signal is subsequently analyzed. In order to select a fragment of speech the 240-element Hamming window is shifted by 180 samples and multiplied by the signal. Subsequently, thanks to the determined signal autocorrelation function within a frame, a prediction filter coefficient is calculated, which is the basis for the calculation of cepstral coefficients. Information concerning the calculation of cepstral coefficients with the linear prediction method can be found in section [3].

The last step involves obtaining the speech signal image with one of the two speech signal parameterization methods.

3. Efficiency of speech signal parameterization methods

The efficiency of speech signal parameterization methods is represented in Fig. 2. The efficiency is the time needed for obtaining an image of a speech signal, using the cepstral method:

- due to logarithmization of the spectrum module and calculation of fast Fourier transform (CC FFT),
- with linear prediction method (CC LPC).

Horizontal axis represents the number of sound samples tested (n). Vertical axis of the spectrogram represents time in milliseconds. The tests were carried out on a PC with 3.00 GHz Intel Pentium 4 processor, 2 GB RAM.

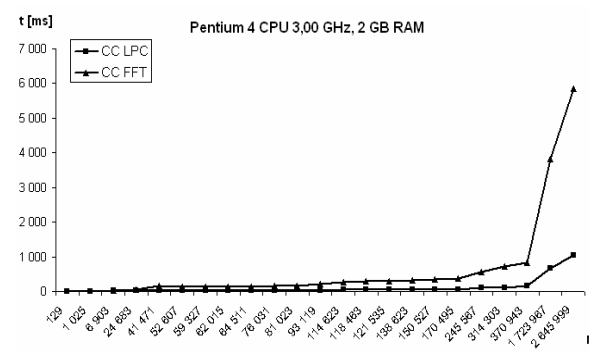

Fig. 2. Efficiency of the cepstral parameterization methods
Rys. 2. Wydajność cepstralnych metod parametryzacji

During implementation of cepstral methods in Java, if the number of the samples is below 314303, the application of the CC FFT method is a sufficient solution that will produce a signal image. If the number of samples exceeds 314303, the CC LPC method becomes a better solution.

2D grid method is a better solution in obtaining speech signal imagery in terms of parameterization of single phonemes and the speed of processing.

In case of cepstral methods, the linear prediction proves to be the most sufficient.

4. Results

The tool was implemented in Java language and it is used for digital signal processing with signal graphic analysis. It provides speech parameterization methods, preparing the signal for speech recognition.

The obtained graphical speech signal in the domain of time and frequency generated by the tool are shown in Fig. 3 and Fig. 4.

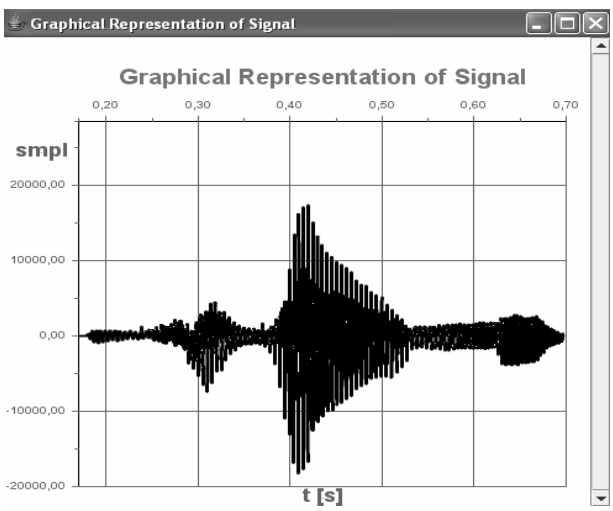

Fig. 3. Discrete time signal of human speech – the word "dzwon"
Rys. 3. Przebieg sygnału w dziedzinie czasu – słowo "dzwon"

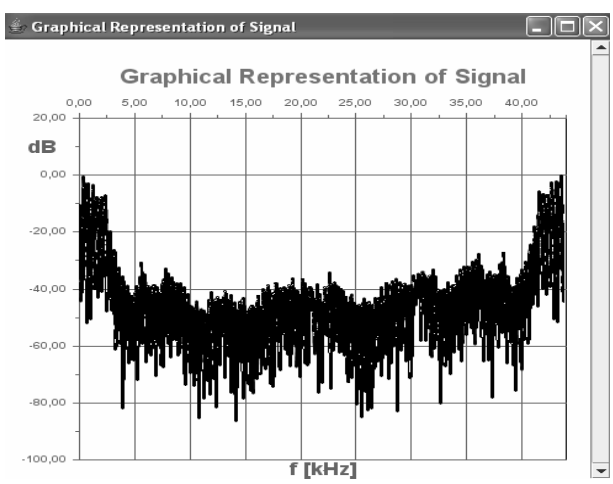

Fig. 4. Normalized power spectrum in Decibels - the word "dzwon"
Rys. 4. Unormowane widmo mocy w dB – słowo „dzwon"

When analyzing Fig. 4, one can notice the apparent symmetry of the results. The symmetry results from DFT symmetry (m-initial value of DFT has the same amplitude as (N-m)-initial value of DFT). In DFT real input series X(m) is an X(N-m) feedback. Left/right mouse click on the image will enlarge or reduce the size of the image.

The image fragment represented as a chart in Fig. 5 was obtained with two fast Fourier transform calculations (CC FFT). Horizontal cells represent discrete fragments of speech signals whereas columns the number of twelve cepstral coefficients calculated for each fragment.

Cepstral Coefficients by FFT

1	2	3	4	5	6	7	8	9	10	11	12
687.0	68.0	46.0	13.0	28.0	2.0	6.0	5.0	8.0	9.0	5.0	15.0
676.0	88.0	21.0	3.0	21.0	4.0	20.0	19.0	27.0	15.0	21.0	24.0
699.0	62.0	35.0	15.0	2.0	5.0	5.0	23.0	12.0	19.0	19.0	6.0
853.0	91.0	32.0	6.0	9.0	24.0	22.0	14.0	16.0	5.0	15.0	14.0
750.0	92.0	3.0	15.0	15.0	49.0	15.0	25.0	14.0	4.0	45.0	7.0
600.0	75.0	40.0	10.0	18.0	24.0	10.0	4.0	11.0	5.0	14.0	4.0
692.0	102.0	28.0	21.0	12.0	22.0	21.0	8.0	5.0	6.0	35.0	23.0
728.0	67.0	34.0	23.0	3.0	8.0	5.0	9.0	5.0	5.0	15.0	6.0
758.0	69.0	56.0	10.0	23.0	14.0	8.0	6.0	11.0	22.0	17.0	27.0
748.0	56.0	58.0	5.0	20.0	15.0	3.0	5.0	11.0	17.0	9.0	11.0
624.0	55.0	25.0	29.0	10.0	16.0	3.0	5.0	4.0	17.0	18.0	11.0
616.0	58.0	22.0	27.0	29.0	22.0	9.0	16.0	7.0	5.0	5.0	19.0
659.0	83.0	51.0	11.0	27.0	14.0	13.0	9.0	11.0	8.0	24.0	6.0
779.0	77.0	44.0	18.0	19.0	10.0	16.0	15.0	7.0	11.0	11.0	9.0
638.0	50.0	45.0	26.0	25.0	5.0	13.0	20.0	11.0	29.0	20.0	6.0
613.0	58.0	21.0	27.0	21.0	11.0	22.0	3.0	9.0	10.0	23.0	13.0
666.0	70.0	39.0	35.0	13.0	8.0	4.0	9.0	4.0	11.0	15.0	15.0
622.0	60.0	20.0	18.0	3.0	11.0	18.0	4.0	16.0	8.0	5.0	6.0
712.0	71.0	45.0	20.0	19.0	9.0	3.0	7.0	8.0	4.0	7.0	32.0
1126.0	152.0	28.0	24.0	13.0	30.0	25.0	29.0	4.0	8.0	5.0	24.0
707.0	89.0	64.0	28.0	36.0	20.0	15.0	3.0	6.0	20.0	5.0	22.0

Fig. 5. Image of the speech signal obtained with cepstral coefficients [2xFFT] – the word "dzwon"
Rys. 5. Obraz sygnału mowy otrzymany metodą współczynników cepstralnych [2xFFT]– słowo „dzwon"

5. Conclusions

The main objective of the work concerning obtaining speech signal imagery has been achieved. In case of parameterization of single phonemes the fastest method proves to be 2D parameterization. As far as the cepstral methods are concerned the most effective seems to be the linear prediction method. Obtaining parameters of speech signals with cepstral methods is very effective. Optimization of the digital signal processing is especially important if the signals are to be processed on a regular PC.

Utilizing Java environment, the tool allows graphical analysis of the signal in the domain of time and frequency. The Java Sound API fulfills the needs of a wide range of application developers. Potential application areas include: communication frameworks, end-user content delivery systems, interactive application programs, content creation and editing.

6. Future work

Currently the author is focusing on the speech signal imagery so that the application can be extended by an additional module – AROA – the speech signal recognition module.

7. References

[1] Dulas Janusz, Skubis Tadeusz: Parametryzacja sygnału stochastycznego za pomocą siatek dwuwymiarowych. Pomiary Automatyka Kontrola, 2002.
[2] Dulas Janusz: Metoda siatek o zmiennych parametrach w zastosowaniu do rozpoznawania fonemów mowy polskiej. Rozprawa doktorska 2002.
[3] Zieliński Tomasz, Gajda Paweł, Stachura Marcin, Wilgat Robert, Król Daniel, Woźniak Tomasz, Grabias Stanisław: Zastosowanie współczynników HFCC jako cech sygnału mowy w automatycznej detekcji wad wymowy. Pomiary Automatyka Kontrola, 2006.
[4] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel: Sphinx-4: A Flexible Open Source Framework for Speech Recognition. SMLI TR-2004-139 Sun Microsystems Inc., November 2004.
[5] Basztura Czesław: Źródła, sygnały, obrazy akustyczne. WKŁ 1988.

Artykuł recenzowany