Section 6. Economics and management

https://doi.org/10.29013/ESR-21-9.10-44-49

Nozima Ragachurina,
senior teacher of Electronics and Radio-engineering Department
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi

ANALYSIS OF SPEECH PHONEMIC AND FORMANT STRUCTURES

Abstract. The article proposes a way to improve the accuracy of the analysis of the formant structure of speech sounds based on the formation of "idealized" phonemes obtained from quasi-stationary fragments of the original phonemes. It is shown that, based on the proposed approach, it is possible to refine the spectrum of any fragment of a phoneme, both for vocalized and non-voiced speech sounds, as well as for any stochastic signals using the standard Fast Fourier Transform (FFT) procedure in sound editors.

Keword: speech, phonemic, Fast Fourier Transform, spectrum, Discrete Fourier Transform.

Introduction

The most natural and popular means of communication between people has been and remains speech, therefore, interest in the development of speech signal processing technologies remains in the focus of attention of specialists in the field of infocommunication systems. This is evidenced by the presence of scientific publications on this topic, for example. Speech itself, by its nature, is a unique signal, and the essence of the process of verbal communication of people has not been fully disclosed. That is why the technologies used for the development of infocommunication systems are based on various approaches, including those based on the analysis of the phonemic and formant structures of speech. When training specialists in the field of speech infocommunication technologies, it is useful to demonstrate the phonemic and formant structures of speech. For these purposes, it would be logical to use audio editors that allow real-time analysis and processing of audio signals. However, an attempt to use them for the spectral analysis of speech sounds was not successful. This article discussed the reasons that did not allow us to demonstrate the formant structure of Russian speech sounds when using sound editors Adobe Audition®, Sound Forge®, Audacity® and the like. For the spectral analysis of sounds in these editors, the standard fast Fourier transform procedure is used, which provides for the division of the studied sound signal into segments of a given sample dimension N, moreover, a multiple of an integer power of 2, i.e. N = 2n. As a rule, values $n \geq 6$ can be set in sound editors, which allows forming segments of speech sounds from 64, 128, … 1024, etc. counts. This limitation on the segment size results in the impossibility of accurately matching the sample size to the duration of individual speech sounds, which is generally a random variable. As a result, certain segments of sounds are subjected to spectral analysis, the size of which is smaller or (most often) larger than the duration of real phonemes. In addition, it was shown that segments of speech sounds are subject to additional distortions resulting from the application of window functions necessary to eliminate the Gibbs effect. These distortions could be significantly weakened by increasing the sample size N so that it contains a large number of periods of the studied phoneme. However, this is rarely possible for two reasons. First, unvoiced speech sounds (short consonants "н", "к", etc.) have a short duration, so it is almost impossible to single out a sound segment containing several identical phonemes. Second, even in vocalized speech sounds, the shape of the phoneme undergoes significant changes in different parts of their sound, which are always accompanied by a redistribution of energy in the spectral region. In addition, of additional interest is the analysis of changes in the formant structure of the phoneme in various areas of its existence: attack, stationary part, decay. Attempts to improve the situation by varying the sample size N, increasing the sampling frequency of audio signals, and using various window functions did not lead to positive results. Therefore, the article concluded that the use of sound editors in the educational process to demonstrate the formant structure of speech phonemes is impossible, since the spectra calculated by them do not reflect the known results. To solve this problem, software should be used that provides the user with complete freedom in choosing the parameters of the sample size of the analyzed fragment and calculates the spectrum using algorithms that do not require a mandatory multiple of the FFT window size of a power of 2. Such conditions can be implemented, for example, in the MATLAB® program, which makes it possible to arbitrarily set the parameters of the discrete Fourier transform. The purpose of the article is to present a way to improve the accuracy of displaying the formant structure of speech sounds using the standard FFT-based spectrum calculation procedure used in audio editors. Statement of the problem aor research, the MATLAB® application program was used, in which all available algorithms for calculating the spectrum of signals are implemented. First of all, it was necessary to resolve the issue of choosing a criterion for assessing the accuracy of spectrum calculation. It is known that in the spectral analysis of deterministic signals, it is possible to analytically calculate the exact values of the coefficients of the Fourier series, which should be used as reference values when analyzing the results of calculating the spectrum of such signals in various ways. Phonemes, on the other hand, are not deterministic signals, since they do not have a stable form, which always changes depending on the place of the sound in the word, the features of the speaker's voice, and a number of other factors. In this sense, the phoneme is often compared to letters written by people with different handwriting. Therefore, the spectrum of each phoneme is unique and unpredictable.

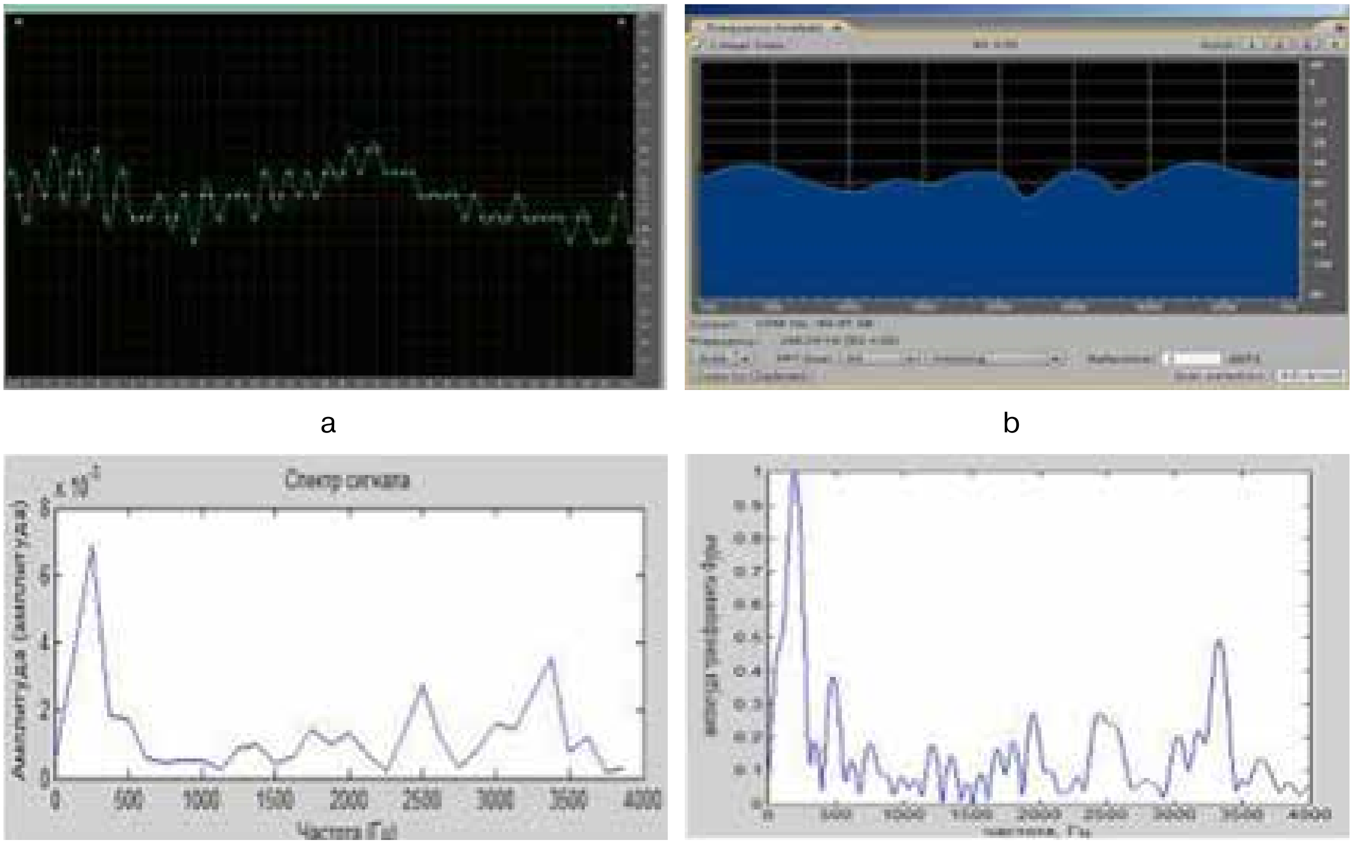
Average spectrum estimates are possible and widely used in the development of infocommunication technologies and speech processing devices. However, in a number of cases, it is precisely the exact assessment of the formant structure of speech sounds that is of practical interest. For example, when solving the problem of identifying the speaker's voice, it is important to know exactly the unique features of the pronunciation of sounds or even their individual parts inherent in a particular person. It is precisely this goal – the preservation of the unique spectral parameters of the studied speech sounds – that the authors of this article set themselves. Considering the above, it was decided to assess the accuracy of spectrum calculation as follows: choose one of the spectrum calculation algorithms as a reference one, having previously determined the conditions for its application, and then compare the control results obtained with its help with the values obtained using other algorithms.

The article made an assumption that the distortion of the spectrum of the studied speech sounds is due to the use of the FFT algorithm, namely: the inability to match the size of the analyzed sample with the duration of the studied phoneme and the subsequent application of window functions to eliminate the Gibbs effect. Therefore, the FFT algorithm cannot be chosen as a reference, but, on the contrary, the results obtained with its help should be compared with the control values obtained with the algorithm chosen as a reference. The further course of reasoning was as follows: in order to eliminate the distortions caused by the Gibbs effect, it is necessary to get rid of the discontinuity points of the first kind in the studied sound segments, i.e. signal voltage surges at segment boundaries. Then there is no need to use window functions that distort the shape of the analyzed segment.

The desired result can be achieved if the sound segmentation is performed manually, highlighting the beginning and end of the studied phoneme at the time moments when the levelogram graph crosses zero. This method of segmentation has one more advantage: it is possible to study the spectral structure of sound at any segment of its existence – attack, stationary part or attenuation. Obviously, the size of the signal sample obtained in this way will not be characterized by a multiple of a power of two, so the use of the standard FFT algorithm for audio editors is not possible. In this case, the spectrum can be calculated using the discrete Fourier transform formula:

$$F(k)=\sqrt{\left(\sum_{i=0}^{N-1}x[i]\cos\left(\frac{2\pi ki}{N}\right)\right)^2+\left(\sum_{i=0}^{N-1}x[i]\sin\left(\frac{2\pi ki}{N}\right)\right)^2}$$

The results of calculation according to formula (1) should be used as control ones, and the DFT calculation method itself should be considered as a reference one. Results of computational experiments To carry out computational experiments, the Adobe Audition® sound editor and the MATLAB® program were used, in which the spectrum was calculated in two ways: using the built-in function FFT, which implements the FFT algorithm with a rectangular window; using a written program that implements the calculation of the DFT according to formula (1).



a

b

c

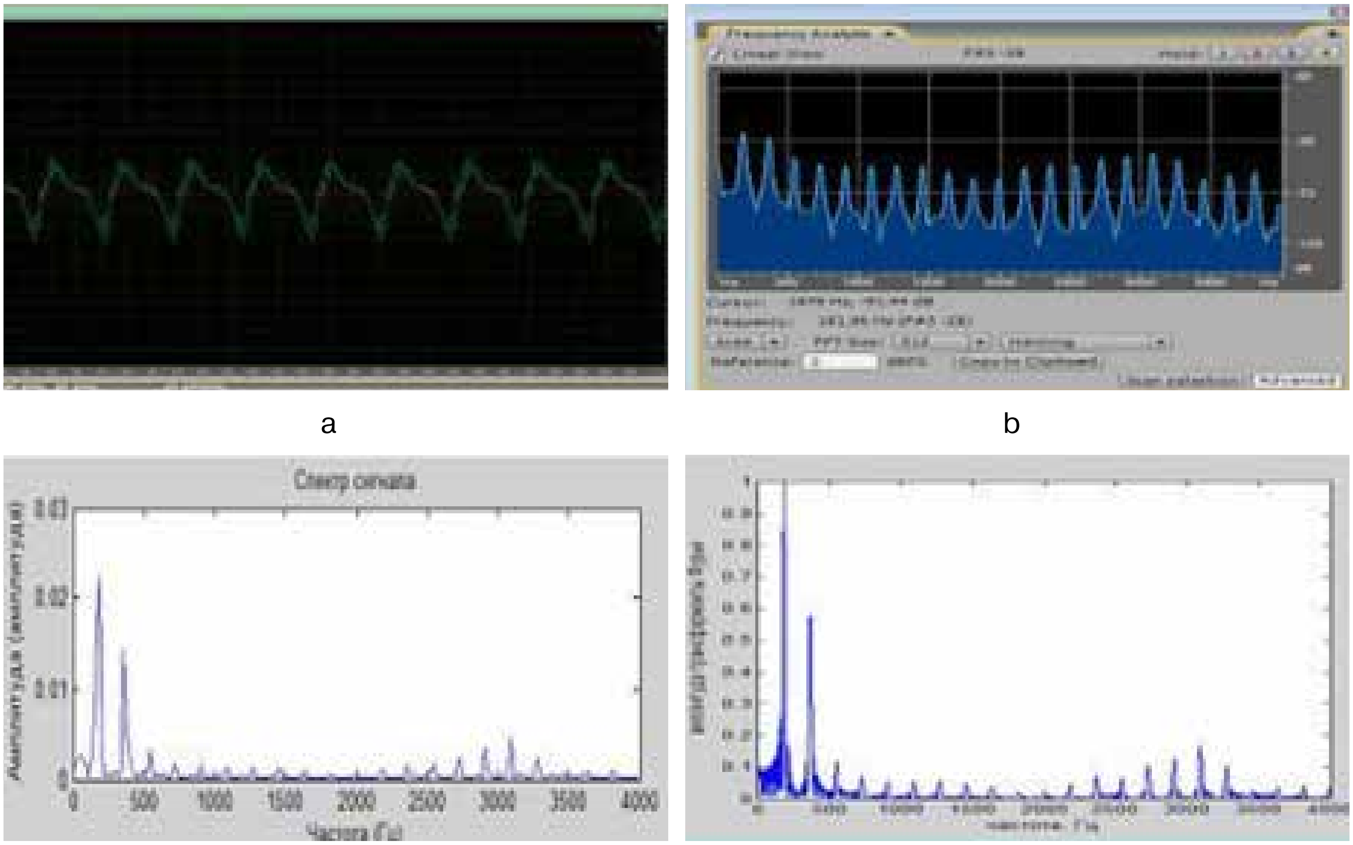
d

Figure 1.

Figure 1 shows the phoneme of the sound “и” selected by the above method with a sampling frequency Fd=8 kHz (a) and the results of its spectral analysis by the FFT method using the Hanning window, N=64 in the Adobe Audition® editor (b), calculated in MATLAB® program using the built-in FFT function, N=64 (c) and using the full DFT formula (d).

Comparison of the spectra in Figure 1, a, b, c shows their significant difference. In Figure 1b, the spectrum looks like a fluctuating function with approximately the same, uniform energy distribution over the entire signal frequency band. The spectra calculated in the MATLAB® program give a completely different picture of the energy distribution – the region around 250 Hz stands out strongly, and local formant regions at frequencies of 2000 Hz, 2500 Hz and 3400 Hz are also visible, which have a clearly lower energy concentration. For the correct interpretation of these spectra, an envelope line should be drawn that smoothly connects the peak values of the spectral components. Obviously, the use of the Adobe Audition® sound editor does not give a satisfactory result and may cause an incorrect assessment of the phoneme energy distribution during the educational process. A similar pattern was observed when calculating the spectrum of phonemes of other speech sounds.

At the next stage of research, it was assumed that in order to improve the accuracy of spectral analysis with the help of sound editors, it is possible to synthesize an “ideal” sound for analysis, formed by repeated copying of one phoneme sample that does not have break points (voltage surges) at the beginning and end. Such an operation is easily implemented using the “insert” option available in all sound editors. Figure 2 shows the levelgram and calculated spectra of 10 phonemes of the sound “и”



a

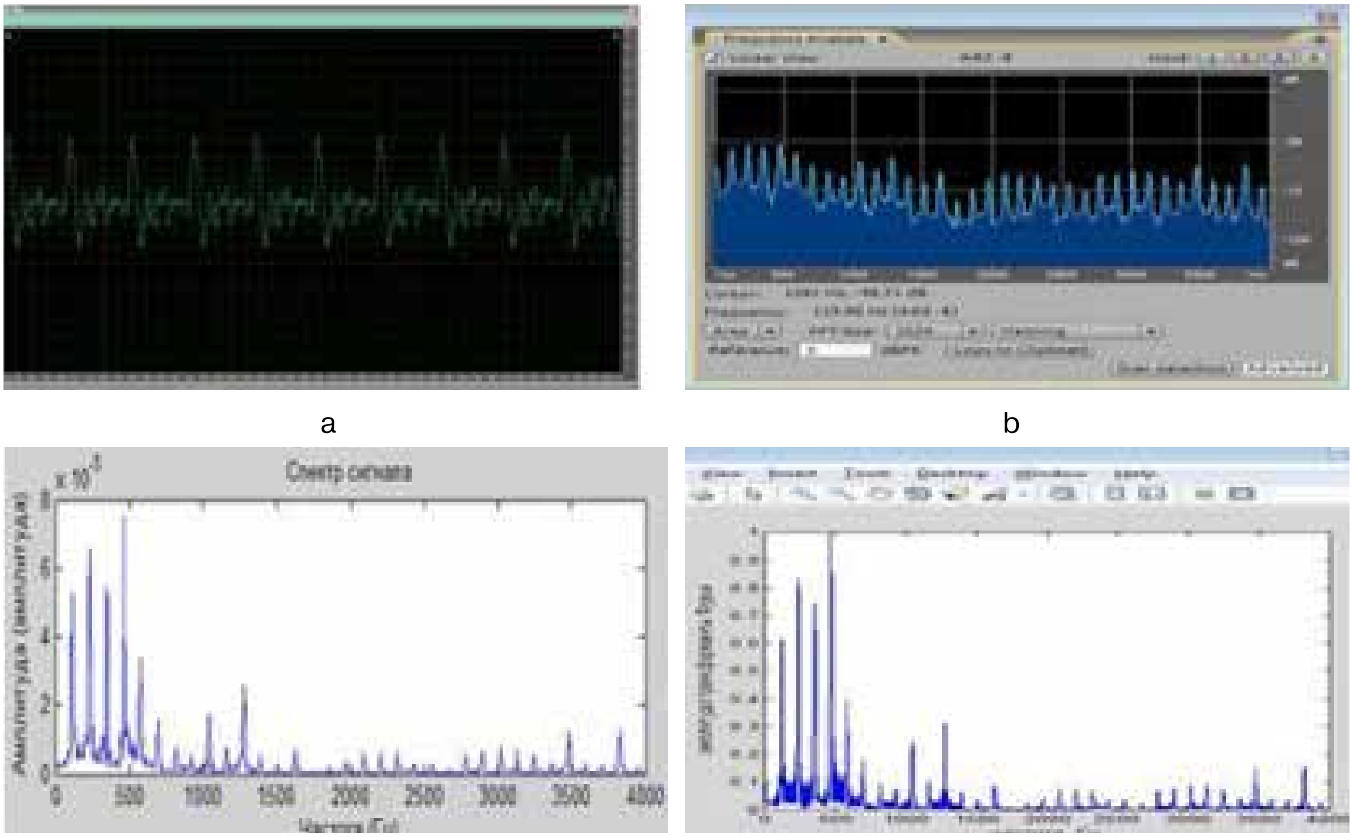
b

c

d

Figure 2.

On Figure 2 shows the results of calculating the spectrum of an audio signal synthesized from 10 periods of the phoneme of the sound "и". The total duration of the received signal was 440 counts. To analyze the spectrum using the FFT method, in both cases, the closest of the large window sizes N = 512 > 440 was chosen. Accordingly, the DFT was calculated exactly for 440 samples. Comparison of the corresponding results of calculating the spectra in Figs. 1 and 2 shows that the synthesized sound gives a clearer picture of the spectrum, which acquires a pronounced periodic structure and facilitates the localization of formant regions. At the same time, the results of calculations for the FFT and DFT performed in the MATLAB® program for the synthesized "ideal" phoneme are very close, which indicates an increase in the accuracy of calculating the spectrum with the FFT, despite the difference in the signal sample size and analysis window size. The appearance of the spectrum in the Adobe Audition® program has also significantly improved – the spectrum envelope has become more expressive and generally repeats the behavior of the spectrum envelope obtained in the MATLAB® program. For greater expressiveness of the energy distribution, one should use not a logarithmic amplitude scale, but a linear one, as in the MATLAB® program. Levelgram and calculated spectra of 10 phonemes of the sound "з" shown below.



a

b

c

d

Figure 3.

In Figure 3 the results of calculating the spectrum for the "ideal" phoneme of the sound "з", consisting of 10 periods of the phoneme, with a total volume of 690 samples, are given. The window size for the FFT was chosen to be 1024 samples. As in the previous case, a significant improvement in the localization of formant regions is noticeable.

Conclusion

The proposed method for calculating formant regions, which provides for the formation of an "ideal" phoneme by repeatedly copying one period of the studied phoneme, makes it possible to increase the accuracy of spectral analysis when using the standard FFT procedure in the general case and when using sound editors in particular. It is also necessary to note the universality of the proposed approach for the spectral analysis of any non-stationary or short-duration signals, if the subject of study is their instantaneous spectrum. In fact, the proposed approach partially eliminates the most important drawback of the Fourier transform – its inherent time-frequency uncertainty.

References:

1. Sidorenko I. A., Kuskova P. A. O spektral'nom analize fonem s ispol'zovaniem zvukovyh re-daktorov // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta, No. 1(144). 2013. vypusk 25/1, serija Informatika, – Belgorod, 2013.– P. 246–250.
2. Zhilyakov E. G., Prohorenko E. I. Chastotnyj analiz rechevyh signalov // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. – No. 2(31). 2006. vypusk 3, serija Informatika i prikladnaja matematika.– Belgorod, 2006.– P. 201–208.
3. Zhilyakov E. G., Firsova A. A. Ocenivanie perioda osnovnogo tona zvukov russkoj rechi // Nauch-nye vedomosti Belgorodskogo gosudarstvennogo universiteta, No. 1(144). 2013. vypusk 25/1, serija Informa-tika, – Belgorod, 2013.– P. 173–181.
4. Babarinov S. L., Budnikova M. A. O raspoznavanii rechi // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta, No. 21(192). 2014. vypusk 32/1, serija Informatika, – Belgorod, 2014.– P. 182–185.
5. Savchenko V. V., Vasil'ev R. V. Analiz jemocional'nogo sostojanija diktora po golosu na osnove foneticheskogo detektora lzhi // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta No. 21(192). 2014. vypusk 32/1, serija Informatika, – Belgorod, 2014.– P. 186–195.
6. Prihod'ko A. I. Determinirovannye signaly. Ucheb. posob. Dlja vuzov.– M: Gorjachaja linija-Telekom, 2013.– 326 p.: il.