The Voice Segment Type Determination using the Autocorrelation Compared to Cepstral Method

OLDŘICH HORÁK
Institute of System Engineering and Informatics, Faculty of Economics and Administration
University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
oldrich.horak@upce.cz

Abstract: - The extraction of the characteristic features of the speech is the important task in the speaker recognition process. One of the basic features is fundamental frequency of speaker's voice, which can be extracted from the voiced segment of the speech signal. This document describes one of the methods providing possibility to distinguish the voiced and surd segments of the voice signal using the autocorrelation, and compare the results to cepstral method.

Key-Words: - autocorrelation, cepstrum, features extraction, fundamental frequency, signal processing, speaker recognition, voice signal

1 Introduction

The speaker recognition is one of the ways how to increase the rate of success in the information system's user identification. It is a biometric method based on speaker's voice tract anatomy parameters. These parameters have direct influence on the sound timbre of the voice.

The voice signal is processed block-by-block. These blocks are short segments of the speech with time duration in tens of milliseconds. The complete voice characteristics consist of several groups of features extracted from voice segments of a specific type. Some features are extractable from the voiced segment of the speech signal only. Therefore, the determination of the segment type can be very important and useful. [1, 2]

One of the basic features is the fundamental frequency of speaker's voice. This characteristic feature can be extracted from the voiced segment, and there are more methods how to find it. The presence of the fundamental frequency in the voice segment can be also used to determinate the type of given segment. The exact value of the frequency is not important in this case. The detection of the fundamental frequency is sufficient to determination of the voiced segment.

Some methods use another process to distinguish the voice segment type. One uses determination by ***the comparison of the energy spread in the given frequency sub-ranges***. Three or more frequency ranges are defined, and the spread of energy in the all of these ranges leads to proper segment type determination. Each segment type has a typical spread of energy in the given frequency sub-ranges. [1, 2, 7, 10]

Other method uses a ***relation of the mean value of the zero-crossing rate to short-time energy*** of the voice signal segment. The voiced segments have higher value of the short-time energy, and lower mean value of the zero-crossing rate. Both of the characteristics have relative values, and are defined without units. [1, 2, 3, 10]

Another methods use a ***statistical processing*** as well [8, 9]. The autocorrelation function applied on the signal segment provides an option to detect the presence of fundamental frequency and to estimate its value. [2, 3]

2 Voice Segment Type Determination

Generally, the determination of the voice segment type is the special case of signal processing. There are two requirements in the evaluation: speed and precision. Of course, these requirements are in opposite relation.

Speed of processing can be important in the case, when the quick determination is required, and the higher error rate is tolerable. On the contrary, if the

high precision is required, the slower methods are necessary to be used to process the signal segment.

2.1 Autocorrelation Method

The autocorrelation method allows determining and evaluating the fundamental frequency of the voice segment. This statistical method uses the sampled signal as a sequence values generated by stochastic process. The sequence is combination of three components in general – trend, periodic component and noise. If the short-time segment is used, the changes of trend can be omitted as in the stationary sequence [1, 3, 4]. Then the periodic component can be detected. The autocorrelation function is defined as the function of two time points "***t***", and "***s***".

$$R(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s} \qquad (1)$$

If the stationary process is expected, the equation can be simpler. The mean value "μ" and variance "σ" are time-independent in this case, so they are the same at both time points. The autocorrelation is then the function of time period "τ" defined as the difference of these time-points:

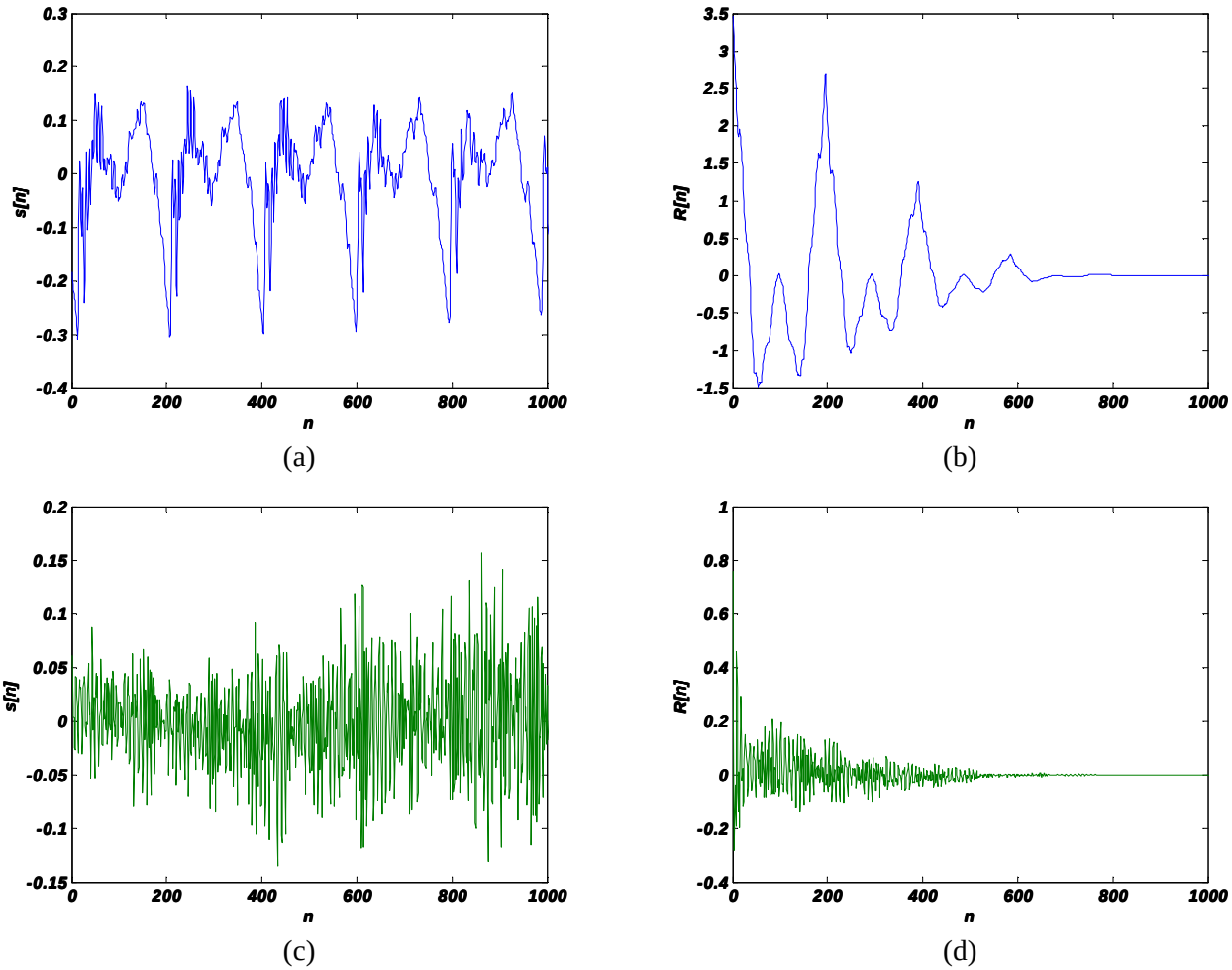
$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2} \qquad (2)$$

It is apparent, that the function (2) is even, and therefore the one-side evaluation is sufficient to be used. In the signal processing, the autocorrelation can be evaluated by the reduced equation (3):

$$R_i = \sum_{j=0}^{N-1} s_j \overline{s_{j+i}} \qquad (3)$$

$$\{s_j\}_{j=0}^{N-1} \qquad (4)$$

The periodic sequence (4) is the representation of the discrete samples of the signal, where "$\overline{s}$" denotes the complex conjugate. In the real case, the complex conjugate equals to real value, and the real value is used in place of them.



(a) (b) (c) (d)

Fig.1 – Voiced (a) and surd (c) segment of the signal, and its one-side autocorrelation function (b, d)

The *Fig.1* shows the recorded signal and its autocorrelation function for voiced (parts *a*, and *b*) and surd (parts *c*, and *d*) segment of the voice.

In the part *b* of the *Fig.1*, the first local maximum corresponds to the shift of the signal by the period of the fundamental frequency. The other maxima correspond to harmonic frequencies. It is observed, if the voiced segment is processed.

The *Fig.1* part *d* shows the course of the autocorrelation function in the case of the surd segment processing. There are none significant maxima detected. This difference can be used to determine the type of the segment.

2.2 Cepstral Method

The cepstral method uses the course of the real cepstrum coefficients "$\boldsymbol{c_n}$" to find the fundamental frequency:

$$c_n = \mathrm{Re}\left\{IFFT\left(\ln\left|FFT\left(s_n\right)\right|\right)\right\} \tag{5}$$

The sequence "$\boldsymbol{s_n}$" represents the sampled values of the voice signal segment. The *Fast Fourier Transform (FFT)*, the *natural logarithm (ln)*, and the *Inverse Fast Fourier Transform (IFFT)* are used to evaluation in general.

In the signal processing, the *Discrete Fourier Transform* can be used in the place of *FFT* to evaluate the spectrum "$\boldsymbol{S_k}$", and the cepstral coefficients "$\boldsymbol{c_n}$":

$$S_k = DFT\left\{s_n\right\} = \sum_{n=0}^{N-1} s_n \cdot e^{-i\frac{2\pi}{N}kn} \tag{6}$$

$$\begin{aligned} c_n &= \mathrm{Re}\left\{IDFT\left[\ln\left|S_k\right|\right]\right\} = \\ &= \mathrm{Re}\left\{\frac{1}{N}\sum_{k=0}^{N-1}\ln\left|S_k\right| \cdot e^{i\frac{2\pi}{N}kn}\right\} \end{aligned} \tag{7}$$

The *Fig.2* shows the real cepstrum of the voiced segment. There is marked the local maximum in the figure. It represents the fundamental frequency. [2]

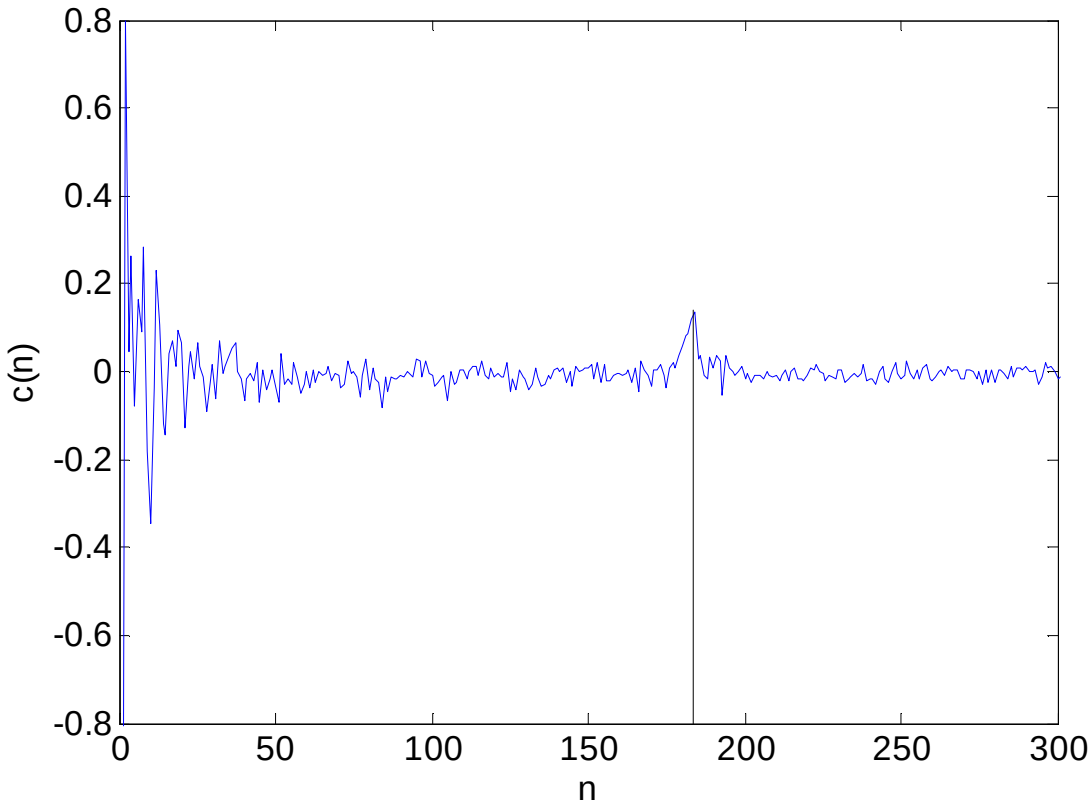


Fig.2 – Cepstrum course for the voiced segment

The value of the fundamental frequency "F_0" can be evaluated from the predefined sampling frequency "f_{smp}" divided by the order of the given cepstral coefficient corresponding to the local maximum "k". It is described by the equation (8):

$$F_0 = \frac{f_{smp}}{k} \tag{8}$$

If there is none significant maximum in the given range, the signal segment is surd, because the fundamental frequency is undetectable (see *Fig.3*).

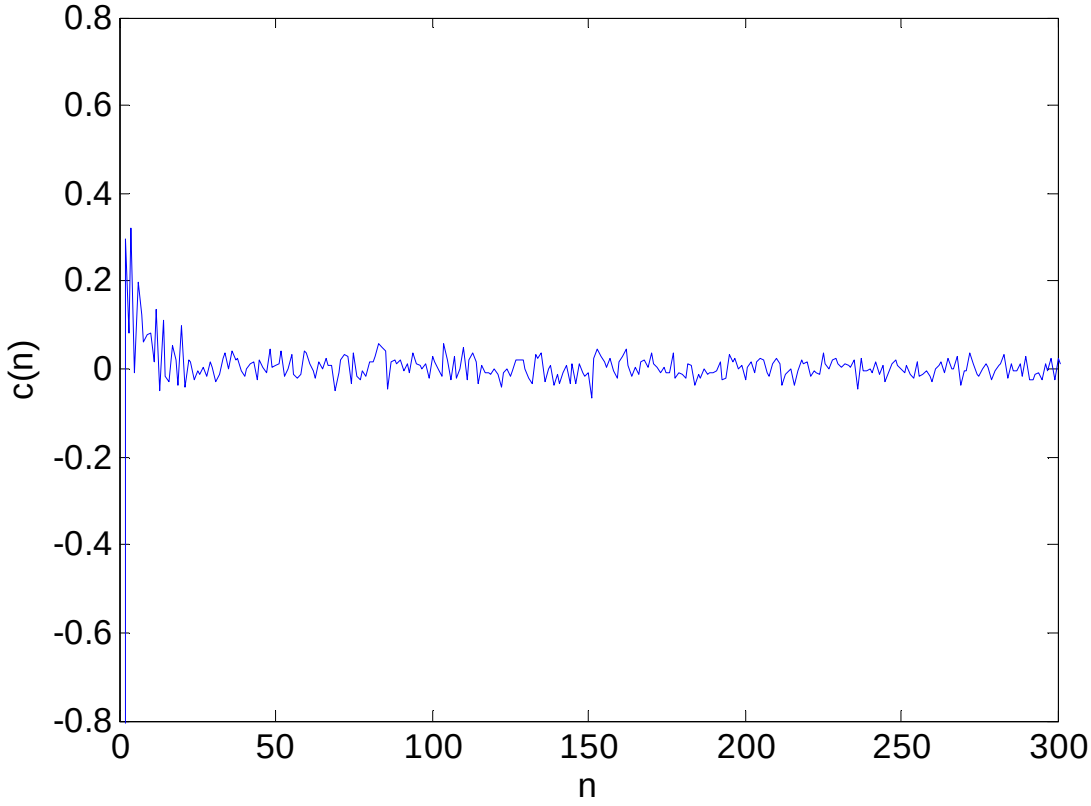


Fig.3 – Cepstrum course for the surd segment

2.3 Other Methods

There are more methods for the voice segment type determination, as they were briefly described in the introduction part. The cepstral method is the most popular, but it can be slow. Other methods can be used for faster determination of the voice segment type without the evaluating of the fundamental frequency.

The determination by ***the comparison of the energy spread in the given frequency sub-ranges*** is described in [1]. This method is based on definition of three or more frequency ranges, where the total energy is calculated. The spread of totals of the energy is typical for both voice segment types. If the spread of energy is compared to experimentally defined model, the proper segment type will be determined. Each segment type has the model of the typical spread of energy in the given frequency sub-ranges. [1, 2, 7, 10]

The method using a ***relation of the mean value of the zero-crossing rate to short-time energy*** of the voice signal segment is described in [1]. The characteristics of the short-time energy and mean value of the zero-crossing rate are calculated for each voice segment. The characteristics have relative values, and they are defined without units. If they are plotted in the chart with the energy on the horizontal axis, the cluster of the voiced segments will be seen in the area of high energies. Surd segments are situated near the vertical axis with low energy and higher value of the zero-crossing rate. [1, 2, 3, 10]

3 Fundamental Frequency Detection

The autocorrelation and cepstral methods are to be used for the fundamental frequency detection. The exact value is not important for the decision, only the fundamental frequency presence.

The MATLAB environment is used for the analysis and evaluation. The voice signal is recorded by the microphone using sound recorder program included as a part of the operating system.

The wave format is imported in the MATLAB environment directly in variables as the sampled values sequence and the sampling frequency. The embedded function can be used to processing.

3.1 Data Preprocessing

The sampling frequency of ***22050 Hz*** is used for all the records. This frequency is one of the basic sampling frequencies provided for the sound recording by the PC's common sound cards and the appropriate software. Other standard sampling frequencies are half and double value of this frequency. The sound can be recorded in one or two channels to have mono or stereo record. The one channel is sufficient for our experiments.

There are six files of cardinal numerals of the Czech language recorded. These words are sound different each other. It is usable for the sufficient variability of the signal. [3, 4]

Each file is as the sequence of values divided in the short segments. The recommended duration is about ***20 milliseconds***. The sample count in the segment equal to power of two is the required value for the better evaluation. The length of ***512 samples*** corresponds to duration of ***23 ms*** at the given sampling frequency (see *Fig.4*).

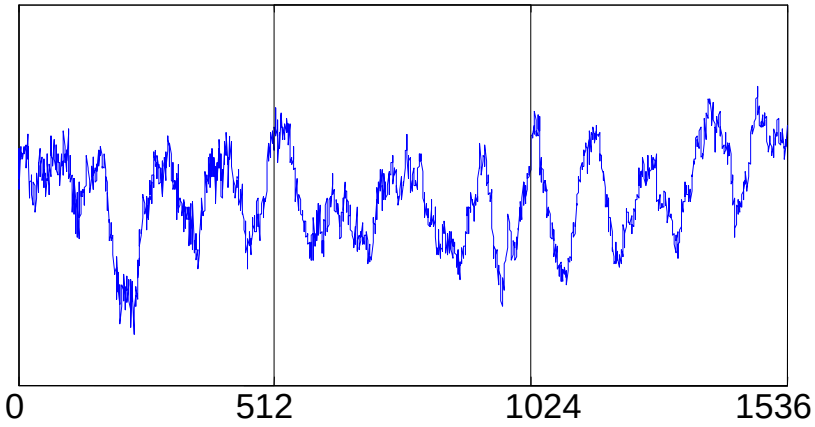


Fig.4 – Signal to segments division

The segments are limited by the *Hamming window* before the processing to avoid border effects (see *Fig.5*). The *Hann window* is optional for the autocorrelation.

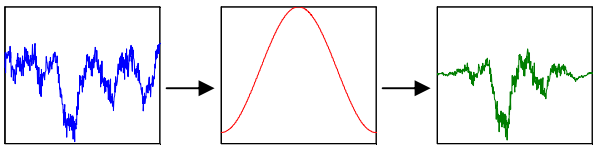


Fig.5 – Using of Hamming window

3.2 Determination by Autocorrelation

The autocorrelation coefficients sequence evaluation is the first step of the determination. The sequence is evaluated from the preprocessed signal segment using (3), and normalized by the value with index ***0***.

The common range of the fundamental frequency is from ***60 Hz*** to ***400 Hz*** for the human voice. It corresponds to autocorrelation coefficient with index from ***150*** to ***300*** for the sampling rate ***22050 Hz***. We have to search for a significant local maximum value in this range of coefficients. The threshold for the maximum is relative value ***0.5***, and the decision is positive for appropriate value found in the given range of coefficients. [1, 2]

3.3 Determination by Cepstrum

The evaluation of the real cepstrum is processed using equation (7). The cepstral analysis expects the significant value of local maximum for the range of the human voice fundamental frequency [4, 5]. As described above, the frequency range is from ***60 Hz*** to ***400 Hz***, and the range of appropriate cepstral

coefficient (8) is from ***50*** to ***350*** for the sampling frequency ***22050 Hz***.

The *Fig.6* shows the composition of cepstra of several signal segments in the given range of cepstral coefficients. There are the maxima rising from the decreasing values on the background. The maximum is detected as a value higher than ***1.5*** multiple of the background values.

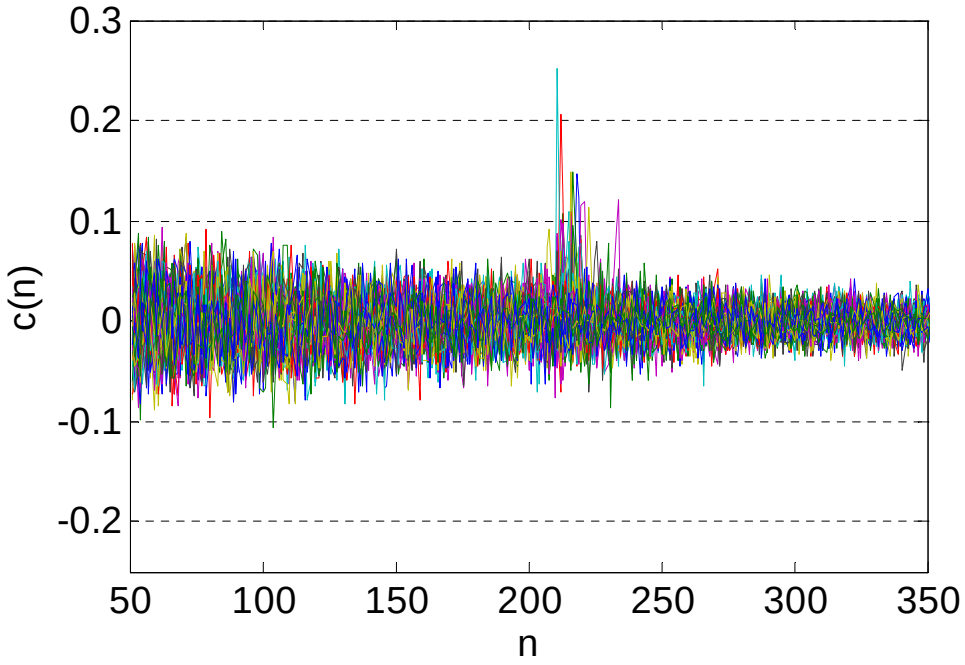


Fig.6 – Composition of Cepstra

3.4 Comparison

The *Figs.7 to 12* show the comparison of the both methods results. There are six signals of cardinal numerals (one to six) in the Czech language processed by autocorrelation, and cepstral method. The "***n***" means the segment number. If the segment is determined as voiced by the autocorrelation, it is filled in the "***A***" bar of the picture. The "***C***" bar shows the voiced segment determined using the cepstral method.

The comparison shows the difference in units of segments. The counter value of error rate is about less than ***10%*** (see *Tab.1* below the *Figs.7 to 12*).

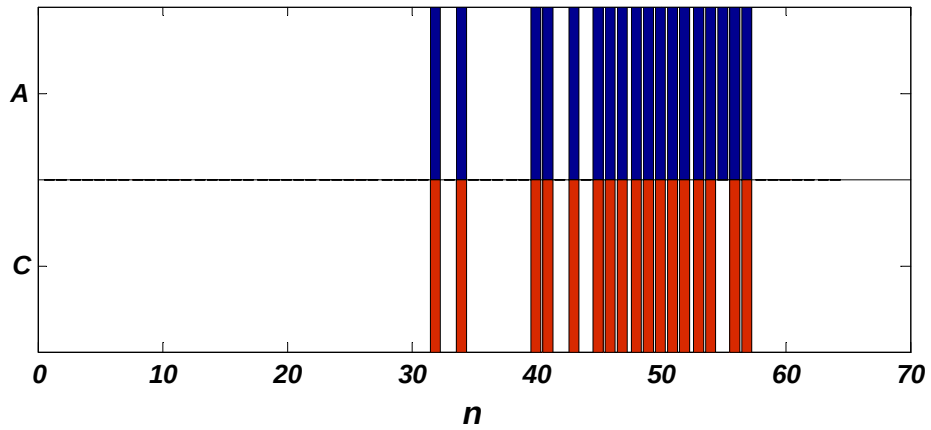


Fig.7 – Comparison for word "jedna"

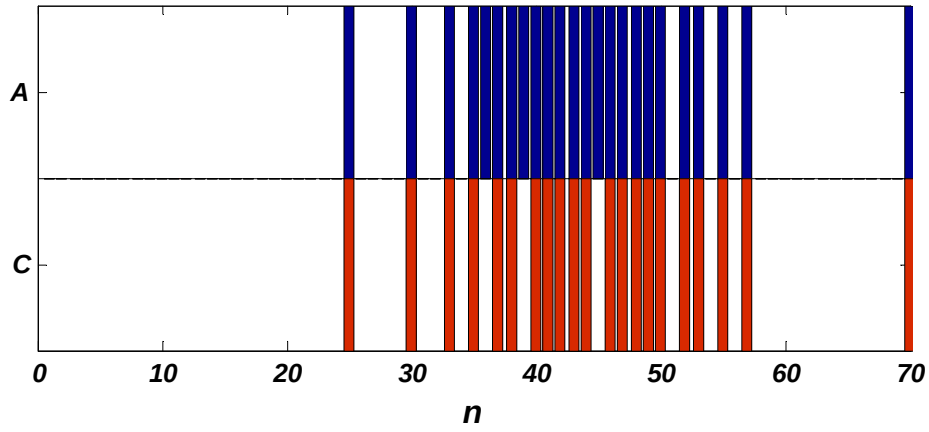


Fig.8 – Comparison for word “dvě”

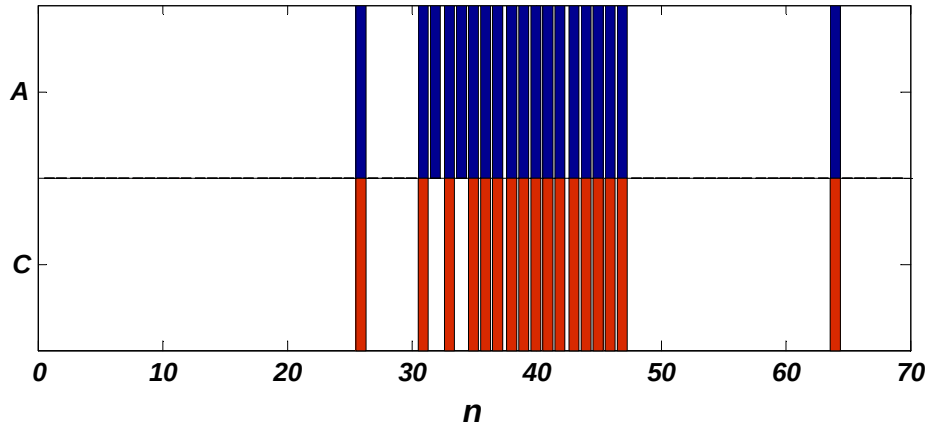


Fig.9 – Comparison for word “tři”

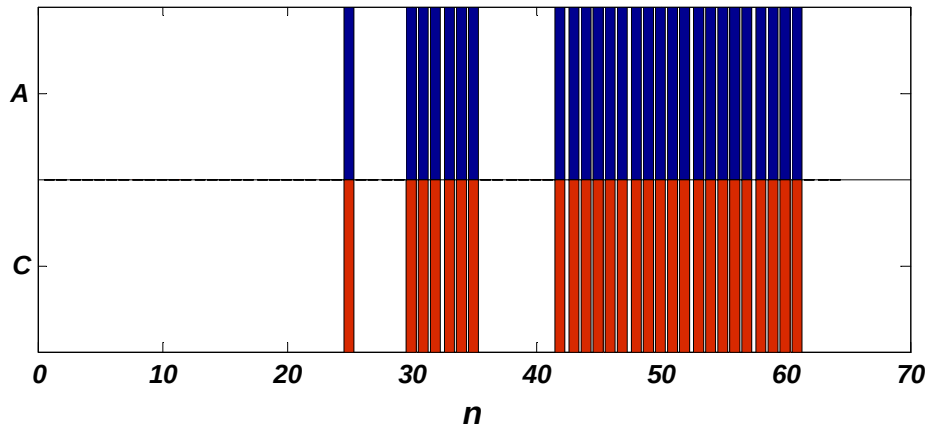


Fig.10 – Comparison for word “čtyři”

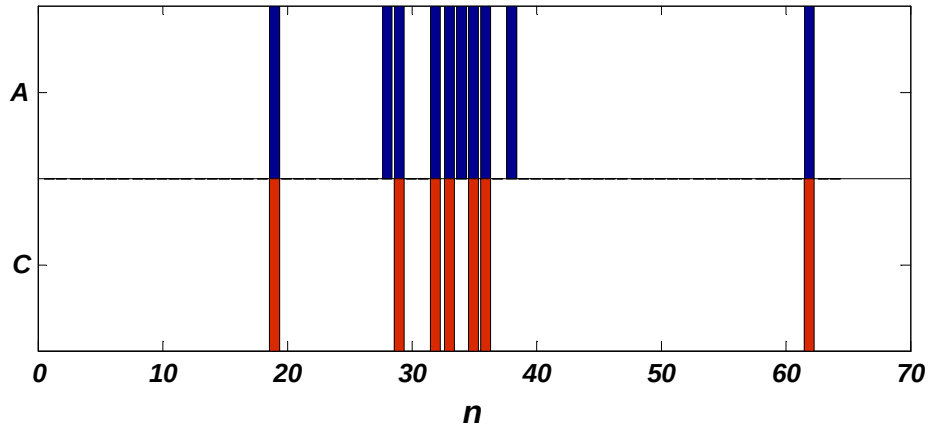


Fig.11 – Comparison for word "pět"

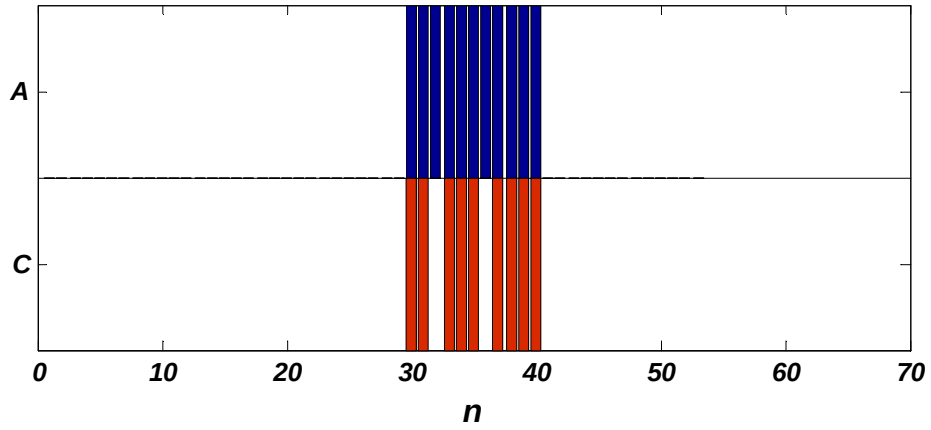


Fig.12 – Comparison for word "šest"

Compared word		Total segments	Autocorrelation		Cepstral method		Difference	
			Voiced	Surd	Voiced	Surd	Absolute	Relative
1	*jedna*	70	18	52	17	53	**1**	**1,4%**
2	*dvě*	70	24	46	21	49	**3**	**4,3%**
3	*tři*	70	19	51	17	53	**2**	**2,9%**
4	*čtyři*	70	27	43	27	43	**0**	**0,0%**
5	*pět*	70	10	60	7	63	**3**	**4,3%**
6	*šest*	70	11	59	9	61	**2**	**2,9%**

Tab.1 – Results comparison

4 Conclusion and Future Work

The results of the experiment show that the counter value of error rate given by segment determination difference is under the common error rate of both methods. It means the determination of the voice segment type using cepstral method can be replaced by the faster autocorrelation method. [4, 6]

It opens the way to select the voiced segments using the autocorrelation first, and then evaluate the fundamental frequency using cepstral method more

precisely. The time intensive cepstral method will not be used for surd segments.

There is possible to select the voiced segment more effectively using the autocorrelation method, and extract the important voice features in shorter time.

The speaker recognition used as the method of information system's user identification is not sufficiently reliable. But, if it is used in the combination with other method, the identification reliability will be increased. More speech signal characteristics and features, than the fundamental frequency only, are necessary to extract from the speaker voice to be possible to use it for the identification. Most of these features are evaluable by the cepstral analysis, but not by autocorrelation. Therefore the evaluation is time intensive, and the next comparisons and experiments need to be done.

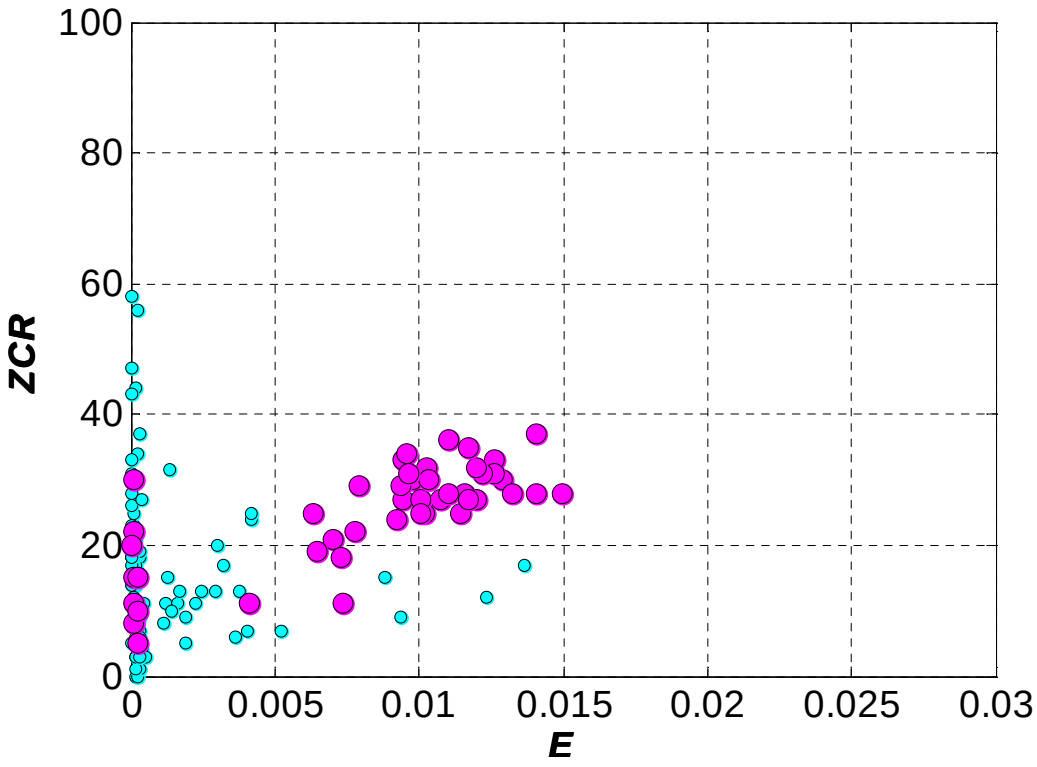


Fig.13 – Voiced and surd segments determination by energy and mean value of the zero-crossing rate

The next step of the complex solution is to find more characteristic features able to be extracted in short time. A long-time processing is hardly usable for the speaker recognition as the user identification process.

The near future work is to compare the methods applied on the longer speech. It will lead to higher count of the signal segments. The optimal length of the segment would be found. *Fig.13* shows the first step of the experimental comparison of the segment type determination by energy and mean value of the zero-crossing rate. It will also be compared to autocorrelation and cepstral method.

The small circles in the picture (*Fig.13*) are surd segments; voiced segments are drawn with the big circles. The type of the segment is determined by the sound listening. The values of energy and zero-crossing rate are evaluated from the signal.

The final comparison is the next work in this research.

References:

[1] H. Atassi, Metody detekce základního tónu řeči. *Elektrorevue*, Vol.4, 2008, ISSN 1213-1539.

[2] J. Psutka, et al., *Mluvíme s počítačem česky*. Praha, Academia, 2006, ISBN 80-200-1309-1.

[3] Y. Tadokoro, et al., Pitch Estimation for Musical Sound Including Percussion Sound Using Comb Filters and Autocorrelation Function, *Proceedings of the 8th WSEAS*

International Conference on Acoustics & Music: Theory & Applications, Vancouver, Canada, June 19-21, 2007, pp. 13-17

[4] C. Moisa, H. Silaghi, A. Silaghi, Speech and Speaker Recognition for the Command of an Industrial Robot, *Proceedings of the 12th WSEAS international conference on Mathematical methods and computational techniques in electrical engineering*, Stevens Point, Wisconsin, USA, 2010, pp. 31-36, ISBN: 978-960-474-238-7.

[5] M. Vondra, Kepstrální analýza řečového signálu. *Elektrorevue*. Vol.48, 2001, ISSN 1213-1539.

[6] M. E. Torres, et al., A Multiresolution Information Measure approach to Speech Recognition, *Proceedings of the 6th WSEAS International Conference on Signal, Speech and Image Processing*, Lisbon, Portugal, September 22-24, 2006, pp. 187-192.

[7] E. Marchetto, F. Avanzini, and F. Flego, An Automatic Speaker Recognition System for Intelligence Applications, *Proceedings of the 17th European Signal Processing Conference (EUSPICO 2009)*, Glasgow, Scotland, August 24-28, 2009, pp. 1612-1616.

[8] J. Sohn, N. S. Kim, and W. Sung, A Statistical Model-Based Voice Activity Detection, *IEEE Signal Processing Letters*, vol. 6, no. 1, January 1999, pp. 1-3.

[9] A. Stolcke, S. Kajarekar, and L. Ferrer, Nonparametric Feature Normalization for SVM-based Speaker Verification, *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2008)*, vol. 104, no. 23, pp. 1577-1580.

[10] J. P. Campbell, Jr. Speaker recognition: a tutorial. *IEEE 85*, 1997, pp. 1437–1462.