UDC 057.087.1:621.391.26

# ANALYSIS OF PROCEDURES FOR VOICE SIGNAL NORMALIZATION AND SEGMENTATION IN INFORMATION SYSTEMS

**Pastushenko M. S.** – PhD, Professor, Professor of V. V. Popovskyy Department of Infocommunication Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

**Pastushenko O. M.** – PhD, Senior Researcher, Serviceman of the Armed Forces of Ukraine, Ukraine.

**Faizulaiev T. A.** – Director of "TAF-87" LLC, Kharkiv, Ukraine.

## ABSTRACT

**Context.** The current task of evaluating formant data (formant frequencies, their spectral density level, amplitude-frequency spectrum envelope, formant frequency spectrum width) in voice authentication systems is considered. The object of the study is the process of digital preprocessing of the voice signal when extracting formant data.

**Objective.** Evaluation of the effectiveness of traditional procedures for digital preprocessing of a user voice signal and development of proposals for improving the quality of formant data extraction.

**Method.** A mathematical model for extracting formant data from an experimental voice signal has been developed to study the influence of normalization and segmentation procedures on the quality of the resulting estimates. By modeling the process of extracting formant data, the results of digital processing of normalized and non-normalized voice signals are compared. The influence of the processed frame duration of the experimental voice signal on the quality of the formant frequencies assessment is estimated. The results are obtained for the experimental phoneme and morpheme.

**Results.** The obtained results show that when processing a voice signal with a sufficient signal-to-noise ratio, normalization procedures are not mandatory when extracting formant data. Moreover, normalization leads to a less accurate measurement of the spectrum width of formant frequencies. It is also unacceptable to use a processed frame duration of less than 40 ms. These results allow us to modify the traditional method of voice signal preprocessing. The use of the modeling method in the study of the experimental voice signal confirms the reliability of the results obtained.

**Conclusions.** The scientific novelty of the research results lies in the modification of the voice signal preprocessing methodology in authentication systems. Eliminating normalization procedures at high signal-to-noise ratios of the voice signal, which occurs in user authentication systems, makes it possible to increase the speed of formant data extraction and more accurately estimate the width of the formant frequency spectrum. Selecting a frame duration of at least 40 ms for the processed signal significantly improves the accuracy of formant frequency determination. Otherwise, the estimates of the formant frequencies will be high. Moreover, when processing phonemes, the processed voice signal cannot be divided into frames. Practical application of research results allows to increase the efficiency and accuracy of the formant data generation. Prospects for further research may be studies of the influence of normalization and framing procedures on other elements of a template of the authentication system user.

**KEYWORDS:** authentication, voice signal, normalization, segmentation, formant data.

## NOMENCLATURE

$i$ is an analyzed element of the voice signal;

$N_o$ is a number of elements in a processed voice signal window;

$s_i$ is a non-normalized voice signal element;

$s_{ni}$ is a normalized voice signal element;

$s_{\max}$ is a maximum value of the absolute values of a processed voice signal sample;

$abs$ is a designation of the absolute value of an element being analyzed;

max is a designation of the maximum value selection function;

$T$ is the duration of a processed section (frame) of a voice signal in seconds;

$df$ is a frequency resolution of spectral analysis in hertz;

$f_d$ is a sampling frequency of an analyzed voice signal in hertz.

## INTRODUCTION

In modern society, the problem of information security of computer systems in various organizations and institutions of all forms of ownership and purposes is becoming more acute. Information protection and reliable operation of computer systems is one of the most important tasks, which largely depend on the methods of authenticating users and processes. In information systems, authentication procedures are becoming a key component in ensuring security. Authentication systems are widely used in computer systems, mobile devices, online services and play a decisive role in confirming user access rights to systems and their resources.

The known types of authentication have their advantages and disadvantages, and the practice of their application shows their low reliability. In this regard, intensive research is being conducted on the use of biometric user features in authentication systems and various access systems. It is known that among biometric authentication systems, voice systems are preferable by the efficiency/cost criterion [1]. Moreover, modern voice systems do not implement all the achievements of digital signal processing, which are widely and effectively used in radio communications and radiolocation. Currently, the Ukrainian state bank PrivatBank is implementing voice authentication systems, which emphasizes the relevance of the research.


DOI 10.15588/1607-3274-2025-4-17

The theoretical basis of voice authentication procedures is formed by research conducted in the field of speech recognition, which began in the middle of the last century. Both in the field of speech recognition and in authentication systems, digital processing of the voice signal begins with the preprocessing stage. Then, in authentication systems, specific procedures for forming a user template are performed. The stage of extracting features from the voice signal and placing them in the template is the second stage of voice signal processing. At the third stage, voice systems compare the current user template with the data placed in the template database, and a decision is made on the user access to resources.

User features extracted from a voice signal typically include: pitch frequency, formant data, cepstral and mel-frequency cepstral coefficients, etc. Formant data play a special role, including: formant frequencies; their spectral density level; amplitude-frequency spectrum envelope; spectrum width of formant frequencies.

A formant is a resonant frequency in the human vocal tract that provides a unique timbre and quality of speech sound. Formant frequencies are created by the structure of the vocal tract, the shape and position of the tongue and lips during speech. Formants are decisive in phonetics and speech processing in all practical applications.

Formants also play a decisive role in a new practical task, namely, voice synthesis.

In this case, modern scientific literature focuses on issues of evaluating formant frequencies, for which frequency filters, spectral analysis, wavelet transform, neural networks, etc. are used. In practical applications, preference is given to spectral methods.

**The object of** study is the process of digital preprocessing of a voice signal during the extraction of formant data**.**

**The subject of the study** is the sampling methods and procedures for extracting formant data from the voice signal of the authentication system user. At the preprocessing stage, the processed signal is traditionally normalized and divided into separate sections (frames) to extract user features.

**The purpose of the work** is to evaluate the effectiveness of traditional procedures for digital preprocessing of a user voice signal and to develop proposals for improving the quality of formant data extraction.

## 1 PROBLEM STATEMENT

It is known that voice signal normalization is an important stage of preprocessing in voice authentication systems, aimed at bringing the signal to a standardized form to improve recognition accuracy. Even simple normalization, which comes down to dividing each element of the processed voice signal array $s_i$ by the maximum value of the absolute values of the processed data

$$s_{ni} = s_i / s_{\max}, \quad (1)$$

where

$$s_{\max} = \max(abs(s_i)) \quad (2)$$

for all analyzed elements within the range of change $i$ from 1 to $N_o$, allows to significantly increase the signal-to-noise ratio. Here $N_o$ is the number of elements in the processed voice signal window. Also, the processed data is converted to the range from $-1$ to $1$.

However, authentication systems typically process voice signals with a sufficient signal-to-noise ratio (20 dB or more) and the feasibility of normalization procedures for extracting formant data is questionable.

In order to extract user features, authentication systems are further recommended to process in a sliding window with some overlap. In this case, the duration of the processed section (frame) is selected in the range from 20 to 40 ms. Note that in most cases, spectral analysis is used to extract user features, including formant data. It is known that the frequency resolution of spectral analysis is determined by the duration of the processed signal, namely

$$df = 1/T, \quad (3)$$

where $T$ is the duration of the processed section (frame) of the voice signal in seconds. For a discrete signal and digital processing, the specified relationship is transformed to the following form

$$df = f_d / N_o, \quad (4)$$

where $f_d$ is the sampling frequency of the voice signal in hertz.

Setting the window duration to 20 ms opens the question of the formant data extraction quality for authentication purposes.

## 2 REVIEW OF THE LITERATURE

The theoretical foundations for the acoustic spectral theory of speech were formulated quite a long time ago by the German scientist H. Helmholtz, which are still used today. Here we will also note the works of the Swedish acoustician G. Fant, who developed the theory of distinctive features – a universal acoustic classification of sounds. In the proposed classification, a special place is given to formant data.

The theoretical foundations of digital processing of speech signals are presented in the works of Rabiner L.R. and co-authors, for example, [2]. This is one of the first textbooks in which the concepts of normalization, segmentation (framing), window functions and frame overlapping were introduced.

We would like to draw attention to the modern fundamental work [3]. This work provides detailed research on the issues of digital processing of voice signals, including the issues considered in the article.

Formant data form the basis of modern templates in authentication systems, but they are also used in a number


DOI 10.15588/1607-3274-2025-4-17

of other speech-processing applications, including medicine. It is known that the first two formants determine the results of speech recognition, while the third and fourth play a significant role in speaker identification [3].

One of the main procedures of preliminary processing of a voice signal is its normalization. A large number of articles is devoted to the problem of normalization. Let us pay attention to the work [4], which considers various mechanisms of normalization in the recognition of English vowel sounds. It was found that the first and second formants of vowel sounds depend on the size of the vocal tract. It was noted that normalization can partially eliminate this effect. At the same time, it was concluded that for better perception of vowel sounds, preference should be given to simpler normalization procedures.

In [5] the results of studies are presented that indicate the connection between perception and pronunciation. A comparative analysis of the original voice signals perception and their normalized representations is performed. The problem was solved using the example of language learning by people who were not native speakers.

Formant frequencies are increasingly analyzed not only for speech recognition or authentication purposes, but also for other purposes. For example, in [6], the length of the vocal tract is estimated and scale-invariant representations of formant templates are created. Normalization is widely used, both when processing individual sounds and in the interests of combining the estimates obtained for different sounds.

The paper [7] investigates the use of vowel formants for text-independent speaker authentication, including methods for extracting and filtering formants. Normalization procedures and digital processing in the form of individual frames are widely used.

Audio processing has become an integral part of modern applications in areas ranging from healthcare to speech-controlled devices. At the same time, segmentation of the analyzed data plays an important role in audio signal processing. Neural networks and deep learning have been increasingly used to solve this problem. Therefore, the work [8] deserves attention. The analysis presented in this paper confirms and establishes the importance of deep learning methods in audio segmentation.

In [9] a model based on non-negative matrix factorization was developed and investigated. Using procedures for segmenting the analyzed signal, the developed model allows dividing the analyzed data into speech, noise and music.

In the article [10] a compact segmentation model is developed that improves the quality of speech translation by framing and pre-training with punctuation. The advantage of the developed model is the increased efficiency of solving the speech translation problem.

## 3 MATERIALS AND METHODS

A mathematical model for extracting format data from a voice signal has been developed to study the influence of normalization and segmentation procedures on the quality of the resulting estimates. The experimental voice signal containing the word “odyn” (“one” in Ukrainian) with a sampling frequency of 64 kHz was analyzed. This signal is presented in Fig. 1.

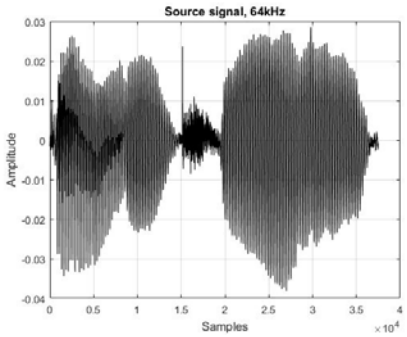


Figure 1 – The analyzed voice signal

We will conduct the study of the proposed signal in three stages. In the first stage, we will analyze the phoneme “o”, which makes up the first part of the analyzed signal. In the second stage, we will analyze the morpheme “dyn”, which makes up the second half of the analyzed signal. Information on the processing of the entire signal will also be obtained.

The main method of evaluating formant data will be spectral analysis of the amplitude-frequency spectrum of the voice signal under study. The spectrum was evaluated using the fast Fourier transform. In this case, we will pay some attention to the voice signal normalization procedures. The normalization procedures will be implemented using the norm method. As a norm, we will use the maximum value of the amplitude of the signal under consideration.

The main studies concern the assessment of the influence of the processed frames duration (sections of the voice signal) on the results of the evaluated formant data. Here, three variants of frame duration are also considered: 20 and 40 ms, as well as the entire analyzed signal.

Currently, when extracting user features from a voice signal, the latter is divided into frames, which are recommended to be selected with a duration of 20 ms. Digital processing is carried out in a “sliding window” that has some overlap.

This approach leads to the fact that spectral analysis must be carried out on a limited amount of initial information. The latter leads to a significant decrease in the resolution of the spectrum, and therefore the accuracy of the formant frequency assessment.

Some attention is paid to the assessment of the influence of the Hamming window function, which is widely used in digital processing of voice signals. The purpose of this function is to eliminate the effect of spectrum “spreading”. This approach will allow to compare traditional approaches to digital voice signal processing and develop practical recommendations for significantly improving the quality of formant data estimates.


DOI 10.15588/1607-3274-2025-4-17

## 4 EXPERIMENTS

The first stage of the research is associated with the analysis of the experimental phoneme “o”, which is presented in Fig. 2. The sampling frequency of the voice signal was 64 kHz.

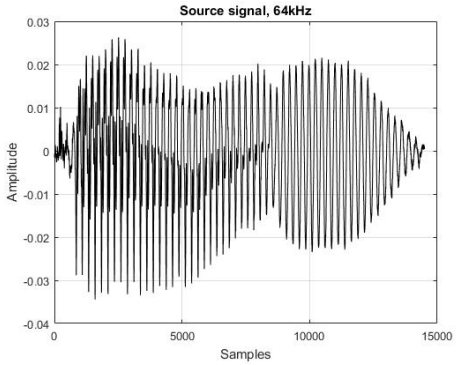


Figure 2 – The analyzed phoneme is “o”

We will begin the analysis with a study of voice signal normalization procedures. Interest in normalization procedures is conditioned by the fact that fundamental and scientific literature has not disclosed in detail the influence of voice signal normalization procedures during its digital processing. The question of the influence of normalization procedures on the quality of generating formant data also remains open.

In addition, the analysis of the Hamming window influence on the formant data of the voice signal is of interest. Before presenting the results of the experimental studies, it is necessary to make a number of comments.

Firstly, the polyharmonic voice signal being processed is recommended to be divided into frames (segments) of 20 ms in known works. This is justified by the fact that in this section the signal being studied will be quasi-stationary. In some works, it is recommended to choose this section of 40 ms.

Secondly, the Hamming window is one of the weighting functions that can be used as one of the characteristics of nonparametric spectral analysis. As stated in a number of works, the purpose of window function processing is to eliminate spectrum spreading.

Thirdly, the real effect during processing of the experimental voice signal is the elimination of anomalous estimates of formant frequencies.

Thus, the procedures of digital processing of the experimental voice signal allowed processing both normalized and non-normalized signals. In addition, to assess the influence of the window function, three duration variants of 20 and 40 ms, as well as a signal of full duration, were studied. Here we will also study the second part of the voice signal under consideration (see Fig. 1), namely the morpheme “dyn”. This will allow us to clarify the features of processing signals including different voice components.

The mathematical model for conducting experimental studies was implemented in the environment of the MatLab computer mathematics system. This approach allowed us to significantly simplify the modeling procedures by using standard MatLab functions. In addition, the use of standard functions allows increasing the reliability of the scientific results obtained.

The mathematical model was based on standard procedures for extracting formant data for different characteristics of the processed voice signal.

## 5 RESULTS

As the results of experimental studies of the phoneme “o” show, the implementation of voice signal normalization procedures essentially has little effect on formant data. At the same time, the spectral density levels of formant frequencies increase by approximately 30 dB. The spectrum envelope, in this case, moves to the positive region. However, the shape of the normalized amplitude-frequency spectrum envelope does not change.

Normalization does not affect the value of formant frequencies. The values of the estimates of the six analyzed frequencies do not change.

The most significant changes occur when assessing the width of the spectrum of formant frequencies. The changes are associated with an increase in the level of spectral power after normalization procedures. Fig. 3 shows the spectral density peak of the second formant frequency of the phoneme “o”. The width of the formant frequency spectrum was measured at a level of 0.8 (dash-dotted line) from the maximum value. The solid curve shows the peak of the non-normalized formant frequency, and the dashed curve shows the peak of the normalized formant frequency.

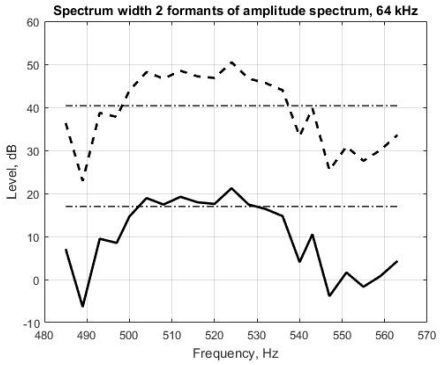


Figure 3 – To the assessment of the formant spectrum width

After the normalization procedures, the spectrum width of the formant frequency increases. The latter is conditioned by the influence of the formant harmonic peak shape and the accuracy of spectrum calculation.

Thus, it is advisable to estimate the spectrum width of formant frequencies for a non-normalized signal; the latter will improve the accuracy of its determination.

Let us proceed to the analysis of the influence of the Hamming window duration on the results of the formant data evaluation. The Hamming window duration corresponded to the duration of the processed section of the voice signal. As indicated above, three variants of the frame duration of the processed voice signal were consid-


DOI 10.15588/1607-3274-2025-4-17

ered: 20 ms, 40 ms and the fully analyzed phoneme “o”. Then the spectrum of the analyzed section was calculated using the fast Fourier transform. Formant frequencies were determined based on the amplitude-frequency spectrum.

When using short frames of 20 and 40 ms, the processing was carried out in a sliding window with some overlap. The need for such a processing option will be explained below. For frames of 20 and 40 ms, the mean estimates were used as estimates of formant frequencies. The results of the processing are presented in Fig. 4.

The following notations are used in this Figure: the dashed-dotted line shows the results when processing 20 ms sections, and the dashed line shows 40 ms sections. The solid line shows the processing of the full phoneme. In this case, average estimates were used in short sections (20 and 40 ms). We will consider the estimates obtained when processing the full phoneme to be true.

Let us analyze the presented dependencies.

With a frame duration of 20 ms, after processing by the Hamming window, the mean square deviation is significantly reduced and anomalous measurements of formant frequencies are eliminated. The accuracy of spectrum calculation was approximately 32 Hz. Note that the low accuracy is caused by the insignificant number of samples in the frame. This also leads to insufficient accuracy in determining formant frequencies. As a rule, in this case we obtain high estimates.

Increasing the frame by two times leads to an increase in the accuracy of determining the formant frequencies, which tend to the estimates obtained when processing the phoneme as a single frame. The accuracy of the spectrum estimate for a 40 ms frame was approximately 16 Hz. In this case, the estimates of the standard deviation are also higher. The accuracy of determining the formant frequencies increases significantly.

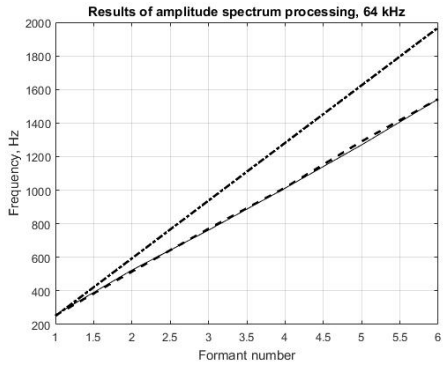


Figure 4 – The estimates of formant frequencies of the phoneme “o”

When processing the entire phoneme “o” as a single frame, the use of the Hamming window function is inappropriate. The use of this function leads to a decrease in the accuracy of the sixth formant estimate. The accuracy of the spectrum resolution in this case was 4 Hz. Errors in determining formant frequencies when processing 40 ms frames and the full frame of the analyzed phoneme are presented in Table 1. As a rule, the estimates obtained in the 40 ms frame are high. It should be noted that with an increase in the number of formants, the values of the estimates of the standard deviation for these frequencies increase.

Table 1 – Errors in determining the formant frequencies of the phoneme “o”

| F1, Hz | F2, Hz | F3, Hz | F4, Hz | F5, Hz | F6, Hz |
|---|---|---|---|---|---|
| 4.2 | 8.6 | 6.4 | 3 | 21 | 1 |

Now let us move on to the analysis of the morpheme (syllable) “dyn”. As before, we will consider the three variants of digital processing of the voice signal considered above. The results of processing the specified morpheme are presented in Fig. 5.

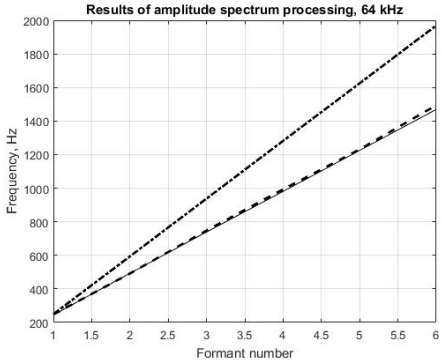


Figure 5 – Formant frequency estimates of the morpheme “dyn”

The analysis of the presented dependencies allows us to draw the following conclusions. As before, the least accurate estimates of formant frequencies are obtained when using a 20 ms frame. They are significantly high. Estimates of the average formant frequencies for a 40 ms frame are also slightly higher than the values obtained when processing the entire morpheme. The errors are presented in Table 2. Estimates of the standard deviation increase with an increase in the formant number.

Table 2 – Errors in determining the formant frequencies of the morpheme “dyn”

| F1, Hz | F2, Hz | F3, Hz | F4, Hz | F5, Hz | F6, Hz |
|---|---|---|---|---|---|
| 3 | 1 | 9 | 12 | 6 | 24 |

Here, attention should be paid to the change in the estimates of the second formant frequency for the 40 ms frame depending on the number of the implementation. This dependence is presented in Fig. 6 and makes it possible to divide the analyzed morpheme into separate sounds.


DOI 10.15588/1607-3274-2025-4-17

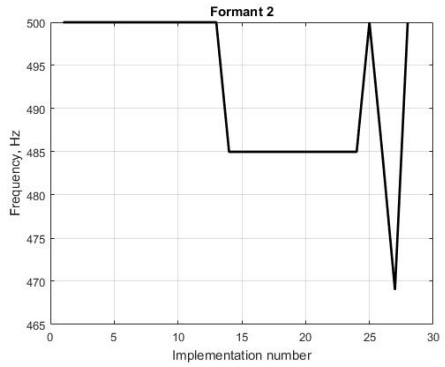


Figure 6 – Estimates of the second formant frequency of the morpheme “dyn”

Table 3 presents estimates of the analyzed formant frequencies for different source information.

When processing the word “odyn”, the estimates of formant frequencies were calculated as underestimated, since the section between syllables was not excluded. This can be seen in Fig. 7, which shows the dependence of the estimate of the first formant frequency on the number of the realization.

Table 3 – Formant frequency estimates

| Type of experiment | F1, Hz | F2, Hz | F3, Hz | F4, Hz | F5, Hz | F6, Hz |
|---|---|---|---|---|---|---|
| “o” | 258 | 524 | 762 | 1012 | 1270 | 1543 |
| “dyn” | 245 | 493 | 739 | 981 | 1225 | 1469 |
| “odyn” | 247 | 494 | 776 | 1015 | 1268 | 1515 |

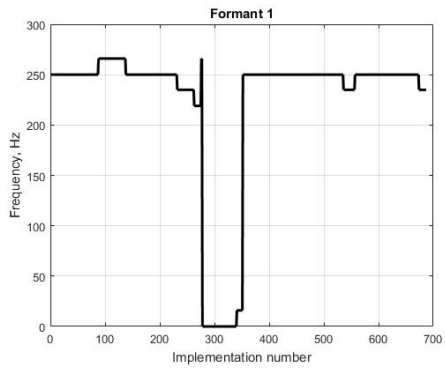


Figure 7 – Estimates of the first formant frequency of the word “odyn”

Fig. 7 shows that some estimates are equal to zero, which corresponds to the section of registration materials between syllables. The latter fact can be used to divide the registration materials into syllables.

## 6 DISCUSSION

First, let us briefly analyze the influence of normalization procedures on the generation of formant data. Along with the known factors of the influence of normalization procedures on the suppression of the noise component and the increase in the signal-to-noise ratio, in essence, the indicated procedures do not lead to a change in the formant data.

An exception is the assessment of the spectrum width of formant frequencies, which after normalization is less accurate. Although with more complex digital processing procedures, normalization can lead to a positive effect, primarily due to an increase in the signal-to-noise ratio [11]. Therefore, if it is necessary to assess the spectrum width of formant frequencies, it is advisable to perform these procedures on a non-normalized voice signal.

More interesting results have been obtained when studying the effect of the processed frame duration on formant data. Literature sources on voice signal processing recommend choosing a frame duration of 20 ms to extract user features. However, as the research results show, in this case we get overvalued estimates of formant frequencies. In this case, the spectrum resolution will also be low.

Doubling the frame length results in a significant increase in the accuracy of determining formant frequencies. More accurate estimates can be obtained by processing the full phoneme “o”. Obviously, when digitally processing phonemes, it is advisable not to divide them into frames.

When processing morphemes, it is advisable to choose a frame length of 40 ms. Using such a frame allows dividing the processed morpheme into individual sounds.

## CONCLUSIONS

The current scientific problem of studying the influence of normalization procedures and the duration of the processed frame on the assessment of voice signal formant data of the authentication system user is considered. The study was conducted using the developed mathematical model and the experimental voice signal.

**The scientific novelty** of the obtained results is in the fact that normalization procedures do not change the formant data when evaluating them. Moreover, the estimates of the spectrum width of formant frequencies are less accurate. The duration of the processed frame of 20 ms is unacceptable, since the estimates of formant frequencies will be significantly high. The frame length of 40 ms can be used to divide the processed voice signal into individual sounds. This will improve the accuracy and reliability of the extracted estimates of formant frequencies. More accurate results of formant data are obtained when processing the complete analyzed signal. Thus, the results obtained allow modifying the traditional method for preprocessing voice signals in authentication systems.

**The practical significance** of the research results is that in some cases it is possible to exclude normalization procedures with a sufficient signal-to-noise ratio. This approach will increase the efficiency of formant data assessment. When choosing the frame length, it is necessary to choose its duration of 40 ms or more. This will allow a more accurate assessment of formant frequencies. When processing phonemes, it is advisable to process the full signal, which will allow a more accurate and efficient assessment of formant data.


DOI 10.15588/1607-3274-2025-4-17

The results of the research can be applied in speech recognition, voice authentication systems, forensics, forensic examination and other applications where formant data assessments are used. Clarification of formant data allows for a better formation of a voice authentication system user template, which has a positive effect on reducing errors of the first and second kind.

**Further research** will focus on assessing the procedures under consideration for other components of the authentication system user template.

## REFERENCES

1. Pastushenko M. O., Pastushenko M. S., Petrachenko M. O. Do pytannja ocinky efektyvnosti biometrychnyh system, *Problemy telekomunikacij,* 2023, № 1(32), pp. 37–44. DOI: https://doi.org/10.30837/pt.2023.1.03
2. Rabiner L. R., Schafer R. W. Digital Processing of Speech Signals. NJ, Prentice-Hall, Inc., 1978, 512 p. URL: https://ie.u-ryu-kyu.ac.jp/~asharif/pukiwiki/attach/Acoustic%20Speech%20Signal%20Processing_Prentice%20Hall%20-%20Digital%20Processing%20of%20Speech%20Signals.pdf
3. Beigi H. Fundamentals of Speaker Recognition. NY, Springer, 2011, 942 p. DOI:10.1007/978-0-387-77592-0
4. Persson A., Barreda S., Jaeger T. F. Comparing normalization against US English listeners'vomel perception, *The Journal of the Acoustical Society of America,* 2025, Vol. 157, № 2, pp. 1458–1482. DOI: //doi.org/10.1121/10.0035476
5. Clopper C. G., Dossey E., Gonzalez R. Raw acoustic vs. normalized phonetic convergence: Imitation of the Northern Cities Shift in the American Midwest, *Laboratory Phonology*, 2024, Vol. 15(1), pp. 1–15. DOI: https://doi.org/10.16995/labphon.10893
6. Anikin A., Barreda S., Reby D. A Practical guide to calculating vocal tract length and scale-invariant formant patterns, *Springer Nature Link,* 2023, Vol. 56, pp. 5588–5604. DOI: 10.3758/s13428-023-02288-x
7. Almaadeed N. Aggoun, A., Amira A. Text-Independent Speaker Identification Using Vowel Formants, *Journal of Signal Processing Systems*, 2015, Vol. 82, № 3, pp. 345 – 356. DOI: https://doi.org/10.1007/s11265-015-1005-5
8. Aggarwal S., Vasukidevi G., Selvakanmani S., Pant B., Kaur K., Verma A., Binegde G. N. Audio Segmentation Techniques and Applications Based on Deep Learning, *Journal of Scientific Programming.* Wiley Online Library, 2022, Vol. 2022, pp. 1–9. DOI: https://doi.org/10.1155/2022/7994191
9. Lebourdais M., Mariotte T., Almudévar A., Tahon M., Ortega A. Explainable by-design Audio Segmentation through Non-Negative Matrix Factorization and Probing, arXiv:2406.13385 [eess.AS], 2024, pp. 1–5. DOI: https://doi.org/10.48550/arXiv.2406.13385
10. Lee J., Kim S., Kim H., Chung J. S. Lightweight Audio Segmentation for Long-form Speech Translation, *arXiv:2406.10549 [eess.AS],* 2024, pp. 1–5. DOI: https://doi.org/10.48550/arXiv.2406.10549
11. Pastushenko M., Krasnozheniuk Ya., Zaika M. Investigation of Informativeness and Stability of Mel-Frequency Cepstral Coefficients Estimates based on Voice Signal Phase Data of Authentication System User, *International Conference Problems of Infocommunications. Science and Technology 6–9 October 2020 (PIC S&T'2020)*. Kharkiv, Ukraine, 2020, pp. 467–472. DOI: 10.1109/PICST51311.2020.9468083

Received 11.08.2025.
Accepted 19.10.2025.

УДК 057.087.1:621.391.26

# АНАЛІЗ ПРОЦЕДУР НОРМАЛІЗАЦІЇ ТА СЕГМЕНТАЦІЇ ГОЛОСОВОГО СИГНАЛУ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

**Пастушенко М. С.** – канд. техн. наук, професор, професор кафедри інфокомунікаційної інженерії ім. В.В. Поповського Харьківського національного університету радіоелектроніки, Харків, Україна.
**Пастушенко О. М.** – канд. техн. наук, старший науковий співробітник, військовослужбовець Збройних Сил України, Україна.
**Файзулаєв Т. А.** – Директор Товариства з обмеженою відповідальністю «ТАФ-87», Харків, Україна.

## АНОТАЦІЯ

**Актуальність.** Розглядається актуальне завдання оцінки формантних даних (формантних частот, рівня їхньої спектральної щільності, огинаючу амплітудно-частотного спектру, ширини спектрів формантних частот) у системах голосової автентифікації. Об'єктом дослідження був процес цифрової попередньої обробки голосового сигналу під час вилучення формантних даних.

**Мета роботи** – оцінка ефективності традиційних процедур цифрової попередньої обробки голосового сигналу користувача та розробка пропозицій щодо підвищення якості вилучення формантних даних.

**Метод.** Розроблено математичну модель вилучення формантних даних з експериментального голосового сигналу для дослідження впливу процедур нормалізації та сегментації на якість одержуваних оцінок. Шляхом моделювання процесу отримання формантних даних порівнюються результати цифрової обробки нормалізованого і ненормалізованого голосового сигналу. Оцінюється вплив тривалості обробленого кадру експериментального голосового сигналу якість оцінки формантних частот. Результати отримані для експериментальної фонеми та морфеми.


DOI 10.15588/1607-3274-2025-4-17

**Результати.** Отримані результати свідчать, що при обробці голосового сигналу з достатнім співвідношенням сигнал/шум процедури нормалізації не є обов'язковими для отримання формантних даних. Більше того, нормалізація призводить до менш точного виміру ширини спектрів формантних частот. Неприпустимим є використання тривалості оброблюваного кадру менше 40 мс. Зазначені результати дозволяють модифікувати традиційну методику попередньої обробки голосового сигналу. Використання методу моделювання щодо експериментального голосового сигналу підтверджує достовірність отриманих результатів.

**Висновки.** Проведені експериментальні дослідження показують доцільність виключення процедур нормалізації при високому співвідношенні сигнал/шум голосового сигналу, що має місце в системах автентифікації користувачів. Такий підхід дозволить підвищити оперативність отримання формантних даних і більш точно оцінювати ширину спектрів формантних частот. Результати експериментального дослідження тривалості оброблюваного кадру голосового сигналу свідчать, що його тривалість не має бути менше 40 мс. В іншому випадку оцінки формантних частот будуть завищеними. Більш того, при обробці фонем можна голосовий сигнал, що обробляється, не розбивати на фрейми. Практичне застосування результатів досліджень дозволяє підвищити оперативність та точність формування формантних даних. Перспективами подальших досліджень може бути дослідження впливу процедур нормалізації та фреймінгу на інші елементи шаблону користувача системи автентифікації.

**КЛЮЧОВІ СЛОВА:** автентифікація, голосовий сигнал, нормалізація, сегментація, формантні дані.

**ЛІТЕРАТУРА**

1. До питання оцінки ефективності біометричних систем / [М. О. Пастушенко, М. С. Пастушенко, М. О. Петраченко] // Проблеми телекомунікацій. – 2023. – № 1(32). – С. 37–44. DOI: https://doi.org/10.30837/pt.2023.1.03
2. Rabiner L. R. Digital Processing of Speech Signals / L. R. Rabiner, R. W. Schafer. – NJ : Prentice-Hall, Inc., 1978. – 512 p. URL: https://ie.u-ryu-kyu.ac.jp/~asharif/pukiwiki/attach/Acoustic%20Speech%20Signal%20Processing Prentice%20Hall%20-%20Digital%20Processing%20of%20Speech%20Signals.pdf
3. Beigi H. Fundamentals of Speaker Recognition / H. Beigi. – NY : Springer, 2011.– 942 p. DOI:10.1007/978-0-387-77592-0
4. Persson A. Comparing normalization against US English listeners'vomel perception / A. Persson, S. Barreda, T. F. Jaeger // The Journal of the Acoustical Society of America. – 2025. – Vol. 157, № 2. – P. 1458–1482. DOI: //doi.org/10.1121/10.0035476
5. Clopper C. G. Raw acoustic vs. normalized phonetic convergence: Imitation of the Northern Cities Shift in the American Midwest / C. G. Clopper, E. Dossey, R. Gonzalez // Laboratory Phonology. – 2024. – Vol. 15(1). – P. 1–15. DOI: https://doi.org/10.16995/labphon.10893
6. Anikin A. A Practical guide to calculating vocal tract length and scale-invariant formant patterns / A. Anikin, S. Barreda, D. Reby // Springer Nature Link. – 2023. – Vol. 56. – P. 5588–5604. DOI: 10.3758/s13428-023-02288-x
7. Almaadeed N. Text-Independent Speaker Identification Using Vowel Formants / N. Almaadeed, A. Aggoun, A. Amira // Journal of Signal Processing Systems. – 2015. – Vol. 82, № 3. – P. 345 – 356. DOI: https://doi.org/10.1007/s11265-015-1005-5
8. Audio Segmentation Techniques and Applications Based on Deep Learning / [S. Aggarwal, G. Vasukidevi, S. Selvakanmani et al.] // Journal of Scientific Programming. Wiley Online Library. – 2022. – Vol. 2022. – P. 1–9. DOI: https://doi.org/10.1155/2022/7994191
9. Explainable by-design Audio Segmentation through Non-Negative Matrix Factorization and Probing / [M. Lebourdais, T. Mariotte, A. Almudévar, M. et al.] // arXiv:2406.13385 [eess.AS]. – 2024. P. 1–5. DOI: https://doi.org/10.48550/arXiv.2406.13385
10. Lightweight Audio Segmentation for Long-form Speech Translation / [J. Lee, S. Kim, H. Kim, J. S. Chung] //arXiv:2406.10549 [eess.AS]. – 2024. – P. 1–5. DOI: https://doi.org/10.48550/arXiv.2406.10549
11. Pastushenko M. Investigation of Informativeness and Stability of Mel-Frequency Cepstral Coefficients Estimates based on Voice Signal Phase Data of Authentication System User / M. Pastushenko, Ya. Krasnozheniuk, M. Zaika // International Conference Problems of Infocommunications. Science and Technology 6–9 October 2020 (PIC S&T'2020). – Kharkiv, Ukraine, 2020. – P. 467–472. DOI: 10.1109/PICST51311.2020.9468083


DOI 10.15588/1607-3274-2025-4-17