Identification and human condition analysis based on the human voice analysis

Oleksandr Yu. Mieshkov*[a], Oleksandr O. Novikov[a], Vsevolod O. Novikov[a], Leonid S. Fainzilberg[b], Andrzej Kotyra[c], Saule Smailova[d], Ainur Kozbekova[e], Baglan Imanbek[f]

[a]Kherson National Technical University, 24 Beryslavske str., Kherson, Ukraine; [b]International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, 40 Acad. Glushkova ave., 03680 Kyiv, Ukraine; [c]Lublin University of Technology, Lublin, Poland; [d]D.Serikbayev East Kazakhstan State Technical University, 69 A.K. Protozanov str., Ust-Kamenogorsk, Kazakhstan; [e]Institute of Information and Computational Technologies, 125 Pushkin str., 050010, Almaty, Kazakhstan;

[f]Kazakh National Research Technical University after K. I. Satpaev, 22 Satpayev str., Almaty, Kazakhstan

ABSTRACT

The paper presents a two-stage biotechnical system for human condition analysis that is based on analysis of human voice signal. At the initial stage, the voice signal is pre-processed and its characteristics in time domain are determined. At the first stage, the developed system is capable of identifying the person in the database on the basis of the extracted characteristics. At the second stage, the model of a human voice is built on the basis of the real voice signals after clustering the whole database.

Keywords: voice signal, fundamental frequency, voice structure

1. INTRODUCTION

Nowadays, there are a number of systems for human condition analysis. The trend in the development of such systems is the creation of complex biotechnical systems that perform human condition analysis, based on a number of technical criteria, characteristics, etc. At the same time this system requires high requirements regarding precision, reliability of results, as well as non-invasiveness.

Human voice signal can be actively used as the creation basis of such a system. It is generally known that any change in human condition to some extent is displayed on human voice. According to these changes it possible to determine the nature of changes that take place in human health. At the same time voice technology is already widely used for the task of personal identification.

Personal identification by voice signal is one of the modern trends in biometrics that is actively developing. Voice identification is the process of personal determination by comparison of voice sample with the templates or standards stored in the database. The task of human voice signal analysis in whole is difficult even for modern computerized systems. Therefore, for the purpose of personal identification human voice signal is parameterized, i.e. interpreted as a small set of information-bearing parameters. In this case, algorithms based on Fourier transforms are usually applied. The results of practical researches of systems that use these parameterization algorithms indicate, that the percentage of speaker identification is greater than[1] 98%. At the same time, when these algorithms are used on real communication channels, identification accuracy rarely exceeds[2] 90%.

* alexunder.meshkov@gmail.com

Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, edited by Ryszard S. Romaniuk, Maciej Linczuk, Proc. of SPIE Vol. 10445, 104453T
© 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2281003

There are also several different extraction methods of human voice characteristics that are used for identification. These characteristics usually include the pitch frequency of voiced sections which are extracted both in time[3-5], and in the frequency domain[6]. Most of the studies are based on frequency domain, however, time domain is used rather frequently since it can take into account one more information parameter – the structure of the signal amplitude distribution[7].

At the same time the use of the voice signal for the task of human condition diagnostics is developing now. In most cases the diagnosis problem involves the formation of the so-called standard voice – digital or analog signal, which is compared to the actual voice signal during the analysis. To simulate the voice signals, some of the authors, as e.g. G.Fant, J.Flanagan, have proposed electric equivalents of vocal apparatus and the "source-filter" model[8]. These analogs were based on human anthropometric parameters, but have not taken into account the individual characteristics of the human body. Some of researches concern determination of correlations between the parameters of human vocal tract and, consequently, the voice signal, and individual anthropometric parameters. These parameters generally include height, weight, age and gender characteristics of a person[9-13]. A number of studies aimed at the methods of certain diseases diagnostics have been found[14,15].

The purpose of the research is to develop a biotechnical system that on the first stage performs personal identification and on the second performs human condition analysis. All operations in this system are based on the human voice signal analysis.

To achieve the settled purpose the following tasks must be solved:

- to consider the structure of the human voice signals and their main characteristics;
- to develop an algorithm for extraction of voice signal structure elements;
- to develop an algorithm for personal identification on the basis of these structure elements;
- to propose a mathematical model of human voice signal that takes into account individual characteristics of human voice, anthropometry and other parameters;
- to verify the possibility of using of this model to the real voice signals for the task of analytic comparison.

1.1. The method applied

Human voice is a complex acoustic signal that is generated by the human vocal apparatus. This signal can be described as a consistent set of separate units called phonemes. Each phoneme is characterized by its amplitude, frequency and spectral characteristics that are distinctive for each phoneme. These characteristics are formed by modulation of airflow that is blown out of lungs due to oscillations of the vocal folds and configuration of articulation apparatus.

For the task of personal identification authors propose to extract phoneme /A/ from the voice signal and to perform analysis of its characteristics. The choice of this phoneme is associated with the fact that this sound is the most informative. Many lung diseases, such as of cardiovascular or nervous system can be analyzed on the basis of analyzing this sound.

During the research it was found that each phoneme in single pronunciation consists of a set of consecutive quasi-periodic oscillations, which are called frames[6]. Also at the beginning and end of the signal the transient processes, known as voice attack and voice damping, are observed. All frames of the signal cannot be used for the task of analysis and personal identification. Therefore, an algorithm was developed for extraction of required number of frames from the voice signal flow. This is achieved by cutting off the voice attack and voice damping at a certain voice level. As a result, this algorithm forms the signal which consists exclusively of information-bearing frames which are suitable for analysis.

For the task of parameterization of this signal the determination of such characteristics as voice fundamental frequency and signal amplitude distribution in time domain is proposed. These characteristics can be determined both for the entire set of frames and for each separate frame. The last variant is the most relevant for the selected characteristics since they both change over time, even within a single sound, from frame to frame. Therefore, another parameterization parameter that is proposed to use is a dynamics of the voice fundamental frequency in time and the signal amplitude distribution in time domain.

To determine these characteristics, it is necessary to divide the analyzed signal into separate frames. For this purpose, the algorithm for searching the division frame point was developed. The approximate position of this point is defined by the values of the average frame duration, and is clarified by the search of signal transition point from the negative to the positive region. As a result this algorithm forms the so-called cloud of separate frames. Each frame has its duration, and,

consequently, the frequency and the structure of signal amplitude distribution in time domain. In general, it supposed that every single frame of derived cloud can be used for the further analysis.

Since each of the received frames has different duration, it may complicate the further analysis process and, in particular, personal identification procedure. Therefore, in the last stage of the signal pre-processing each frame is scaled to a clearly defined duration. The structure of the signal remains unchanged. As a result, the frequency of each primary frame and the amplitude distribution of scaled frames are recorded in a special database.

The database is composed as a two-dimensional array in which sample amplitudes of each scaled frame and its primary frequency are recorded. A serial number of the speaker corresponds to each entry. Along with the described array original acoustic recordings of all speakers in *.wav format are saved in the database. This is made for the possibility of recovering the database in case of data loss.

1.2. Personal identification on the basis of extracted characteristics

For the task of personal identification, it is necessary to compare the input signal with the acoustic signals stored in the database on the basis of extracted characteristics. The input signal is processed by the same algorithm and the received frame cloud averages. After this a two-dimensional feature space is built. The taken features are signal fundamental frequency and signal amplitude distribution in time domain. The last feature is defined as a root mean square deviation between the input signal and the database one[16]. Each input signal, divided into frames, is represented in this space as a point with the coordinates:

- the difference between fundamental frequency of the input frame and the database one;
- the RMS deviation between the input signal and the database one.

Both characteristics are normalized by maximum value before the feature space formation. According to the fact that the database contains a large number of speakers, and thus voice frames, the comparison of average input signal frame will form a cloud of points in the given space. Next, the distance from the average input signal frame to each database frame is determined[15]. In this case, the distance threshold value is settled in the system. With the distance exceeds this value, the system decides the absence of the speaker in the database. Therefore, an obligatory condition for the identification procedure is the presence of the speakers' acoustic material in the database. If the distance from the input signal to the database frame does not exceed the specified threshold, the developed system determines this frame as such that belong to an input speaker with a certain probability. The value of this probability is determine by the percentage of the frames of the speaker among the frames that does not exceed the settled threshold (see Fig.1).

During the experimental research for base speakers set (75 female, 75 male voices) the weight coefficients values and the threshold value for the optimal performance of the developed system in the identification mode. Performance of the developed system also has been tested using the test speakers set (30 female, 30 male voices) and has shown absolute accuracy in speakers' identification for the short-time period (one week). However, research on long-time period (two or three weeks) has shown that the majority of basic speakers' signals ceases to be relevant. Therefore, they must be updated.

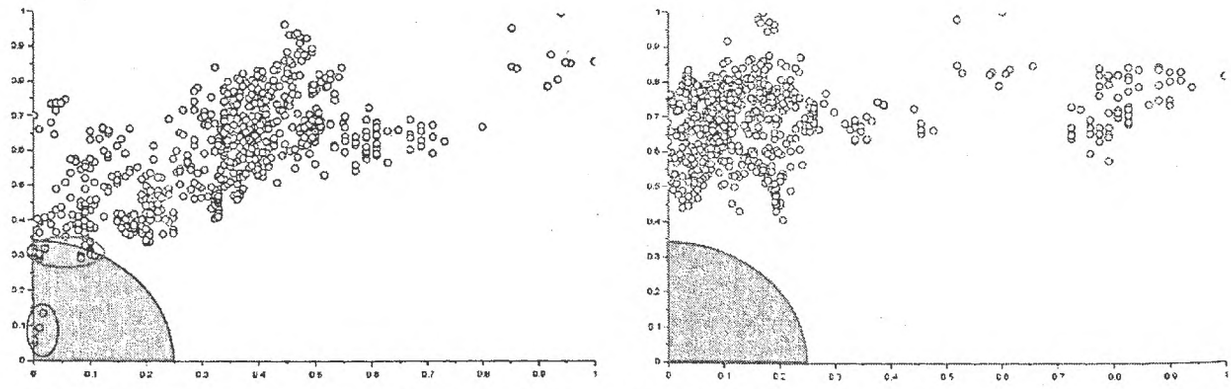

Figure 1. An example of personal identification and false identification in the developed system.

1.3. Mathematical model of human voice signal

The above results suggest that changes in the human voice signal may be caused by the change of its anthropometry and changes in the human physical condition. Therefore, conclusions about the changes in the general human condition can be made on the basis of changes in human voice signal. For this task, it is necessary to form so-called human voice standard signal using a special mathematical model.

Paper[7] presents the developed a mathematical model of human voice signal, according to which the human voice fundamental frequency and the structure of signal amplitude time distribution are dependent on the height, weight, age and persons' gender. The dependence has been determined by multiple regression method based on these parameters. This model has shown quite adequate results for human individual voice standard signals construction, that are signals which are based on the dynamics of human acoustic signals taken in different moments. However, if this model is used to create the standard signal based on the signals of different speakers. Often, the obtained signal has been significantly different from the real signal.

Therefore, an alternative model has been developed. This model that takes into account the closest signals by the following parameters:

- anthropometry (persons' age, height, weight);
- characteristics of the human voice signal (the fundamental frequency and the structure of the signal).

On the basis of these two parameters two corresponding feature spaces are formed. In each of these spaces the developed system extracts the closest signals to the input one by KNN-graph. Both sets of signals are combined into a single cluster. For each cluster voice signal, the percentage of distance measure has been defined. This percentage has been taken into account in the equations that determine the main parameters of standard voice signal:

$$\begin{aligned} F_0 &= \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3 + \ldots + \alpha_n F_n \\ Y_i &= \alpha_1 Y_{i1} + \alpha_2 Y_{i2} + \alpha_3 Y_{i3} + \ldots + \alpha_n Y_{in} \end{aligned} \quad (1)$$

where $F_1, F_2, \ldots, F_n$ – the fundamental frequencies of cluster voice signals; $Y_{i1}, Y_{i2}, \ldots, Y_{in}$ – i-th amplitude sample of cluster voice signals; $\alpha_1, \alpha_2, \ldots, \alpha_n$ – the percentage of cluster voice signal in standard voice signal; n – number of cluster voice signals.

Standard human voice signals, built on the base of this model has shown more efficient results in comparison with the standard signals, built on the base of the regression model (see Fig. 2).

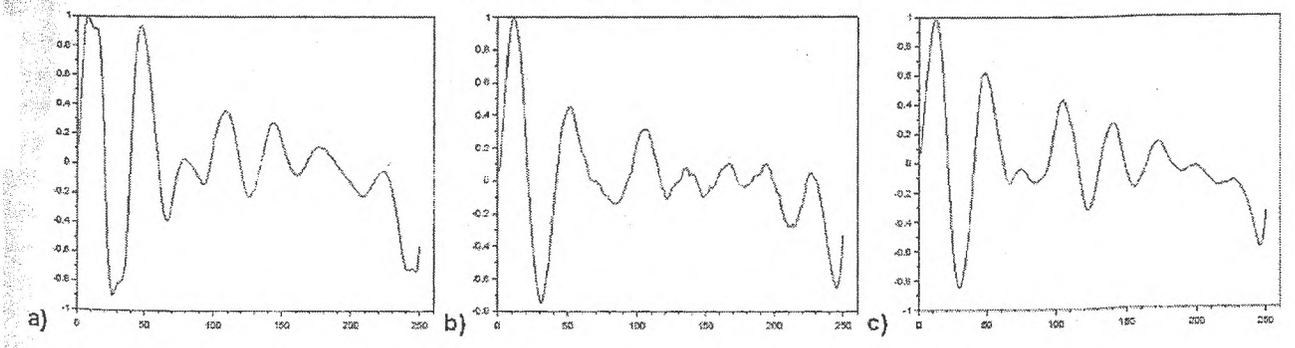

Figure 2. Comparison of real human voice signal frame (a) with the standard human voice frames, built on the base of the multiple regression model (b) and the alternative percentage model (c)

2. THE DEVELOPED ALGORITHMS AND RESULTS OBTAINED

Acoustic recordings of human voice signals have been performed with the audio editor Audacity 2.1.2 in *.wav format (sample frequency 22050 Hz, bit depth 32 bit, channels Mono). Several types of microphones have been used, among them a usual headset, cardioid and general orientation microphones.

All developed algorithms have been implemented as Sci-Lab 5.5.2 scripts and shown in see Fig. 3. The voice signal database is organized as a catalog system and has been divided by the gender into two bases. Also as a separate

subdirectory the branch of individual speakers' catalogs is allocated. In these catalogs, acoustic recordings are stored which were used for the tasks of identification and human condition analysis. It allows to keep track of changes in human voice signal dynamics, as well as to improve the developed voice signal standards and the system performance. A proper Sci-Lab script is responsible for updating the database and human voice standards recovery.

For the task of human condition analysis by voice signal first it is necessary to submit a gradation of human conditions (normal, a slight deviation, pathology) and to determine the parameters of voice signals that will be used for human condition analysis. Next, an algorithm for input and database signal comparison by selected parameters is devleped. Then it is necessary to determine the threshold value of input signal parameters deviations from standard ones.

At the current stage of the study authors have used the developed system principally for the task of personal identification. For the task of human voice signal analysis, it is proposed to use the described parameters – voice fundamental frequency and the amplitude distribution in time domain. However, at this stage we have not explored sufficiently the correlation between these parameters and the general human condition yet.

2.1. Possible application of the developed algorithms

The developed system can be applied in many spheres of life. For example, it can be used for individual condition monitoring at home. This should be very important for some persons who are limited in motion or try to monitor their health continuously, especially elder people. In this case, we hope that in future such system can be integrated into communications and will have a direct connection with some special medical institutions, which can give a certain help to the person if necessary.

The developed system can be also used in medical facilities as an independent or recommendation diagnostics procedure. It also can substitute the system of the medical cards that is often used in these facilities to keep patients' records and medical history. In such a case the developed system can simplify the registration of the patients (with the identification part) and their diagnostics (with the analysis part).

In prospect, it is also advisable to use the developed system by people of certain professions while performing their duties. In such a case the developed system can permit or submit the person in performing these duties according to either the persons' identification or its' condition analysis. This will be very actual for persons of such professions drivers, pilots, rescuers etc.

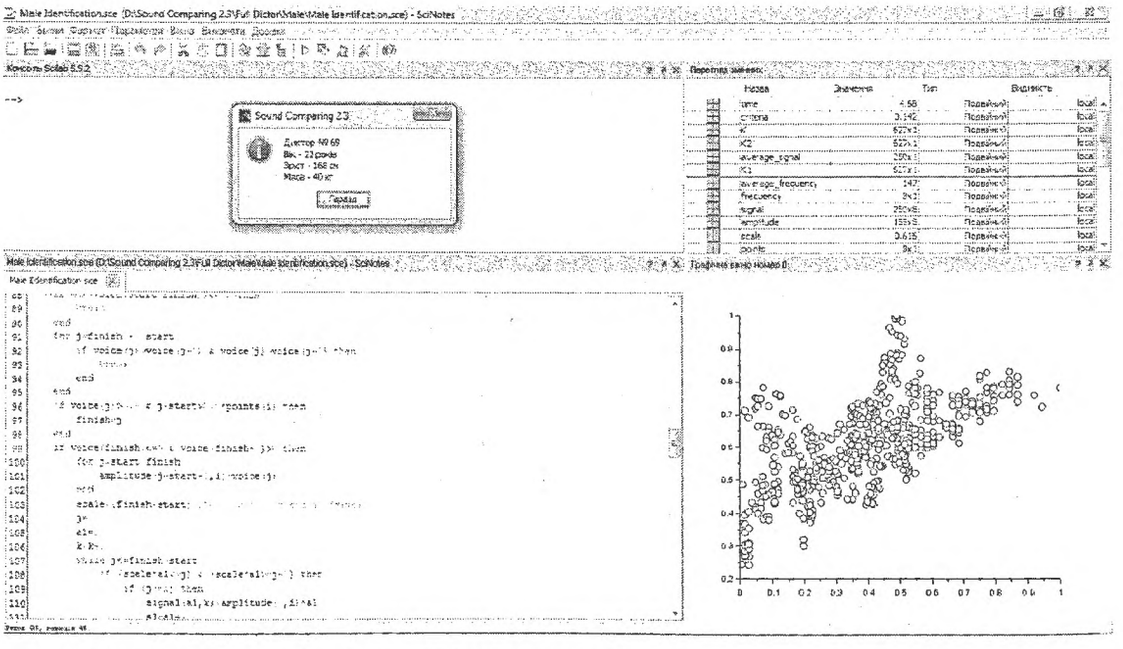

Figure 3. The example of the developed system in Sci-Lab 5.5.2 environment.

3. CONCLUSIONS

Human voice signal is a complex signal that is dependent on the individual characteristics of human body and its physical condition. This signal can be represented as a set of characteristics. Among these characteristics, authors have chosen the voice fundamental frequency and the structure of signal amplitude distribution in time domain. A set of algorithms for voice characteristics extraction from the speech material has been developed. Based on these parameters, the identification procedure is arranged on the nearest neighbor method. What is more, algorithms for standard voice signal formation based on the clustering of the speakers database and mathematical models of chosen characteristics have been developed.

The algorithms presented were implemented as Sci-Lab scripts. The system has been tested using several sets of speakers and it has shown the expediency of using this approach for the task of personal identification.

Authors have marked the main points of using the human voice signal for the task of human condition analysis. The perspective directions of research are shown, as well as the planned scopes of applications of the developed system.

REFERENCES

[1] Matsui, T., and Furui, S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's," IEEE Transactions on speech and audio processing 2(3), 456-459 (1994).

[2] Sorokin, V.N. and Tsyplikhin, A.I., "Verifikatsiya diktora po spektralno-vremennym parametram rechevogo signala," Informatsionnye protsessy 10(2), 87-104 (2010)

[3] Pervushin, Ye.A., "Obzor osnovnykh metodov raspoznavaniya diktorov," Matematicheskie struktury i modelirovanie 3(24), 1-54 (2011).

[4] Pervushin, Ye.A. and Lavrov, D.N., "Algoritm izvlecheniya priznakov rechevogo signala vo vremennoy oblasti dlya zadachi raspoznavaniya diktorov," Vestnik Omskogo Universiteta 2, 182-185 (2011).

[5] Kamińska, D. and Pelikant A., "Zastosowanie multimedialnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej," IAPGOŚ 3, 36-39 (2012).

[6] Kolokolov, A.S., "Obrabotka signala v chastotnoy oblasti pri raspoznavanii rechi," Problemy w Upravlenia (3), 13-18 (2006).

[7] Mieshkov, O.Y. and Novikov, O.O., "Development of Universal Program Complex for Human Condition Analysis Based on the Analysis of Human Voice," Theoretical and Applied Aspects of Cybernetics. Proceedings of the 4th International Scientific Conference of Students and Young Scientists, 294-305 (2014).

[8] Flanagan, D., [Analiz, sintez i vospriyatie rechi], Svyaz, Moscow, 1-400 (1968).

[9] Titze, I.R., [Principles of voice production], Prentice Hall, Englewood Cliffs, 1-354 (1994).

[10] Novikov, A.A. and Mieshkov, A.Yu., "Elektricheskiy analog golosovogo apparata cheloveka," Biomeditsinskaya inzheneriya i elektronika 2(2), 81-89 (2012).

[11] Kreiman, J., and Sidtis, D., [Foundations of voice studies: An interdisciplinary approach to voice production and perception], John Wiley & Sons, Chichester, 1-516 (2011).

[12] Tsanas, A., Little, M.A., McSharry, P.E., Spielman, J. and Ramig, L.O., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on Biomedical Engineering 59(5), 1264-1271 (2012).

[13] Titze, I.R., "Vocal fold mass is not a useful quantity for describing F0 in vocalization," Journal of Speech, Language, and Hearing Research 54(2), 520-522 (2011).

[14] Kvasov, A. N., [Model golosoobrazovaniya i analiz rechevogo signala v norme i pri patologii], Ph.D thesis, Tomskiy Gosudarstvennyy Universitet Sistem Upravleniya i Radioelektroniki, Tomsk, (2007).

[15] Trail, M., Fox, C., Ramig, L.O., Sapir, S., Howard, J. and Lai, E.C. "Speech treatment for Parkinson's disease," NeuroRehabilitation 20(3), 205-221 (2005).

[16] Mieshkov O.Y. and Novikov O.O., "Automated system for identification and human condition diagnostics based on its voice signal analysis," Proc. of 18-th International conference System Analysis and Information, 35-38 (2016).