RECOGNITION AND IDENTIFICATION BY TIMBRE OF A LIVING CREATURE'S VOICE BASED ON TRAINABLE NEURAL NETWORKS IN REAL-TIME MODE AND THEIR IMPLEMENTATION IN THE INTELLIGENT SYSTEM «NEUROCYBER»

Anton Gafurov, *Oleg* Gafurov[a]

National Research Tomsk State University, 634050, Tomsk, Russia

Abstract. The article considers physical basis for formation and propagation of sound and its psychophysical perception by living creatures. The authors suggest an algorithm of living creatures' identification by the voice timbre and the algorithm's practical realization by artificial neural networks, by initialization of a conversation by the robot in order to reduce ambiguity of responses and to improve the quality of recognition in real-time mode.

1 Introduction

The use of artificial neural networks allows solving the problem of non- linear object control by creating an adaptive control system with trainable neuro-emulator. The problems of identification, synthesis of control systems, and their analysis and hardware implementation are successfully solved because of properties of generalizations and universal approximation of artificial neural networks, which are common to various nonlinear dynamic objects. Training is a process of development of desired reaction to external signals in a control system through multiple system effect and external adjustment. A "teacher" who knowns the desired reaction of a control system to certain effects carries out external adjustment. Thus, during the process of training «teacher» provides the system with additional information whether its reaction is true or false.

The method of recognition and identification by timbre of a living creature's voice based on trainable neural networks in real-time mode and its implementation in the intelligent system "NeuroCyber'' relates to the field of speech analysis technology. In particular, the method relates to the system of formation of robots' adaptive behavior in real-time mode, to system of formation of unauthorized access guarding of material or information resources based on biometric information about the speaker. The technical result is an increase of reliability of recognition and identification by timbre of a living

[a] Corresponding author: author@email.org

creature's voice; improvement of the system of formation of robots' adaptive behavior in real time mode when computational and operating systems must perceive signal of external environment and operate on their own without help of "a person-translator"; and improvement of the systems of high noise immunity recognition with noisy observations. Use of combinations of very different by distinctive features acoustic signal attributes as a parametric description of timbre of a living creature's voice allows achieving technical result. In addition, use of the mechanism of automated training, additional training, retraining of neural networks in real time mode, formation of a new object identifier based on words-commands or recognition of familiar subject in changing external conditions also contributes to achieving technical result [1, 2].

2 Physical basics of sound waves and oscillatory processes

Transfer of semantic and emotional information between two subjects (in the usual sense, living subjects) occurs based on acoustic waves caused by oscillatory modulated processes of throat ligaments (alveolus) of human beings and animals.

Oscillation means some periodic or approximately periodic process, in which the value of a physical quantity is repeated exactly or approximately at equal or approximately equal time intervals.

Robots and robotic systems can perform oscillatory processes in a very wide range including radio waves and sound waves; throughout the range, sound waves occupy a very narrow region. Information transmission between two robots can take place in any range, either in the range of sound unheard for ear of a human being or animal, or in the range of ultrashort radio waves.

Basic concept of oscillatory processes and sound waves propagation in atmospheric air under normal atmospheric pressure and average temperature.

Wave equation:

$$\frac{d^2s}{dx^2} + \frac{d^2s}{dy^2} + \frac{d^2s}{dz^2} = \frac{1}{v^2}\ \frac{d^2s}{dt^2} \tag{1}$$

v –wave propagation velocity, or

$$\Delta S - \frac{1}{v^2}\ \frac{d^2s}{dt^2} = 0 \tag{2}$$

where $\Delta = \frac{d^2s}{dx^2} + \frac{d^2s}{dy^2} + \frac{d^2s}{dz^2}$ Laplace operator

In atmosphere (ideal gas), wave frequency is equal to frequency of wave oscillation source. Different people and living creatures have individual vocal apparatus; this fact enables us to distinguish them by voice and interpret correctly.

A person with normal hearing ability can perceive as sound only elastic waves with frequency of not less than 16 Hz and not more than 20 000 Hz. The sensitivity of the human ear is not the same to the waves of different frequencies; it is maximal for waves with frequencies of 1.5 – 3 kHz. These is because of structural features of the human organ of hearing. The upper and lower frequency limits of different animals vary: γ = 38 kHz for dogs, γ max> = 100 kHz for bats and whales. Waves with frequencies γ <16 kHz are called infrasound, with frequencies γ> 20 kHz - ultrasonic, and with γ> 10 to the ninth power kHz – hypersonic.

In this paper, the term "sound" is used to describe the feeling, which a sound wave makes for the human organs of hearing, i.e. sound is a physical process of elastic waves propagation in the medium (physical acoustics), this paper describes psychophysiological processes caused by this physical process (physiological acoustics).

The robot receives and distributes sound in accordance with laws of physical

acoustics, but human beings and animals - in accordance with structure of their external and internal ear and other psychophysiological constraints and processes.

Simulation of these processes is that the robot and robotic systems must correctly interpret psychophysical information transmitted to them by man or animal, moreover they are trained to recognize the information on their own or under the guidance of a dominant Figure – a master (Mowgli principle).

Physical acoustics uses numerical characteristics – (γ) frequency of sound wave (frequency spectrum), intensity of sound.

Intensity or sound force is physical quantity **I** equal to modulus of the average value of vector of the sound wave energy flux density (Umov's vector):**I** = | **U** |, $I = \boldsymbol{\omega} * \boldsymbol{U}$, where **U** –group velocity of the wave, **ω** – the average value of the volumetric energy density

$$\boldsymbol{\omega} = \frac{1}{T}\int_0^T \omega * dt \tag{3}$$

where **T** – is time required to complete one oscillation

In case of sinusoidal wave, velocity **U** coincides with the phase velocity **V**, and

$$\boldsymbol{I = A^2\omega^2}$$

$$\boldsymbol{\omega = \frac{1}{2}\delta}$$

$$\boldsymbol{I = \frac{1}{2}\delta V^{A^2\omega^2}}$$

The SI unit of sound intensity is the watt per square meter (W/m^2).

Physiological acoustics uses concepts of height, timbre and volume to characterize sound sensations.

Sound height is a quality of periodic or almost periodic sound depending on the sound frequency and judged by ear.

For humans, we introduce the concept of sound volume depending on the root-mean-square sound pressure **P** and on the ear sensitivity, which is not the same for sounds of various intensity and frequency.

The hearing threshold depends on the sound frequency, reaches minimal value of about 2×10 to the negative fifth power N/m^2 at frequencies γ = 500 – 1000 Hz and makes about 200 N/m^2. Auditory-sensation area limited by two threshold curves on Figure 1 are projected along the coordinate axes in logarithmic frequency scale and root-mean-square sound pressure **P**.

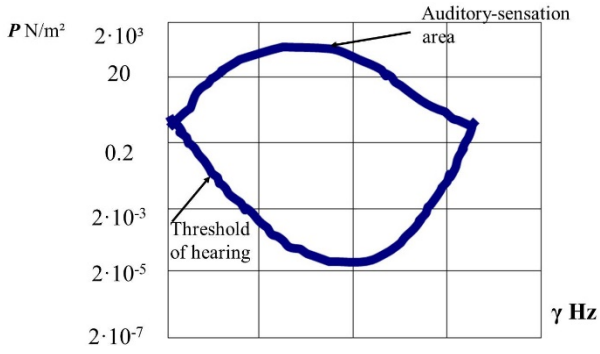

Fig.1. Auditory-sensation area.

Sound pressure is the product of the medium wave resistance ρυ and velocity of oscillations of its particles ds/dt.

Root-mean-square sound pressure $P_0 = \sqrt{\rho v I}$

Values of intensity and root-mean-square pressure for different sounds in air under normal pressure and **20° C**

(***ρ v = 418 kg/m²sec)*** (Figure 1 and Table 1).

Table 1. Values of intensity

Sound	Distance from the source in m^2	Level of sound pressure in dB	Sound intensity in W/m^2	Root-mean-square pressure in N/m²
Threshold of hearing when*v=1000 Hz*	-	0	10^{-12}	2×10^{-5}
Low sound	1	40	10^{-8}	2×10^{-3}
Loud sound	1	70–80	10^{-5}–10^{-4}	0.06–0.02
Fortissimo of symphony orchestra	10	100	10^{-12}	2
Airplane engine noise	5	120	1	20

Calculations in the «NeuroCyber» are made on the distance of 1 meter from the source to the robot, when a person stands in front of the microphone under normal atmospheric pressure and external temperature **20° C** (***ρ v = 418 kg/m²sec).***

3 Neural networks: basic concepts

Neural network methods are methods based on the use of different types of neural networks (NN). NN consists of elements called formal neurons, which imitate the work of neurons in the cerebral cortex. Each neuron transforms a set of signals coming to its input into an output signal. Figure 2 shows an example of formal neuron work.

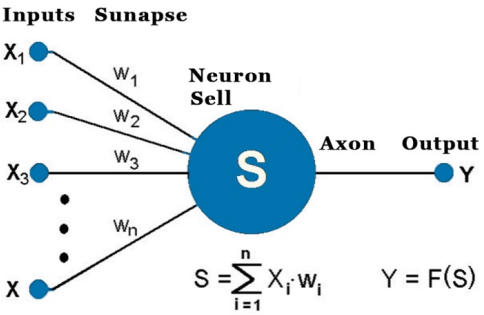

Fig. 2. Formal neuron.

Figure 3 shows neural network structure.

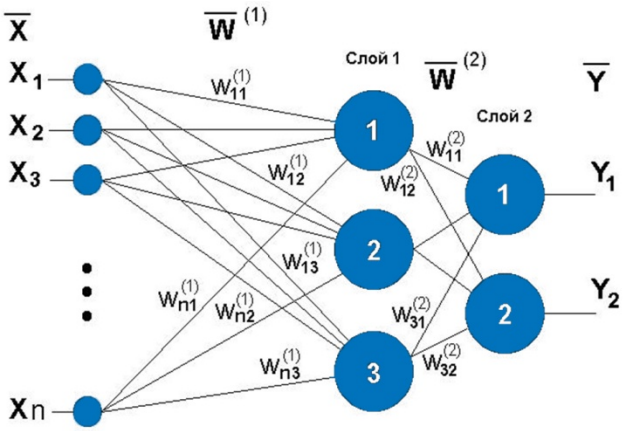

Fig. 3. Neural network.

Connection can exist between individual neurons. Connections between neurons encoded by weight coefficients play a key role in function of the nervous system. On the Figure, connections between neurons are signed by Latin letter W. Index on the top means belonging of a weight coefficient to a layer. One of the advantages of the nervous system is the possibility of parallel operation of all the elements, which considerably increases the efficiency of problem solving in general. This feature of the nervous system is successfully used in image recognition systems. The neural network has inputs X and outputs Y. It represents a system, which generates output state depending on the input state. Presence of weight coefficients that can be determined algorithmically, allows the nervous system to have the most important property – the ability to learn.

During the process of image recognition, a person unconsciously attracts a huge stock of contextual knowledge accumulated throughout the life. Application of neural network methods based on trainable neural networks implemented in the intelligent neuro-informational system "NeuroCyber" potentially can provide the possibility to simulate processes occurring in the human brain during images recognition. In the first approximation, the human brain can be represented in the form of an extremely complex neural network [3, 4].

4 Known methods of speaker identification and recognition

There is a method of automatic identification of a person by means of specific features of pronunciation of a passphrase by this person. Patent of the Russian Federation №2161826, IPC G 10 L 17/00, published on 10.01.2001, consists in the following: the speech signal is divided into voiced zones, time intervals are set in voiced zones - in the area of maximal intensity of the speech signal, and at the beginning of the first and at the end of the last voiced zones. Parameters of the speech signal are determined for the set time intervals and are compared with the etalons formed with mathematical expectations and admissible parameter spread. For this purpose, time intervals are set at the end of the first voiced zone, at the beginning of the last voiced zone and at the beginning and at the end of the other voiced zones. Duration of time intervals is set as multiple of a period of a pitch of speech signal; estimations of correlation coefficients of speech signal parameters are determined and included in the number of positions to be compared with etalons; correlation coefficients of speech signal parameters are additionally considered when forming etalons. Decision on identification of personality

is made based on obtained speech signal parameters and corresponding statistical characteristics.

The disadvantage of this method of identification of personality is low noise immunity of the method. Implementation of the method requires separation of exact position of the boundaries of the main voice tone in the input speech signal, which is almost impossible in conditions of acoustic noise (noise from the street, office, etc.).

There is a method of identification of a speaker based on the US patent №6389392, IPC G 10 L 17/00, published 14.05.2002. The method consists in comparison of the input speech signal of an unknown speaker with etalons representing speeches of speakers known in advance, and at least one speaker is presented by at least two etalons. Successive segments of the input signal are compared with etalon successive segments; thus, proximity measure of the compared segments of the input speech signal and etalon signal is obtained.

Composite result of comparison of etalon signal with input speech signal is formed for each etalon of a speaker known in advance, who has at least two etalons. Composite result is formed based on selection of the closest segment of compared etalon for each segment of the input speech signal by the used proximity measure. Further identification of an unknown speaker is made based on composite results of comparison of the input speech signal with etalons. This speaker recognition method is limitedly applied in practice, as it is difficult to realize in real conditions a requirement for availability of at least two etalons of a known in advance speaker. Besides, this method does not provide a high level of reliability of speaker recognition when working in conditions of acoustic noise from offices, streets and vehicles. This is because of the fact, that purely per-segment parametric description of speech signals is heavily influenced by additive acoustic noise and the natural speech variability. Moreover, low reliability of the method in conditions of noise is connected with the fact that the closest by proximity measure segment of compared etalon is searched for each segment of the input speech signal. This leads to the fact that among closest segments there is a great number of presence of the found next segments of a large number of similar pure noise segments corresponding to segments of speech pauses both in etalon and in the input speech signal.

5 Recognition and identification by timbre of a living creature's voice based on trainable neural networks in real-time mode

Principle of sound representation involves digitization – dividing into separate intervals (indications), inside which the amplitude is assumed constant.

Digitization step determines the frequency range of the converted sound; digitization frequency **2 F** is enough to transmit signal with treble frequency **F**.

People and animals hear the same sounds completely different because of the physiology of structure of their hearing and sound-reproducing apparatuses. The robot hears and understands sounds in the most correct way; moreover, the range of its hearing is almost unlimited. With the correct setting, the robot can be a great assistant to humans, it can hear and understand the language of animals, insects, dangerous directed radiation, can register and identify bugs, and can serve as a self-emitting constant source of information.

The proposed method provides a steady recognition of voice of a human being or other living creature after appropriate acquaintance and its identification with the help of a teacher.

The goal is achieved by providing continuous scanning of the surrounding space and continuous filling of the first ring buffer with digitized signal; receiving signal spectrum

in quasi-logarithmic scale by applying recursive filter comb to the signal and filling of the second ring buffer with the signal spectrum; determination in the second ring buffer availability and borders of a speech fragment based on adaptive estimation of the noise environment; translation of the spectral components of the speech fragment in a linear analysis buffer.

In order to adapt to the noise environment, determination of availability and borders of a speech fragment in the second ring buffer is made by analysis of spectral components exceeding over adaptive thresholds of noise environment and correspondence of energy and duration limits to etalon records of database. If during the analysis there is a result «not a living creature's voice», thresholds of noise environment are updated by means of a low frequency filter of the first order.

In the case of determination of availability of a living creature's voice in the digitalized record of the scanned window, maximum is defined in a window range containing the voice; and fragment of final predetermined length is cut in order to form a sample, which is supplied to neuro emulator.

Parameterization of the sample is provided by calculation of the fast Fourier transformation (FFT), calculation of instantaneous frequency, signal phases, energy and averaged amplitudes. This multi-dimensional sampling is supplied to neuro emulator.

Recognition is made according to proximity degree or the sample identification by neuro emulator; in case of misrecognition, signal "Do not know" or "Foe" is generated.

Further, if there is a need to identify the voice under the same circumstances, a teacher (human) gives the word command «Get acquainted» or enters the identifier forcibly. Thus, etalon formation occurs, i.e. voice and its identifier are recorded to the database; this sample along with the previous one is supplied to neuro emulator and retraining occurs.

In the case of misrecognition of already familiar voice because of changes in external noise or emotional state of a human being, re-recording of the sample is made forcibly by mans of teacher's commands with the appropriate additions "Alexander is angry", "Dog is in pain". Retraining of neuro emulator is made the similar way.

As regards the device, in order to solve the task of recognition of a speaker, the device contains the source of speech signal; calculating unit of Fourier transformation; unit for determination of parametric description of a speech signal in the form of extractor of the beginning/end of a speech signal; segmentator of a speech signal into a sequence of segments; calculator of a signal power spectrum in the segment and a driver of parametric descriptions of the input speech signal, which are connected in series; neuro emulator unit; unit for making decision on recognizable voice; and a unit for etalons memorization.

5.1 The results of application of different methods of human voice identification

The authors have tried the following groups of methods during the process of development of a mechanism of human voice recognition in real-time mode:

1) Neural network methods.

2) Difference methods for analysis of spectral and cepstral signal coefficients.

Table 2 shows the test results of these methods during analysis of records of various speakers' voices in WAV format. Records of three speakers were analyzed: Artem Soldatenko (a1.wav, a2.wav, a3.wav, a4.wav, a5.wav), Oleg Gafurov (g1.wav, g2.wav, g3.wav, g4.wav, g5 .wav), Alexander Melkozerov (mlkz1.wav, mlkz2.wav).

Table 2. Difference methods. Method of average error. Method of cross-correlation.

Speakers name	File.wav	Neural network	Analysis of spectral coefficients		Analysis of cepstral coefficients	
			Method of average error	Method of cross correlation	Method of average error	Method of cross correlation
Artem Soldatenko	A1.wav	1	0	1	0	1
	A2.wav	1	0.06	0.02	0.005	0.02
	A3.wav	0.3	0.55	0.03	0.006	0.03
	A4.wav	1	0.7	0.04	0.011	0.04
	A5.wav	0.3	0.42	0.01	0.01	0.01
Oleg Gafurov	G1.wav	1	0.6	0	0.03	0
	G2.wav	1	0.7	0.01	0.04	0.01
	G3.wav	1	0.88	0.02	0.03	0.017
	G4.wav	1	0.95	0.13	0.04	0,13
	G5.wav	0.5	0.62	0.01	0.01	0.01
Alexander Melkozerov	Mlkz1.wav	1	0.76	0.02	0.007	0.02
	Mlkz2.wav	1	0.4	0	0.05	0

Neural network methods

The training sample consisted of files a1.wav, g1.wav and mlkz1.wav. Amplitude window with size of 1024 samples was cut out of each file; these windows were inputs of the network.

Correct answers were obtained in 80% of experiments.

Difference methods. Method of average error. Method of cross-correlation.

The same persons were tested as in the previous experiment, and the sample consisted of files a1.wav, g1.wav and mlkz1.wav. Amplitude window with size of 1024 samples was cut out of each file. Then, 1000-band spectrogram in the range of 4000 – 8000 Hz was built on the window using fast Fourier transformation, and 1000 cepstral coefficients were built using the discrete cosine transformation.

Recognition coefficient based on spectral characteristics was unstable and made up $\approx 20\%$

Recognition coefficient based on cepstral coefficients made up less than 10%.

Thus, it is obvious that difference methods of characteristics analysis are not suitable for solving tasks of voice recognition, so the neural network method of amplitudes analysis was selected.

5.2. Results of subjects' recognition and identification in real time mode

The developed technology of subjects' adaptive recognition and identification has been implemented in the system "NeuroCyber" in the subsystem "NeuroAudio" as a component for control of neuro robots. These neuro robots in images of Kuzya, Liza and Bear cub and have been repeatedly demonstrated and participated in appropriate exhibitions in the real exposition mode.

For program testing, 150 subjects were selected and voices of the living creatures (cats, dogs, cows, etc.) were recorded. Test technique was as follows. In the situation of acquaintance, neuro robot Lisa (in the image of a woman secretary), has reacted to

appearance of the subject in her field of vision and initiated the conversation on her own without human intervention.

- Hello, what is your name?

From the follow-up response (human voice), the initial segment of the audio file with parameterization was singled out and transmitted to neuro emulator of voice recognition. The whole series of audio files was recorded as the identifier to the database. If it is recognized and identified as one of the known names (Irina, Oleg, etc.), then later, name recorded by the speaker is used in the dialogue, i.e. as the neuro robot's own voice. In case of non- recognition, the recorded file is repeated. In case of a secondary approach, a neural network is used, which has been created in the recent period of time. Or the neuro robot uses additional information by calling all subjects under the common identifier (Alexanders, Olgas, etc.) and creates a neural network trained by parameterized characteristics of voices and individuals. Moreover, if necessary, the robot asks the question, which requires correct and clear answer:

- Are we familiar with you?

Depending on the word command "Yes" or "No", the process of additional training or network starts or the process of recognition by the program "NeuroCyber". Permanently, in case of re-addressing, correction of audio files recording takes place. There is formation of separate files from amplitudes and parameters for different emotional state of the subject under the common identifier, but with additions like "Father is angry", "Mouse is in pain". Respectively, henceforth, the neuro robot comments their state by voice. The test results have shown almost 90% of correct answers.

6 Conclusion

This design project resulted in development of programs of human-robot voice interface with recognition and identification of speech and emotional state by timbre of a speaker's voice based on training of the neural network component.

At the present time, «Neurorobot» company has been the first in the world to create prototypes of neuro robots able to recognize words and complete expression, to determine the emotional component of living beings, as well as to identify individual subjects by the timbre (frequency). Parameterization of digital images of words, phonemic images of alphabet of a particular language and emotional components of the living beings' voices is formed in the training sample for the most typical representatives of human beings and animals; neural network is trained, and the further recognition is made based on it.

Potential partners who might be interested in these technologies include military industrial complex, oil and natural gas industry, factories of computer technology production, aerospace industry, and public health services.

Acknowledgements

The research presented in this paper was supported by the TSU Academic D.I. Mendeleev Fund Program (project No. 8.2.31.2015) and the Russian Foundation for Basic Research (grant No 16-29-04388/17). The authors are grateful to Tatiana B. Rumyantseva from Tomsk State University for her assistance in preparing the paper.

Reference

[1] A.O. Gafurov, The 58th Republican scientific conference of young scientists and graduate students dedicated to the 70th anniversary of the Kazakh National University, 122 (2004)
[2] O.M. Gafurov, A.O. Gafurov, All-Russia scientific and technical conference «Science. Defense. Industry», 34 (2004)
[3] G.A. Kuharev, *Biometric Systems: Methods and Tools for Identification of a Human Person* (Polytechnics, SPb., 2001)
[4] O.M. Gafurov, V.I. Syryamkin, A.O. Gafurov, S.S. Stolyarova, Telecommunications and radio engineering **71**, 1565 (2012)