UDC 004.934.8:004.52

ARCHITECTURE AND TRAINING ALGORITHM FOR NEURAL NETWORK TO RECOGNIZE VOICE SIGNALS

Molchanova V. S. – PhD, Associate Professor, Department of Informatics, SHEU "Pryazovskiy State Technical University", Mariupol, Ukraine.

Mironenko D. S. – PhD, Head of department, Department of Informatics, SHEU "Pryazovskiy State Technical University", Mariupol, Ukraine.

ABSTRACT

Context. Typically, interaction between user and mobile devices is realized by touchings. However, many situations, when to implement such interaction is too awkward or impossible, exist. For example, with some diseases of musculoskeletal system, motility of movements may be impaired. It leads to inability to use device efficiently. In that case, a task of looking for alternative ways of person-device interaction becomes relevant. Voice interface development can be one of the most prospective tasks in that way.

Objective. The goal of the study is to develop a project of neural network architecture and internal components for voice-controlled systems. Resulting interface have to be adapted for processing and recognition Ukrainian speech.

Method. An approach, based on audio signal analyzing by sound wave shape and spectrogram, is used for making got via microphone data, appropriable for processing. Using neural network makes possible sounds classification by generated audio signal and information of its transcription. The neural network structure is completely adapted to peculiarities of Ukrainian phonetics. It takes into account the nature of the sound wave, generated during sound pronunciation, as well the number of sounds in Ukrainian phonetics.

Results. Experiments were carried out aimed to choosing optimal neural network architecture and training sample dimension. The root-mean-square deviation of neural network error was used as the main criterion in assessing its effectiveness. A comparative analysis of effectiveness of the proposed neural network and existed on the market speech recognition tools showed improvement in the relative measures of recognition by 9.26%.

Conclusions. Obtained in the research results can be used for full-featured voice interface implementation. Despite the fact that the work is focused on recognition Ukrainian speech, the proposed ideas can be used during developing transcribing services for other languages.

KEYWORDS: voice interface, audio signal, signal amplitude, spectrogram, neural network, training set, standard deviation.

NOMENCLATURE

$A(t)$ is a dependence of sound signal amplitude from time over a continuous time period;

$A'(t_k)$ are discrete values of audio signal amplitude;

e_i is an experimental error of learning a neural network a i-th iteration;

E is a permissible error of neural network learning;

f_j is a function of converting the sound wave characteristics at definite time moment to a sound reference, which generates sound wave with corresponding characteristics;

h is a neuron of a hidden layer of a neural network;

H is a set of hidden layer neurons;

$H(t)$ is a dependence of sound signal frequency from time during some continual time period;

$H'(t_k)$ are discrete values of sound wave frequencies;

i is a neuron of an input layer of a neural network;

I is a set of hidden layer neurons;

J is an index of a word, sound, or speech fragment in a predefined alphabet R;

L is a size of training sample;

n is a size of output alphabet;

N is a number of neurons in the neural network;

N_I is a number of neurons in a input layer of the neural network;

N_H is a number of neurons in a hidden layer of the neural network;

N_O is a number of neurons in a output layer of the neural network;

o is a neuron of an output layer of a neural network;

O is a set of output layer neurons;

P is a training sample;

p_j is probability of matching the sound wave to an element;

R is an alphabet of sounds and words;

r_j is a single, predefined sound, word, or speech fragment;

t is a time;

t_k is a discrete time;

T is a test sample.

INTRODUCTION

Voice recognition methods in conjunction with speech-to-text technologies is a very important tool for creating voice interfaces. Quality voice interface can act as a full-fledged alternative to traditional interactive one for hands-free systems. Such services are especially useful for users, suffering from musculoskeletal disorders. Due to impaired movements coordination, it is difficult and often almost impossible for them to interact device touching screen. Other way for using voice interface is in situations when user's hands are busy with executing another action.

The object of research is a is a process of speech -to-text converting.

The subject of research is neural networks application in development transcribation services adapted on pronunciation features in Ukrainian, able to recognize separate speech fragment.

DOI 10.15588/1607-3274-2020-3-9

The purpose of the paper is to develop a method for recognition voice commands and to implement based on it service, which provides a voice interface for devices running Android. As a main mathematic tool is supposed using neural networks. This service will let use phone, tablet, etc. fully, without necessity to touch its screen by fingers. The service will be represented as an additional add-in, which can recognize voice commands in natural human language and translate them into device control commands.

To achieve the goal, it is necessary to solve a number of tasks:

1) To analyze main characteristics of sound waves generated when sounds and words are pronounced in Ukrainian;

2) To consider existing approaches to recognizing sounds by nature of their sound waves from the viewpoint of possibility to adapt them for recognition voice commands in Ukrainian;

3) To develop architecture of the neural network and to carry out its training;

4) To test the proposed neural network and to compare its efficiency in recognizing voice commands in Ukrainian to existing voice transcribation tools efficiency.

1 PROBLEM STATEMENT

Let some alphabet $R = \{r_1, r_2, ..., r_n\}$, is given, each its element r$_i$ can correspond to a separate sound, word or speech fragment.

As sound signal has arrived, sound wave is generated. According to [1–3], dependences of sound wave amplitude and frequency from time are main characteristics of sound wave which identify uniquely the nature of sound. It lets identify sounds basing on sound wave data and/or spectrogram for a certain time period (1, 2), as well as perform the inverse transformation digital data of the sound wave into analog sound.

$$f_1(A'(t_k), t_k) \rightarrow \{p_j\}_n, \quad (1)$$

$$f_2(A'(t_k), H'(t_k), t_k) \rightarrow \{p_j\}_n. \quad (2)$$

The functions f$_1$ and f$_2$ implement determination probability of correspondence signal, which has generated the investigated sound wave, and each sound from the alphabet *R*. The first one uses only its amplitude during a certain short time period, the second one has an extra parameter, which let use extra information about the signal frequency. The sound classification process is implemented by artificial neural networks.

The final decision, whether the sound signal corresponds to some element r_j is taken based on corresponding to it maximum p_j (3):

$$r_j \rightarrow p_j : p_j = p_{\max} = \max(p_j),\ p_{\max} >> 1 - p_{\max}. \quad (3)$$

Designing the set R, it is appropriate to include one element corresponds to an empty result, i.e. to a situation, when no one sound from set has been identified. Restriction $p_{\max} >> 1 - p_{\max}$ at (3) let avoid false positives, when found probability maximum isn't quite large to classify sound generated sound wave to corresponding class.

2 REVIEW OF THE LITERATURE

The idea to develop a service, which is capable to recognize voice commands and speech in a natural language, isn't essentially new. Developments in speech recognition and voice controlling have been ongoing for over 20 years.

The Dragon Dictate Naturally Speaking system [4] was one of the first software systems capable to perform human speech recognition. In some cases, its recognition accuracy reached 95%. However, high recognition rates were achieved only for speech in English with certain pronunciation rate. In 1997, an attempt to adapt this system for recognition speech in Russian was made. Thus, the Gorynych system was developed [5]. The Gorynych system supported possibility to dictate text and to control some Windows functions with a voice. Meanwhile, speech recognition quality rarely exceeded 30%, which isn't an acceptable result. No attempt to adapt this system to speech recognition in Ukrainian has been made.

This direction of researches became especially popular when mobile devices running on Android and IOs operating systems, appeared. Nowadays, next voice assistants are the most popular: Siri (Apple) [6], Alexa (Amazon) [7, 8], Google Voice Assistant [8], etc. However, they are focused on voice input of text messages, which afterwards usually send via standard messengers, and keywords, which are used for searching information on the Internet. These tools provide satisfactory results within declared capabilities. Meanwhile, they are sensitive to pronunciation quality as well as to speech timbre. An essential disadvantage of these systems is necessity for permanent Internet connection. This disadvantage is caused by necessity of voice processing on the server side of application. In addition, in this case, whether confidentiality of the transmitted information will be retained and possible further ways of using it remain unclear. Another essential disadvantage of these services is limited set of available languages. So, for example, none of above mentioned voice assistants is able to recognize voice commands in Ukrainian. In addition, these systems are only voice assistants. None of them is a fully-featured voice interface.

The technologies of converting speech into a command or text used for developing voice assistants can also be used for creation full-fledged voice interfaces. Moreover, developers provide complete API for this: Google Speech API, YandexToolkit API etc. In addition, there are specialized platforms, which main task is to convert voice to text, for example, PocketSphinx. Development of an original service based on these technologies will provide a full-fledged voice interface for devices managed by

DOI 10.15588/1607-3274-2020-3-9

Android. However, it won't provide to fix the rest disadvantages of existing voice recognition services. Therefore, it is useful to develop an original application, which core will be focused on linguistic features of Ukrainian.

As a result of review literature sources devoted to the problem of converting speech into text [9–12], a number of tasks, which have to be solved sequentually were identified, as well as results which have to be obtained in course of solving each of them (Fig. 1).

At this stage the most interest represents the task of recognizing voice signals and converting them to text.

There are 3 main approaches to speech recognition algorithms implementation: hidden Markov models, dynamic programming, and artificial neural networks.

The approach based on hidden Markov models [13] needs long-term system debugging on large sets of test samples. This approach is quite simple from implementation viewpoint. In addition, enlargement set of recognizable words does only a little effect to computational complexity increase. However, it doesn't guarantee high accuracy of result, because to estimate error value reliably isn't always possible

The approach based on using dynamic programming [14] presupposes comparing two speech segments and determination difference indicator between them. Known in advance pattern is used as a first segment, identifiable – as a second. Using dynamic programming in this approach lets perform optimization and determine the template, which most accurately matches the recognized one. This approach lets get good results for low time and computational costs for small data samples, upon a condition, if a recognizable pattern matches to one element is in set. However, even slight increase of test data sample or outputs variants leads to significant complication of the calculation model.

The most powerful tool for solving speech recognition problem is artificial neural networks [15]. This approach provides not only individual words and sounds recognition, but continuous speech. Using neural networks represents the most interest for development speaker-independent speech recognition systems. Nevertheless, in a cause of complexity of neural network structure determining and its proper training, using neural networks would be recommended only if two previous approaches proved ineffective.

Taking into account specifics of the project being developed (necessity for quick adaptation to specifics of each person's pronunciation and possible minor diction disturbances characteristic for disabled), approach based on using neural networks is the most prospective.

3 MATERIALS AND METHODS

Received from microphone sound signal is a sound wave with continually changing frequency $H(t)$ and amplitude $A(t)$. Amplitude determines the sound volume, and frequency – its tone. At the same time, digital sound processing and recognition can be performed only for discrete data sets $A'(t_k)$, $H'(t_k)$. Thus, for further full-fledged signal processing, it is necessary to perform transformations (4, 5):

$$A(t) \to A'(t_k). \tag{4}$$

$$H(t) \to H'(t_k). \tag{5}$$

The procedure to replace continuos dependence of sound charasteristics to discrete ones, is called sampling, in doing so sampling frequency determines quality of result discrete signal. Usually the sampling rate is in range of 8–48 kHz.

As patterns were used discretized sound signals, gained in pronouncing all kinds of sounds and their typical combinations, characteristic for Ukrainian phonetic (Table 1).

Fig. 2 shows typical fragments of sound waves generated while some sounds of Ukrainian phonetic are being pronounced.

Shapes of sound waves, as well as spectrograms, make it possible to come up with information about amplitude $A'(t_k)$ and frequency $H'(t_k)$ of sound wave at each point of discrete time scale. This data can be used by the neural network to classify sounds.

Nowadays, multilayer neural networks are the most popular and efficient tool in speech recognition.

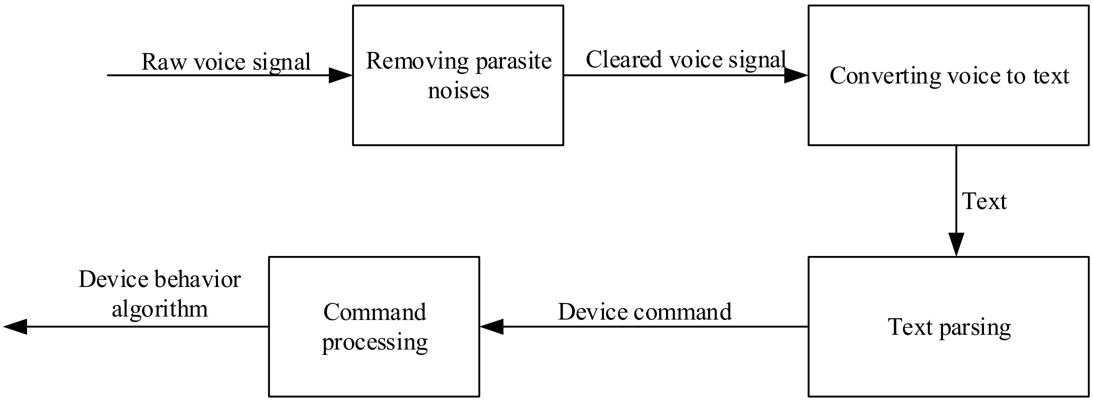
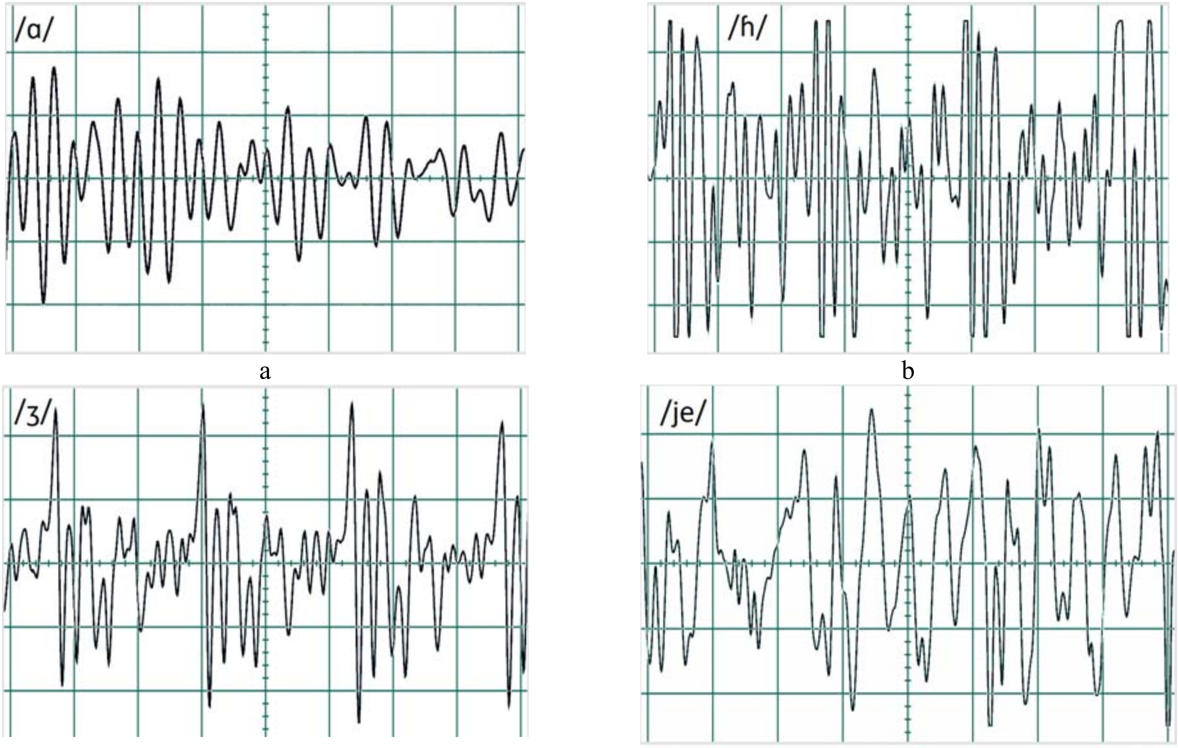


Figure 1 – A structured scheme of a typical voice command recognition and procession system

© Molchanova V. S., Mironenko D. S., 2020
DOI 10.15588/1607-3274-2020-3-9

Table 1 – Transcriptions of some Ukrainian alphabet letters and its combinations

№	Letter	Transcription
1	А а	/ɑ/, /ɐ/
2	Б б	/b/
3	В в	/ʋ/, /w/, /ʍ/, /u̯/
4	Г г	/ɦ/
5	Ґ ґ	/g/
6	Д д	/d/
7	Е е	/ɛ/, /e/
8	Є є	/je/, /ʲe/
9	Ж ж	/ʒ/
…	…	…



a

b

c

d

Figure 2 – Examples of sound waves shapes for some sounds of Ukrainian phonetic

To solve this task, it is assumed to use a three-layer neural network, which contains an input layer, one hidden layer and an output layer. Number of neurons are in input layer is defined as N_I, in output layer – N_0, in hidden layer – N_H. Denote by $I = \{i\}_{N_I}$, $O = \{o\}_{N_0}$, $H = \{h\}_{N_H}$ elements are in input, output, and hidden layers, respectively. Number of all neurons in the neural network is labeled as N, limit permissible learning error value as E.

Speech recognition will be performed based on the sound waveform data. To the input of neural network, data is provided sound wave frequency at each discrete time instant.

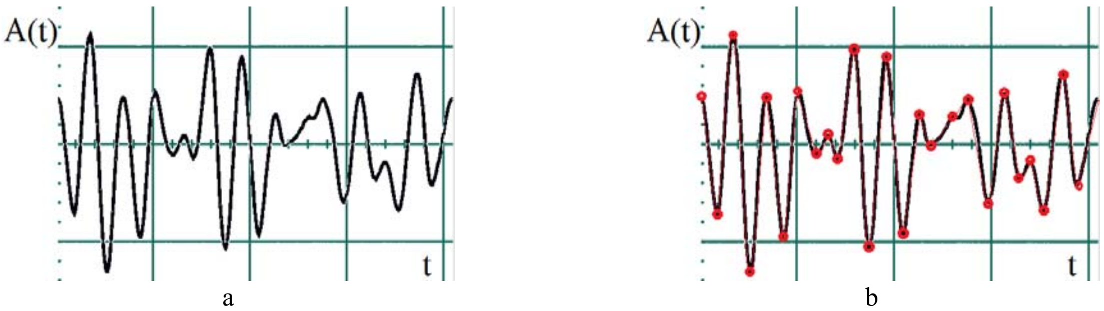
This approach implementation will lead to necessity to use input data vectors with length of 1000 or more elements. Algorithms used for the corresponding neural network learning will be demanding on computing resources, and got neural network will not always be able to provide result required accuracy.

To optimize neural network structure, the assumption was made that it is appropriable to take in attention only extreme values of the sound waves amplitudes $A(t)$ and time moments t, when they're got detected (Fig. 3). Later, information about sound wave amplitudes between extreme values can be obtained by linear approximation.

Consequently, to inputs of the neural network, pairs $(t, A(t))$ will be supplied, and its number will be reduced to 20–30.

Number of output layer neurons corresponds to number of sounds which have to be recognized. Ukrainian phonetics involves 38 different sounds. Thus, number of neurons in the neural network output layer needed to recognizing every individual sound in Ukrainian phonetic is 38. As an activation function of the output layer a linear function is chosen.

DOI 10.15588/1607-3274-2020-3-9



a b

Figure 3 – Linear approximation of sound wave sections between it's extremums

Number of hidden layer neurons depends on some different factors: as training sample size, number of input layer neurons, number of output layer neurons, etc. An initial approximation of this value N_{H_O} can be defined as the average between number of input and output layers neurons. Afterwards a learning error e_i is calculated. Further optimization of neural network structure is performed by reducing (increasing) number hidden layer neurons and constructing a learning curve. The most optimal is considered solution, which provides learning error the closest to acceptable neural network learning error.

$$N_{H_O} = \frac{N_i + N_O}{2},\ e_0 = E(N_{H_0}),$$
$$N_{H_i} = N_{H_{i-1}} \pm 1,\ e_i = E\left(N_{H_i}\right),$$
$$\left|e_i - e\right| \to \min . \qquad (6)$$

As an activation function of hidden layer neurons hyperbolic tangent is chosen.

Created neural network was trained by the direct error propagation method. Each element in training sample P^L contains a vector, which dimension corresponds to number of neural network inputs, and a single integer value, which determines corresponding output of the neural network. Each element of training sample vector is represented as a pair of values $(A(t),t)$.

For three-layer neural network training sample size L is determined by relationship (7):

$$2 \cdot \left(N_i + N_H + N_0\right) \le L \le 10 \cdot \left(N_i + N_H + N_0\right). \qquad (7)$$

Thus, a training sample P^L is formalized by the expression

$$P^L = \left\{ \left(\begin{matrix} \left(A(t_1),\ t_1\right),\ \left(A(t_2),\ t_2\right), \ldots, \\ \left(A\left(t_{N_i}\right),\ t_{N_i}\right) \end{matrix} \right),\ o_k \in \{O\}_{N_0} \right\}_L . \qquad (8)$$

Final assessment of developed neural network efficiency is executed on test samples T, which hasn't been used in neural network training. Test sample for the proposed in this paper neural network is formalized by expression (9)

$$T = \left(\left(\begin{matrix} \left(A(t_1), t_1\right),\ \left(A(t_2), t_2\right), \ldots, \\ \left(A\left(t_{N_i}\right), t_{N_i}\right) \end{matrix} \right), o_k \in \{O\}_{N_0} \right). \qquad (9)$$

It should be noted that during test process, the neural network uses ready output values only to estimate error, but not to improve result.

4 EXPERIMENTS

Experiments on the developed neural network were carried out using the original application Voicer, which implements the neural network proposed in the paper. The application is adapted for running on any devices managed by Android. Android Studio was chosen as the development environment, because of nowadays it is the most popular tool for developing Android applications. The neural network is implemented with TensorFlow Mobile library.

The application Voicer has a friendly user interface and let recognize the voice command received from the device's microphone without necessity to delve into the structure and principles of the neural network.

During the experiment was being carried out, various neural network architectures were analyzed, as well as options for number of samples in training set. As determining criterion for choosing neural network structure was admitted the standard deviation MSE. The module of difference between its value and permissible neural network error has to be minimal. It should be noted that we shouldn't try to minimize MSE to 0 value, in this case the effect of retraining neural network is possible and as a subsequent, incorrect result will be got on test samples.

5 RESULTS

Preliminary calculations have let limit number of neurons in the hidden layer (6) in range from 20 to 38 and training sample dimension (7) in range from 2×N to 10×N. However, not each of possible architectures lets create a neural network efficient to solve assigned task. As an indicator of the neural network efficiency mean-squared error was used. We've tested each of available architectures and estimated its efficiency. Results are assembled in table 2. Table columns contain dimension of test sample set, rows – number of hidden layer neurons and in the cells are mean square error values for respective neural network and test sample set.

DOI 10.15588/1607-3274-2020-3-9

Presented in Table 2 experimental results show that the best quality in recognizing sounds in Ukrainian was got if training sample set dimension was in range from 4 × *N* to 8 × *N*, and number of hidden layer neurons was in range from 24 to 31. For illustrative purposes, this area is highlighted in the table and data, contained in, are presented in diagram (Fig. 4). Separately, minimum mean square error values for each number of elements in training sample set were highlighted.

Examined alternative tools for speech recognition and compared their efficiency to efficiency of developed one. To pursuit of testing process some commonly known voice assistants as Siri, GoogleVoice Assistant, Alexa were used. The initial sample of words is constructed in such a way that it uses all sounds characteristic of pronunciation in Ukrainian. Each word was pronounced 100 times by different voices, with different intonations and timbre. Information about number of correct recognitions is summarized in table 3.

Table 2 – Results of experiments

L N_H	2×N	3×N	4×N	5×N	6×N	7×N	8×N	9×N	10×N
20	7.39	4.56	1.90	10.27	15.4	15.78	16.34	18.23	23.98
21	5.62	4.21	1.76	9.81	12.1	15.26	17.39	20.35	25.35
22	4.62	3.76	0.52	7.92	8.87	14.83	15.82	13.50	18.76
23	3.28	5.12	0.31	3.96	5.12	13.67	10.05	5.52	13.34
24	1.40	1.22	0.08	2.39	4.63	4.36	4.52	7.84	12.18
25	5.06	5.57	0.26	1.59	1.89	7.12	5.65	12.96	14.73
26	5.09	3.58	0.28	1.12	1.77	5.65	5.95	13.14	18.23
27	6.28	5.75	0.25	1.28	0.91	2.09	3.39	12.96	20.34
28	7.25	2.85	0.54	0.69	0.15	0.77	1.34	12.67	14.36
29	8.45	2.58	1.24	0.92	0.28	0.98	1.73	8.93	12.34
30	10.63	5.84	2.83	2.63	0.81	0.24	3.56	7.65	15.28
31	14.48	7.36	3.25	2.09	1.95	0.56	7.82	12.84	18.56
32	12.54	4.56	5.60	4.89	3.12	0.33	4.23	9.23	12.90
33	15.94	12.34	8.96	6.98	5.86	1.98	7.12	34.19	18.76
34	16.75	15.11	11.7	12.03	12.98	2.09	15.74	20.65	21.63
35	20.73	18.39	12.1	15.09	19.27	4.51	10.09	23.12	23.45
36	23.86	20.00	11.3	18.82	21.43	10.94	9.33	8.95	28.32
37	23.12	21.98	14.0	16.31	20.60	12.25	17.52	12.51	20.52
38	15.94	16.93	14.2	18.59	21.85	23.15	25.12	23.18	18.74

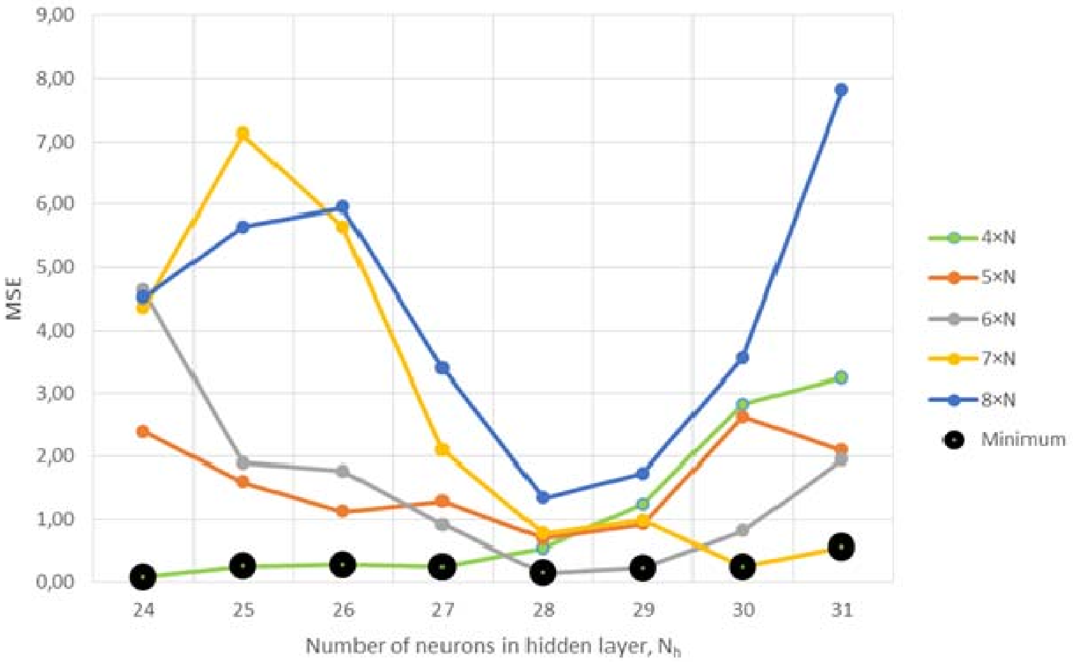


Figure 4 – Dependence of mean square error of the neural network from number of neurons are in hidden layer and dimension of training sample set

© Molchanova V. S., Mironenko D. S., 2020
DOI 10.15588/1607-3274-2020-3-9

Table 3 – The number of correct recognition Ukrainian words got by various tools

Word	Transcription	Siri	Google Voice Assistant	Alexa	Cortana	Proposed tool	Relative improvement in recognition quality
Дощ	[д о ш ч]	70	77	65	68	91	15.38%
М’який	[м й а к и й]	73	75	54	66	88	14.77%
Сьогодні	[с' о г о д н' і]	77	78	56	37	89	12.36%
Прикмета	[п р и к м е т а]	65	70	67	76	87	12.64%
Підсолоджувач	[п' і д с о л о дж у в а ч]	82	75	45	55	80	–2.50%
Мережа	[м е р е ж а]	77	90	57	46	90	0.00%
Відпустка	[в' і д п у с т к а]	80	74	45	36	88	9.09%
Боротьба	[б о р о д'' б а]	75	70	65	50	83	9.64%
Місяць	[м' і с' а ц']	79	81	74	55	92	11.96%
Average		75,33	76.67	58.67	54.33	87.56	9.26

6 DISCUSSION

In accordance with table 2, we can conclude that there are several cases, when the best result is achieved. So, for example, for a neural network, which number of hidden layer neurons is in range from 24 to 27 neurons, optimal dimension of training sample is 4 × N. If number of hidden layer neurons is 28 or 29, as optimal dimension of training sample set will be $6\times N$, and if number of hidden layer neurons is 30 or 31, as optimal dimension of training sample set will be $7\times N$. In addition, the minimum values of the neural network mean square error 0.08, 0.15 and 0.24 are highlighted in the table 2.

Further experiments are reduced to comparing implemented neural network efficiency to efficiency of ready-made tools available on market. It should be noted that any instruments, able to recognize Ukrainian phonetics, are not currently available on market. However, an attempt to adapt tools, used for recognizing sounds in Russian, for recognition in Ukrainian was carried out. Ready tools usually recognize not individual sounds but whole words. These words are being formed by a sequence of sounds, therefore, the neural network proposed in the paper is able to cope with this task successfully. A spoken word is considered recognized incorrectly if its transcription differs from sequence of sounds got by neural network.

On the basis of comparing efficiency of attempts to adapt ready-made speech recognition tools to phonetic features in Ukrainian and tool, proposed in the paper, we can conclude advantages of the last one. So, in considered examples, relative improvement in quality of recognition is 12.3%. This is due to the initial orientation developed neural network to recognition speech in Ukrainian.

CONCLUSIONS

Got result is satisfactory, so neural network, proposed for recognizing sounds in Ukrainian, can be used to develop a full-fledged system for voice control on Android devices.

The scientific novelty of the obtained results is that the method for optimizing data of sound waves formed by pronunciation of sounds and their combinations in Ukrainian was firstly proposed. The proposed method is based on using only extreme values of sound waves characteristics and obtaining intermediate data by linear approximation. It reduced the training sample dimension and increased the data processing speed without losing quality of the result.

The practical significance of obtained results is that digitized sound waves, are generated during pronouncing separate sounds in Ukrainian, can be subjected to further intellectual analysis in order to search elements of certain device commands and their parameters. It is undoubtedly a very important element in building a full-fledged interface with voice control. However, considering this problem isn’t within the scope of this study.

Prospects for further research are to train the proposed neural network to recognize whole words and collocations in Ukrainian and in other national languages.

REFERENCES

1. Tohyama M., Tsunehiko K. Fundamentals of Acoustic Signal Processing. Boston, Academic Press, 1998, 321 p. DOI: 10.1121/1.429575
2. Giannakopoulos T., Pikrakis A. Introduction to Audio Analysis: A MATLAB Approach. Fl, Academic Press, 2014, 288 p.
3. Lerch A. An introduction to audio content analysis. Applications in signal processing and music informatics. Hoboken, Wiley, 2012, 259 p. DOI: 10.1002/9781118393550
4. Poulter C. Voice recognition software – Nuance Dragon naturally speaking, *Occupational Medicine*, 2020, Vol. 70, Issue. 1, pp. 75–76. https://doi.org/10.1093/occmed/kqz128
5. Kumar A., Paek T., Lee B. Voice recognition software – Nuance Dragon naturally speaking/ *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 5–10 May 2012: proceedings.* Los Alamitos, IEEE, 2012, pp. 2277 –2286.
6. Natale S. To believe in Siri: A critical analysis of AI voice assistants, *Communicative Figurations Working Papers*, 2020, Vol. 32, pp. 130–146.
7. Pham C., Yuto L, Yasuo T. A platform for integrating Alexa Voice Service into ECHONET-based smart homes, *Proceedings of International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 19–21 May 2018: proceedings.* Taichung, IEEE, 2018, pp. 895–902. DOI: 10.1109/ICCE-China.2018.8448893
8. Guzman A. L. Making AI Safe For Humans: A Conversation With Siri, *Socialbots and Their Friends: Digital Media and the Automation of Sociality/ edited by R. W. Gehl.* Routledge, 2017, Chapter 4, pp. 70–85.
9. Khilari P., Bhope V. P. A review on speech to text conversion methods, *International Journal of Advanced Research in Computer Engineering & Technology(IJARCET),* 2015, Volume 4, Issue 7, pp. 3067–3072.
10. Deepa V. J., Mustafa Alfateh, Sharan R. A Novel Model for Speech to Text Conversion, *International Refereed Journal of*

DOI 10.15588/1607-3274-2020-3-9

Engineering and Science (IRJES), 2014, Volume 3, Issue 1, pp. 239–245.

11. Trivedi Ayushi, Pant Navya, Shah Pinal, Sonik Simran and Agrawal Supriya Speech to text and text to speech recognition systems-Areview, *IOSR Journal of Computer Engineering*, 2018, Volume 20, Issue 2, pp. 36–43. DOI: 10.9790/0661-2002013643.
12. Saksamudre Suman K., Shrishrimal P. P., Deshmukh R. R. A Review on Different Approaches for Speech Recognition System, *International Journal of Computer Applications*, 2020, Volume 115, No. 22, pp. 385–396. DOI: 10.5120/20284-2839
13. Chavan Rupali S., Sable Ganesh S. An Overview of Speech Recognition Using HMM, *International Journal of Computer Science and Mobile Computing*, 2013, Vol. 2, Issue 6, pp. 233–238.
14. Furtună, T. F. Dynamic Programming Algorithms in Speech Recognition, *Revista Informatica Economicănr,* 2008, Vol. 2(46), pp. 94–98.
15. Graves A., Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947
16. Maas M. Lexicon-Free Conversational Speech Recognition with Neural Networks, *NAAC,* 2015, pp. 156–163. DOI: 10.3115/v1/N15-1038
17. Ossama A.-H., Mohamed A. R., Jiang H. etc. Convolutional Neural Networksfor Speech Recognition, *ACM transactions on audio, speech, and language processing*, 2014, Vol. 22, No. 10, pp. 1533–1545. DOI: 10.1109/TASLP.2014.2339736
18. Lekshmi K. R., Sherly E. Automatic Speech Recognition using different Neural Network Architectures – A Survey, *International Journal of Computer Science and Information Technologies*, 2016, Vol. 7 (6), pp. 242–248.
19. Aha D. W., Kibler D., Albert M. K. Instance-based learning algorithms, *Machine Learning*, 1991, No. 6, pp. 37–66. DOI: 10.1023/A:1022689900470

Received 14.07.2020.
Accepted 27.08.2020.

УДК 004.934.8:004.52

АРХІТЕКТУРА ТА АЛГОРИТМ НАВЧАННЯ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ РОЗПІЗНАВАННЯ ГОЛОСОВИХ СИГНАЛІВ

Молчанова В. С. – канд. техн. наук, доцент кафедри інформатики ДВНЗ «Приазовський державний технічний університет», Маріуполь, Україна.

Міроненко Д. С. – канд. техн. наук, завідувач кафедри інформатики ДВНЗ «Приазовський державний технічний університет», Маріуполь, Україна.

АНОТАЦІЯ

Актуальність. Зазвичай взаємодія користувача з мобільним пристроєм, наприклад, телефоном або планшетом реалізується за допомогою торкань. Однак можливий цілий ряд ситуацій, коли здійснення такого способу людино-машинної взаємодії виявляється скрутним або навіть неможливим. Наприклад, при деяких захворюваннях опорно-рухового апарату можливе порушення моторики рухів, що в свою чергу призводить до неможливості повноцінно використовувати пристрій, помилок, втрати часу. У такій ситуації актуальним стає пошук альтернативних шляхів взаємодії користувача з системою. Розробка голосових інтерфейсів є одним з найбільш перспективних напрямків даної роботи.

Мета дослідження полягає в розробці методу оптимізації даних звукових хвиль і їх застосуванні при навчанні нейронної мережі для розпізнавання голосових сигналів, утворених вимовою звуків українською мовою.

Метод. Для реалізації проекту запропонованої у роботі системи, використовується підхід, заснований на аналізі аудіосигналу за формою утвореною їм звукової хвилі і спектрограми, а також застосуванні штучних нейронних мереж у процесі подальшої класифікації та виділення окремих, характерних для української мови, звуків. Нейронна мережа являє собою тришаровий персептрон, структура якого повністю адаптована під особливості української фонетики. Враховується характер звукової хвилі, яка утворюється під час вимови того чи іншого звуку, а також кількість різноманітних звуків в українській фонетиці.

Результати. Проведено ряд експериментів, спрямованих на вибір оптимальної архітектури нейронної мережі і розмірність навчальної вибірки. В якості основного критерію при оцінці ефективності нейронної мережі використовувалося середньоквадратичне відхилення її помилки. В процесі тестування було визначено кілька варіантів комбінацій параметрів нейронної мережі, при яких досягалися найкращі результати. Порівняльний аналіз ефективності запропонованої в роботі нейронної мережі й існуючих на ринку інструментів розпізнавання голосу показав поліпшення відносних показників розпізнавання на 9,26%.

Висновки. Отримані в роботі результати досліджень і архітектура нейронної мережі можуть бути використані під час реалізації повноцінного голосового інтерфейсу для мобільних пристроїв під управлінням операційної системи Android. Незважаючи на те, що робота орієнтована на розпізнавання мовлення українською мовою, ідеї які використовуються для її реалізації можуть бути використані при транскрібації голосу на інших мовах..

КЛЮЧОВІ СЛОВА: голосовий інтерфейс, аудіосигнал, амплітуда сигналу, спектрограмма, нейронна мережа, навчальна вибірка, середньоквадратичне відхилення.

УДК 004.934.8:004.52

АРХИТЕКТУРА И АЛГОРИТМ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ ГОЛОСОВЫХ СИГНАЛОВ

Молчанова В. С. – канд. техн. наук, доцент кафедры информатики ГВУЗ «Приазовский государственный технический университет», Мариуполь, Украина.

Мироненко Д. С. – канд. техн. наук, заведующий кафедры информатики ГВУЗ «Приазовский государственный технический университет», Мариуполь, Украина.

АННОТАЦИЯ

Актуальность. Обычно взаимодействие пользователя с мобильным устройством, например, телефоном или планшетом реализуется посредствам касаний. Однако возможен целый ряд ситуаций, когда осуществление такого способа человеко-

DOI 10.15588/1607-3274-2020-3-9

машинного взаимодействия оказывается затруднительным или даже невозможным. Например, при некоторых заболеваниях опорно-двигательного аппарата возможно нарушение моторики движений, что в свою очередь приводит к невозможности полноценно использовать устройство, ошибкам, потери времени. В сложившейся ситуации актуальным становится поиск альтернативных интерфейсов взаимодействия пользователя с системой. Разработка голосовых интерфейсов является одним из наиболее перспективных направлений данной работы.

Цель исследования состоит в разработке метода оптимизации данных звуковых волн и их применени при обучении нейронной сети для распознавания голосовых сигналов, образованных произношением звуков на украинском языке.

Метод. Для реализации проекта предложенной в работе системы, используется подход, основанный на анализе аудиосигнала по форме образуемой им звуковой волны и спектрограммы, а также применении искусственных нейронных сетей в процессе последующей классификации и выделении отдельных, характерных для украинской речи, звуков. Нейронная сеть представляет собой трехслойный персептрон, структура которого полностью адаптирована под особенности украинской фонетики. Учитывается характер звуковой волны, образуемой при произношении того или иного звука, а также количество разнообразных звуков в украинской фонетике.

Результаты Проведен ряд экспериментов, направлен на выбор оптимальной архитектуры нейронной сети и размерность обучающей выборки. В качестве основного критерия при оценке эффективности нейронной сети использовалось среднеквадратическое отклонение ее ошибки. В процессе тестирования были определены несколько вариантов комбинаций параметров нейронной сети, при которых достигались наилучшие результаты. Сравнительный анализ эффективности предложенной в работе нейронной сети и существующих на рынке инструментов распознавания показал улучшение относительных показателей распознавания на 9.26 %.

Выводы. Полученные в работе результаты исследований и архитектура нейронной сети могут быть использованы при реализации полноценного голосового интерфейса для мобильных устройств, работающих под управлением операционной системы Android. Несмотря на то, что работа ориентирована на распознавание речи на украинском языке, используемые при ее реализации идеи могут быть использованы при транскрибации речи на других языках.

КЛЮЧЕВЫЕ СЛОВА: голосовой интерфейс, аудиосигнал, амплитуда сигнала, спектрограмма, нейронная сеть, обучающая выборка, среднеквадратическое отклонение.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Tohyama M. Fundamentals of Acoustic Signal Processing / M. Tohyama K. Tsunehiko. – Boston : Academic Press, 1998. – 321 p. DOI: 10.1121/1.429575
2. Giannakopoulos T. Introduction to Audio Analysis: A MATLAB Approach / T. Giannakopoulos, A. Pikrakis. – Fl: Academic Press, 2014. – 288 p.
3. Lerch A. An introduction to audio content analysis. Applications in signal processing and music informatics / A. Lerch. – Hoboken : Wiley, 2012. – 259 p. DOI: 10.1002/9781118393550
4. Poulter C. Voice recognition software – Nuance Dragon naturally speaking/ C. Poulter // Occupational Medicine. – 2020. – Vol. 70, Issue 1. – P. 75–76. https://doi.org/10.1093/occmed/kqz128
5. Voice recognition software – Nuance Dragon naturally speaking/ A. Kumar, T. Paek, B. Lee // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 5–10 May 2012 : proceedings. Los Alamitos: IEEE, 2012. – P. 2277–2286.
6. Natale S. To believe in Siri: A critical analysis of AI voice assistants/ S. Natale // Communicative Figurations Working Papers. – 2020. – Vol. 32. – P. 130–146.
7. Pham C. A platform for integrating Alexa Voice Service into ECHONET-based smart homes / C. Pham, L. Yuto, T. Yasuo // Proceedings of International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 19–21 May 2018: proceedings. Taichung : IEEE, 2018. – P. 895–902. DOI: 10.1109/ICCE-China.2018.8448893
8. Guzman A. L. Making AI Safe For Humans: A Conversation With Siri / A. L. Guzman // Socialbots and Their Friends: Digital Media and the Automation of Sociality/ edited by R. W. Gehl. – Routledge, 2017. – Chapter 4. – P 70–85.
9. Khilari P. A review on speech to text conversion methods / P. Khilari, V. P. Bhope // International Journal of Advanced Research in Computer Engineering & Technology(IJARCET). – 2015. – Volume 4, Issue 7. – P. 3067–3072.
10. Deepa V. J. A Novel Model for Speech to Text Conversion / V. J. Deepa, Mustafa Alfateh, R. Sharan // International Refereed Journal of Engineering and Science (IRJES). – 2014. – Volume 3, Issue 1. – P. 239–245.
11. Speech to text and text to speech recognition systems-Areview / [Ayushi Trivedi, Navya Pant, Pinal Shah et al] // IOSR Journal of Computer Engineering. – 2018. – Volume 20, Issue 2. – P. 36–43. DOI: 10.9790/0661-2002013643.
12. Saksamudre Suman K. A Review on Different Approaches for Speech Recognition System / Suman K. Saksamudre, P. P. Shrishrimal, R. R. Deshmukh // International Journal of Computer Applications. – 2020. – Volume 115, No. 22. – P. 385–396. DOI: 10.5120/20284-2839
13. Chavan Rupali S. An Overview of Speech Recognition Using HMM / Rupali S Chavan, Ganesh. S Sable. // International Journal of Computer Science and Mobile Computing. – 2013. – Vol. 2, Issue. 6. – P.233 – 238.
14. Furtună T. F. Dynamic Programming Algorithms in Speech Recognition/ T. F. Furtună // Revista Informatica Economicănr. – 2008. – Vol. 2(46). – P. 94–98.
15. Graves A. Speech recognition with deep recurrent neural networks / A. Graves, A. Mohamed, G. Hinton // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. – 2013. – P. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947
16. Maas M. Lexicon-Free Conversational Speech Recognition with Neural Networks / M. Maas // NAAC. – 2015. – P. 156–163. DOI: 10.3115/v1/N15-1038
17. Ossama A.-H. Convolutional Neural Networksfor Speech Recognition / A.-H. Ossama, A. R. Mohamed, Jiang H etc. // ACM transactions on audio, speech, and language processing. – 2014. – Vol. 22, No. 10. – P. 1533–1545. DOI: 10.1109/TASLP.2014.2339736
18. Lekshmi K. R. Automatic Speech Recognition using different Neural Network Architectures – A Survey / K. R. Lekshmi, E. Sherly // International Journal of Computer Science and Information Technologies. – 2016. – Vol. 7 (6). – P. 242–248.
19. Aha D. W. Instance-based learning algorithms / D. W. Aha, D. Kibler, M. K. Albert // Machine Learning. – 1991. – № 6. – P. 37–66. DOI: 10.1023/A:1022689900470

DOI 10.15588/1607-3274-2020-3-9