ANALYSE AND DEVELOP THE SOFTWARE OF AUTOMATIC SEARCH FOR AN ANONYMOUS PERSON IN THE VOICE DATABASE

Yana Bielozorova

Abstract*: Objective of the proposed development is to investigate efficiency and develop a program system for formation of the voice databases of information messages involvants. Automatic search for involvants in the database by means of similarity of the voice characteristics.*

Keywords*: Voice database, probability density, main tone, phonogram.*

Objectives

Sound recording materials are considerable part of evidences used during criminal investigations and proceedings related to corruption, bribery, extortion, racketeering, kidnapping etc. But there are special digital phonograms which contain messages being recorded, for example, in different monitoring and field services. They form a separate type of expert sound recording materials. It can be explained by the fact that the phonogram digital copy is usually used during the expert examination of the materials and examination of originality and authenticity of phonograms is not performed. Another feature of such phonograms is short-duration of the recorded messages.

Such messages are combined and stored in the voice databases. Very often the messages are anonymous. In many cases such messages are made by the same persons but there can be considerably long period between messages made by the same anonymous author. Very often these messages cause significant property and moral damage. On the other hand, sometimes anonymous messages can be related to terrorist activity or have serious social consequences. In view of these reasons sometimes it is required to identify anonymous person (involvant). And during operative investigations, if they are

required, at initial stage it is very important to determine whether the messages of the involvant are of repeated nature. And in this case assistance for investigation and retrieval can be performed in expert way. The first task of such expert examination is to identify voices of involvants with similar parameters of signals of their oral speech, contained in the database. This can reveal both recidivism and potentially suspected persons and thus facilitate investigation process. The main tasks of such expert examination include:

- personal identification of the involvant, suspected in delivery of anonymous message or his available voice tokens, obtained during operation by means of parameters of his speech signals;
- identification of the involvants' voices with similar parameters of signals of their oral speech, contained in the database in order to identify persons, potentially suspected in delivery of information (or false information).

Appropriate expert departments can ensure completion of these tasks. However, the database of these messages even within the same country can contain thousands of records, and experts are not able to perform this task without automation of the process. In order to facilitate the involvants identification task it is required to develop automated system of quick search for suspected persons in the available voice database for further personal identification through the physical parameters of his (her) oral speech signals.

Model of voice characteristics identification

Spectrum analysis of audio data based on Fourier transform mathematical apparatus is the key factor in majority of modern systems for solving the tasks of speaker identification by means of voice characteristics. It is caused by several factors. On the one hand - by known neurophysiological rules of audio information processing by primary auditory receptors. On the other hand - by absence of more effective analysis methods and, to a certain extent, by historical traditions in this field.

However, in spite of creation of rather effective systems for voice characteristics identification and development of systems of identification and text entering of

voice information, there is no enough clearness in the principal theoretical and practical issues concerning speech technology till now.

In spite of a lot of investigations in this field and use of powerful computers for last twenty years there is no principal breakthrough in the field of physical and mathematical concepts for effective processing of voice information (comparable with effectiveness of acoustical apparatus).

In opinion of many specialists, mainly, it is caused by absence of effective mathematical tool for analysis of voice audio information.

Concept implemented in the experimental system of instrumental identification of speaker voice characteristics is described below. This concept is based on presentation of speech fragments as a set of multifractal structures. In order to determine parameters of multifractal structures the wavelet analysis with special basis in form of two-parameter Morlet wavelet is used.

Let’s consider the fragment of voice audio file shown in Fig. 1 (fragment of phoneme [a]).

In the years since H. Helmholtz [Helmholtz, 1863] there has been known the evident fact, mentioned several times by investigators (particularly the classic work of G. Fant [Fant, 1960]), that majority of phonemic structures of speech can be formed on the basis of similar geometric components (“atomic” structures) of acoustic wave. Usually these structures are limited by time intervals which are equivalent to the main speech tone frequency. Geometric similarity of these structures is approximate, but in many cases it is visually evident (see Fig.1). Time intervals seized by these structures are opposite to the frequency value of the main tone and located within range from 2 to 15 ms. The present atomic structures, considered independently, are not perceived by ear due to their short duration of sound.

It is evident that following the works of Mandelbrot these structures can be interpreted as atomic components of multifractal [Mandelbrot, 1972, Mandelbrot, 1982, Mandelbrot, 1999, Mandelbrot, 1969]. Provided that there was created a mathematical model ensuring effective determination of parameters of “atomic” structures and multifractal in whole, the correct

description and solution of all the main tasks of voice and speech identification at the phonemic level may be expected.

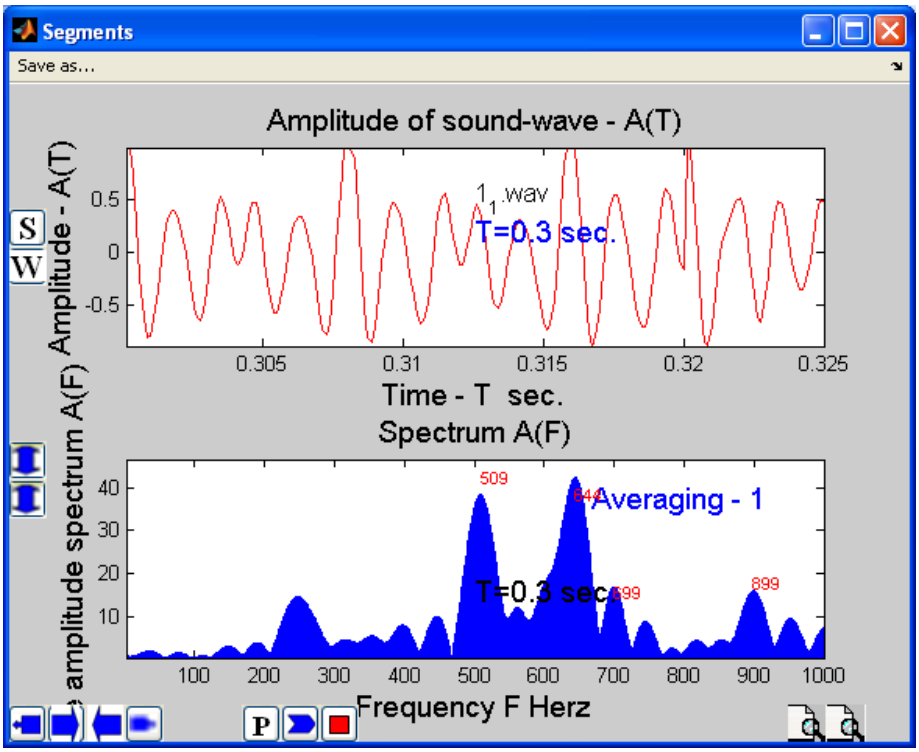

Figure 1. Fragment of phoneme [a]

Till now, actually, the single methodological direction for building the mathematical models for calculation of phoneme parameters and voice characteristics is the application of some of modifications of spectrum analysis based on Fourier transform. However, the application of this spectrum analysis method has number of serious disadvantages (known in the years since Helmholtz). Analysis of formant spectrum structure in small time intervals (approximately 20 ms) has considerable disadvantages caused by precision of formant localization. It is the sequence of Heisenberg principle for information technologies and features of orthogonal discrete Fourier transformation with fixed frequency step [Mandelbrot, 1969]. Thus, for the most interesting intervals within the range of 10 to 20 ms such analysis requires the frequency step from 50 to 100 Hz correspondingly.

It is known that wavelet analysis is an alternative to Fourier spectrum analysis [Mandelbrot, 1969]. However, many attempts to use wavelet analysis methods for processing of voice information has not produced important results up to now.

On the one hand, the obstacle is the complexity of physical interpretation of investigation results for the majority of wavelet bases. On the other hand, application of wavelet bases which are similar to utterances (for example, Morlet basis) is complicated due to high computational complexity for two-parameter bases.

Let's consider utterances in audio data as a discrete time series of acoustic wave amplitude. We'll set the task to extract the characteristics of the self-similar structures in the obtained time series for atomic utterances extracted in the previous section. For identification of similar structures the wavelet analysis methods are used [Mandelbrot, 1969]. For this purpose complex Morlet wavelets are selected [Mallat, 1999].

$$C_{mor}\left(t_i, T_k, F_b, F_c\right) = \left(\pi F_b\right)^{0,5} \exp\left(2j\pi F_c t_i\right) \exp\left(-\left(t_i - T_k\right)^2 / F_b\right) \quad (1)$$

F_b – parameter the width of the wavelet, F_c – the central frequency of the wavelet, t_i – digital time samples, T_k – timing corresponding central part of the time window, j - complex unit.

Suppose $A\left(t_i\right)$ – is the value of acoustic wave amplitude of utterance in audio file at the time moment t_i. Let's consider the time slot of utterance with δT interval lower than 20 ms. Width parameters of complex Morlet wavelet F_b is selected as constant for all transformations and based on experimental investigations. Its value was selected from the condition of practical attenuation of Morlet wavelet absolute values at $t_i - T_k$ values equal to $\delta T / 2$. Let's determine convolution of Morlet wavelet for every utterance with fragment of the time series of acoustic wave amplitude in the following form:

$$C(T_k, F_b, F_c) = (1/N)\, abs\left(\sum_{t_{ij}=0}^{N_m} C_{mor}(t_i, T_k, F_b, F_c) \otimes A(t_i) \right) \quad (2)$$

$C(T_k, F_b, F_c)$ – the value of the wavelet transform coefficient module, N – the number of discrete samples in the interval δT time window.

If complex Morlet wavelet width parameter F_b is fixed, the module value shall be the frequency function F_c of Morlet wavelet and position of the time slot in time – T_k. Typical diagram of space skeylogrammy $C(T_k, F_b, F_c)$ in function F_c and T_k for utterances, which are under consideration, is shown in Fig. 2. This diagram is three-dimensional Morlet spectrum for phoneme fragment [a] shown in Fig. 1.

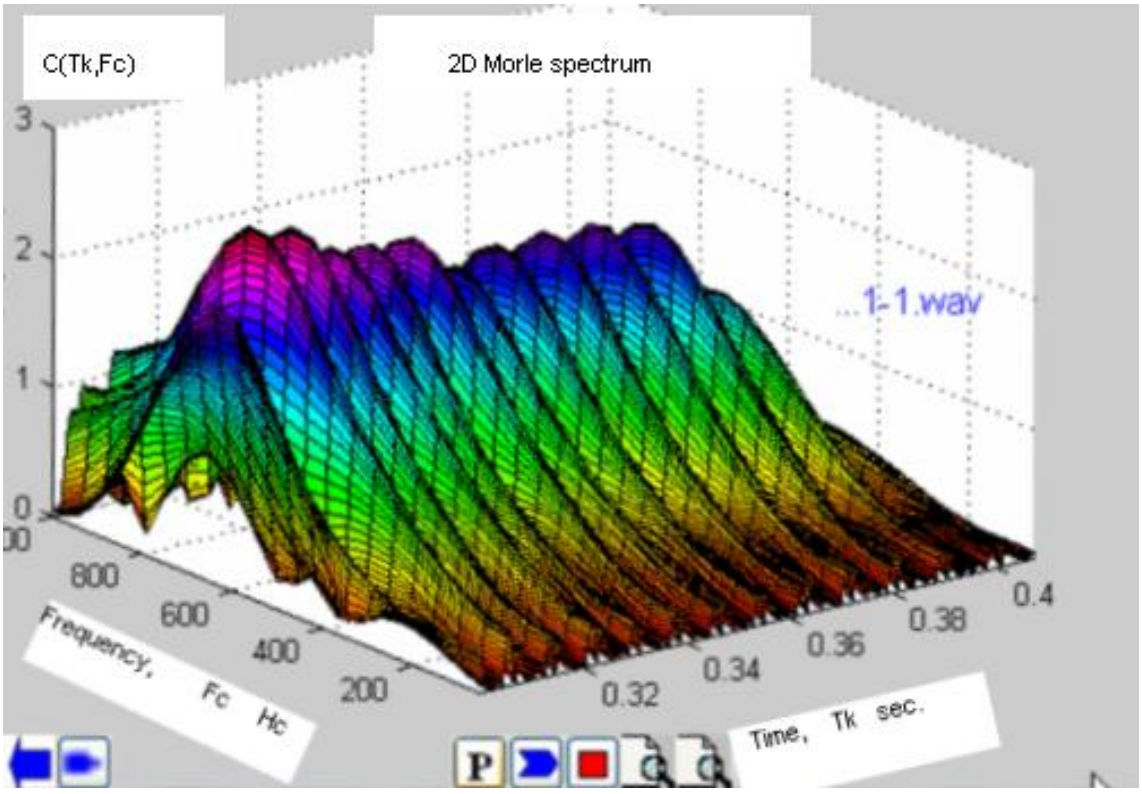

Figure 2. Utterance skeylogrammy

Time and frequency presentation of the utterance as a space skeylogrammy based on Morlet basis has a number of important features which allow to significantly increase the efficiency of identification of self-similar structures. Particularly, local maxima of wavelet transformation are rather informative for analysis of atomic components of multifractals in audio data.

It is important to mention the following investigation moments for further analysis. Frequency step F_c, further on is considered as independent of sizes of transformation window. It will be the step equal to 1 Hz. Frequency resolution according to Heisenberg principle [Mallat, 1999] doesn't impose considerable restrictions to formant structure analysis in small time intervals. The fact is that the distance between formant maxima in speech spectrum in small time intervals corresponds to the frequency of the main speech tone (approximately 85-600 Hz). And similarity frequency resolution of two maxima in transformation interval of 20 ms is approximately 50Hz. If transformation frequency step is 1 Hz the precision of distance assessment between formant maxima in small time intervals is of the same order. This precision of assessments doesn't contradict Heisenberg uncertainty principle for information technologies.

It is also important to mention, that the use of Morlet wavelet doesn't pose the problems of physical interpretation of frequency transformations. Because this wavelet is a short time Fourier transform with a Gaussian function [Mallat, 1999].

Analysis of skeylogrammy shows that arrangement of skeylogrammy ridges under the time parameter in Fig.2 strictly corresponds to local extremes of acoustic wave amplitude in the time domain. At that, these local extremes correspond to acoustic wave amplitude bursts, caused by the main tone frequency.

The important factor of high similarity of Morlet basis with the self-similar structures in utterances is the higher level of skeylogrammy smoothness comparing to, for example, analogous Fourier transform. The higher level of function smoothness ensures rather effective possibility for mathematical analysis of ridges parameters.

At that, frequency formant maxima can be very different from their maxima in Fourier spectrum.

Distances between local skeylogrammy maxima in frequency size are the assessment of the main tone speech frequency in this approach. The important factor for stability and reliability of assessments of the main tone frequency for this methodology is a possibility to estimate the main tone frequency not only by local maxima, but also by correlation between fragments of maxima areas. In small intervals these areas are approximately self-similar structures. During the analysis of such self-similarity it is possible to cut at random the structures of the same size of the time slot and not be guided by skeylogrammy maxima [Solovyov, 2014].

Thus, for example, let's consider utterance of 1 s duration, at a sampling frequency of 44100 Hz. We'll analyze its spectral characteristics and parameters of the main tone frequency using the time slot of 20 ms based on the studied model. Let's accept a minimum possible discrete step of the analysis window displacement equal to the reciprocal value of the sampling frequency. Then, a number of assessments of the window spectra and parameters of the main tone frequency for 1s will be approximately 44080. These values are correlated between each other. But such a number of statistics for the short-duration utterances allow obtaining the reliable and stable assessments of speech characteristics of short-duration messages.

The developed model allows implementation of the effective identification of utterance atomic components in frequency-time area. It is possible to build the stable and reliable frequency characteristics and assessments of the main tone frequency and spectral characteristics of short-duration utterances.

Experimental program modulus based on developed approach

Experimental model of the program product, created on the basis of investigations, for realization of these tasks was approved in the expert departments of the Ministry of Justice and the Ministry of Internal Affairs of Ukraine.

This approach was realized and approved using the developed experimental program modulus for search for involvants of information messages in the voice database [Rubalsky et al, 2014].

Experimental program system (EPS) for search of involvants of information messages has a form of a search engine. Thus, the search results are not identical to identification of person by voice, because they are just results of ranking by the similarity degree of separate parameters of voice signals.

Using digital records of information messages the EPS performs an automatic calculation of parameters of voice characteristics and further ranking of these characteristics in the voice database.

The EPS uses a ranking method by four different criteria. They include:

- calculation of similarity of two-dimensional probability density functions curves for the main tone frequency (MTF) and arrangement in the spectrum of seven formants, extracted from the speech recorded in the phonogram;
- calculation of similarity of probability density functions curves for each of these signs separately;
- calculation of similarity degree for the absolute maxima of formats spectra, extracted from the speech recorded in the phonogram.

Peculiarity of the used method for spectrum calculation is that the calculation of spectrum characteristics is performed with resolution ability of 1 Hz (non-orthogonal transformations in a small time slot). Maxima of the first seven formants are distinguished in the spectrum of each time slot.

For calculation of similarity of two-dimensional probability density functions curves, the function Ff – voice characteristics function should be calculated. This function is a non-linear empirical function of formant maxima amplitude and frequency.

For all utterances the two-dimensional probability density is determined by means of two coordinates – MTF Fb and voice characteristics function Ff. Thus, a combined method for assessment based on MTF and formant spectrum characteristics is the fundamental element of the system. Similarity degree of

these characteristics of two voices is determined by absolute differences between the two-dimensional probability densities.

It is clear, that projections of two-dimensional probability densities per each of one-dimensional axes of coordinates provide distribution of MTF F_b and distribution of voice characteristics function F_f, which also can be presented in a form of separate dependences. It allows making calculation of similarity of probability density functions curves for each of these signs separately.

For ranking by the densities of distribution of the main tone F_t and voice characteristics functions F_f, the value equivalent to Kolmogorov fitting criterion is used as a similarity degree [Mallat, 1999]:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{Xi} \le x \tag{3}$$

Where $I_{Xi} \le x$ indicates whether the value $X_i(-\infty, x]$ on condition

$$I_{Xi} \le x = \begin{cases} 1, & X_i \le x \\ 0, & X_i > x \end{cases} \tag{4}$$

For calculation of similarity degree of absolute maxima of formants spectra, extracted from the speech, recorded in the phonogram, the following criterion is used

$$S_i = abs\left\{\left[\max P_1(x_1, x_2) - \max P_2(x_1, x_2)\right]\right\} \tag{5}$$

where $P_1(x_1, x_2), P_2(x_1, x_2)$ – are two-dimensional probability densities for two different records with arguments both as the main tone frequencies and voice characteristics functions, compared separately.

Results of search for voices with similar characteristics are presented as rank tables.

It should be mentioned, that all calculations of voice characteristics are made for the record fragments containing speech signals. Division into pauses and speech fragments in the system is made automatically based on two criteria – level of normalized audio signal and availability of record fragment which contains the speech sound [Solovyov et al, 2014, Solovyov, 2013, Solovyov, 2013]. System for sound extraction is additional and integrated into the EPS. It performs general division of the fragments in which the speech sounds are present, based on the special method, similar to the methodology of hidden Markov chains.

Conclusion

An experimental program system ensuring automatic search in the database of persons with voices which are the most similar to the involvant's voice, recorded in the phonogram pattern was created.

However, further improvement of the system in order to create the prototype requires carrying out of many investigations and approval of the program system. In particular, it requires:

- accumulation of large databases with different language groups presented;
- investigation of efficiency of voice characteristics ranking in the databases;
- correction and adaptation of the program product to conditions of different real tasks of recording of information messages involvants' voices;
- creation of multi-language versions of program product localization.

Bibliography

[Helmholtz, 1863] Helmholtz H. von, Die Lehe von Tonempfindungen. Brannschweig, Vieweg, 1863.

[Fant, 1960] Gunnar Fant. Royal Institute of technology Stockholm MOUTION & GO. 'S-GRAVENHAGE 1960

[Mandelbrot, 1972] Mandelbrot B. Statistical Methodology for Non-Periodic Cycles: From the Covariance to R/S Analysis. Annals of Economic Social Measurement 1, 1972.

[Mandelbrot, 1982] Mandelbrot B. The Fractal Geometry of Nature. New York: W. H. Freeman, 1982.

[Mandelbrot, 1999] Mandelbrot B. A Multifractal Walk Down Wall Street. Scientific American, 1999.

[Mandelbrot, 1969] Mandelbrot B.B. Robustness of the rescaled range R/S in the measurement of non-cycling long-run statistical dependence // Water Resources Research. 1969. V. № 5. P. 967-988.

[Mallat, 1999] Mallat S., A wavelet tour of signal processing, Courant Institute, New York University, 1999, 671 pp.

[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 (sity. Lviv, 05 - June 6, 2014). - Lviv: - p.p.. 110 - 112.

[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.

[Solovyov, 2013] V. Solovyov. Using the fractal dimension of audio files in the problem of segmenting the audio file (Russian). Visnik of the Volodymyr Dahl East Ukrainian national university. Vol 5 (194). – p. p. 165-169, 2013

Authors' Information

Yana Bielozorova – Senior Lecturer of Software Engineering Department, National Aviation University, Kyiv, Ukraine.

E-mail: bryukhanova.ya@gmail.com

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