speech recognition, cepstral coefficients,
LPCC, MFCC, isolated word,

Robert PROKSA*

VISUALIZATION OF STAGES OF DETERMINING CEPSTRAL FACTORS IN SPEECH RECOGNITION SYSTEMS

The article presents two methods of determination of cepstral parameters commonly applied in digital signal processing, in particular in speech recognition systems. The solutions presented are part of a project aimed at developing applications allowing to control the Windows operating system with voice and the use of MSAA (Microsoft Active Accessibility). The analysed voice signal has been visually presented at each of the crucial stages of developing cepstral coefficients.

1. INTRODUCTION

Operating system manual control with the use of peripherals such as the keyboard or the mouse is a commonly applied solution. There are more and more frequent attempts improving control over the computer, in particular in cases in which physically handicapped or blind persons wish to use the computer. One of the solutions to the problem is the use of voice commands. Systems enabling voice recognition continue to meet with such problems as adverse impact of background noises during voice acquisition, utterance speed change, voice intensity change depending on the mood, psychical condition or plain cold. The implemented isolated words recognition system ought to compensate for the impact of these adverse factors on its performance.

1.1. DESIGN OF THE SPEECH RECOGNITION SYSTEM

Isolated words recognition systems comprise three basic blocks (Fig. 1). The first block is responsible for voice signal acquisition and conversion to a form enabling feature extraction [6, 7, 8]. Within block two, the signal is converted to a form enabling quick and easy storage and classification of the voice word patterns obtained. Block three is responsible for the decision on the signal classification in a correct group or specific pattern from the pattern base [3].

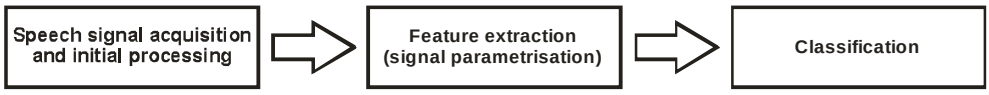

Fig. 1. Speech recognition system design general diagram

2. SPEECH SIGNAL PARAMETRISATION STAGES

Speech signal conversion into a parametric form is necessary due to a great complexity of the speech signal digital form. Parameters obtained based on a full signal need to accurately reflect the speech signal physical features which we are interested in and to allow for the possibly most accurate classification, which in turn affect recognition of the words uttered. For this purpose, a number of parameters obtained based on time course, signal spectre or linear predictive coding [3].

At present, parameters most frequently used in the speech recognition process are the so-called cepstral coefficients which may be determined in two ways. The first of them involves the application of linear predictive coding (Fig. 2) (LPCCs, Linear Prediction Cepstral Coefficients). The other involves the application of Fourier transform (Fig. 3) (MFCCs - Mel-Frequency Cepstral Coefficients) [10, 11].

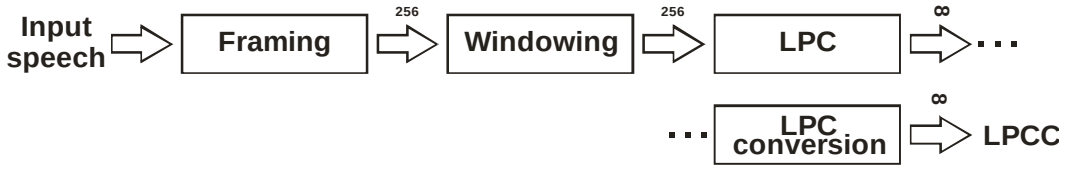

Fig. 2. Block diagram of the LPCC evaluation algorithm (numbers among arrows mean vector lengths)

* Institute of Informatics, University of Silesia, Będzińska 39 St., 41-200 Sosnowiec, Poland, robert.proksa@us.edu.pl

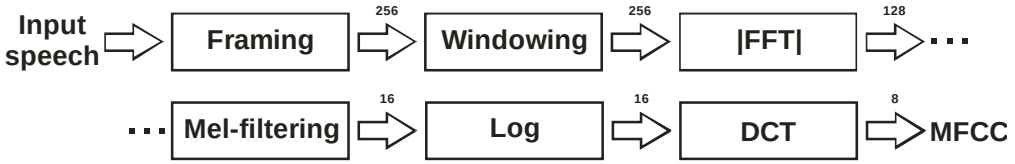

Fig. 3. Block diagram of the MFCC evaluation algorithm (numbers among arrows mean vector lengths)

In both cases, the analysed speech signal ought to be converted into a correct form by means of using a FIR pre-emphasis filter and free from unnecessary data by means of determining word boundaries [6]. Fig. 4 shows the word signal of 'Property' following detection of its boundaries. This is the output signal by means of which further stages of determining cepstral coefficients will be presented.

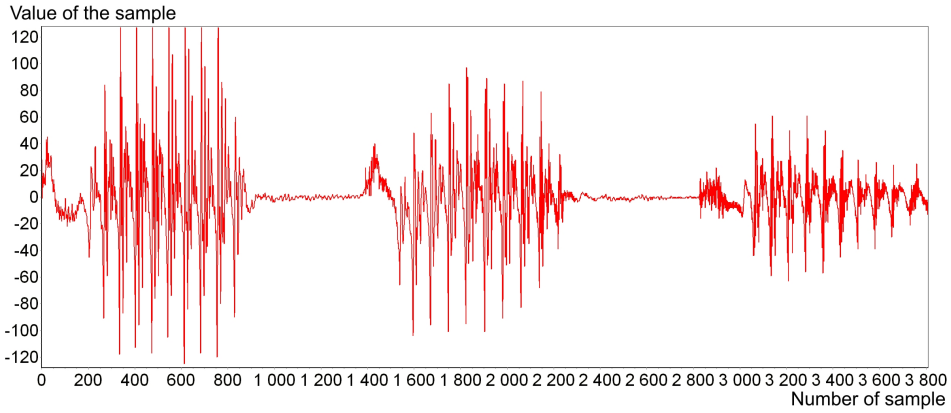

Fig. 4. 'Property' word signal

The next step is the division of the signal into frames and the application of the windowing function (Fig. 5). This is performed in order to protect the instantaneous spectre against the occurrence of interference during Fournier transform. In order to get rid of additional harmonics in the instantaneous spectre, time course fragments are smoothed over at frame ends. The windowing operation may be recorded with the following formulas [9, 12]:

$$y_t = \hat{y}_t(n) \cdot w(n) \tag{1}$$

$$w(n) = 0,54 - 0,46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \tag{2}$$

where:

$w(n)$ - Hamming window,

y_t - signal after windowing

$\hat{y}_t(n)$ - (n) the signal sample value for the frame,

N - frame length,

n - sample number in the signal frame.

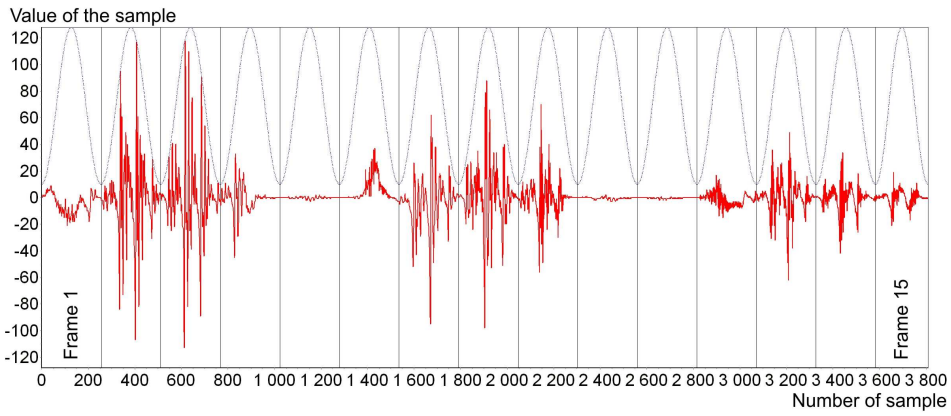

Fig. 5. 'Property' word signal after application of Hamming window (dotted line indicates window shape). Horizontal lines separate individual signal frames 256-sample-long).

3. LPCCS

Determination of cepstral coefficients with the first method requires LPCCs (Fig. 2). During their determination, we additionally obtain autocorrelation coefficients (AC) and reflection coefficients (REF). The optimal method of determination of the foregoing coefficients is the use of the two-stage Levinson-Durbin algorithm.

The linear predictive coding technique is based on the assumption that a signal sample may be presented as a linear combination of p previous samples meaning that its value may be 'predicted' based on p preceding speech signal values. Assuming that we know the origin of the y signal, i.e. $\{y(n-p), \ldots, y(n-1)\}$, we may develop a predicate described with the following formula:

$$\hat{y}(n) = \sum_{i=1}^{p} \alpha_i y(n-i) \tag{3}$$

where:

$\hat{y}_t(n)$ - (n) th signal sample value for the frame,

α_i - prediction coefficients.

The algorithm used to determine the LPCCs which we are interested in is as follows [5]:

1) Determination of autocorrelation coefficients with the following formula:

$$r(k) = \sum_{n=0}^{N-1-k} y(n)y(n+k), \text{ where } k = 0,1,2,3,\ldots,p \tag{4}$$

2) Determination of predictive coefficients of the p-th power:

a) Setting initial values:

$$E_0 = r(0) \quad ; \quad a_{11} = k_1 = r(1)/E_0 \quad ; \quad E_1 = E_0(1-k_1^2) \tag{5, 6, 7}$$

b) Iterative execution of the following 6 steps $m>1$ until $m=p-1$:

$$q_m = r(m) - \sum_{i=1}^{m-1} a_{i(m-1)} r(m-i) \quad ; \quad k_m = \frac{q_m}{E_{(m-1)}} \quad ; \quad a_{mm} = k_m \tag{8, 9, 10}$$

$$a_{im} = a_{i(m-1)} - k_m a_{(m-i)(m-1)} \text{ for } i = 1,\ldots,m-1 \quad ; \quad E_m = E_{m-1}[1-k_m^2] \quad ; \quad m=m+1 \tag{11, 12}$$

where:

$r(k)$ - autocorrelation coefficients ($0 \le k \le p$),
$a_i = a_{ip}$ - predictive coefficients ($1 \le i \le p$),
k_m - reflection coefficients ($1 \le m \le p$),
p - number of predictive coefficients,
E, m - auxiliary parameters.

Using the formula (3) and having appropriate LPC and original signal values, one may reproduce the signal seen in the following figure (Fig. 6). The signal was generated based on prediction regarding every fourth signal sample, with the use of the three preceding samples ($p=3$) of the original signal.

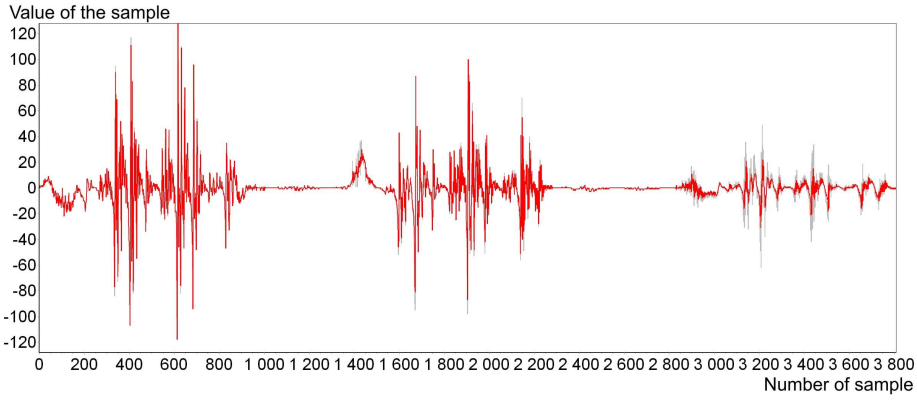

Fig. 6. 'Property' word signal generated with the use of linear predictive coding for $p=3$. Bright colour indicates the original signal

In order to obtain LPCCs one needs to use recurrent dependency allowing to determine cepstral coefficients directly from predictive coefficients a_i [5]:

$$c(1) = a_1 \tag{13}$$

$$c(n) = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c(k) a_{n-k} \quad for\ 1 < n \leq p \tag{14}$$

$$c(n) = \sum_{k=n-p}^{n-1} \frac{k}{n} c(k) a_{n-k} \quad for\ n > p \tag{15}$$

where:

c(n) - n-th LPCC.

Cepstrum *c(n)* has infinite length; thus, system recognition systems use from 8 to 12 coefficients. So generating eight LPCCs for each frame of the 'Property' word examined, we obtain the signal shown below. (Fig 7).

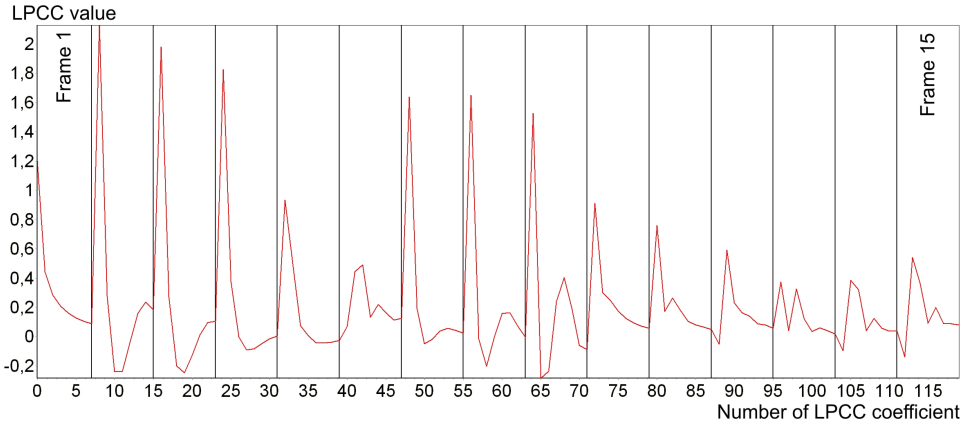

Fig. 7. Graph showing LPPC values for individual frames of the signal examined

4. MFCCS

The other method of determining cepstral coefficients is based on the discreet Fourier transform (DFT) and Mel-scale filtration. MFCC determination stages have been shown in Fig. 3.

The first step is to covert the signal from time form to frequency for in order to determine instantaneous spectres of the signal. This is carried out with the discreet Fourier transform [1, 2]:

$$y'(k) = \sum_{n=0}^{N-1} y(n) e^{\frac{-2j\pi kn}{N}} \quad k = 0,1,\ldots,N-1 \tag{16}$$

where:

$y'(k)$ - k-th signal value after DFT,
$y(n)$ - n-th signal sample value after windowing,

or the fast Fourier transform algorithm (FFT) involving recurrent DFT division into smaller (N/2) tasks [1]:

$$y'(k) = \sum_{n=0}^{\frac{N}{2}-1} y^{**}(n) e^{\frac{-2\pi jkn}{\frac{N}{2}}} + e^{\frac{-2\pi jk}{\frac{N}{2}}} \sum_{n=0}^{\frac{N}{2}-1} y^{*}(n) e^{\frac{-2\pi jkn}{\frac{N}{2}}} \tag{17}$$

where:

$y^{**}(n)$ - values of even signal samples after windowing,
$y^{*}(n)$ - value of uneven signal samples after windowing.

As a result of calculations involving complex number we receive the actual part $Re(y'(k))$ and the imaginary $Im(y'(k))$ of the transform which are used to determine the signal amplitude spectre (Fig. 8) based on the following dependency [5]:

$$\left|y'(k)\right| = \sqrt{\mathrm{Re}^2\left(y'(k)\right) + \mathrm{Im}^2\left(y'(k)\right)} \quad k = 0,1,2,\ldots,N/2 \tag{18}$$

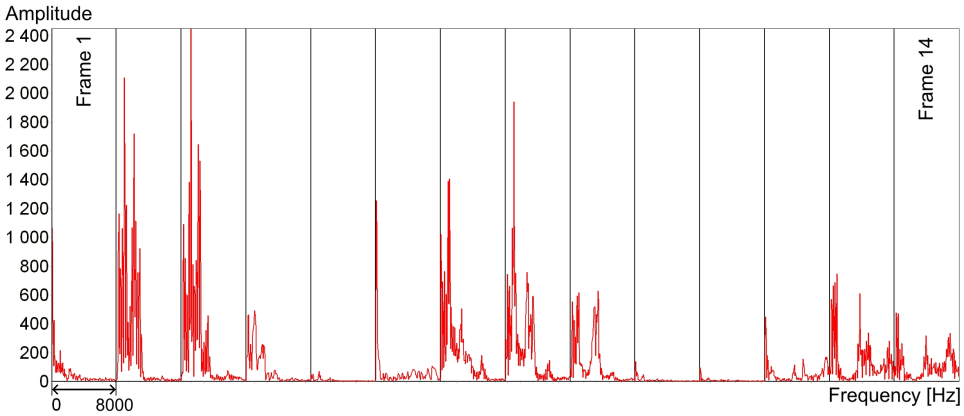

Fig. 8. Spectre of the signal examined

The signal spectre is multiplied by the triangular Mel filters bank the frequency response of which is supposed to simulate the behaviour of the human ear. The relation between the Mel scale and the Hertz scale is expressed with the following dependency (Fig. 9) [11]:

$$F_{mel} = 2595\log_{10}\left(1 + \frac{F_{Hz}}{700}\right) \tag{19}$$

The filters are distributed in a way ensuring a 150mel distance between filter central frequencies and 300mel width of them. The number of the filters is usually 16 or 24 (Fig. 10). In order to obtain information on the location of a frequency expressed in Hertz for a given signal sample after DFT, the following formula is used:

$$F_{Hz} = \frac{F_p}{N} \cdot k \qquad \mathrm{k} = 0,1,\ldots,N-1 \tag{20}$$

where:

F_p - signal sampling frequency during acquisition.

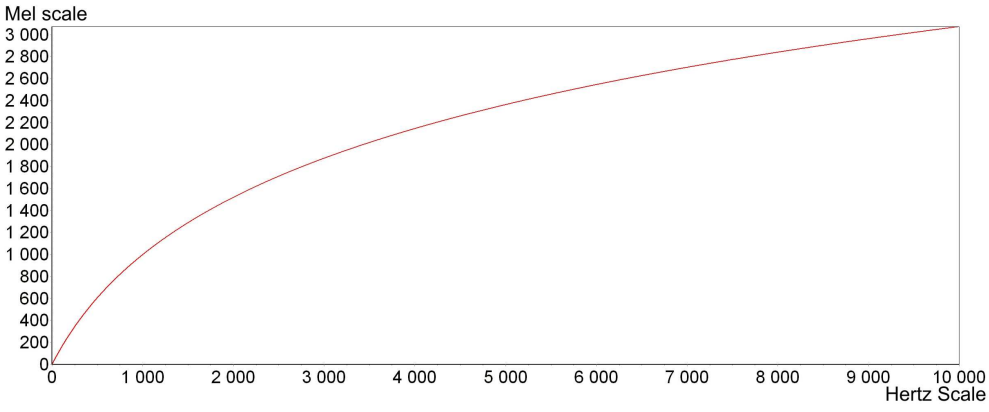

Fig. 9. Graph showing dependencies between Mel-scale values and Hertz-scale values

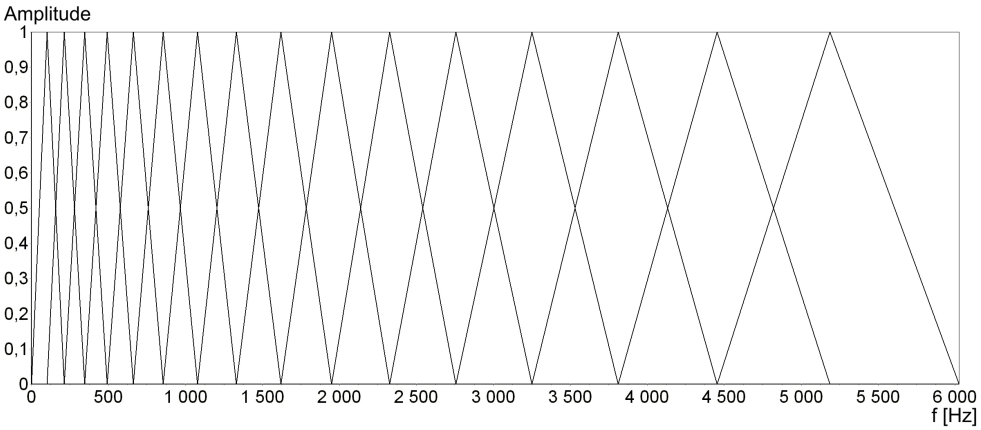

Fig. 10. Distribution of 16 Mel-scale filters [4, 9, 11]

The filtration operation may be presented with the following formula [10] (Fig. 11):

$$y''(m) = \sum_{k=b_m-\frac{\Delta_m}{2}}^{b_m+\frac{\Delta_n}{2}} \left| y'(k) \right| \cdot F_{\Delta_m}(k) \qquad 1 \le m \le M \tag{21}$$

$$F_{\Delta_m}(k) = \begin{cases} 1+\dfrac{2(k-b_m)}{\Delta_m} & k \le b_m \\ 0 & k > b_m \end{cases} \tag{22}$$

where:

$y''(m)$ - frame signal after filtration,
b_m - location of the m filter central frequency in the DFT field,
Δm - width of the m, filter band where: $\Delta_m = b_{m+1} - b_{m=1}$,
M - number of filters (16),
$F_{\Delta_m}(k)$ - k - m filter value where k is the DFT index.

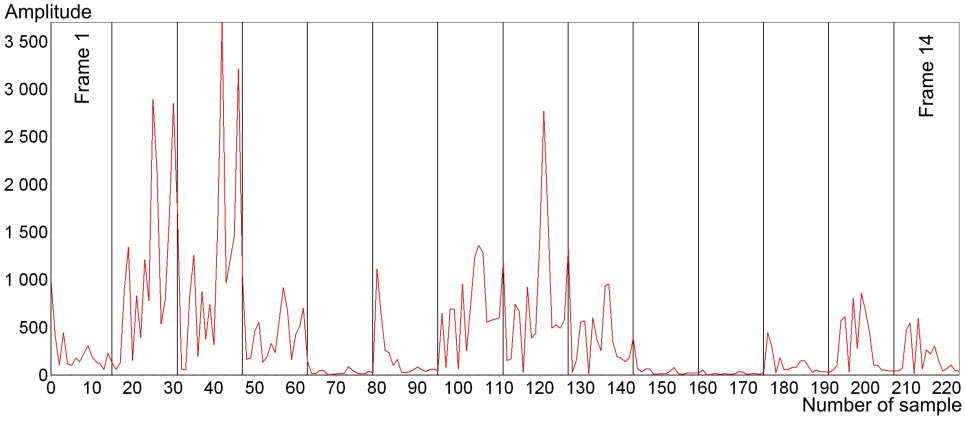

Fig. 11. Signal examined after Mel-scale filtration

The last step of determining MFCCs is finding the algorithm (Fig. 12) and reverse conversion of the signal into the time form with the reverse Fourier transform. These operations may be replaced with one dependency using the discreet cosine transform reducing the algorithm calculation complexity [4]:

$$c(k) = \sum_{n=1}^{M} \log(y''(n)) \cdot \cos\left(\frac{\pi(n-0,5)k}{M} \right) \tag{23}$$

where:

$c(k)$ - k - th Mel–cepstrum coefficient for a specific frame.

In this way, we obtain the MFCCs shown in the following diagram (Fig. 13) in which 8 coefficients per each signal frame were determined.

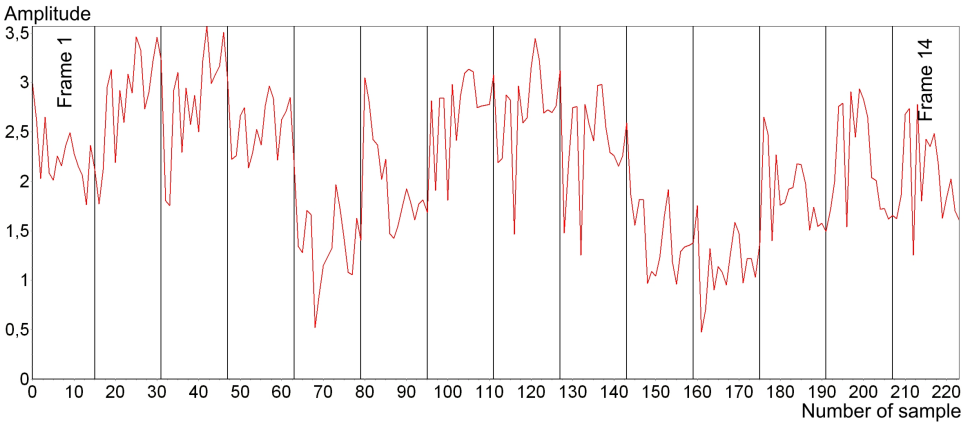

Fig. 12. Signal after finding the algorithm

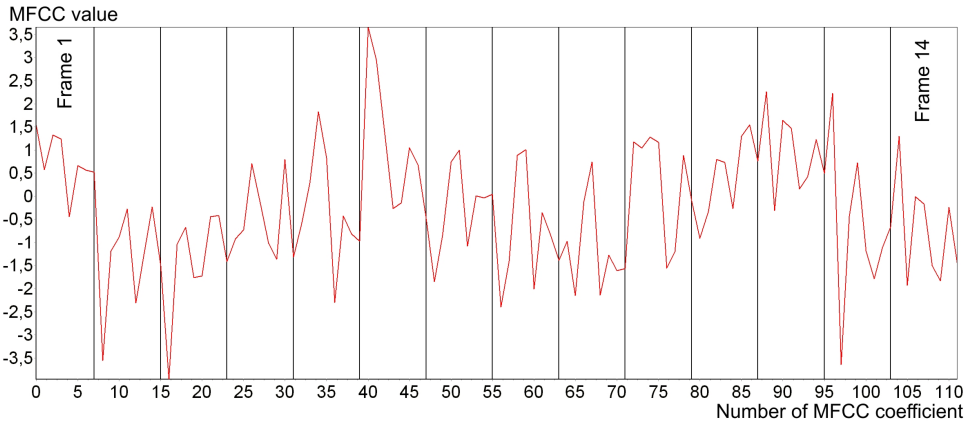

Fig. 13. Graph showing MFCC values for individual frames of the signal examined.

5. CONCLUSIONS

Cepstral coefficients are now the primary element of speech systems operation. As the research conducted shows, the application of MFCCs provides better classification results as compared to LPCCs. Still, from the algorithmical point of view, determination of MFCCs is far more difficult and complex with respect to calculations. Application of the FFT algorithm reduces the calculation complexity yet entails the need to divide the signal into frames of dimensions being a power of two (256-sample-long most frequently) and the need to remove incomplete frames. With current computation capacities, determination of MFCCs in real time does not constitute a significant problem, however.

BIBLIOGRAPHY

[1] AHMED N., RAO K. R., 1985: Orthogonal Transform for Digital Signal Processing, Springer-Verlag, New York.
[2] AUGUSTYN G., Rekursywno adaptacyjna dyskretna transformata Fouriera jako nowe narzędzie analizy sygnałów, IX Międzynarodowe sympozjum reżyserii i inżynierii dźwięku ISSET, 2001.
[3] DUSTOR A., IZYDORCZYK J., Rozpoznawanie mówców, Przegląd telekomunikacyjny, rocznik LXXVI, nr 2-3/2003, pp. 71-76.
[4] KAMM T., HERMANSKY H., ANDREOU A. G., Learning the Mel-scale and Optimal VTN Mapping, Johns Hopkins University, Center for Language and Speech Processing, 1997 workshop (WS97), 1997.
[5] PORWIK P., Isolated word descriptors as control parameters of the computer applications, Journal of Medical Informatics&Technologies, Vol.10, 2006, pp.35-46.
[6] PORWIK P., PROKSA R., Endpoints detection level for isolated words recognition, IMM 2008, Warsaw, Poland.
[7] PORWIK P., PROKSA R., Word extraction method in human speech processing, Journal of Medical Informatics&Technologies. 2008, Vol 12, pp. 209-216.

[8] PROKSA R., Metody detekcji granic słowa dla zaszumionych sygnałów mowy, Systemy Wspomagania Decyzji 2008, Zakopane, Poland.
[9] SIGURDSSON S., PETERSEN K. B., LEHN-SCHIØLER T., Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music, Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR), 2006.
[10] THRASYVOULOU T., BENTON S., Speech parameterization using the Mel scale Part II, 2003.
[11] TYCHTL Z. PSUTKA J., Speech Production Based on the Mel-Frequency Cepstral Coefficients, In EUROSPEECH'99, 2335-2338.
[12] ZHANG X., GUO Y., HOU X., A Speech Recognition Method of Isolated Words Based on Modified LPC Cepstrum, Proceedings of the 2007 IEEE International Conference on Granular Computing, 2007, p. 481.