# PRACTICAL APPROACH TO SPEECH IDENTIFICATION

**Serhii Zybin, Yana Bielozorova**

***Abstract****: The rules for building a speaker identification system was described. The approaches to creation of such system was analyzed. The ways of creating such system were proposed.*

***Keywords****: speaker identification system, wavelet, coding and information theory, pattern analysis.*

## Introduction

Parameters of involvant's individual voice characteristics are the basis of every voice search system. Parameters of individual voice characteristics in modern automatic systems of speaker identification are usually determined on the basis of two main factors - the main tone frequency F0 and spectral characteristics [Solovyov, 2014]. Efficiency of such systems mainly depends on the methods for determination of these factors and stability of spectral characteristics and the main tone frequency F0.

Speaker Recognition Evaluation (SRE) carried out by the National Institute of Standards and Technologies (NIST) since 1996 is the most objective and competent source of information as to capabilities of the modern systems and methods of automatic speaker identification.

SRE allows to obtain data on real capabilities of identification methods and systems including in comparison with others and to choose the most perspective trends of development.

For the period of the last ten years the test results show considerable progress in this field. However, modern systems of automatic speaker identification are essentially behind the effectiveness of the speaker identification implemented by human acoustical apparatus.

Traditionally NIST publishes impersonal test results by means of which it is impossible to determine what identification method or system is found to be the best one. Absence (with few exceptions) of the test results data at the test participants' websites usually points out the poor results or restricted developments which are

often performed for authorities ensuring the state security or for commercial purposes.

At the same time achieved level of developments and progress in outlined perspective directions in this field allow to proceed with the stage of collective developments of such systems within the frames of the EU countries.

Complex long-term investigations of scientific and research groups in this area in Ukraine allowed creating an experimental model of modern system for search of involvants in the voice database [Rubalsky et al, 2014]. Further, there is presented a physical and mathematic model based on which the investigations were performed, and experimental model was developed.

## Model of voice characteristics identification

We will consider fragments of speech in audio data as discrete time series of the amplitude of a sound wave. Let us consider the problem of determining the characteristics of self-similar structures in a time series. These can be various geometrically similar structures, visually observed when examining the graphs of changes in the amplitude of the sound wave. We will consider self-similarity as a geometric similarity associated with transformations of compression, extension, both along the time axis and along the amplitude coordinate. To reveal approximately similar structures, we use the methods of wavelet analysis. For this purpose, we will use the complex Morlet wavelet [Bielozorova, 2019], [Solovyov et al, 2014], [Solovyov, 2013].

Figure 1 shows an illustration that combines fragments of an audio recording of speech and a scalograms built on the basis of the model under consideration. The features of the scalograms construction were considered in the work [Bielozorova, 2019], [Solovyov et al, 2014], [Solovyov, 2013].

The analysis shows that the location of the tops of the scalograms in terms of the time parameter in Fig. 1 strictly corresponds to the local maxima of the amplitude of the sound wave in the time domain. In this case, the local maxima correspond to bursts of the amplitude of the sound wave due to the frequency F0.

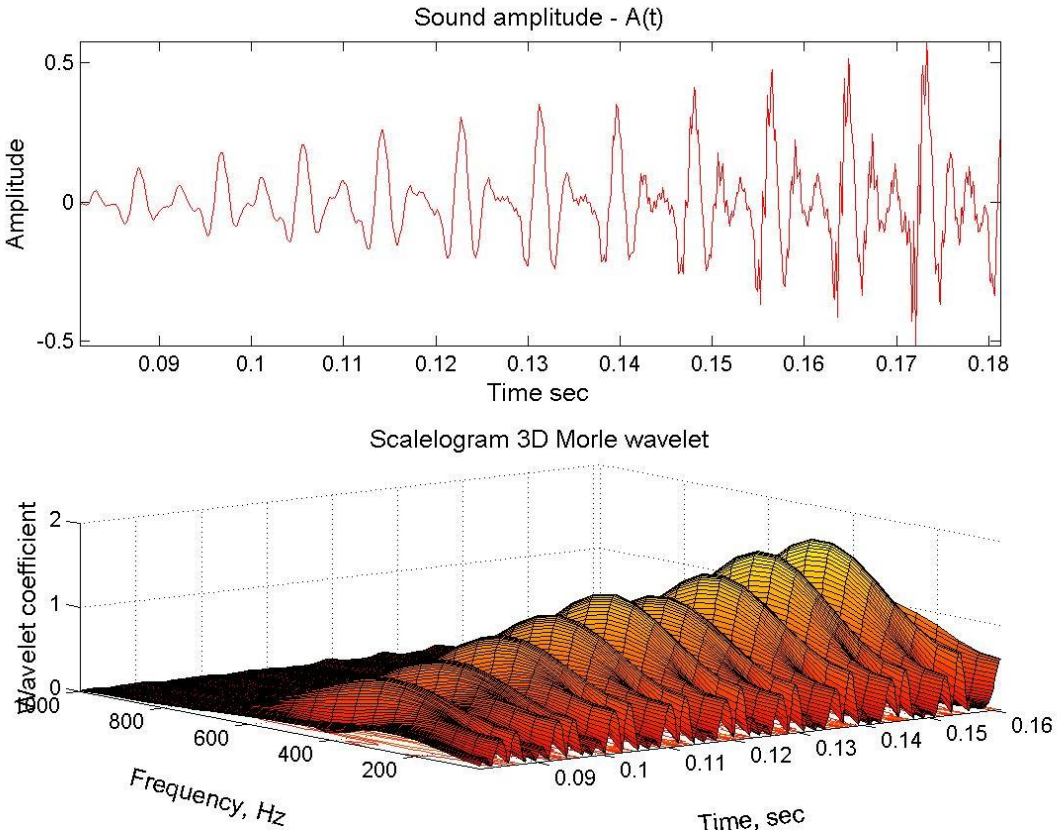



Fig.1. Scalogram of a voice signal using the complex Morlet wavelet

An essential feature of the characteristics of the ridges is the shape of the ridge. Studies show that after rational normalization of the description function, the value of the modulus of the wavelet transform coefficient at fixed parameters: the position of the time window in time and the width of the wavelet at the peak, these functions have a high degree of geometric similarity. At the same time, the shape of the normalized peaks is individually different when the characteristics of the voice differ.

The reason for the similarity of structures of the type of peaks in the frequency domain on the scalogram for the same characteristics of the voice is the specificity of the affinity of the Morlet wavelet to the temporal structure of the amplitude of the sound wave in the regions of local maxima. The Morlet wavelet transform effectively separates these structures with this approach. The similarity under consideration has a very transparent physical interpretation. The structures of the amplitude of the sound wave in the region of local maxima corresponding to the frequency F0 have a fairly pronounced geometric symmetry with respect to the amplitude of the local

maximum. In this case, the Morlet wavelet, due to its affinity, makes it possible to reveal this symmetry in the form of pronounced extrema of the scalograms.

In fig. 2 and fig. 3 two-dimensional slices of the spatial scalogram in frequency and in time are presented. The smoothness of these dependencies illustrates the thesis about the reduction of the mathematical complexity of identifying self-similar structures in the time-frequency domain with the considered approach.

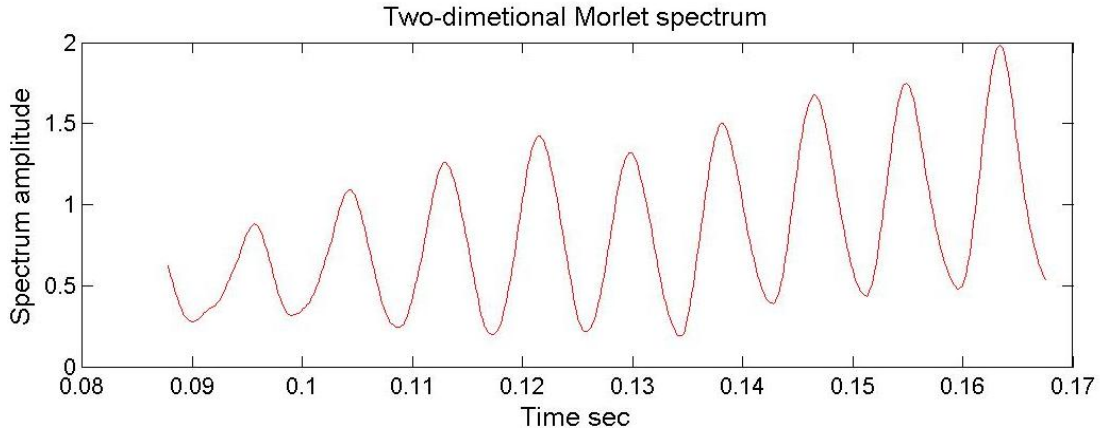



Fig.2. Two-dimensional time scale chart

To extract speech characteristics, the following approach was used:

1. Divide signal into short overlapping frames (frame-20msec, overlapping-10msec)

2. Calculate spectrogram for each frame by using complex wavelet Morle (as fig.2). A feature of the applied method for calculating the spectrum is that the calculation of the spectral characteristics is carried out with a resolution of 1 Hz (The wavelet frames were used in the work, which allow obtaining a higher frequency resolution of signal, instead of the standard wavelet decomposition with a frequency resolution of 50 Hz).

3. Extracts two-dimensional frequency scale chart (as fig.3) from 7 local maxima of spectrogram for each frame (A feature of this stage is the selection of two-dimensional slices only for the positions of the time maxima of the two-dimensional slices of the scalograms in time. That is, the frequency dependences of the wavelet coefficients are considered only at the peaks).

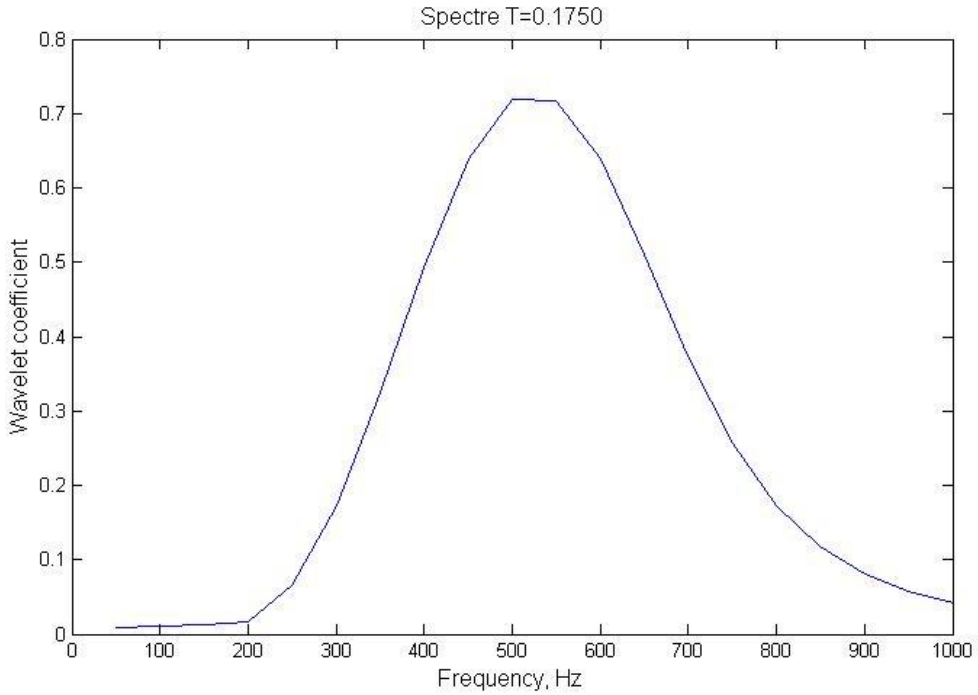



Fig.3. Two-dimensional frequency scale chart

The frequency range of the frequency F0 is divided into intervals in the range from 100 to 1500 Hz. For all fragments of speech, the probability density is determined in terms of frequency F0 and the function of voice characteristics Fc (see Fig. 1). Thus, the system is based on a combined estimation method based on the frequency F0 and spectral characteristics. Proximity of the two characteristics of the voice are determined by the absolute differences between two-dimensional probability densities. Of course, the projection of two-dimensional probability densities on each of the one-dimensional coordinate axes gives the frequency distribution F0 and the distribution of the function of voice characteristics Fc, which can also be represented as separate dependencies. This allows you to calculate the proximity of the curves of the probability density functions for each of these features separately.

The analysis of these dependences showed that at different values of the magnitudes of the maxima in frequency, when combining the graphs of frequency dependences in the position of the maxima, they have a high degree of similarity after the corresponding normalizations for the same characteristics of the voice. The final acceptance of the hypothesis about the identity of the characteristics of the

voice of two speech fragments is made after testing the hypothesis about the identity of the distributions of frequency dependences.

In the identification process, a comparison of the parameters of the distributions selected in the preparatory stage will be performed.

There are 2 main approaches to building systems based on the method under consideration:

1.The correlation of the normalized temporal spectrograms (point 2 of the preparatory stage for the selection of characteristics) of each frame of one file with the frames of another file is estimated (by searching for the Pearson's correlation coefficient or calculating the maximum of the correlation function). In the case of high correlation characteristics of the normalized time spectrograms, the characteristics of the frequency scalograms of these frames are compared (point 3 of the preparatory stage) and a decision is made on the similarity of the speakers in 2 files.

2.Evaluation of the similarity of speakers is carried out on the basis of a separate consideration of the probability density distribution of the pitch frequency and the function of voice characteristics as described above

The Experimental studies of speech files and speech fragments based on the developed model have shown the effectiveness of the approach in solving problems of identifying voice characteristics.

## Scientific novelty of investigations and development

In view of importance and commercial nature of investigations and developments of leading designers in the field of voice characteristics identification, there is not enough information about detailed principles of modern system operation in open access publications. However, based on the analysis of the many years investigations, it is possible to suppose that the methodology of investigations and developments, described in above sections, is a complex new scientific approach to tasks of voice identification.

Scientific novelty of the approach under consideration is an application of complex of the following mathematical methods and technologies of digital processing of acoustic audio information:

1 complex discrete two-parameter Morlet transformation;
2 non-orthogonal discrete transformations with discrete frequency step of 1 Hz;
3 application of the wavelet transforms maxima approach for determination of the features of frequency characteristics;
4 determination of connection of frequency characteristics features with the frequency F0 and other parameters of voice characteristics;
5 analysis of voice information in small time intervals – 10-30 ms.

## Essence of the innovative nature of the project

Confidential nature of the modern most advanced developments in the field of identification of the voice characteristics creates essential problems both for accelerated evolution of these systems and for wide application in the intergovernmental voice search databases.

The present development, if the practical efficiency is proved, can be distributed in any EU country with appropriate version of language localization.

Innovative nature of the proposed project consists of posing the task for creation of prototypes of more open modern systems for identification of voice characteristics within the frame of voice databases for specialized structures.

The second important moment ensuring innovative nature of the project from our point of view is new technologies and models for processing of discrete voice information.

## Bibliography

[Solovyov, 2014] V. Solovyov. Spectral analysis and speech technology (Russian) // V. Solovyov, O. Rubalskiy / Journal of Kiev National University T. Shevchenko, Military special sciences, - Kiev, Vol, 42, p. p. 145-151, 2014.

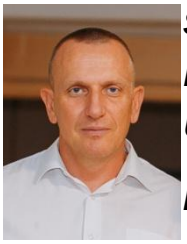
[Rubalsky et al, 2014] O. Rubalsky, V. Solovyov, A. Shablja, V. Zhuravel. New tools to identify the person for voice of Data (Russian) Protection and Security of Information Systems: Proceedings of the 3rd International Scientific Conference (city. Lviv, 05 - June 6, 2014). - Lviv: - p.p.. 110 - 112.

[Bielozorova, 2019] Bielozorova Y.: Analyse and develop the software of automatic search for an anonymous person in the voice database//ITHEA International Journal "Information Technologies & Knowledge" Volume 13, Number 2 -2019 – p.p. 152-164

[Solovyov et al, 2014] V. Solovyov, Y. Byelozorova: Multifractal approach in pattern recognition of an announcer's voice. Polish Academy of Sciences University of Engineering and Economics in Rzeszów, Teka, Vol. 14, no 2, p.p. 164-170, 2014.

[Solovyov, 2013] V. Solovyov. Using multifractals to study sound files (Russian). Visnik of the Volodymyr Dahl East Ukrainian national university. Vol 9 (151). – p. p. 281-287, 2013.

## Authors' Information



***Serhii Zybin*** *– DSc, Associate Professor, Head of Software Engineering Department,National Aviation University, Kyiv, Ukraine.*

***E-mail****: zysv@ukr.net*

***Major Fields of Scientific Research****: game theory, cloud computing, cyberspace, distributed nets, software engineering, information security*

***Yana Bielozorova*** *– Senior Lecturer 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*