УДК 004.934

METHOD OF SEGMENTATION OF SPEECH SIGNAL INTO PHONEMS

Ibragimova Sayyora Nomoz qizi
snibragimova@mail.ru
PhD student at the Research Institute for the development of digital technologies and artificial intelligence
17A, Buz-2, Mirzo Ulugbek, 100125, Tashkent, Uzbekistan;

Annotation. This article discusses the advantage of the wavelet transform over traditional signal processing methods. A method of phonemic segmentation of a speech signal is considered by analyzing the change in the energy of a speech signal at each level of the wavelet transform. A step-by-step algorithm of the proposed method is described. Based on the proposed algorithm, the result of the segmentation process into phonemic units is given.

Keywords: Signal segmentation, phonemes, speech signal, speech signal recognition, wavelet transform, discrete wavelet transform.

1 Introduction

In the modern global world, the study and knowledge of foreign languages acquires a new status. The use of new information technologies in teaching is one of the important aspects of improving and optimizing the educational process, enriching the arsenal of methodological tools and methods that allow you to change the forms of work and make the lesson of the Uzbek language interesting and memorable.

The modern educational paradigm based on computer learning tools is based on learning with an intelligent interface, that is, learning a language through interactive communication using computer tools. An important feature of a computer in the process of teaching a foreign language is that it can be the "interlocutor" of the student, that is, in a communicatively oriented interactive mode and in a certain way, for example, graphically, by analyzing and synthesizing a speech signal, it makes up for the lack of natural speech, shows the analysis of the correct pronunciation of foreign words.

For a computer to become an "interlocutor", you need to teach it to recognize speech. It is known that in recent years, systems for automatic recognition of speech signals have been rapidly developing abroad. However, the issues of automatic recognition of speech signals of the state language of the Republic of Uzbekistan are not sufficiently considered. One of the most important issues of automatic speech recognition systems is the processing of speech signals and the extraction of informative features.

To simplify the task of speech recognition, various restrictions are introduced using the grammatical rules of the language or limited to a smaller task, such as recognizing only isolated words. Recognition of isolated words underlies various types of recognition based on words, syllables or phonemes. Maintaining a large dictionary of patterns increases the requirements for computer memory, and the recognition process, which requires comparison with each pattern, becomes very long and less reliable as the dictionary grows. One way to solve these problems is to use approaches based on phoneme recognition. The division of various sounds in the composition of a speech signal - into minimal phonetic units of the language - into segments corresponding to phonemes, increases the recognition efficiency [7-8].

The possibility of successful application of wavelet transforms in speech processing and recognition problems is related to the characteristics of the speech signal. Wavelet analysis can represent a signal as a sequence of symbols with different levels of detail, which allows you to determine the local characteristics of the signal and classify them by intensity.

The purpose of this work is to develop a segmentation model of a speech signal based on a discrete wavelet transform and to determine the short-term energy of a speech signal.

The results of the proposed algorithm are modeled using the Matlab package program.

2 Formulation of the problem

Let's consider the speech signal $f(t)$, in this case, it is spoken in the Uzbek language. The task is to build such an algorithm that segments the speech signal into phonemic units for their further recognition. If segmentation of speech into phonemic elements is performed for the purpose of speech recognition or solving problems related to recognition processes, then the boundaries of the segment for the phonetic element should be set in such a way as to increase the degree of correct definition of the phonemic image, regardless of the acoustic sound of the object being determined.

3 Solution method

Establishing interphonemic boundaries by traditional methods will be very problematic. The wavelet transform helps to solve the problem in phonetic segmentation. The fact is that at interphonemic transitions, the speech signal undergoes significant changes at many scales of study, and, accordingly, is characterized by an increase in wavelet coefficients for detail levels . Drawing a conclusion from the above, we can say that the problem of finding interphonemic transitions is reduced to finding moments of increase in wavelet coefficients at a significant number of zoom levels. The decomposition of a speech signal with a length of N samples has the form [1-4, 12]:

$$f(t) = \sum_{k=0}^{\frac{N}{2^n}-1} s_{nk}\,\varphi_{nk} + \sum_{i=1}^{n}\sum_{k=0}^{\frac{N}{2^i}-1} d_{ik}\,\psi_{ik} \quad (1)$$

where n – is the level of detail, s_{nk} and d_{ik} – these are the approximating and detailing coefficients of the wavelet decomposition, respectively, at the n-th level, φ – scale function, ψ – basic (mother) wavelet.

The frequency range below 125 Hz is not considered, because it does not contain information important for the segmentation task. This is due to the nature of human speech, covering the interval 150-4000 Hz. Thus, 6 levels of decomposition are sufficient (Table 1) [5-6].

Table 1

Level of detail	Wavelet frequency range	
	Daubechies 16	Meyer
1	2000-4000 Hz	2756-5512 Hz
2	1000-2000 Hz	1378-2756 Hz
3	500-1000 Hz	689-1378 Hz
4	250-500 Hz	345-689 Hz
5	125-250 Hz	172-345 Hz
6	86-172 Hz	

The task of segmenting a speech signal into phonemic ones includes the following main steps:

1. Signal preprocessing. All readings are divided by the maximum value, to set uniform threshold values for any input signals.

2. Splitting the signal into frames. The input signal is split into frames of 512 samples at a sampling rate of 16 kHz with an overlap of 25% to 50%.

3. Hamming window preprocessing. Each frame is covered with a Hamming window to eliminate defects at the edges. The Hamming function has the form:

$$w(n) = 0.3836 - 0.46164\cos(\frac{2\pi n}{N-1})$$

4. Application of the wavelet transform. A wavelet transform is applied to each frame. Decomposition up to the 6th level of decomposition is used.

5. Calculation of short-term energy. At each level of detail, we calculate the short-term energy:

$E_j = \sum_{i=0}^{\frac{N}{j}-1} d_{j,i}^2$, where j – decomposition level number.

6. Building a number sequence: $\widetilde{e_{ij}} = \frac{\sum_{k=0}^{n_j-1} d_{j,i+k}^2}{E_j}$

where i – sliding window number, $n_j = \frac{n}{2^j}$ – sliding window size at j-th level.

7. Definition of hissing sounds and pauses. To do this, we introduce a threshold:

$$p = \frac{\max_i(\widetilde{e_i}) * \text{average}_i(\widetilde{e_i})}{10} + \min_i(\widetilde{e_i}) * 10,$$

where $\text{average}_i(\widetilde{e_i})$ – arithmetic means of a sequence $\{\tilde{e}_i\}_{i=1}^{N/512}$. The boundary between the hissing (pause) and other sounds is the boundary of the i-th window, for which the condition

$$(\tilde{e}_i - p)(\tilde{e}_{i+1} - p) < 0$$

8. Search for local maxima and minima to determine interphonemic boundaries. At the next stage of segmentation, the boundaries are located on the signal sections, which correspond to $\tilde{e}_i > p$. Looking for local maxima and minima of the sequence $\{\tilde{e}_i\}$ in the vicinity of two points in the areas under study. The right borders of the windows corresponding to these maxima and minima are the boundaries of the assumed phonemes.

Thus, we have defined an algorithm for segmenting a speech signal into phonemic units using a discrete wavelet transform for its further recognition.

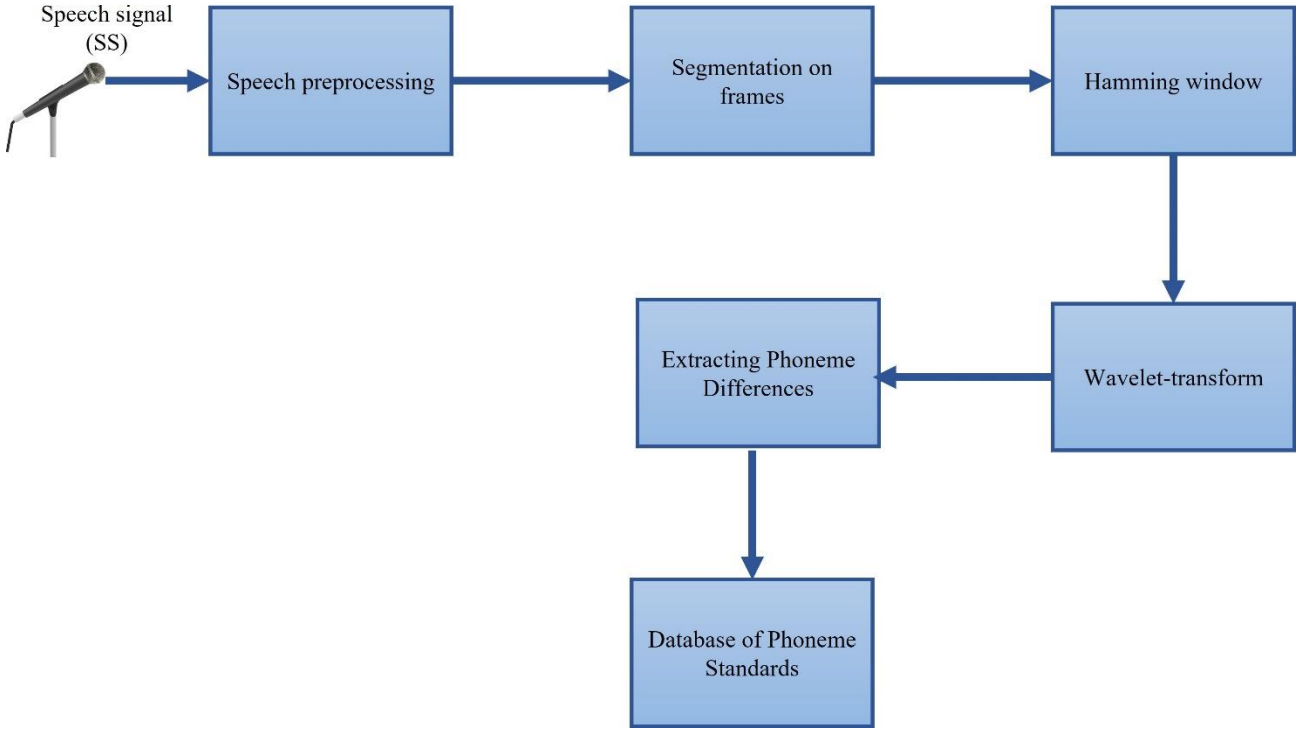

Fig. 1. Formation of a frame standard for a dictionary

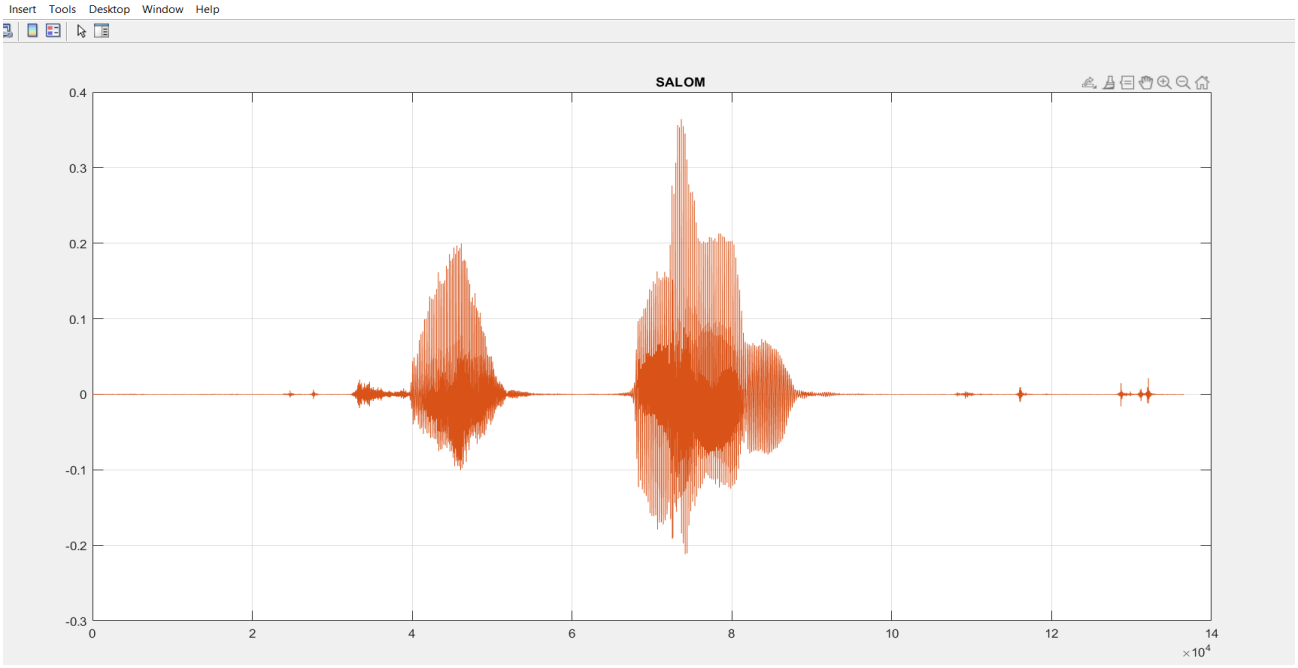

Fig. 2. Graphical view of the original signal

4 Experiments and results

In this work, a discrete wavelet transform (DWT) was used over the original signal, where the word "SALOM" is written. Using DWT, the original speech signal is decomposed into 6 levels of detail. The speech signal, digitized with a sampling frequency of 44100 Hz, is divided into overlapping windows of 16ms, which corresponds to 512 samples (Fig. 2).

The results obtained using the wavelet transform are presented graphically in the form of a scale diagram shown in Figure 3.

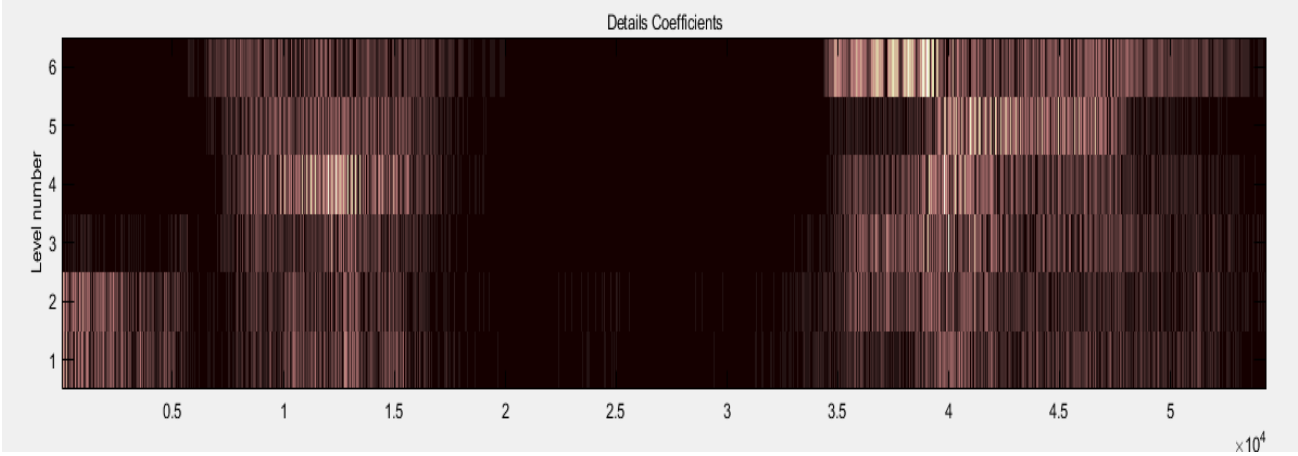

Fig. 3. The scalogram of the coefficients obtained with the Daubechies-10 mother wavelet.

On fig. 4 shows a visual segmentation of the speech signal of the word "SALOM" based on a discrete wavelet transform and the above algorithm.

Fig. 4. Segmentation of the word "SALOM" into phonemes

5 Conclusion

The proposed segmentation method into phonemic units is based on a discrete wavelet transform. The proposed speech signal segmentation algorithm based on the discrete wavelet transform is effective, as evidenced by the impressive results in signal segmentation into phonemic units using the MATLAB environment.

References

[1] Gorshkov Yu.G. Research complex for frequency-time analysis of a speech signal using wavelet technology. Bulletin of MSTU im.Bauman. Series Instrumentation, - 2011, No. 4, Pp. 78-87.

[2] Novikov L.V. Fundamentals of wavelet analysis of signals. Tutorial, 1999, p. 152.

[3] Smolentsev N.K. Fundamentals of wavelet theory. Wavelets in Matlab/H.K. Smolentsev. — M.: DMK Press, 2005, p. 304.

[4] Dobeshi I. Ten lectures on wavelets. - Izhevsk: Research Center "Regular and Chaotic Dynamics", 2001, p.464.

[5] Sorokin V.N. Segmentation and recognition of vowels / V.N. Sorokin, A.I. Tsyplikhin // Information Processes, 2004, V. 4, No. 2, pp. 202–220

[6] Vishnyakova O.A. Automatic segmentation of a speech signal based on discrete wavelet transform / O.A. Vishnyakova, D.N. Lavrov // Mathematical Structures and Modeling, 2011, Vol. 23, pp. 43–48.

[7] Kipyatkova I.S., Automatic processing of colloquial Russian speech / I.S. Kipyatkova, A.L. Ronzhin, A.A. Karpov. - St. Petersburg: GUAP, 2013, P. 314.

[8] Rabiner, L. Fundamentals of speech recognition / L. Rabiner, B.-H. Juang // Prentice Hall PTR, Englewood Cliffs, NJ 07632, 1993, p 507.

[9] Novoselov S. A. Isolation and preprocessing of signals in systems of automatic recognition of speech commands / S. A. Novoselov. Abstract Candidate's thesis those. Sciences. - Vladimir: VlGU, 2011. p 20.
[10] Zheltov P.V., Semenov V.I. A method for determining the boundaries between vowels and consonants of speech using fast continuous wavelet transform // Dynamics of Scientific Research - 2011: Proceedings of the VII Intern. scientific-practical. conf. Przemysl: Nauka i studia, 2011, pp. 12–17.
[11] Zheltov P.V., Semenov V.I. Some problems of speech recognition // Computer technologies and modeling: Sat. scientific tr. / KSTU im. A.N. Tupolev. Kazan, 2008. Issue. 1. pp. 33–37.
[12] Ibragimova S.N. The advantage of the wavelet transform in processing of speech signals // Technical sciences. 2021. Vol 4, Issue 3, pp 37-41.
[13] Matyokubov O.K., Madaminov F.Q. Analysis of signal spectra in the program ADP MATLAB // Actual problems of cooperation of higher and secondary special, vocational education institutions, Proceedings of the II scientific-practical conference. May 20-21, 2016, pp.558-560.
[14] Raximov T.O., Matyokubov O’.K., Yangiboyeva M.R. Signallarni veyvlet almashtirish yordamida filtrlash va siqish jarayonlarini tahlili // International scientific conference «global science and innovations 2019: central asia» Nur-sultan, Kazakhstan, may 2019. Pp. 58-61.