UDC 629.039.58

USE OF THE NEURAL NETWORK SYSTEM OF SOUND SIGNALS AS A METHOD OF RECOGNITION OF THE HUMAN SPEECH. STRUCTURE OF SYSTEMS OF THE SYNTHESIS AND RECOGNITION OF SPEECH. PROBLEMS AND THEIR SOLUTIONS

Turmanov A.
Lanzhou Jiaotong University

The system of synthesis and post-processing of the human voice is analyzed. The spectra and physical characteristics of speech sounds are studied. The ways of speech recognition are considered based on the system of neural networks.
Keywords: synthesis, sound spectrum, speech recognition, neural network.

Computers are increasingly used in various areas of our lives. As a result, new problems related to the input of information have arisen in front of the humanity. The input of data through a keyboard or a mouse which is usual for us takes a considerable amount of time and energy. It would be simpler to use the instrument that is inherent to us by nature – our voice. The possibility of voice control of systems will greatly simplify the data input and open big prospects in the development of technologies, incl. computer ones. Speech recognition systems would allow to control the work by voice and to input any text.

In the countries of the former Soviet Union, the serious experience has been accumulated in this respect. Thus, «the results of Soviet times researches corresponded to the global level, but they were of the scientific and applied nature, not setting a goal of the successful commercial use of these results. There was a competition of several scientific schools: Leningrad, Novosibirsk, Georgia, Belarus, Ukraine, etc. – both at the scientific field and especially in the last years of this renaissance, in the sphere of creating the prototypes of command speech recognition systems» [1, p. 84]. To date, there is a number of software products for the recognition of Russian speech: «Gorynych», «Dictograph», «Perpetuum», «Dragon», etc. The areas of application of these systems are extensive: cartography, telephony and Internet, security systems, medicine, etc. These systems are especially relevant for disabled. The process of practical use of speech recognition systems at the modern level is complicated by the following problems: speaker's slips of the tongue, gibber and mistakes, accent, other individual features of the diction (human factor), extraneous noises, machine errors in the recognition. In order to understand the essence of these problems in more details, let us focus on the very process of speech recognition.

An artificial neural network is a mathematical model of parallel calculations that internally connects a system of interacting simple processors (artificial neurons). Each process of such a network only works with the signals that it occasionally receives, and the signals which it sends to other processors. Connected to a large enough network with the controlled interaction, such simple processors together are able to perform quite complex tasks. This model is called artificial neural networks (ANN).

Thus the process of speech recognition can be divided into two main phases: digitization and decoding. In the first phase, the input audio signal is recorded and divided into fragments. In the decoding phase, the information obtained is analyzed based on the use of different models and algorithms (see Figure 1).

In the proposed system, we neglected the McGurk effect [2, p. 746–748]. The essence of it is as follows: the perception of speech is multi-modal, which is the compilation of information from several senses. In the McGurk experiment, the

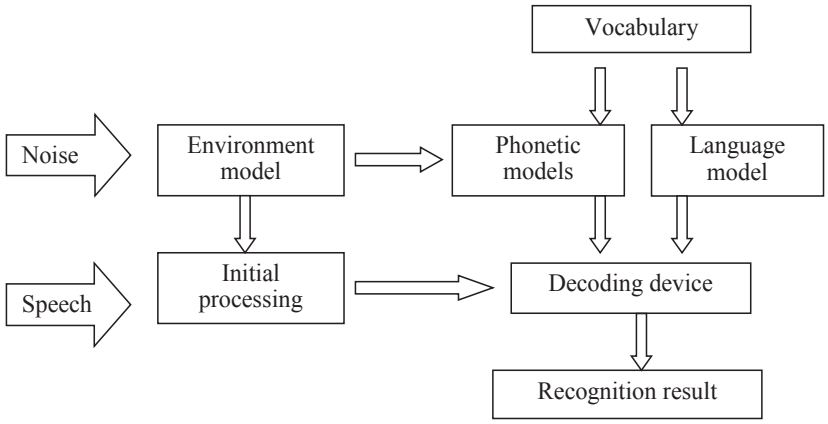

Fig. 1. Structure of the speech recognition system based on the neural network system

subjects saw a man on the screen whose articulation of the lips corresponded to the syllables «ga-ga». At the same time, the acoustic signal «ba-ba» sounded. Most subjects heard a completely different sound – da-da, while they did not realize the discrepancy between auditory and visual simulations. This effect is very stable.

There are many methods of synthesizing the human speech, the main ones are: the compilation synthesis – the compilation of a unique knowledge base from individual samples of a sound (phoneme), pronounced by a certain person, and the formant-voice synthesis where a person's speech path is modeled with the certain accuracy. Apparently, the first type of synthesis requires a very extensive work on creating a sound database, and the self-learning of this synthesis model is extremely difficult. The formant-voice model allows the self-learning, but because of the complexity of modeling the speech path, a person has the lower recognition accuracy. Nevertheless, integrating into the neural network model, the synthesized sounds acquire the intelligible character of recognition, thus the first one is preferable for the research.

Now, we face the task of collecting and isolating the physical characteristics of sounds. For this purpose, we have assembled a circuit consisting of a microphone, an amplifier and a program (Gold wave), which shows the spectrum of a sound.

During the research it has been revealed that four frequencies predominantly occur in the formation of speech, which arise resonantly in the cavities of the speech path. These frequencies are called ***formants***.

Spectral studies of the voice reveal a change of the content in speech sounds of these or those certain frequency parts. As a result of these studies, the fact of presence of formant frequencies containing the basic speech information has been revealed. The monitoring of changes of these frequencies, as well as changes of the amplitude of a sound signal makes it possible to extract lexical elements – phonemes and allophones from the signal.

Figure 2 correctly shows the formant composition of vowels «и [i:]» and «у [u:]» when pronouncing the sequence of these sounds. When transiting from the vowel «и [i:]», there is a displacement of the formant frequency f2 from 2400 Hz to 784 Hz, as well as the simultaneous weakening of formants f3 and f4.

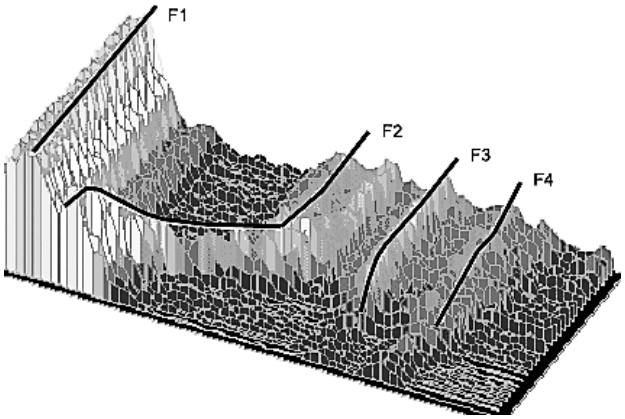

Fig. 2. Spectrum of sounds «и [i:]» and «у [u:]»

It is clearly visible that both the amplitude and the frequency of formant parts of the sound can vary during the articulation. In this case, the number of the formants themselves in speech sounds is constant and always equals to 4.

The next object of research is noise sounds, it is difficult to distinguish formant parts in them. This can be seen in Figure 3, where the spectrum of sound is «x [h]» (which is a turbulent noise).

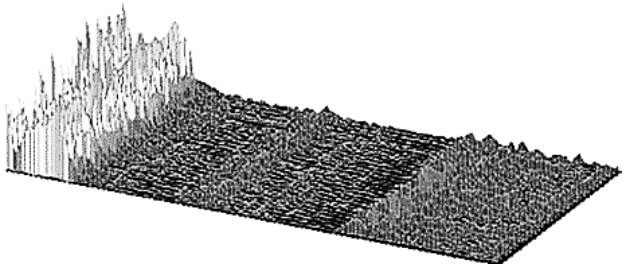

Fig. 3. Spectrum of sound «x [h]»

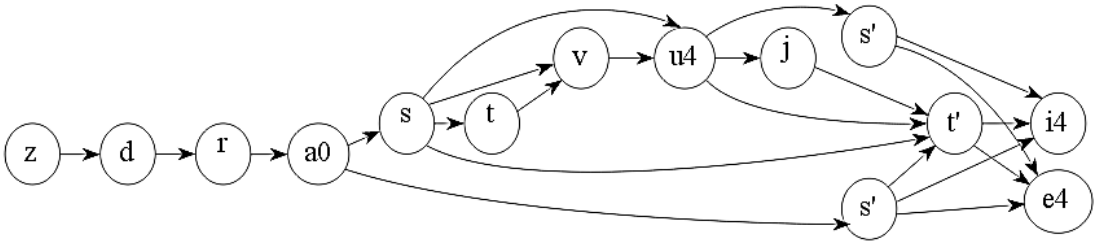

Fig. 4. Scheme of word recognition algorithm

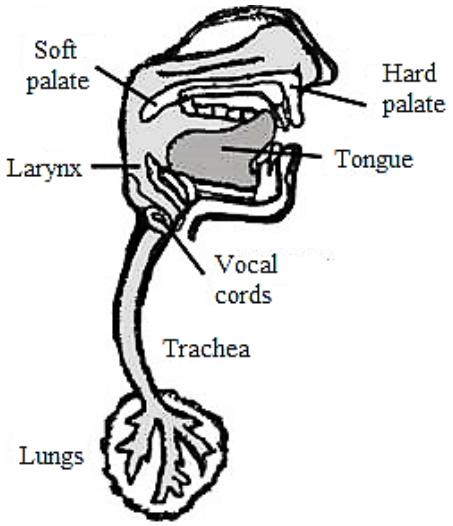

Fig. 5. Speech system of a human

Ahead of the events, let's say that the method of a speech recognition system presented by me performs the spectral analysis, which makes it possible to extract the most informative components from the speech signals: formant frequencies and noise.

Figure 4 shows a particular example of the implementation of the word recognition algorithm. Thus several dozen variants of the pronunciation are possible for the word «zdravstvuite [hello]». Although the computer «hears» this word in different variants, it «learns to understand correctly» due to a phonetic recognition model.

Thus, we propose a model for recognizing the human speech based on a neural network model. This method is highly efficient, as it allows you to recognize the speech with the adequate accuracy and high quality, and to increase the speed of the user's work with the machine systems.

This method can be effectively implemented in the system of numbers recognition, continuous

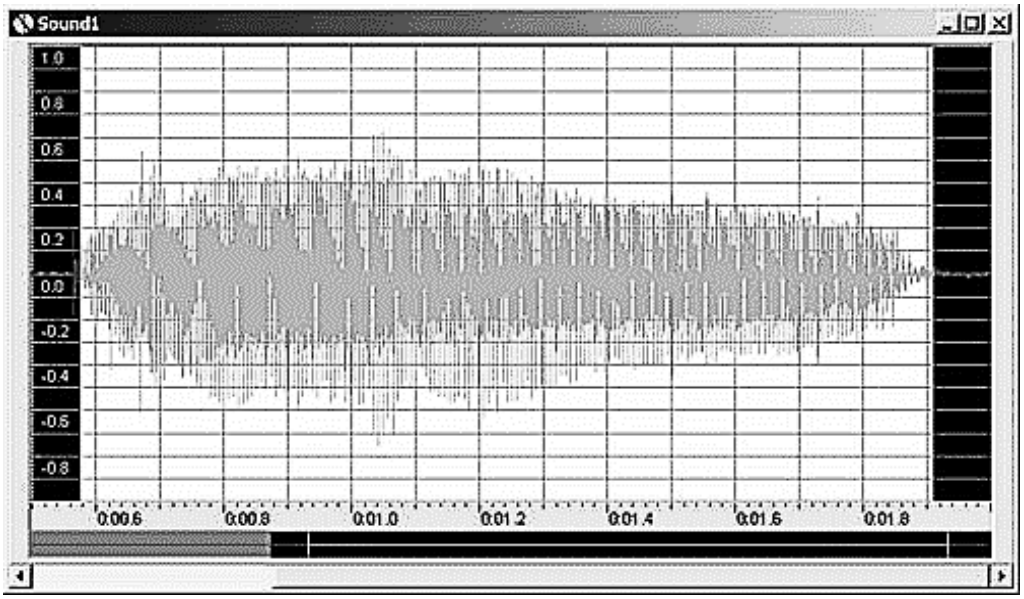

Fig. 6. Oscillogram of the sound «a [a:]»

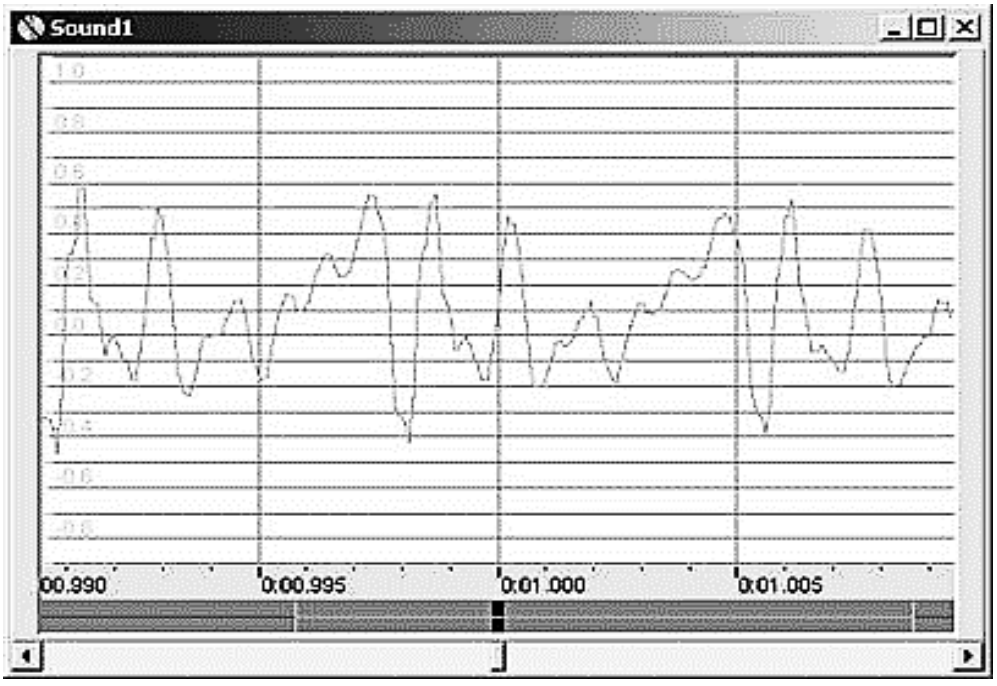

Fig. 7. Extended fragment of the Oscillogram

speech and individual voice commands. This will lead to a more widespread use of a speech recognition technology in many areas of our lives and make it easier for people to perform some complicated tasks.

For decades, many scientists diligently worked to create a system for the synthesis and speech recognition. All these attempts began with the study of anatomy and principles of the speech, as well as auditory organs of a man, for the purpose of modeling them with the help of computers and special electronic devices. This topic of research is still relevant, because there is not yet a system that meets all the necessary requirements: high speed, recognition without errors and stability to the external noise. In this article, we propose ways to solve the above problems with an optimal method.

First, let's consider how speech appears. In Figure 5, a simplified scheme of the human speech system is shown [3]. The main «generator» of the path is the lungs. The acoustic signal together with the air from larynx enters simultaneously into the nose and mouth. The mouth plays a major role in the formation of sounds, while the nose is a resonator, i.e. it enhances the vibrations of certain frequencies. Each person is unique, thus the same sound is pronounced differently due to anatomical peculiarities. The fact is that apart from a frequency fundamental tone, there are always so-called harmonics in the voice, which are sounds of other frequencies, different from the fundamental frequency. For a thorough study of this feature, we visualize the spectrum of vibrations using GoldWave software.

Using a standard studio microphone, we record sounds (more precisely letters). On the working plane of the software we get the output data in the form of an oscillogram (Figure 6), stretch along the time axis (Figure 7) to see the smallest details. Let's analyze this oscillogram. The size of an input signal that enters the microphone takes both negative and positive values with respect to the time. When only one frequency is present in the input signal (that is, if the sound has been «clean»), the shape of the output signal would be sinusoidal [4]. However, as it has been mentioned above, the spectrum of the sound of the human voice consists of a set of frequencies, as a result of which the shape of the oscillogram is far from sinusoidal. Thus, the research task has been formed: to propose a method for constructing a recognition system that determines the corresponding sound (a letter in the alphabet) under the peculiarities of waves. First, the signal is digitized in a converter. Based on the experiment, it has been found out: the frequency of the sound signal lies in the range of 300–4000 Hz, according to the Kotelnikov theorem, the conversion frequency should be not less than 8000 Hz [5]. For the purity of experiment, I chose the frequency of 42 000 Hz, for further sifting of extraneous noises. As a result, we get a digitized signal.

The next step is the collection of information in the form of signal spectra for the further use in the knowledge base. Using the utility **Spectrogram**, we record the sounds of letters, (Figure 8) the spectrum of the letter «ю [yu]» is given below.

The spectrogram shows the change in the spectrum with the respect to time: along the horizontal axis – time, along the vertical axis – frequency of the signal. The sound «ю [yu]» has not been chosen by chance: it consists of two sounds «и [i:]» and «у [u:]». This spectrogram clearly reveals individual phonemes of sounds and a pause between them. To distinguish this spectrum into separate segments, an analyzer is required, using the artificial neural networks it is possible to isolate these elements from speech.

The artificial neural network is a mathematical model of parallel calculations that connects a system of interacting simple processors (artificial neurons). Each process of such a network works only with signals that it occasionally receives, and signals it sends to other processors (Figure 9). Simply said, the input signal will be matched and compared to the sound base prepared by us in advance, since the sound base will be quite extensive, the probability of error will be quite low. This method is remarkable as each user can «train» the system for him/herself within several minutes and then use it without problems. The system is adjusted

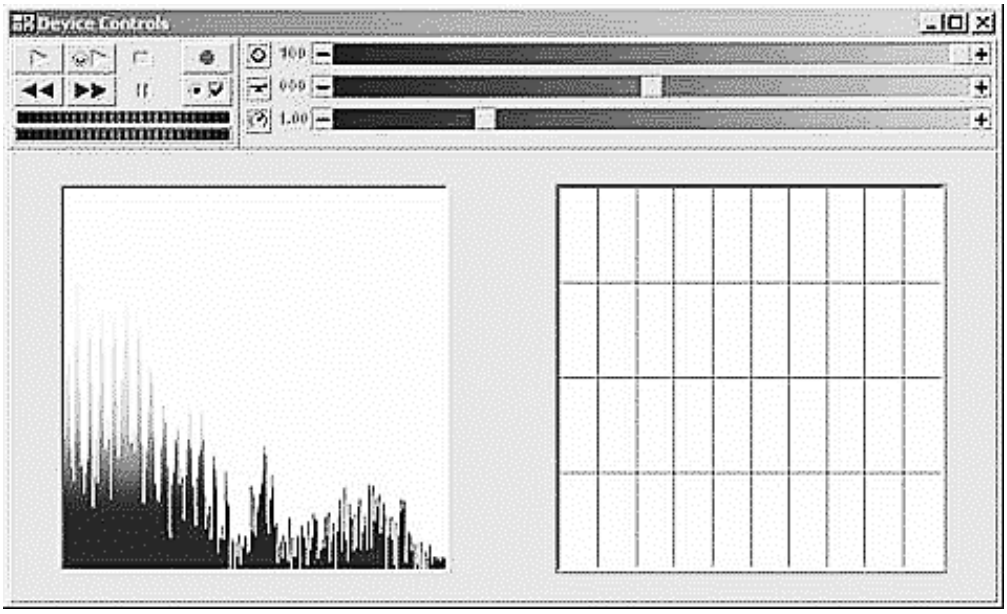

Fig. 8. Spectrogram of the sound «ю [yu]»

to the physical peculiarities of human voice: tone, timbre, tempo, obtained by the spectral analysis of the sound signal.

The analysis of the acoustic peculiarities of human speech made it possible to identify a number of parameters that make it possible to use them in the speech recognition system most effectively. As a result of the analysis of spectra and oscillograms, it has been concluded that the best option for a speech recognition system on the quality/resource intensity criteria is the formation of a base knowledge-recognitions on the grounds of creation of a neural network system. The proposed method of the neural network system gives a solution to several problems: the resistance to interferences, the speed, and it reduces the probability of errors. Thus, a method for highly effective speech recognition has been developed by isolating the basic physical parameters of the person's voice.

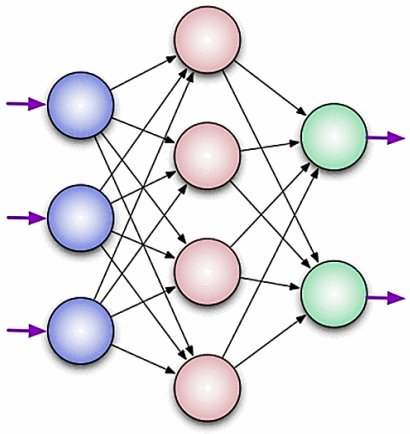

Fig. 9. Scheme of a simple neural network. The blue circle indicates the input elements, green – the output data

References:

1. Hitrov M.V. Recognition of Russian speech. «Speech technologies». – No 1. 2008. – P. 84.
2. McGurk H., MacDonald J. Hearing lips and seeing voices. «Nature», Vol. 264(5588). – 1976. – Pp. 746–748.
3. Barabash Yu.L. Collective statistical solutions in case of recognition. – Moscow: Radio and Communication, 1983. – 224 p.
4. Gladun V.P. Partnership with the computer. – K.: «Port-Royal», 2000. – 128 p.
5. LeCun Y., Bottou L., Orr G., Muller K. Efficient BackProp // Neural Networks: Tricks of the trade. – Springer Verlag, 2014. – P. 5–50.

Турманов А.
Ланьчжоуский транспортный университет

ИСПОЛЬЗОВАНИЕ НЕЙРОСЕТЕВОЙ СИСТЕМЫ ЗВУКОВЫХ СИГНАЛОВ КАК МЕТОД РАСПОЗНАВАНИЯ ЧЕЛОВЕЧЕСКОЙ РЕЧИ. СТРУКТУРА СИСТЕМ СИНТЕЗА И РАСПОЗНАВАНИЯ РЕЧИ. ПРОБЛЕМЫ И ИХ РЕШЕНИЯ

Аннотация
Анализируется система синтеза и постобработки голоса человека. Исследуется спектры и физические характеристики звуков речи. Рассматривается пути распознавания речи на основе системы нейросетей.

Ключевые слова: синтез, спектр звука, распознавание речи, нейросеть.