A METHOD OF INCREASING THE ACCURACY OF SEGMENTATION OF A SPEECH SIGNAL BASED ON ITS FRACTAL CHARACTERISTICS

Yana Bielozorova

Abstract: *The paper considers the analysis of approaches to the segmentation of the speech signal into vocalized and not vocalized fragments. The necessity of improving the accuracy of segmentation due to a more accurate description of the process of speech signal representation is shown. A rational method for calculating the fractal dimension to increase the accuracy of the speech signal segmentation process has been determined.*

Keywords: *voice signal, Morlet wavelet, segmentation, special structures, fractal dimension, speaker identification.*

DOI: *https://doi.org/10.54521/ijita29-02-p03*

Introduction

Segmentation of the speech signal can be defined as the process of finding boundaries (with a specific characteristic) in a conversation between words, syllables or phonemes [Al-Mamie et al., 2009, Makowski et al., 2014].

The main purpose of speech signal segmentation is to serve other speech analysis problems, such as speech signal synthesis, training data for speech signal recognition, identification of vocalized fragments for person identification or for the production and labeling of speech databases. With this, speech signal segmentation is an important subtask for various fields of speech analysis.

Related works

The traditional approach to solving this problem consists in manual segmentation of the speech signal, which is most often performed by specialized phoneticians. However, this method is based on listening to fragments and creating a judgment due to visual images of various characteristics of the speech signal, which makes it quite time-consuming, but highly accurate compared to the methods of automatic segmentation of the speech signal [Chefir et al., 2001].

Speech recognition systems require the decomposition of the speech signal into some basic units such as words, phonemes or syllables. The word is the most natural unit of segmentation, it is more likely to carry the characteristics of a person in contrast to phonemes or syllables. That is why, identification of a person by listening to individual phonemes or syllables has not found opportunities for use at the present time [Sharma et al., 1996].

Phonemes are the smallest segmental unit of sound used to form meaning. The same phoneme in different words has a different meaning. There is an overgeneralization of phonemes. So the mixture of phoneme and word gives rise to the next level of the basic unit of language, which is called syllables [Van Hemert, 1991].

The realization of a phoneme is strongly influenced by neighboring phonemes. Phonemes are very context dependent. Consequently, the acoustic variability of basic phonetic units due to context is extremely large and poorly understood in many languages [Lee et al., 2003].

Strategies for automatic speech signal segmentation can be grouped by different perspectives, but one common classification is the division into blind and self-segmentation methods.

1. Blind segmentation. This type of segmentation does not have prior knowledge of linguistic characteristics (spelling, full phonemic notation of a character or its fragment). In order, to solve problems of this kind, methods are used that are based on static analysis and do not require complete knowledge of the audio material. The method of blind segmentation is carried out in 2 stages. The first stage depends entirely on the acoustic properties of the voice signal, while the second stage (or, as it is also called, bottom-up signal processing) focuses on determining the parameters of voice perception, often using MFCC, LP or MFP [SaiJayram et al., 2002].

2. Self-segmentation. This type of segmentation refers to the so-called auxiliary methods. The idea of the method is that during operation it uses only some part of the external data of the voice signal to divide the signal into certain segments. Self-segmentation can include recognition with dynamic time transformation (DTW), or artificial neural networks (ANN), or with hidden Markov models (HMM) [Tanqueiro, 2017], [Toledano et al., 2003], [Mporas et al., 2008], [Go´mez et al., 2011], [Siniscalchi et al., 2007]. The method of self-segmentation by hidden Markov models (HMM) has been widely used in voice recognition. This method has successfully

supplanted other techniques due to the low computational complexity and high recognition speed.

Table 1. Existing methods of speech signal segmentation

Researchers	The essence of the segmentation method	Segmentation accuracy
The use of wavelet - analysis		
B. Zio´łko, S. Manandhar, R. C. Wilson, and M. Zio´ łko [Zio´łko et al., 2006].	The speech signal was segmented due to the selection of phonemes	High efficiency
S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Ua-trongjit, and K. Likit-Anurucks [Ratsameewichai et al., 2002].	Divided the speech signal into components (low-frequency and high-frequency), and further worked within the limits of words or phonemes	Accuracy of 96.0%, 89.9%, 92.7% and 98.9% for initial consonants, vowels, final consonants and silences respectively

M. Tolba, T. Nazmy, A. Abdelhamid, and M. Gadallah [Tolba et al., 2005].	Segmentation is focused on finding boundaries between consonant and vowel parts of the speech signal	Segmentation accuracy is about 88.3%
Artificial neural nets		
Y. Suh and Y. Lee [Suh et al., 1996].	Proposed the technique of using the phoneme segmentation method using a multilayer perceptron	84% accuracy was obtained for 5 ms and 87% accuracy for 15 ms speech signal
Black area blocking method		
M.M. Rahman, F. Khatun, and M. A.-A. Bhuiyan [Rahman et al., 2015].	Proposed an automatic segmentation method with a dynamic threshold	Average accuracy 90.58%
The method of changing short-term energy		

E. A. Kaur and E. T. Singh [Kaur et al., 2010].	Used short-term energy to segment the speech signal into pauses, words and syllables	The proposed method was implemented and analyzed for speech signals of different languages
Hybrid method of speech signal segmentation		
M. Kalamani, S. Valarmathy, and S. Anitha [Kalamani et al., 2017].	Segmented the speech signal by detecting speech boundaries (using the method of threshold values)	Segmentation accuracy 95.33%
Markov models		
P. Bansal, A. Pradhan, A. Goyal, A. Sharma, and M. Arora [Bansal et al., 2014].	proposed phonetic segmentation and speech analysis at the phonetic level	Segmentation accuracy 78,14%
J. Dines, S. Sridharan, and M. Moody [Dines et al., 2002].	They proposed segmentation based on learning strategies	Segmentation accuracy 95,4%

A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman [Stolcke et al., 2014].	Segmentation based on statistical models for boundary correction (due to additional information about the structure of the speech signal)	Segmentation accuracy from 93.9% to 96.8% (intervals of 20 ms were analyzed)

Taking into account the performed review of speech signal segmentation methods, it can be concluded that there are a number of limitations of these methods and "floating" characteristics of segmentation accuracy, which poses the task of developing a method of its segmentation based on speech signal analysis methods.

Materials and methods

According to existing studies, the process of forming a speech signal is a process which constant oscillation of individual parts of the speech tract. Given the fact that these parts practically do not change during speech, self-similar or multifractal structures are formed in the speech signal. Thus, at the beginning of the speech, we will observe sharp changes describing the fractality of the speech signal. To develop a speech signal segmentation method, it is necessary to evaluate the variability of these characteristics during the transition from speech to pauses and make a decision about the possibility of their use. We will use the fractal dimension as a fractal characteristic that allows us to describe the characteristics of a speech signal of this type.

There are many methods of calculating fractal dimension. For use in linguistic information identification tasks, it is necessary to evaluate these methods and choose a rational approach for their calculation. Let's consider each of them in more detail.

The Yard Stick Method. First, a fixed window size is selected r, which divides the speech signal into fragments covering its profile. Thus, the signal will be represented as $N(r)$. When calculating the fractal dimension by this method for each fragment n the length of the curve is calculated L_i, and everyone r_i is matched with the obtained length. The obtained series (r_i, L_i) is built on a logarithmic scale for both coordinates, and the line is fitted to the graph using the method of least squares. Based on the resulting graph, the fractal dimension is calculated, as $D = 1 - \alpha$, where α - regression coefficient.

The Hausdorff method. The method is based on covering the signal profile with square grids. When the grid width r the number of grid elements changes $N(r)$ will change too. Relations between $N(r)$ and r are presented as: $N(r) = kr^{-D}$. Given that the width of the square grid will be $r_1, r_2, r_3, \dots, r_k$, the number of elements will be accordingly $N(r_1), N(r_2), N(r_3), \dots, N(r)$. Calculation of the fractal dimension according to this method is performed as follows. A graph of the obtained series is constructed $(r_i, N(r_i))$ on a logarithmic scale for both coordinates, and a linear regression method is used for data analysis. As a result, we get the regression coefficient α. The fractal dimension is calculated as $D = -\alpha$.

The variational method. For calculation using this method, the profile is covered with rectangles with width r. For each rectangle, a reference point is selected and the deviation is calculated H_i between the highest and the lowest position. If the width r very small, H_i approximately equal to the length of the curve $V(r)$. Thus, the measure is equal to $V(r) =$

$\sum \frac{rH_i}{r^2} = \sum H_i/r$. A graph of the obtained series is constructed $(r_i, V(r_i))$ on a logarithmic scale for both coordinates and a linear fit is performed. The fractal dimension is calculated as $D = 2 - \alpha$, where α - the slope of the obtained function.

The method of structural function. The structure function method is also called the augmentation approach. The profile is considered as a sequence of height function $z(x)$. For any two points from a distance r in sequence $z(x)$ the function of the structure is determined $S(r)$, which is the arithmetic mean of the square of the height difference. Relationships between sequences $z(x)$ and structure function $S(r)$: $S(r) = E[z(x + r) -$ *zx2=cr4−2D*, where r called the interval scale. When calculating the fractal dimension, different scales are chosen r, and function values are obtained $S(r)$. and function values are obtained $(r_i, S(r_i))$ on a logarithmic scale for both coordinates. The fractal dimension is calculated as $D = 2 - \alpha$ /2, where α - the slope of the obtained function.

The root mean square method. The basic principle is similar to the structural function method. The study showed that the scale function ratio $z(x)$ the following formula corresponds to fractal characteristics

$$z(x) - z(x_0) = \zeta |x - x_0|^{2-D} \# \qquad (1)$$

Let $x_0 = 0$ to $z(0) = 0$, We can then calculate the variance or correlation moment of the function sequence $z(x)$

$$S(r) = D(r)^{1/2} = cr^{2-D} \# \qquad (2)$$

where r is an interval scale, and $r = x - x_0 = x$. Equation (2) shows that the relationship between the correlation moment $S(r)$ and interval scale r is the power exponent, and the power is a function of the fractal dimension. Fractal dimensionality is calculated as follows for each interval scale $r_i(i = 1,2,\dots,n)$. The dispersion is calculated $S(r)$. A graph of the obtained series is constructed $(r_i, S(r_i))$ on a logarithmic scale for both coordinates. The fractal dimension is calculated as $D = 2 - \alpha$, where α is the slope of the line.

The Hurst method. For the height function $z(x)$, at a given scale r, the average value is calculated in the form: $\overline{z_r} = \frac{1}{r}\sum_{x=0}^{r} z(x)$, cumulative deviation: $z(x,r) = \sum_{x=0}^{r}[z(x) - \overline{z_r}]$, maximum difference $R(x) = \underset{0\le x\le r}{max} z(x,r) - \underset{0\le x\le r}{min} z(x,r)$, and standard deviation $S(r) = \sqrt{\frac{1}{r}\sum_{x=0}^{r}[z(x) - \overline{z_r}]^2}$. Hurst's research found that a statistical law R/S is equal to

$$R/S = cr^H\# \tag{3}$$

where, c it is constant, H its Hurst index. Taking the logarithm on both sides of equation (3), we get

$$\ln(R/S) = \ln(c) + H * \ln(r)\# \tag{4}$$

After determining the maximum difference $R(r)$ and standard deviation $S(r)$ with different scale r, on the basis of equation (4), the Hurst index H can be calculated using the method of least squares. Calculation of fractal dimension has the form: $D = 2 - H$.

The Higuchi method. The original formulation of the calculation method is explained by Higuchi in [Sapozhkov, 1963]. Higuchi fractal dimension (HFD) with X is calculated as follows: For each $r \in \{1, \dots, r_{max}\}$ and $m \in \{1, \dots, r\}$ the length is determined $L_m(r)$ on

$$L_m(r) = \frac{N-1}{\left\lceil \frac{N-m}{k} \right\rceil r^2} \sum_{i=1}^{\left\lfloor \frac{N-m}{r} \right\rceil} \left| X_N(m+ir) - X_N\big(m+(i-1)\big)r \right| \# \qquad (5)$$

Length $L(r)$ is determined by the average value r length $L(r) = \frac{1}{r}\sum_{m=1}^{r} L_m(r)$. The slope of the best-fit linear function through the data points $\left\{\left(\log\frac{1}{r}, \log L(r)\right)\right\}$ is defined as the Higuchi fractal dimension.

Experiments

Let's determine which of the fractal dimension calculation methods is the best for estimating the fractal dimension of the speech signal. 2 test speech signals, presented in Figure 1, were used for preliminary adjustment of the methods. Moreover, Fig. 1a) is a vocalized fragment, and Fig. 1b) is a not vocalized fragment or pause.

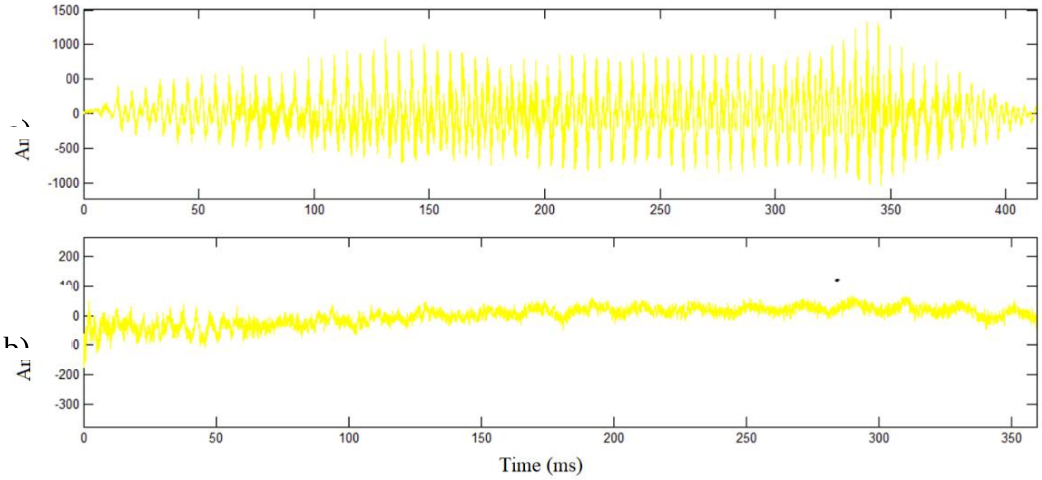

Figure 1: Test speech signals a) vocalized fragment, b) not vocalized fragment

For the calculation of each of the options, a rational value of the indicator was selected r – it is chosen in such a way as to ensure the level of significance of the correlation coefficient 0.03. This was achieved by changing the indicator r with a step of 0.1 and calculating the value of the fractal dimension of the corresponding method $D(r)$, taking into account the coverage of the speech signal curve according to the established indicator of the correlation coefficient.

The results of calculating the fractal dimension by each of the methods presented above are given in Table 2.

Table 2. Comparison of fractal dimension calculation methods for speech signal (shown in Figure 1)

The method of calculating the fractal dimension	Vocalized fragment		Not vocalized fragment	
	Fractal dimension	Correlation coefficient	Fractal dimension	Correlation coefficient
The Higuchi method	1.75	0.995	1.221	0.981
The Hurst method	1.72	0.968	1.181	0.979
The root mean square method	1.657	0.979	1.119	0.982
The method of structural function	1.701	0.997	1.199	0.998
The Yard Stick Method	1.422	0.993	1.011	0.992
The variational method	1.353	0.996	1.154	0.999
The Hausdorff method	1.68	0.998	1.197	0.995

Previous studies for test fragments of speech and pauses showed poor correlation indicators when calculating the fractal dimension using the Higuchi, Hurst and root mean square methods, the highest correlation indicators are provided when using the structural function method and the Hausdorff method.

The next stage of the research was to analyze the performance of the methods for a larger set of test data with settings according to the test fragments. 300 prepared vocalized and non-vocalized fragments were used in the study. The results of the study are presented in figure 2.

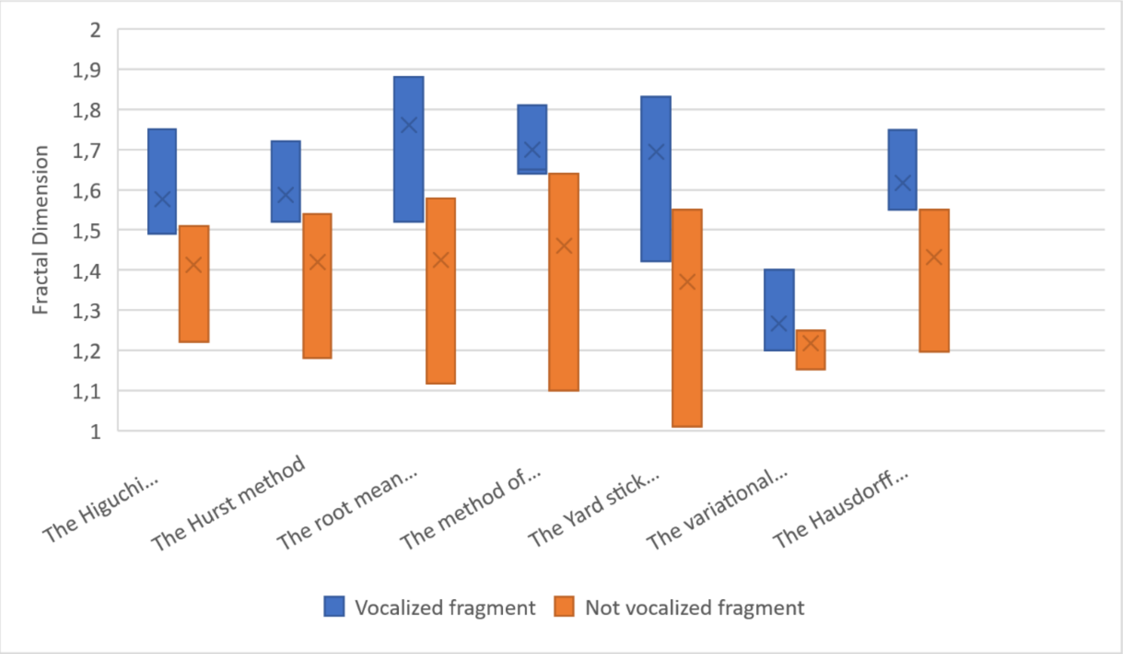

Figure 2: Comparison of fractal dimensionality calculation methods for fragments of the speech signal

Based on the analysis of the results, it can be concluded that the Hausdorff and structural function methods are the most suitable for use in speech signal segmentation, because they provide a smaller overlap of the fractal dimension ranges for vocalized and non-vocalized fragments, the variational method cannot be considered in the tasks of segmentation in communication due to the small range of variation of the fractal dimension under the given conditions, the "Yard Stick" and mean square methods have a wide range of overlapping values of the fractal dimension, which will lead to a decrease in the accuracy of segmentation when using these methods. Higuchi and R/S analysis methods can be satisfactorily used for speech signal segmentation, but their quality will be significantly lower than that of Hausdorff and structural function methods.

In summary, the Hausdorff and structure function methods are more suitable for computing the fractal dimension in the speech signal segmentation task. Given the higher computational complexity of the structural function method for speech signal segmentation, we choose the Hausdorff fractal dimensionality calculation method.

Results

To study speech signals and approximate the graphs of time series of sound wave amplitudes by certain sums (cell-type breakdowns), we turn to the Hausdorff dimension. The Hausdorff fractal dimension Dx is defined as follows

$$S(p){\sim}p^{2-Dx}\,\mathrm{at}\,p \to 0\# \tag{6}$$

where $S(p)$ - the full area of the complex, with the scale of division p.

From a practical point of view, when trying to calculate Dx on the basis of (6), a number of problems arise. This is due to the fact that real time series will always have a minimum scale p_0 and at the same time, the transition to the asymptotic representation in (6) is quite slow.

Unlike the conventional time series considered in most signal processing tasks, speech signals have significant differences. One of the main differences is that the amplitude of a sound wave is well described as a sum of harmonic oscillations (both from a physical and mathematical point of view). This approach allows you to significantly reduce the number of readings, which are necessary to conduct a real analysis of fractal dimension.

The next, important point is the possibility of evaluations in a number of tasks of language technologies of the order of the minimum fractal scale. Based on previously conducted research, it is known that all important speech information is contained in a certain frequency range, namely up to 4500 Hz, thanks to which it is possible to perform a qualitative assessment of minimally rational fractal scales of speech fragments.

As an example, consider a speech signal with a sampling frequency of F_s and a bit rate of r bits. This means that F_s is in the range of 8000 Hz to 44100 Hz and r is in the range of 8 to 24 bits. Therefore, in order to systematically cover the graph of the investigated sound wave, it is necessary to use some minimum values similar to the size of the rectangle $a * b$. Where a will be defined as the minimum possible change in the amplitude of the sound wave for a specific bit rate of the sound file. As an example, for an 8-bit sound file, $a = 2/256$. This means that all sound files are converted to wav. format, which means that the amplitude value is considered in the range from -1 to +1 and is represented as a floating-point number. For the general case

$$a = 2/2^r\# \tag{7}$$

The minimum value of the side of the rectangle for the time axis b_0 will be equal to $1/Fs$. And since the size is not essential for this case, we will take the minimum size $b_0 = 1$. We use the following method of calculating the fractal dimension according to Hausdorff: each time window of the speech signal will be represented as a set of rectangles of size $a \times b$ covering the graphic representation of speech signal. Let the representation scale :

$$p = k \cdot b\# \tag{8}$$

where $k = 1,2,3, \ldots$. is the representation scale factor.

It is known that the calculation of the fractal dimension according to Hausdorff is performed as

$$D = 2 - \lim\left[\ln\left(N(p)\right)/\ln(p)\right]\# \tag{9}$$

where $\ln\left(N(p)\right)$ is the natural logarithm of the scale-dependent representation of the number of rectangles $N(p)$, which include at least one value of the amplitude of the speech signal, $\ln(p)$ is the natural logarithm of the scale of the representation.

We determine the fractal dimension D on the basis of [Soloviov, Bielozorova, 2013, Zybin, Bielozorova, 2020]

$$D = 2 - Dx\# \tag{10}$$

Depending on the sampling frequency, this time interval corresponds to the number of counts $N = F_s * 2/100$.

Let's introduce the minimum fractal scale $k \geq 3$. After numerous studies of different fragments on different audio files, the value of the fractal dimension was estimated, which showed a large variability of the value of Dx at minimum values of $k \geq 2$. Along with this, the value of the fractal dimensionality estimate will change slightly within some time intervals of 20ms.

For a specific realization of the time window of the pause of the audio file, we will change the scale. To do this, we will plot the graph of the dependence $\ln\left(N(p)\right) = f(\ln(p))$. The next step, after constructing the graph, is to approximate the first points of the graph using a linear relationship

$$f = c * \ln(p) + c_0\# \tag{11}$$

where c, c_0 are approximation coefficients.

Previous studies have shown the need to approximate the first few points of the speech signal, so the size of the approximation of 10 points was adopted, which will be adjusted during a more detailed study.

Within the framework of this method, the value is equal to Dx

$$Dx \ = abs(c)\# \quad (12)$$

It should also be noted that when calculating the fractal dimension based on the method described above, using the minimum fractal scale $k = 1$, the values of the fractal dimension differ significantly. So, in the proposed method of estimating the fractal dimension based on Hausdorff, its calculated value does not correspond to the definition of Hausdorff dimension. But, after numerous studies of sound files based on a modified approach, the effectiveness of using parameter (11) in language technology tasks was proven.

The calculation of the fractal dimension made it possible to construct figure 3, which shows the value of D - the fractal dimension for a fragment of a speech signal for the entire time window.

An overview of the graph shows a sufficiently high level of tracking the fractal dimension of the change in the amplitude values of the speech signal, which can be used for segmentation of the speech signal.

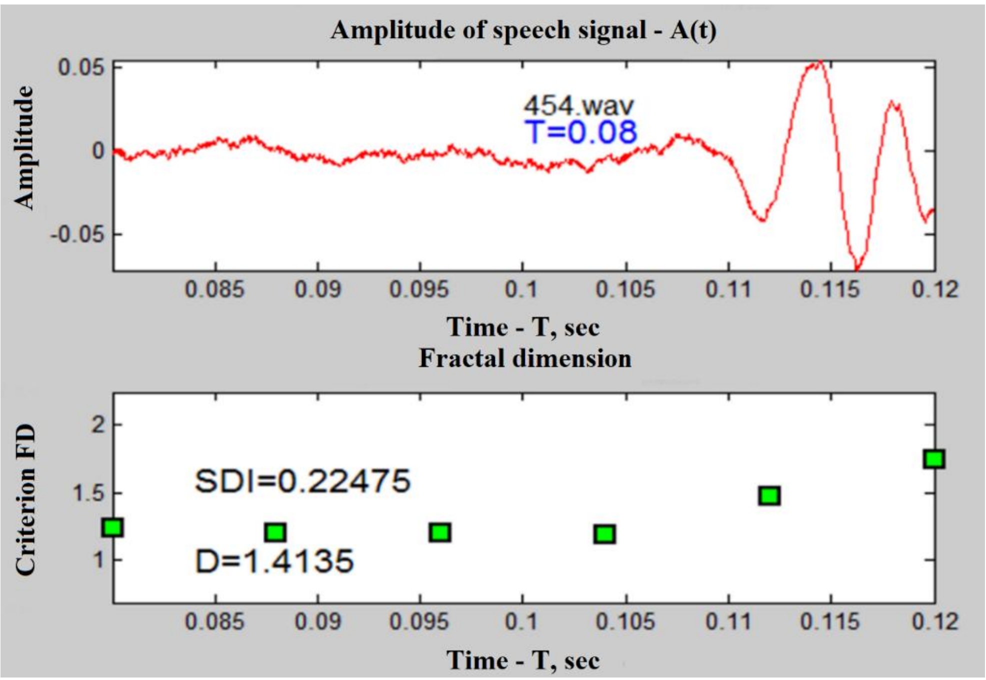

Figure 3: Fractal dimension during the transition from pause to speech

On the basis of the performed research, the following method of speech signal segmentation using fractal dimension is proposed.

1. division into time frames;
2. preliminary approximation of the speech signal (dependency 11);
3. determination of the fractal dimension with a given time window in each time fragment (dependency 10);
4. distribution of time fragments into vocalized and non-vocalized fragments according to the established threshold of the fractal dimension.

The proposed method for calculating the fractal dimension can be used in many areas of speech signal analysis. As an example, consider the use of the proposed method in the task of speech segmentation. Given the structural complexity of the task, the following research methodology was proposed (Figure 4).

When conducting a study of the effectiveness of segmentation, the following were chosen as the signs of making the decision "vocalized/non-vocalized fragment":

1. root mean square deviation of the fractal dimension;
2. range of time window sizes.

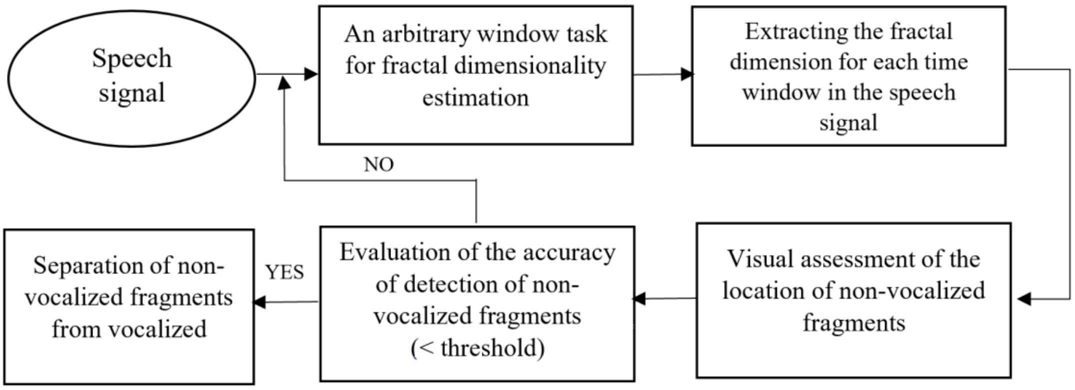
Speech signal
An arbitrary window task for fractal dimensionality estimation
Extracting the fractal dimension for each time window in the speech signal
NO
Separation of non-vocalized fragments from vocalized
YES
Evaluation of the accuracy of detection of non-vocalized fragments (< threshold)
Visual assessment of the location of non-vocalized fragments

Figure 4: Methodology of speech signal segmentation studies

Given the dependence of the fractal dimension on the size of the window, at the next stage, we set an arbitrary value of the window to estimate the fractal dimension. At the next stage, we calculate the fractal dimension for each window in the speech signal. Next, the speech signal is divided into pauses and speech fragments (visually). The subjectivity of the method of dividing pauses and fragments of speech at this stage of research should not have a significant impact on the following conclusions and results, given the significant differences in the fractal dimension of pauses and fragments of speech.

To carry out a comparative analysis of speech signal segmentation methods, the following methods were implemented: frame-by-frame; singular; deviation from the average; black area blocking; analysis of the spectral shape and the own method proposed in paper. The study involved 50 people (men and women with different linguistic backgrounds) who pronounced the same set of words. In the created test set, automatic marking of vocalized fragments and pauses was performed. When the comparison was made, all methods used their standard settings recommended by the developers. Each method performed segmentation of the marked speech signal, when the marking and the decision of the segmentation method coincided, the value 0 or 1 was set for each observed fragment of this method. When determining errors of the 1st and 2nd kind, an analysis was performed and a summary of the correspondence of the values of the previous marking 0 and 1, set by each method segmentation relative to the total number of pauses and vocalized fragments. The results of the study are presented in table 4.

Thus, in terms of effectiveness, the developed speech signal segmentation method performs its functions better, and can be recommended for use in the speech analyzing systems.

Table 4. The results of the study of the effectiveness of speech signal segmentation methods

Method	Error of the 1st kind	Error of the 2st kind
Frame by frame	0,177	0,1
Singular	0,133	0,1
Deviation from the average	0,180	0,17
Blocking of the black area	0,156	0,11
Analysis of the spectral form	0,141	0,1
The proposed method	0,108	0,1

Conclusion

The conducted review of speech signal segmentation methods showed a number of limitations of these methods and "floating" characteristics of segmentation accuracy, which requires analysis and development of a speech signal segmentation method. It was determined that the similarity of structures in the speech signal is possible due to their scaling. Considering this, it is shown that fractal and wavelet analysis can be the main approaches in the problem of speech signal identification. A comparative analysis of methods for calculating the fractal dimension was performed and it was determined that for the task of segmenting the speech signal into vocalized and non-vocalized fragments, the method of calculating the fractal dimension according to Hausdorff is the most effective, on the basis of which the method of segmentation of the speech signal is proposed. As a result of the research conducted on the basis of the proposed method of segmentation of the speech signal, as well as the modified assessment of the fractal dimension of fragments of speech signals, stable characteristics of the increase in the value of the modified fractal dimension for fragments of speech signals containing speech were established. The fractal dimension for pauses in 99% was within $1{,}04 \leq D \leq 1{,}45$, and the fractal dimension of speech fragments was not observed less than $D = 1{,}55$ for a time frame of 20 ms.

Bibliography

[Al-Mamie et al., 2009]. Al-Manie M.A., Alkanhal M.I., Al-Ghamdi M.M., Mastorakis N., Croitoru A., Balas V., Son E. and Mladenov V. Automatic speech segmentation using the arabic phonetic database. In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, no. 10. World Scientific and Engineering Academy and Society. 2009.

[Makowski et al., 2014]. Makowski R. and Hossa R. Automatic speech signal segmentation based on the innovation adaptive filter. International Journal of Applied Mathematics and Computer Science, vol. 24, no. 2. 2014. pp. 259–270.

[Chefir et al., 2001]. Cherif A., Bouafif L. and Dabbabi T. Pitch detection and formant analysis of arabic speech processing. Applied Acoustics, vol. 62, no. 10. 2001. pp. 1129–1140.

[Sharma et al., 1996]. Sharma M. and Mammone R. Subword-based text-dependent speaker verification system with user-selectable passwords. in Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 1. IEEE, 1996, pp. 93–96.

[Van Hemert, 1991]. Van Hemert J.P. Automatic segmentation of speech. IEEE Transactions on Signal Processing, vol. 39. no. 4. 1991. pp. 1008–1012.

[Lee et al., 2003]. Lee Y.-S., Papineni K., Roukos S., Emam O. and Hassan H. Language model based arabic word segmentation. in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. pp 399-406.

[SaiJayram et al., 2002]. SaiJayram A., Ramasubramanian V. and Sreenivas T. Robust parameters for automatic segmentation of speech. In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference. vol. 1. IEEE. 2002. pp. I–513.

[Tanqueiro, 2017]. Tanqueiro H.M.M. Utilitzacio´ didiomes. International Journal of Computer Science and Mobile Computing. Vol.6 Issue.4. 2017. pp. 308-315.

[Toledano et al., 2003]. Toledano D.T., Go´mez L.A.H. and Grande L.V. Automatic phonetic segmentation. IEEE transactions on speech and audio processing. vol. 11. no. 6. 2003. pp. 617–625.

[Mporas et al., 2008]. Mporas I., Ganchev T., and Fakotakis N. A hybrid architecture for automatic segmentation of speech waveforms. in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. 2008. pp. 4457–4460.

[Go´mez et al., 2011]. Go´mez J.A. and Calvo M. Improvements on automatic speech segmentation at the phonetic level. in Iberoamerican Congress on Pattern Recognition. Springer. 2011. pp. 557–564.

[Siniscalchi et al., 2007]. Siniscalchi S.M., Schwarz P., and Lee C.-H. High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring. in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP’07. vol. 4. IEEE. 2007. pp. 4–869.

[Zio´łko et al., 2006]. B. Zio´łko, S. Manandhar, R. C. Wilson, and M. Zio´łko, “Wavelet method of speech segmentation,” in Signal Processing Conference, 2006. 14th European. IEEE, 2006, pp. 1–5.

[Ratsameewichai et al., 2002]. S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Ua-trongjit, and K. Likit-Anurucks, “Thai phoneme segmentation using dualband energy contour,” ITC-CSCC: 2002 Proceedings, pp. 111–113, 2002.

[Tolba et al., 2005]. M. Tolba, T. Nazmy, A. Abdelhamid, and M. Gadallah, “A novel method for arabic consonant/vowel segmentation using wavelet transform,” International Journal on Intelligent Cooperative Information Systems, IJICIS, vol. 5, no. 1, pp. 353–364, 2005

[Suh et al., 1996]. Y. Suh and Y. Lee, “Phoneme segmentation of continuous speech using multi-layer perceptron,” in Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol. 3. IEEE, 1996, pp. 1297–1300.

[Rahman et al,. 2015]. M.M. Rahman, F. Khatun, and M. A.-A. Bhuiyan, “Blocking black area method for speech segmentation,” Editorial Preface, vol. 4, no. 2, 2015.

[Kaur et al., 2010]. E. A. Kaur and E. T. Singh, "Segmentation of continuous punjabi speech signal into syllables," in Proceedings of the World Congress on Engineering and Computer Science, vol. 1. Citeseer, 2010, pp. 20–22.

[Kalamani et al., 2017]. M. Kalamani, S. Valarmathy, and S. Anitha, "Hybrid speech segmentation algorithm for continuous speech recognition." International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 308-315

[Bansal et al., 2014]. P. Bansal, A. Pradhan, A. Goyal, A. Sharma, and M. Arora, "Speech synthesis-automatic segmentation," International Journal of Computer Applications, vol. 98, no. 4, 2014.

[Dines et al., 2002]. J. Dines, S. Sridharan, and M. Moody, "Automatic speech segmentation with hmm," in Proceedings of the 9th Australian Conference on Speech Science and Technology, 2002.

[Stolcke et al., 2014]. A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 5552–5556.

[Sapozhkov, 1963]. Sapozhkov M.A. Speech signal in cybernetics and communication. Moscow: Radio and Communications. 1963. 452 p.

[Soloviov, Bielozorova, 2013]. Soloviov V.I., Bielozorova Ya.A. The use of fractal dimension of audio files in the task of segmentation of a sound file/ Scientific journal. Eastern Ukrainian National University named after Volodymyr Dahl. – 2013. – № 5(194) ch.2. – pp. 165 – 168.

[Zybin, Bielozorova, 2020]. Zybin Serhii, Bielozorova Yana. Practical approach to speech identification // International Journal "Information models & analyses". – 2020. – Vol.9. – № 3. – pp. 224 – 231.

Authors' Information

Yana Bielozorova *– PhD, Associate Professor of Software Engineering Department, National Aviation University, Kyiv, Ukraine.*

E-mail: *bryukhanova.ya@gmail.com*

Major Fields of Scientific Research*: Speech Recognition Models, Wavelet analysis, Software Architecture*