THE FAST FOURIER TRANSFORM AND CEPSTROGRAM BASED APPROACH TO THE ASSESSMENT OF HUMAN VOICE STABILITY

Krzesimowski Damian

***Abstract**: The topic of the paper is to assess the stability of the human voice on the basic of results of Fast Fourier Transform and cepstrogram, which are subjected to statistical analysis. The presented results are the first part of the study on the usefulness of these analyses in voice quality assessment and identification of persons and voice commands in a noisy environment. The purpose of the study is to select the indicators, characteristic for the voice of the one particular person. In the paper is described the data selection algorithm for testing purposes from a single recording of a human voice, and the results of statistical analysis of the stability of expression of voiced vowel. Proposed algorithm allows extracting voiced elements of a desired length from the recordings with the exception of noise and silence. Then, to assess stability of waveforms, the recording is divided into several to several tens of fragments a length of few tens milliseconds. Each fragment is analysed independently of the other, and the result is a measure of the error of inference algorithms for identifying the person and voice commands. Particular attention is given to comparing the results of statistical analyses, after the splitting into blocks of the same duration and the same number of micro phonemes in the waveform. Due to fact, that in the studies have used a frequency analysis, it is possible to determine the stability of both the fundamental frequency and formants using the same statistic apparatus.*

***Keywords**: voice, FFT, cepstrogram, data selection, signal processing.*

***ACM Classification Keywords**: F.2.1 Analysis of Algorithms and Problem Complexity – Numerical Algorithms and Problems – Computation of transforms, G.3 Mathematics of Computing – Probability and Statistics – Statistical computing*

Introduction

The topic of the presented research is the usefulness of Fast Fourier Transform and cepstrogram for assessing voice quality of a healthy person, and identify the persons and voice commands. The purpose is to define the vectors, which are characteristic for the individual human voice using signal analysis and statistics. Research should be used to identify people, voice commands and voice quality assessment such people as opera singers or speakers. It will be possible to apply the developed algorithms in speech therapy and the voice training for speakers or singers. The basis for the research will be own database of recordings at least 100 persons of both sexes at different ages.

The survey plan is established as follows:

1. development of algorithms for the selection of voiced recordings of the human voice with the elimination of noise components;
2. execution of Fast Fourier Transform, spectrographic and cepstrographic analyses on selected parts of recordings;
3. statistical study of the results of the mentioned above numerical analysis for a single recorded signal divided into sections of predetermined length;

4. selecting from the resulting figures coefficients which for one recording (one person) will not deviate beyond a preset threshold error;
5. comparison of numerical results with the results for step 4 for at least 100 people for verification of found coefficients.

In the paper is presented:

1. data selection algorithm for testing from a single recordings of a human voice;
2. Fourier analysis, spectrographic and cepstrographic for one recording;
3. statistical analysis of the mentioned above numeric result studies for one recording.

The subject of the described stage of research is to determine the stability of the human voice for the one person. Unmodulated voiced sound, typically pronounced a few seconds by a person with a healthy speech executive and decision-making apparatus, the listener perceives as stable. That is, changes cannot be detected at the lowest level by the listening [Lombardi et al, 2009, Larrouy-Maestri et al, 2012]. For the research purposes was introduced the concept of micro-phoneme, as the smallest indivisible entity, repeated periodically while generating voice. Duration of the micro-phoneme is possible to calculate on the basis of the fundamental frequency and ranges from 4ms to 10ms. The shape of micro-phoneme includes both fundamental frequency sine wave, formants, and noise associated with the imperfection of the speech executive apparatus. The question posed by the author, on which tries to answer in this paper, is as follows: how stable during several microseconds is the human voice and how it affects the results of the analysis of the signals?

Data preparation

It was decided, that to develop of voice pattern would be conducted recording of voiced vowel, as in many similar researches [Bala et al, 2010, Fang and Gowdy, 2013]. A vowel used in the study is the vowel "a" pronounced a few seconds. On the next stages of research will be allowed the opportunity to change the frequency of voice generated by a recorded person. This is a very important statement because sound in the process of speaking is modulated, and an object of the present stage of research is to obtain a vector identifying a person, regardless of the frequency of sound generated by it. Work began with the recording of sound "a" pronounced by 1944ms by a man at the age of 31 years. The voice was recorded using the built-in condenser microphones device Tascam DR-40 with parameters: sampling frequency of 96kHz, the accuracy of 24 bits per sample, the format of uncompressed WAV/BWF. Recording formed the basis for the development of algorithms for selecting input signal for analysis. The waveform of the recording is shown in Fig. 1.

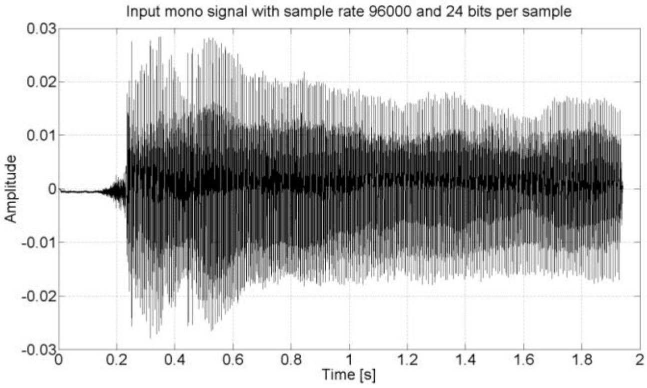

Fig. 1. Voice recording as a basis for the development of algorithms for data selection.

Subjectively established recording length, subjected to analysis, is 1000ms. It was also decided to eliminate the elements of silence and noise at the beginning and end of the recording. Also take into account the possibility of discontinuous speaking, that is, recording containing voiced elements divided by with silence or noise. Partly based on the published results of the author [Krzesimowski, 2012].

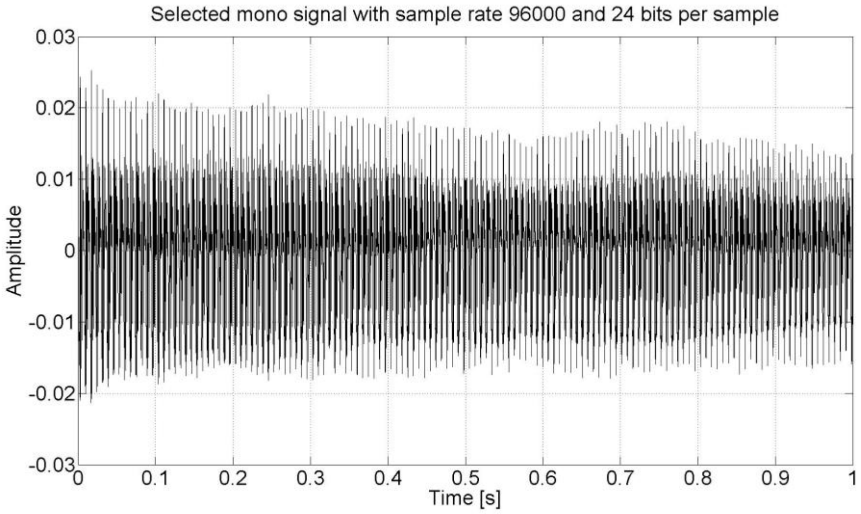

Fig. 2. A fragment of recording with a length of 1000ms selected from the recording shown in Fig. 1.

Material selection is based on the determination of the maximum value of the recording and division recordings on blocks with a length of 6400 samples. For the sound recorded at a frequency of 96000Hz are obtained sections of 67ms length. Then, the maximum value of the whole recording is compared to the registered in block. If this value is less than the 16.7%, block is rejected, otherwise the block is appended to the new vector. The result is the recording devoid of elements of silence and low-amplitude noise. The length of the recording thus obtained may be greater than expected 1000ms, that is why occurs the cutting of a predetermined length. For this purpose, it is determined the midpoint of data, and next, intervals of length 500ms counted from that point. Selection is relative to the centre to eliminate the modulation at the beginning and end of the recording. The waveform of the recording after the described selection is shown in Fig. 2.

Studies so far of the voice [Reilly et al, 2004, Grimaldi and Cummings, 2008], including studies of the author [Krzesimowski and Ciota, 2010], were based on analysis of voice in the form exactly as shown in Fig. 2. This time, it was decided to approach the analysis in a different way, not by analysing data from the whole recording, but by analysing the variability in the data based on the same recording treated as many independent recordings. This means, that waveform of signal analysis is not important at this stage, but its variability described in a statistical manner. This variability is then compared to all the recordings in order to extract the characteristic vectors of the person. The first step to getting so understood result is the division of the current data for a few to several dozen. At this stage of the study the recording was divided into 25 parts, the length of which is determined by the expression (1).

$$PL = \frac{(RL - 1)}{NP} \tag{1}$$

where: PL is the length of the current part, RL is the length of the recording and NP is amount of parts. Samples are cut down with an added margin with length dependent on the length of recording after the elimination of silence, and the number of the desired fragments in accordance with the expression (2).

$$ML = \frac{PL}{20} \tag{2}$$

where: ML is the length of the margin and PL is the length of the current part. Example charts of two sections of the recording are shown in Fig. 3 and Fig. 4. Each fragment is 44ms length.

In Fig. 3 are mapped the 5 full micro-phonemes, two micro-phonemes at the end and the beginning of the recordings are cut off. In Fig. 4 are mapped 6 full micro-phonemes, but the latter is cut off.

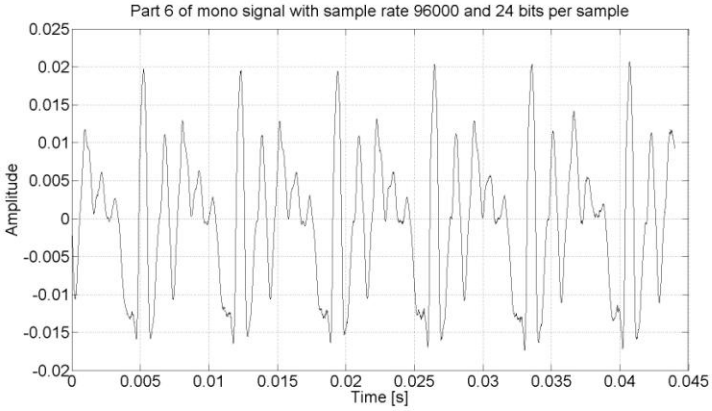

Fig. 3. Fragment number 6 after the split into blocks.

Given the fact, that a different person generates voiced sounds with different frequencies, the number of full micro-phonemes in so-characterized blocks may be different. In addition, because studies have used the frequency characteristics, it is possible that unacceptable errors can occur. Therefore it was decided to extract the blocks in such a way, that they contain only full micro-phonemes without cropping. For this purpose, the algorithm of trimming blocks was developed with respect to the local maximum value, and the counting of occurrences of that value. The fact is used of occurrence of one clear maximum in each micro-phoneme.

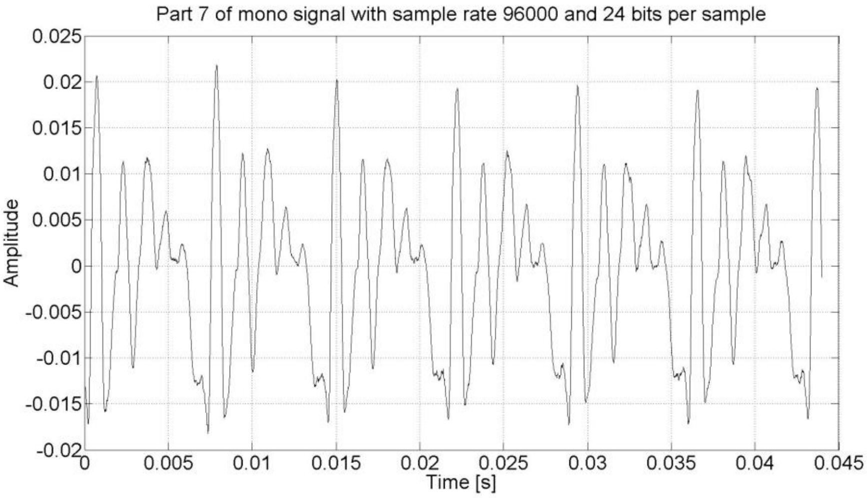

Fig. 4. Fragment number 7 after the split into blocks.

In the first step of cutting the maximum value is determined and the minimum value of block, and then is calculated the counter limit according to the expression (3).

$$CL = \frac{\left(|Max| + |Min|\right)}{10} \tag{3}$$

where: CL is limit for counter, Max is the maximum value of the current block and Min is the minimum value of the current block. Position the first values, which will satisfy the condition *max – limit*, is saved as the initial flag cut. Then all the positions of values are stored that meet condition, far from the position of the last value not less than 100 samples. After checking the entire block, as the final flag, is taken the position of value that starts the last section of maximum. In this way, ensured is sufficiently accuracy to extracting the full micro-phonemes, with designated a maximum error the peaks of the amplitude of 0.0005. So obtained examples of two blocks are shown in Fig. 5 and Fig. 6.

So obtained data blocks have different lengths related to the modulation of voice while recording. Moreover, in this example blocks contains a different number of micro-phonemes, 5 in Fig. 5 and 6 in Fig. 6. Therefore, the algorithm has been expanded so that the number of micro-phonemes was the same in the blocks. For this purpose, after cutting the recording in accordance with the previous point, its length is measured. This number is a reference value for the next block, and it is necessary to determine the 5% margin of error in estimation of length. If the next segment of recording is in this range, then is saved to the file, and another segment is collected. If one of the following sections of the recording is longer than the designated margin, it is trimmed in accordance with the criterion of maximum signal to the desired range. Alternatively, if a fragment length is shorter than the designated range, the length becomes a reference value. Thus, the appointed interval time is new, and splitting the entire recording begins again.

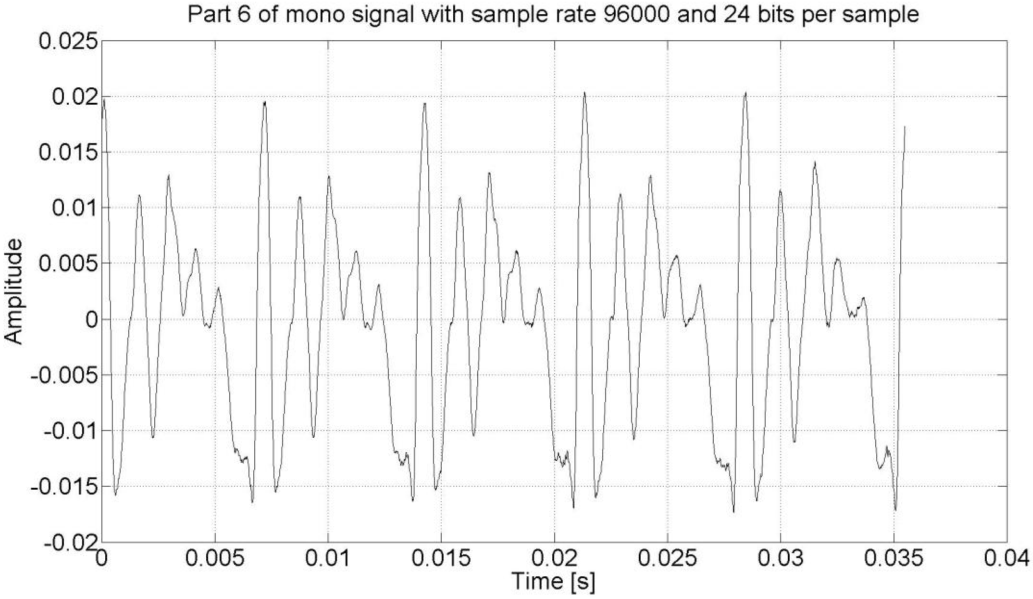

Fig. 5. Example section number 6 consisting only full micro-phonemes.

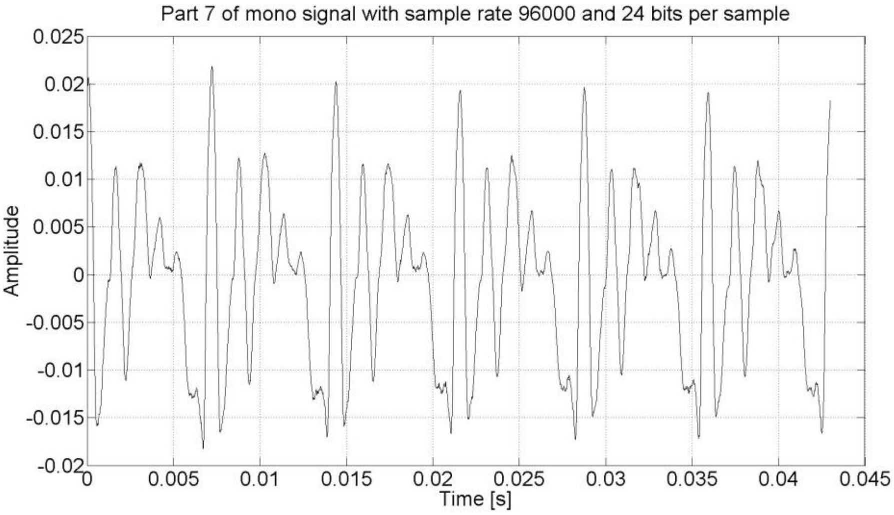

Fig. 6. Example section number 7 consisting of only full micro-phonemes.

Defined in this way algorithm allows to obtain fragments of recordings of the same number of micro-phonemes with an accuracy of 5% of their duration. In addition, in this way are eliminated the errors caused by trimming the voice recording in the first stage, if there were breaks in the sounds voiced pronouncing. Sample parts of recording after applying the final stage of selection are presented in Fig. 7 and Fig. 8. It is worth emphasizing, that the input and output data, at each stage of the division, are the sounds – the recorded voice that can be played. There is no interference in the waveform, operations affect only the duration of the recordings.

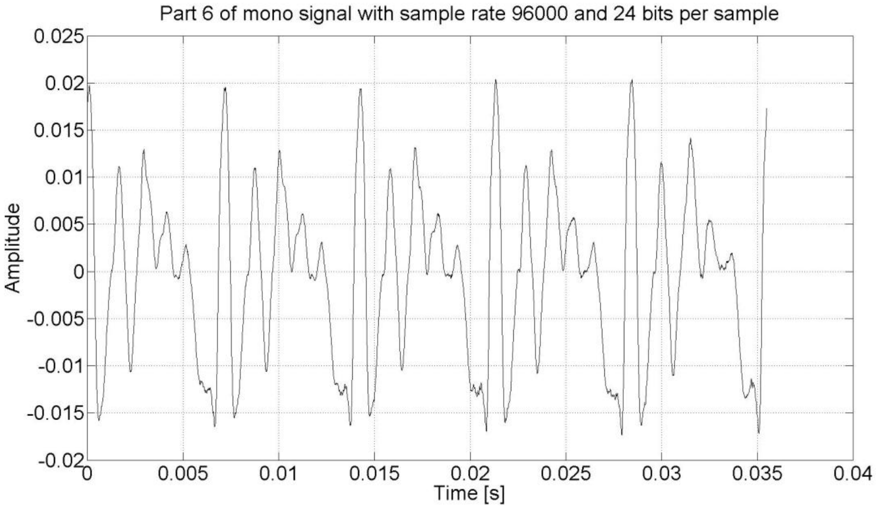

Fig. 7. Section number 6 composed exclusively of full micro-phonemes with a length within the limits of tolerance relative to the remaining blocks.

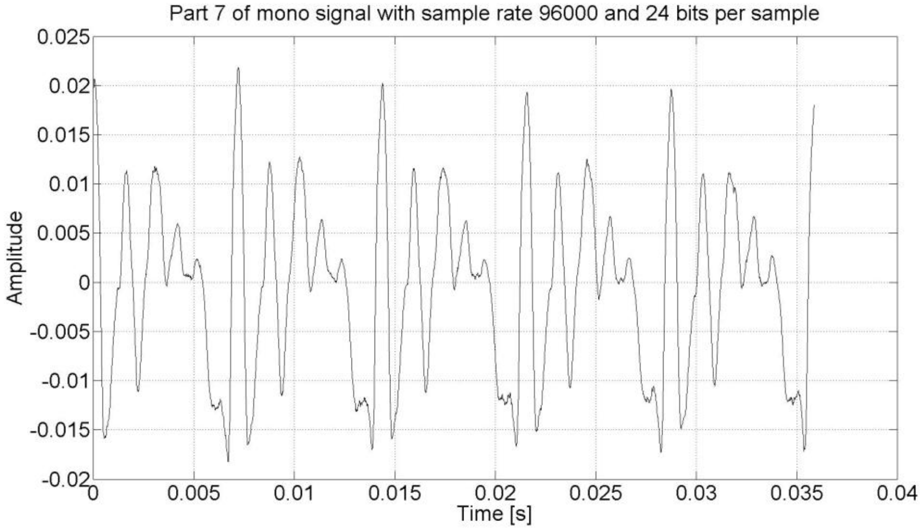

Fig. 8. Section number 7 composed exclusively of full micro-phonemes with a length within the limits of tolerance relative to the remaining blocks.

Analysis

The next step is to perform, for each block, signal analyses, mentioned above Fast Fourier Transform, spectrographic and cepstrographic. To get a complete view of the changes, the analysis were made of segments from each of the stages of division. It was noted small changes in signal waveform analysis.

These changes are undoubtedly related to the way in which was selected research material. Thus proved, that the method of selecting of the test material has an influence on the final results, regardless of the nature of an analysis. It was decided to estimate the size of the observed changes in the waveform using the statistical apparatus. Number of analysis of signals is equal to the number of blocks for which the input signal is divided. The values are stored as vectors in external files, possible to open in most computational programs.

In the same way are stored results for other analyses. Statement of vectors for each piece of recording in the matrix represents the starting point for statistical analysis. For the analysed sample of voice was determined:

1. standard deviation;
2. median;
3. arithmetic mean;
4. maximum value;
5. covariance;
6. correlation coefficient.

Two recent analyses were not carried out for a spectrogram, which results from the final vectors of different lengths for different blocks. In Fig. 9, Fig. 10 and Fig. 11 are presented as an example of the correlation coefficient for the Fast Fourier Transform relative to each of the sections of the recording of all three stages of the selection of material.

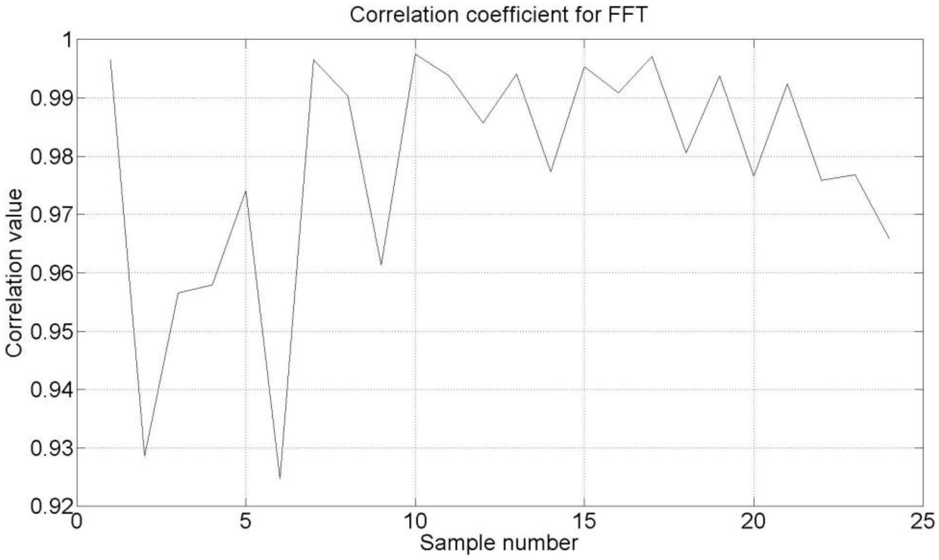

Fig. 9. Chart of the correlation coefficient for the Fourier transform performed for first stage of the selection of material.

At one stage the selection of data achieved are 16 characteristics of a statistical analysis of signal analysis. For described audio samples were obtained in a total of 48 plots, which are compared with each other. In the paper are described observations related to changes in the shape of the characteristics and values for each analysed separately fragments. There is no description of changes associated with the statistical values obtained from the analysis of signals.

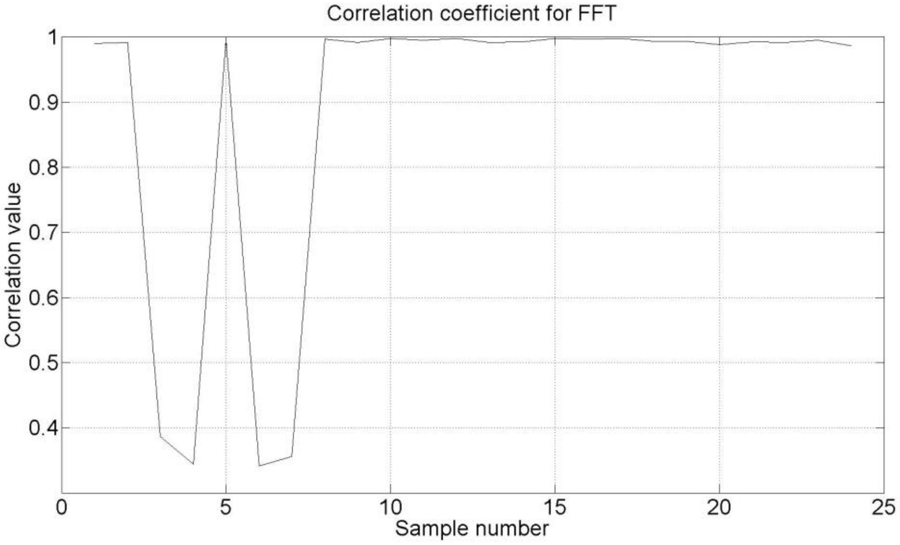

Fig. 10. Chart of the correlation coefficient for the Fourier transform performed for second stage of the selection of material.

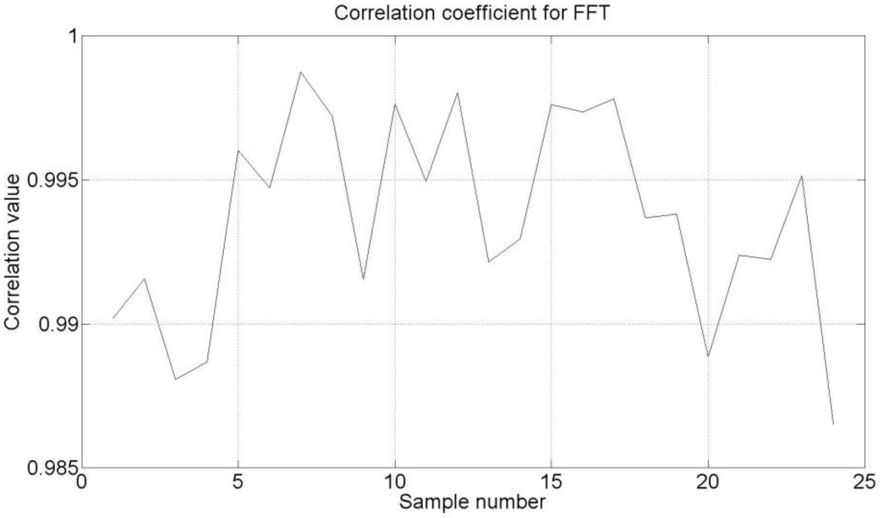

Fig. 11. Chart of the correlation coefficient for the Fourier transform performed for third stage of the selection of material.

In the case of the arithmetic mean and standard deviation was noted, that the values are 20% greater for the blocks after the first cutting step as compared to other steps. The shape of the waveform after the first and third stage of selection is on a close approximation the same. However, significant differences were noted in the waveform after the second selection phase, for the samples with number 4 and 7. These samples contain a different number of micro-phonemes than the other; the waveform of block number 7 has already been presented in Fig. 6. These same differences between blocks caused similar changes in the waveforms of the arithmetic mean and standard deviation for cepstrogram. Here as well noted a significant boost to values for samples number 4 and 7 after the second stage of selection. The shape of the waveform for the first and third cutting step, however, is different. In the case of the graph of the arithmetic mean and standard deviation for a spectrogram not be noticed significant changes associated with in the length of the analysed fragments, whereas the waveform after the second and third stages are almost identical.

There has been noticed a significant differences in the waveforms of the maximum and median Fourier transform between the first and second and third selection step. In addition, the lengths of the analysed blocks affect the value of the maximum for sample number 4 and 7. Apart from these two exceptions, the waveform of results of the second and the third stage is the same. An identical relationship has been observed for the waveform of maximum and median of cepstrogram and spectrogram. Here too the biggest differences are related to the different number of full micro-phonemes in blocks after the selection. The most interesting from the point of view of the usefulness of the described research have proved waveforms of covariance and correlation coefficient. For a chart of covariance for Fourier transform strong correlation was observed depending on the length of the analysed sections of the data. Again, the most abrupt changes have been observed in the waveform on the border of fragments with number 4 and 7. In addition, the change between the blocks of the same length and blocks having the same number of micro-phonemes proved to be negligible. On the other hand the differences between the first, the second and the third cutting steps has been observed for the waveform of covariance for cepstrogram. After the first stage, the amplitude of the waveform is characterized by significant changes. Waveforms in the second and third stage of selection are almost identical, and much more smooth.

The correlation coefficient has been treated in a particular way, as the most important parameter associated with the stability of the voice. In the case of the waveform for Fourier transform for the first stage of the selection values are in the range from 0.925 to 0.997 (approximately 7.5% difference). After the second stage, a significant influence the amount of micro-phonemes on the waveform has been observed, which can be seen in Fig. 10. Because they are compared with each other the current and the next block, the difference in their lengths has to affect the final result. After the third selection step values of the correlation coefficient ranges from 0.987 to 0.997, the difference between these values is approximately 1%. In the case of the cepstrogram no significant changes had been noticed in relation to the amount of micro-phonemes in the sample. However the differences have been noted in the waveforms for blocks after the first and subsequent stages of selection. Values of the correlation coefficient after the first cutting step are from 0.25 to 0.49, after the second step from 0.31 to 0.47, while the third stage from 0.31 to 0.5.

Conclusion and future work

The graphical results allowed formulating conclusions regarding the extracting methods for research material and sensitivity to changes in blocks for signal analyses. On the basis of the correlation coefficient has been found, that the material cut into equal blocks is a solution introducing an error. It results from the fact, that a block may include various amount of full and fragments of other micro-phonemes. From the parts of statistical analyses can be concluded, that the material selection can be terminated at this stage, but only in the case of the Fourier transform. The results of spectrographic and cepstrographic analyses proved to be sensitive to inaccurate trimming the research material. Such data should not be subjected to detailed analysis. Error appointed on the basis of the correlation coefficient for the Fourier transform is about 7.5%, there is need for further steps of selection of material. The second stage cuts allow storing only full micro-phonemes without their fragments. The result is the difference in the duration of the analysed fragments, which have significantly influenced the results of all analyses related to the Fourier transform. However, has not been noted the influence the length of the analysed data in the waveform of the spectrogram. To obtain meaningful final results, characterized by as little as possible error associated with the selection of data, has been developed third stage of cuts. This results in data, that contains the same number of micro-phonemes in each new block, and thus for the unmodulated voice is the time of each piece almost the same. Error determined on the basis of the correlation coefficient for the Fourier transform of the thus-obtained fragment of approximately 1%. This number can be considered as the value of volatility of the human unmodulated voice. This is a numeric value of changes, that may be seen in the waveforms of analysed fragments (compare Fig. 3 and Fig. 4) and which is associated with the stability of the speech executive apparatus. Moreover, thanks to the presented method obtaining the results of signal analysis is possible to use described frequency analysis, which are sensitive to different parameters related to the input data.

As the article explains how to evaluate the stability of the voice for one person, the next stage of research is to repeat the same steps for at least 100 people of all ages and gender. Still recorded will be voiced vowel, pronounced a few seconds without modulation. After estimating the stability of the voice for each person, on the basis of analysis of signals used here will be extracted factors, that do not change with the modulation of the voice of the person. Therefore will be referred to the similarity waveforms of micro-phonemes without their duration. This will allow the develop a model of the waveform of voiced sounds for the person, which will allow for the unambiguous identification of the person taking into account the error determined at the first, described in the paper, the research stage.

Bibliography

[Bala et al, 2010] A.Bala, A.Kumar, N.Birla. Voice command recognition system based on MFCC and DTW. In: International Journal of Engineering Science and Technology. Vol. 2 (12). 2010

[Fang and Gowdy, 2013] E.Fang, J.N.Gowdy, New algorithms for improved speaker identification. In: International Journal of Biometrics. Vol. 5, No 3-4. 2013

[Grimaldi and Cummings, 2008] M.Grimaldi, F.Cummings. Speaker identification using instantaneous frequencies. In: IEEE Transactions On Audio, Speech And Language Processing, Vol. 16 No 6. 2008

[Krzesimowski and Ciota, 2010] D.Krzesimowski, Z.Ciota. Signal processing of voice in case of patients after stroke. In: Electrical Review. R. 86 No 11a. 2010

[Krzesimowski, 2012] D.Krzesimowski. Preliminary processing of the human voice recordings. In: Conference Archives PTETiS. Vol. 31. 2012

[Larrouy-Maestri et al, 2012] P.Larrouy-Maestri, Y.Lévêque, D.Schön, A.Giovanni, D.Morsomme. The evaluation of singing voice accuracy: a comparison between subjective and objective methods. In: Journal of Voice. Vol. 27. 2012

[Lombardi et al, 2009] C.P.Lombardi, M.Raffaelli, C.DeCrea, L.D'Alatri, D.Maccora, M.R.Marchese, G.Paludetti, R.Bellantone, Long-term outcome of functional post-thyroidectomy voice and swallowing symptoms. In: Surgery. Vol. 146. 2009

[Reilly et al, 2004] R.B.Reilly, R.Moran, P.Lacy. Voice Pathology Assessment based on a Dialogue System and Speech Analysis. In: Proceedings of American association for artificial intelligence fall symposium on dialog systems for health communication. 2004.

Authors' Information

Krzesimowski Damian – Kielce University of Technology, Department of Applied Computer Science, al. 1000-lecia PP 7, Kielce 25-314, Poland; e-mail: damiank@tu.kielce.pl

Major Fields of Scientific Research: signal processing and numerical methods, mobile systems, automated systems