Research Article Open Access

Yedilkhan Amirgaliyev, Minsoo Hahn, and Timur Mussabayev*

The speech signal segmentation algorithm using pitch synchronous analysis

DOI 10.1515/comp-2017-0001
Received Nov 03, 2016; accepted Feb 07, 2017

Abstract: Parameterization of the speech signal using the algorithms of analysis synchronized with the pitch frequency is discussed. Speech parameterization is performed by the average number of zero transitions function and the signal energy function. Parameterization results are used to segment the speech signal and to isolate the segments with stable spectral characteristics. Segmentation results can be used to generate a digital voice pattern of a person or be applied in the automatic speech recognition. Stages needed for continuous speech segmentation are described.

Keywords: speech signal segmentation; pitch frequency; speech parameterization; signal smoothing; FIR filter

1 Introduction

To cope with the tasks related to automated speech synthesis and recognition, speech signals need to be segmented. The process of speech signal segmentation can be performed on different levels by using different algorithms: subphonemic, phonemic, syllabic, word level, syntagmatic level, *etc* [1]. Here, the accuracy and stability of segmentation results largely affect the efficiency of speech recognition process, as well as the genuineness of synthesized speech.

The effectiveness of segmentation algorithms is evaluated by the following criteria: accuracy of determining the borders of segments, robustness, resistance to ambient noise and operation speed. The main advantage of the automatic segmentation method, as compared to the manual segmentation, is the possibility to achieve high repeatability of segmentation results. This means that the borders between two standard segments in various locations of a signal will be identified in the same way over the entire signal.

In this study, we will discuss a new approach to the problem of speech signal segmentation by applying cofunctional modality, namely, a signal from a human's vocal cords vibrations. In a general sense, here modality refers to the mode of existence of any object, or the course of any phenomenon, or a mode of understanding, assertion about the object, phenomenon, or event [2].

Initially, the authors' task was to develop fully automated efficient algorithms for multilevel segmentation of continuous speech signals, with a view to create phonetically balanced speech databases and their subsequent use in the systems of speech synthesis and recognition. The period of vocal cords vibration was selected as the smallest segment. To perform segmentation of a speech signal in accordance with the frequency of vocal vibrations, we decided to use a laryngophone, since the standard algorithms operating only with the vocal microphone signal fail to provide the required accuracy and needed the subsequent manual adjustments. At the same time, the process of manual adjustment is time-consuming and can also lead to errors.

The major feature of a speech signal (or language, if viewed in a broader sense) is its multi-level nested hierarchical structure. In this sense, a language resembles a Russian nesting doll "Matryoshka". The first basic external level is the level of a speech signal, where its main acoustic properties can be analyzed. Upon the analysis of acoustic properties of a speech signal, we can move to the next level – subphonemic. At this level we can perform the above described process of identifying relatively stable plots of the speech signal with a relatively uniform formant structure, which represent subphonemic segments. At the next level, the resulting sequence of subphonemic segments is grouped into individual phonemes, *i.e.* we can identify higher-level phonemic segments consisting of low-level subphonemic segments. Further transitions in depth of the structure follow the same way: from phonemes to syl-

Yedilkhan Amirgaliyev: Institute of Information and Computing Technologies, Kazakhstan, Almaty; Email: amir_ed@mail.ru
Minsoo Hahn: Korea Advanced Institute of Science and Technology (KAIST), South Korea, Daejeon; Email: mshahn2@kaist.ac.kr
***Corresponding Author: Timur Mussabayev:** Kazakh National University named after Al-Farabi, Kazakhstan, Almaty; Uniline Group LLP; Email: tmusab@yandex.ru

lables, from syllables to words, from words to their various inherent and group characteristics (morphological, lexical, syntactic), from inherent and group characteristics of the words going up to the level of semantic relationships, and from the level of semantic relationships to the formation of a sequence of integral patterns.

In a general sense, the problem of automatic continuous speech recognition can be reduced to the solution of individual tasks of sequential speech signal segmentation at various levels listed above, followed by the classification of speech segments derived at a certain level and transition to the next, higher level. The logical outcome of the recognition process is the formation of a sequence of interrelated patterns, which are the subject of study in the theory of pattern recognition [3]. It appears that high-quality recognition results can be generated only by synthesizing the available pattern information from the results of segmentation and classification produced at the previous levels. It should be noted that the majority of modern systems dedicated to continuous speech recognition are lacking semantic functionality with access to the sequence of correlated patterns or meanings. In most cases, this level is replaced by various statistical algorithms (hidden Markov models, n-gram models, *etc*.). The result is that we currently have inefficient systems of continuous speech recognition, due to the lack of the understanding element – what these systems are attempting to recognize. The same thing happens with man: if he does not understand the meaning of what is being communicated to him, the probability of various errors occurring in his recorded statement increases significantly.

Further improvement of the results of continuous speech recognition will benefit apparently from the development of innovative, more efficient algorithms for multilevel segmentation and classification of speech signal at lower levels, as well as from development and application of the algorithms of semantic analysis at higher levels. An example of the semantic approach to improve recognition results can be offered by the model “Text-Meaning” and “Meaning-Text” [4].

2 General description of algorithms and methods

The process of speech signal segmentation involves several major stages.

The first stage is the generation of a phonetically balanced set of natural language texts of the required volume. The text is generated in automatic mode in such a way as to include the largest possible number of various phonetic combinations, including those rare. To generate a phonetically balanced text, we used our original Kazakh language corpus, as well as the procedure for phonetic transcription, which provides a phonemic representation for any text in the Kazakh language. The algorithm of formation of a phonetically balanced text takes an iterative approach, which ensures that sentences with the greatest number of rare phoneme combinations are included in the set.

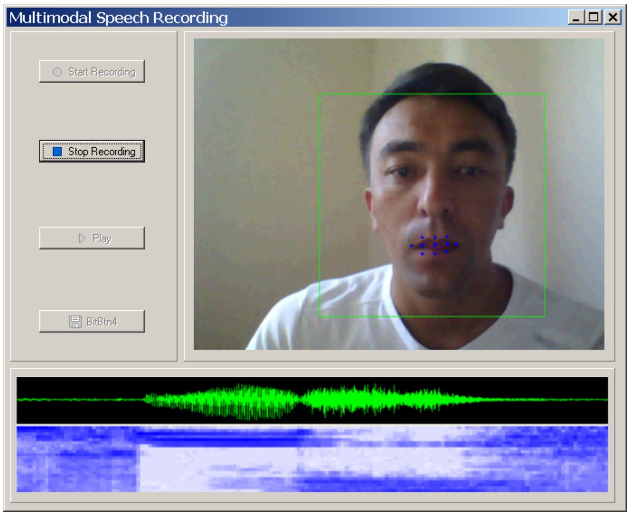
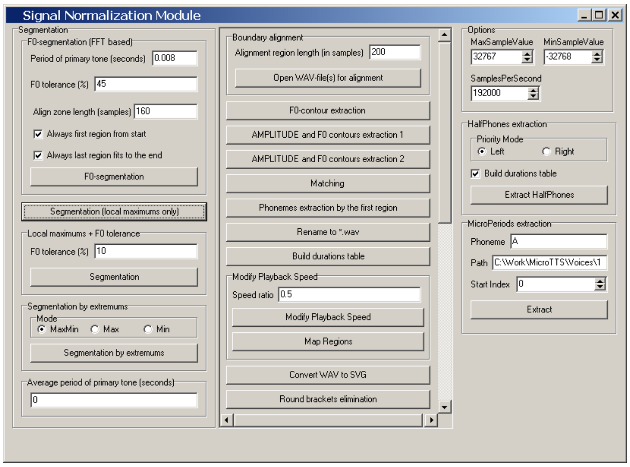


Figure 1: Module interface of the multimodal speech signal recording.

The second stage involves recording speech signal from the text voiced by a speaker and forming his/her personal digital voice portrait. This recording is performed in a multimodal way: a speech signal is recorded with the microphone, a vocal vibrations signal is recorded with the laryngophone and the digital camera is used for recording the video stream that reflects the speaker’s lip movements. Upon the completion of recording process, we receive a triple-stream synchronized signal with each stream saved in a separate file. To form a valid segmented speech signal for each speaker, at least several hours of recording is needed. This process can be divided into several sessions, 45 minutes each, separated with rest breaks. Figure 1 shows the module interface of the multimodal speech signal recording.

The third stage includes the process of speech and laryngophone signals normalization. Principally, normalization implies bandpass signal filtering. In particular, the bandwidth of a speech signal is [154, 22 050] Hz and that of laryngophone signal - [154, 300] Hz. Bandpass filtering allows for eliminating the uninformative component of signals, which is mainly represented by noise. Filtering process is performed with a bandpass FIR filter. This stage also


Signal Normalization Module
Segmentation
F0-segmentation (FFT based)
F0-segmentation
Segmentation (local maximums only)
Local maximums + F0 tolerance
Segmentation
Segmentation by extremums
Segmentation by extremums
Boundary alignment
Open WAV-file(s) for alignment
F0-contour extraction
AMPLITUDE and F0 contours extraction 1
AMPLITUDE and F0 contours extraction 2
Matching
Phonemes extraction by the first region
Rename to *.wav
Build durations table
Modify Playback Speed
Modify Playback Speed
Map Regions
Convert WAV to SVG
Round brackets elimination
Options
HalfPhones extraction
Extract HalfPhones
MicroPeriods extraction
Extract

Period of primary tone (seconds) 0.008

F0 tolerance (%) 45

Align zone length (samples) 160

☑ Always first region from start

☑ Always last region fits to the end

F0 tolerance (%) 10

Mode: ● MaxMin ○ Max ○ Min

Average period of primary tone (seconds) 0

Alignment region length (in samples) 200

Speed ratio 0.5

MaxSampleValue 32767

MinSampleValue -32768

SamplesPerSecond 192000

Priority Mode: ● Left ○ Right

☑ Build durations table

Phoneme A

Path C:\Work\MicroTTS\Voices\1

Start Index 0

Figure 2: Interface part of a software module for normalization and pitch frequency markup for the speech and laryngophone signals.

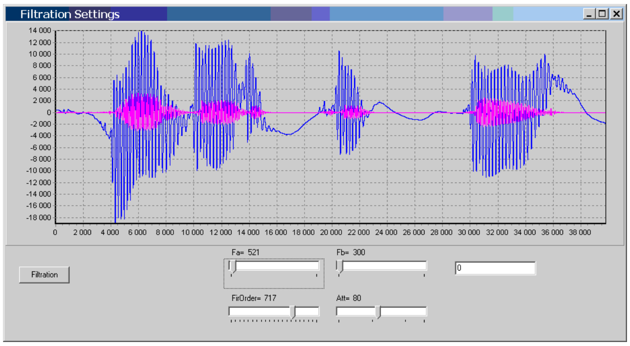
involves the automatic labeling (markup) of speech and laryngophone signals for the pitch frequency by placing corresponding tags on the signal. Figure 2 demonstrates the interface part of a software module for normalization and markup by the pitch frequency for the speech and laryngophone signals. Markup by the pitch frequency is performed by finding local maxima in a smoothed laryngophone signal. Figure 3 demonstrates the interface module of setting filtering parameters using the FIR filter. In Figure 3, the original speech signal is colored blue (on the chart) and the filtering result – pink.

The fourth stage of the segmentation process is the process of parameterization of speech and laryngophone signals using the generated set of task-specific mathematical models and algorithms designed to parameterize various aspects of speech production process. At this stage, the user can select from a list of available algorithms, configure the parameters for a given algorithm and start the process of parameterization. Upon the results of the parameterization, a separate file is formed, where all parameterization results are saved.

Parameterization can be implemented in two modes: synchronously with the pitch frequency and on the basis of the floating frame with a fixed step. Figure 4 shows the interface part of the module for parameterization of speech and laryngophone signals.

The fifth stage is the process of automated segmentation of the speech signal into acoustically homogeneous parts with dimension less than a single word.

Figure 5 demonstrates the interface module of automated speech signal segmentation into acoustically homogeneous speech segments.



Fa= 521

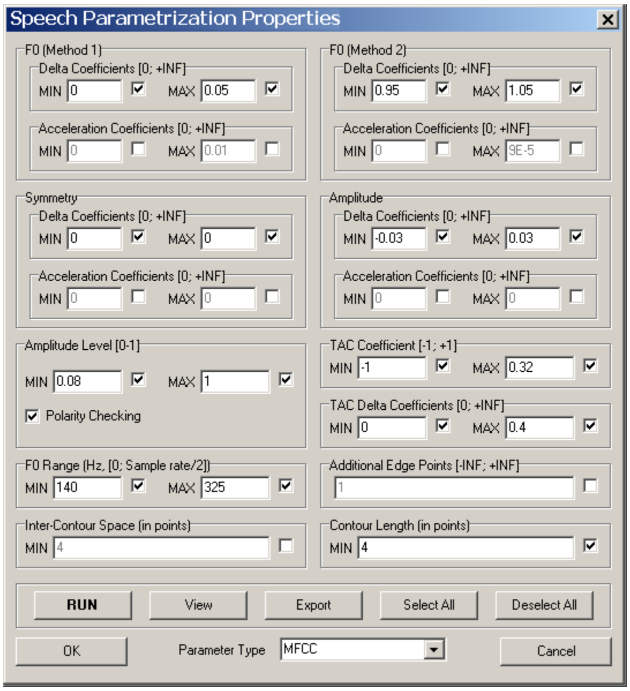
Fb= 300

FirOrder= 717

Att= 80

0

Figure 3: Module interface for setting the filtering parameters using the FIR filter.



F0 (Method 1)

Delta Coefficients [0; +INF]: MIN 0 ☑ MAX 0.05 ☑

Acceleration Coefficients [0; +INF]: MIN 0 ☐ MAX 0.01 ☐

F0 (Method 2)

Delta Coefficients [0; +INF]: MIN 0.95 ☑ MAX 1.05 ☑

Acceleration Coefficients [0; +INF]: MIN 0 ☐ MAX 9E-5 ☐

Symmetry

Delta Coefficients [0; +INF]: MIN 0 ☑ MAX 0 ☑

Acceleration Coefficients [0; +INF]: MIN 0 ☐ MAX 0 ☐

Amplitude

Delta Coefficients [0; +INF]: MIN -0.03 ☑ MAX 0.03 ☑

Acceleration Coefficients [0; +INF]: MIN 0 ☐ MAX 0 ☐

Amplitude Level [0-1]: MIN 0.08 ☑ MAX 1 ☑

☑ Polarity Checking

TAC Coefficient [-1; +1]: MIN -1 ☑ MAX 0.32 ☑

TAC Delta Coefficients [0; +INF]: MIN 0 ☑ MAX 0.4 ☑

F0 Range [Hz, [0; Sample rate/2]]: MIN 140 ☑ MAX 325 ☑

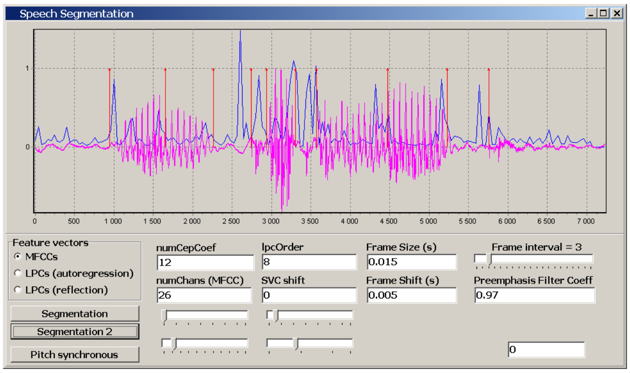
Additional Edge Points [-INF; +INF]: 1 ☐

Inter-Contour Space (in points): MIN 4 ☐

Contour Length (in points): MIN 4 ☑

Parameter Type MFCC

Figure 4: Interface part of the module for parameterization of speech and laryngophone signals.



Feature vectors: ● MFCCs ○ LPCs (autoregression) ○ LPCs (reflection)

numCepCoef 12

lpcOrder 8

Frame Size (s) 0.015

Frame interval = 3

numChans (MFCC) 26

SVC shift 0

Frame Shift (s) 0.005

Preemphasis Filter Coeff 0.97

0

Figure 5: Interface module for automatic speech signal segmentation into acoustically homogeneous speech segments.

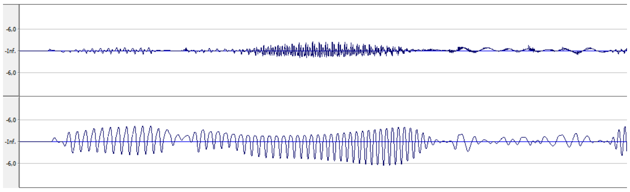


Figure 6: Aggregated speech signal and pitch signal.

The borders of the derived segments are indicated with arrows.

The sixth stage is the process of automatic clustering of speech segments derived in the previous stage. Clustering is performed by using parametric representations of these segments. It is possible to configure the clustering parameters. According to clustering results, every segment is attributed to one or another cluster. Each cluster has its own serial number, as well as its centroid – a segment that features the most typical characteristics within this cluster. Thus, each cluster can be defined as a condition, where the number of cluster of acoustically homogeneous segments is taken as the number of condition. The total set of clusters and their constituent segments, centroids within each cluster and parameters of the segments make individual digital voice pattern. The following algorithms was used for the clustering: k-means, mean shift, DBSCAN [5–7].

3 Receiving a speech signal and its decomposition

At first, the authors recorded a reference speech of female voice type discretized by a frequency of 44 KHz and a bit depth of 16 bits. The reference was recorded using the microphone and the laryngophone receiving the sound directly from vibrations of the vocal cords.

First, we need to synchronize a speech signal with pitch frequency over time. The acoustic correlate of a tone is the so-called pitch frequency (PF) defined as the frequency of vocal cords vibration. From the acoustic point of view, PF is the first harmonic of a speech signal [8]. The height of pitch depends on the frequency of vocal vibrations (the more frequently cords vibrate, the higher the vocal pitch frequency is), which is determined by the tension, length and total weight of the vocal cords [9]. To synchronize the speech and the pitch signals, we can use the technique of simultaneous signal recording with the microphone and the laryngophone. In our case, the pitch frequency sets the period of repetitions of acoustic vibrations, which can be particularly helpful in speech signal classification.

The next step is the decomposition of recorded signals for subsequent processing and applying different filters to every signal. In fact, any speech signal recorded in the working environment can contain a lot of interfering information: various noises, the presence of high and low frequencies with high-frequency oscillations, *etc*. It should be noted that the base pitch frequency is characteristic of any speaker and is determined by individualities of the larynx structure. The average for a male voice ranges from 80 to 210 Hz, for female — from 150 to 320 Hz [10]. Therefore, frequencies located above or below this range are considered to be redundant for the task of PF isolating. The presence of this redundant information can distort the classification of the original speech signal. This information can be eliminated by using a filter with finite impulse response (non-recursive filter, FIR filter).

4 Smoothing speech signals

If we need to view the approximate dynamics of a signal, it should be made “smooth” by removing minor included noise and eliminating small fluctuations in the signal. Such treatment is called signal smoothing. This type of processing can involve an operation called the moving average, which represents the elementary form of the FIR filter. That is, we take a certain area before and after the considered point and, given the numerical values of the measurements included in this area, calculate the average value.

This is determined by the ratio:

$$g_i = \frac{1}{K+1}\left(f_{i-K} + f_{i-K+1} + \ldots + f_i\right), \tag{1}$$

where N measurement points of a digital signal $\{f_1, f_2, \ldots, f_N\}$: f_i; $i = 1, 2, 3, \ldots, N$.

To find the moving average around the considered point i, we take the arithmetic average from K preceding and subsequent points including the point i. Besides, pay attention to the fact that the value of smoothing can not be calculated at the first and the last points i on the x-axis. The area, where the calculation is possible, is defined as follows:

$$i = 1 + K, 2 + K, \ldots, N - K. \tag{2}$$

Using the summation sign, the ratio (1) is written as:

$$g_i = \frac{1}{K+1}\sum_{j=-K}^{0} f_{i+j}\,(i = 1 + K, 2 + K, \ldots, N) \tag{3}$$

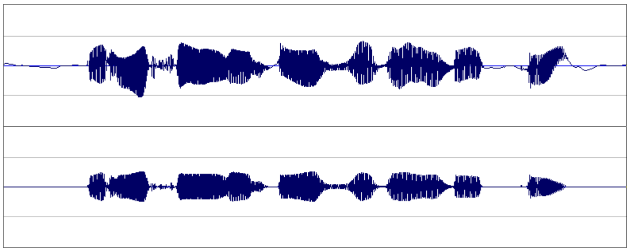


Figure 7: Original pitch signal (top), the same signal processed by the FIR filter (bottom).

When we perform smoothing, the weight of each point should be evaluated according to its importance. This can be written by the following expression:

$$g_i = \sum_{j=-K}^{K} w_i f_{i+j} \, (i = 1+K, 2+K, 3+K, \ldots, N-K) \qquad (4)$$

In order to avoid distortion of the averaged function value, we assume the following condition:

$$\sum_{j=K}^{K} w_i = 1 \qquad (5)$$

where, w_i is a function that attributes weights to the points. Gaussian distribution is generally used as the weight function [11]. Figure 7 demonstrates the results of processing the original pitch signal with the FIR filter.

It should be mentioned that the use of the FIR filter for smoothing the received signals was selected for a good reason. Actually, there are two basic types of digital filters: those with finite impulse response (FIR filters) and with infinite impulse response (IIR filters). When it comes to speech signal processing, this classification applies to the pulse characteristics of filters. By varying the weights of coefficients and the number of units in the FIR filter, we can implement almost any frequency response [12]. FIR filters have a number of useful properties:

- FIR filters are robust;
- Implementation of FIR filters does not require feedback;
- The phase of FIR-filters can be made linear [13]

When the original audio signal and the pitch frequency are smoothed, the speech signal should be preprocessed, in order to derive a set of spectral vectors that characterize the signal. Why is it done? Since speech is a non-stationary process (*i.e.*, its spectral characteristics are relatively unsteady), it is traditionally analyzed in short sections (10 — 30 ms), where the spectral correlation characteristics remain roughly constant. The fundamental assumption taken by the current speech recognition software implies that the speech signal is considered stationary and the size of a segment is taken randomly and of a single dimension. The typical size of one segment is 25 ms.

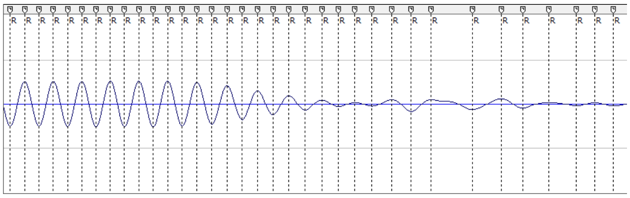


Figure 8: Semi segments of the pitch signal isolated by sampling frequency.

The author of this article has suggested forming these segments by identification of positive extreme values in the pitch signal and drawing the segment borders through them. After that the selected segments should be synchronized with the speech signal over time. This technique can provide the increased accuracy of speech signal parameterization and selection of the optimum segment size. Finding extreme values is a simple mathematical task and is not reviewed in this article. We will mention only that the search for extreme values was carried out using simple enumerative search aided by the comparative nearest neighbor method on the y-axis values.

Using the above described method, the optimum segments of the speech signal have been formed with an average time value of 9 ms, which depends on the sampling frequency of a pitch signal. The result can be seen in Figure 8.

5 Parameterization of a speech signal

The largest part of speech processing methods is based on the assumption that speech signal properties are changed slowly over time. This assumption leads to methods of short-term analysis, where the speech signal segments are isolated and processed as if they were short sections of individual sounds with different characteristics. This procedure is iterated as often as required. The result of processing each segment is a number or a set of numbers [14]. To classify the speech signal into vowels and consonants, methods of processing in the time domain can be applied. Here, domain refers to the previously isolated speech signal segments synchronized with the sampling frequency of the fundamental pitch. Short-term functions of the average number of zero transitions and the signal energy were used to process the temporal segment of a speech signal. These methods represent the most important parameters in voice classification and are often used in automatic speech recognition systems.

6 Function of the average number of zero transitions

When processing signals in discrete time, we suggest that if two consecutive readings have different signs, there is a zero transition. The frequency of zero occurrences in a signal can provide an elementary characteristic of its spectral properties. Let us consider the method of its calculation. We define the average number of zero transitions:

$$Z_n = \sum_{m=-\infty}^{\infty} \left| sgn\left[x(m)\right] - sgn\left[x(m-1)\right] \right| w(n-m), \tag{6}$$

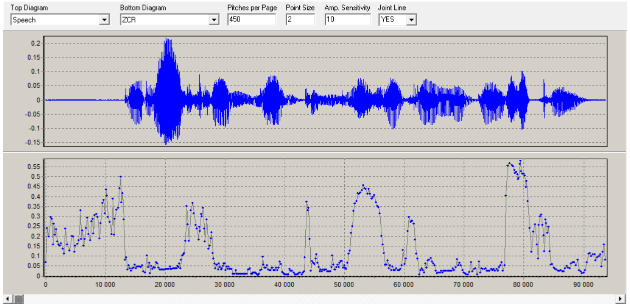
where,

$$sgn\left[x(n)\right] = \begin{cases} 1, x(n) \geq 0, \\ -1, x(n) < 0 \end{cases} \tag{7}$$

and

$$w(n) = \begin{cases} \frac{1}{2N}, 0n \leq N-1, \\ 0, elsewise. \end{cases} \tag{8}$$

The model of speech production suggests that the energy of voiced segments of a speech signal is aggregated at frequencies below 3 kHz, due to the decreasing spectrum of the excitation signal, whereas for unvoiced segments, most of the energy lies in the high frequency range. Since high frequencies lead to the greater number of zero transitions and low frequencies – to the lesser number, there is a strong link between the number of zero transitions and energy distribution over frequencies [12]. It should be noted that the number of zero transitions differs also for the time segments that characterize vowels and consonants. Consonants are sounds, in the pronunciation of which the air encounters an obstacle. Slit and occlusion are the two basic ways of forming consonants. The type of obstruction determines the nature of a consonant. The presence of noise is another distinctive feature of consonants. Therefore, consonants inherently belong to high frequencies with a greater number of zero transitions. When pronouncing vowels, the air passes freely through the mouth cavity without encountering obstacles, so vowels can be attributed to low frequencies with a smaller number of zero transitions [15]. In Figure 9, we can observe that vowels have "smooth" and relatively low amplitudes on the graph that implements the function of the average number of zero transitions. On the opposite, consonants have "uneven" and relatively high oscillatory amplitude.



Top Diagram: Speech
Bottom Diagram: ZCR
Pitches per Page: 450
Point Size: 2
Amp. Sensitivity: 10
Joint Line: YES

Figure 9: Speech signal parameterized with a function of the average number of zero transitions.

7 Short-term signal energy

As noted above, the amplitude of a voice signal varies significantly over time. In particular, the amplitude of unvoiced segments of a speech signal is far smaller than that of voiced segments. Such variations in the amplitude are well described by the function of short-term signal energy. In general, the function of energy can be determined as

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)\, w(n-m)\right]^2, \tag{9}$$

This expression can be rewritten as

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)\, h(n-m), \tag{10}$$

where,

$$h(n) = w^2(n). \tag{11}$$

In this case, $x^2(n)$ signal is filtered by a linear system with impulse response $h(n)$.

To describe rapid variations in the amplitude, it is desirable to have a narrow window (short impulse response); however, a too narrow window can lead to improper averaging and, hence, to insufficient smoothing by the energy function [14]. Therefore, the size of a speech segment to be processed by this function was not selected at random but synchronized with the pitch frequency. Figure 10 demonstrates the result of using the short-term signal energy function. In contrast with the function of the average number of zero transitions, vowels displayed on the chart of the signal energy function will have high amplitudes.

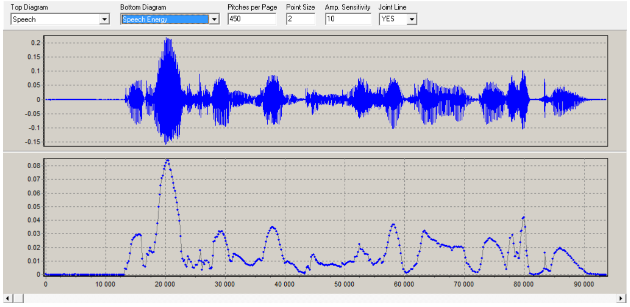

Top Diagram: Speech
Bottom Diagram: Speech Energy
Pitches per Page: 450
Point Size: 2
Amp. Sensitivity: 10
Joint Line: YES

Figure 10: Speech signal parameterized with the energy function.

8 Segmentation of a speech signal

The speech signal consists of quasi-stationary plots corresponding to the phonic and sibilant phonemes interspersed with plots of relatively rapid changes in the spectral characteristics of a signal (interphoneme transitions, obstruent and occlusive phonemes, speech-pause intraword transitions). An important role in the speech analysis within the fixed plots is played by the spectral signal characteristics determined by the transfer function of the vocal tract that vibrates in the process of articulation. It can be stated that a speech signal is characterized by nonlinear fluctuations of various scale. If we consider phonemes as structural units of the speech, the task of segmentation is reduced to finding interphoneme transitions [16].

To define the borders of quasi-stationary segments, we can use the results of previous parameterization. As demonstrated in the graphs of short-term functions of the average number of zero transitions and the signal energy, the difference in the amplitude height and its "smoothness" can be noticed. The borders are determined by the method of iterating amplitude values with the subsequent finding of the maximum difference value between two nearest amplitude values on the y-axis. If this differential value is found, we can assume that there is a rapid change in the spectral characteristics of a speech signal. It can be noted as well that smoothing the parameterization results by a filter with finite impulse response prior to the start of speech signal segmenting can eliminate the unnecessary minor fluctuations in the signal and improve the accuracy of segmentation. Figure 11 presents a graph of the segmented speech signal based on the parameterization results of the short-term signal energy function.

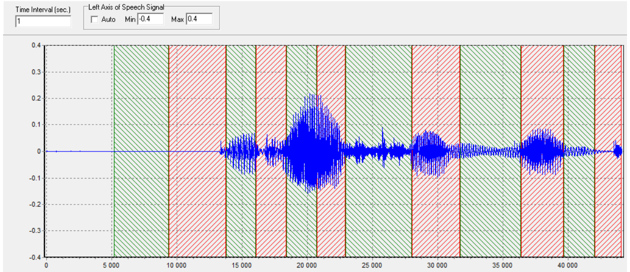

Time Interval (sec.): 1
Left Axis of Speech Signal: □ Auto Min -0.4 Max 0.4

Figure 11: Segmented speech signal.

9 Conclusion

According to the results produced in this study, the use of pitch frequency (PF) received from the laryngophone and synchronized with the speech signal over time can be helpful in calculating the optimum segment size with respect to time, as well as in implementing the automated parameterization and segmentation of a speech signal including various algorithms. In the future, the results of parameterization and segmentation of a speech signal can aid in a more accurate classification of the original speech signal into vowel and consonant phonemes by applying cluster analysis.

Acknowledgement: This work is supported by the Science Committee of RK, under grant 0115PK00552 and grant 0115PK02741

References

[1] Linguistic encyclopedic dictionary. Article "Segmentation" 1990, http://tapemark.narod.ru/les/436b.html. (in Russian)

[2] Averintsev S.S., Arab-Ogly E.A., Ilyichev L.F., et al., Philosophical encyclopedic dictionary, 2nd ed., M.: Sov. encyclopedia, 1989 (in Russian)

[3] Vapnik V.N., Chervonenkis A.Ya., Theory of pattern recognition, Moscow, 1974 (in Russian)

[4] Glushkov V.M., Amosov N.M., Artemenko A. I., Encyclopedia of Cybernetics. Volume 2, K.: Chief editorial board of Ukrainian Soviet encyclopedia, 1974, 46–48 (in Russian)

[5] Jain A.K., Murty M.N., Flynn P.J., Data Clustering: A Review, ACM Computing Surveys, 1999, 31(3) 265–323

[6] Cheng Y., Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 1995, 17(7) 790–799

[7] Sharma L., Ramya K., A Review on Density based Clustering Algorithms for Very Large Datasets, IJETAE, 2013, 3(12) 398–403

[8] Kodzasov S.V., Krivnova O.F., General phonetics, RSUH, Moscow, 2001, 106 (in Russian)

[9] Kedrova G.E., Potapov V.V., Egorov A.M., Emelianova E.B., Physical characteristics of speech sounds, 2002, http://fonetica.philol.msu.ru/nn/n15.htm. (in Russian)

[10] Ashby M., Maidment J., Introducing Phonetic Science, Cambridge University Press, 2005

[11] Yukio Sato, Signal processing. First view, Dodeka, 2002, 29-32

[12] Kester W., Digital signal processing, Chapter 6, 2010

[13] Shishkov A. N., Digital signal processors, 2011, http://frela-mk.narod.ru/olderfiles/1/COS_i_CSP.pdf (in Russian)

[14] Rabiner L.R., Schafer R.V., Digital signal processing, M.: Radio and communications, 1981, 112–121 (in Russian)

[15] Phonetics, http://russkiy-na-5.ru/articles/157, (Accessed: 17.06.2016), (in Russian)

[16] Vishnyakova O.A., Lavrov D.N., Automatic segmentation of the speech signal based on the discrete wavelet transform. Mathematical structures and modeling, 2011, 23, 43–48 (in Russian)