Features of Speech Commands Recognition Using an Artificial Neural Network

Gulmira K. Berdibayeva
Electronics and Robotics
Almaty university of power engineering and telecommunications
Almaty, Kazakhstan
email: horli@mail.ru

Andrey N. Spirkin
Information-measuring technology and Metrology
Penza State University
Penza, Russia
email: spirkin.andre@yandex.ru

Oleg N. Bodin
Information-measuring technology and Metrology
Penza State University
Penza, Russia
email: bodin_o@inbox.ru

Oksana E. Bezborodova
Information-measuring technology and Metrology
Penza State University
Penza, Russia
email: oxana243@yandex.ru

2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) | 978-1-7281-7691-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/USBEREIT51232.2021.9455111

***Abstract*—The article is devoted to the features of speech commands recognition using an artificial neural network. The existing methods of speech recognition are considered. The algorithm of operation of the speech recognition system is presented. The stages of speech signal processing are described, indicating widely used methods. It is shown that the use of neural network analysis increases the efficiency of speech signal processing. As a type of neural network architecture, we used a recurrent bidirectional deep learning neural network Bidirectional Long Short-Term Memory and a method for obtaining mel-frequency cepstral coefficients. The choice of this neural network is due to the fact that the network uses both a priori and a posteriori information. To train the neural network in the study, a ready-made open database of speech commands Google Speech Commands Dataset was used. An experimental study of speech command recognition using BiLSTM artificial neural network was carried out using the Matlab software package and the AudioLabeler additional application.**

Keywords—speech recognition methods, digital signal processing, voice control, speech command recognition, artificial neural network, neural network training

I. INTRODUCTION

The problem of verbal communication between man and machine has been relevant for many years. The first speech recognition device [1] appeared in 1952. It could recognize the numbers spoken by a person. In the early eighties, speaker-dependent speech recognition systems appeared [2]. In them, the sound image of the team was stored as an integral standard. Dynamic programming methods were used to compare the unknown speech signal and the command reference. These systems were good at recognizing small sets of 10-30 commands and understood only one speaker. These systems required a complete reconfiguration to work with another speaker.

The speech technology sector is recognized as one of the fastest growing in the world. According to a report by Markets and Markets [8], the global speech technology market will grow from $ 3.7 billion in 2019 to $ 12 billion by 2022.

The purpose of the article is to study the features of recognition of speech commands and the development of a speaker-independent method of recognition of speech commands based on an artificial neural network

II. MATERIALS AND METHODS

Currently due to advances in the field of information technology, speech recognition has made quite a lot of progress. Modern developments of speech recognition systems are aimed at increasing their speaker independence. Companies such as Google, Apple and Yandex are leaders in this area. All of the aforementioned services use a combination of Hidden Markov Models and neural networks as the main tool for speech recognition. At the stage of preprocessing of the speech signal, mainly cepstral coefficients on the Mel scale are used, or Mel-frequency cepstral coefficients, which make it possible to compactly describe the signal spectrum [3, 4].

When developing a human-machine interface, the easiest to implement a speech recognition system is to use ready-made solutions offered by Google and Yandex. Speech recognition in Google Speech and Yandex Speech Kit programs are performed on high-performance remote servers which contain word libraries. Therefore, this option is applicable only in cases if constant access to the Internet is provided.

The speech recognition system has the following operation algorithm (Fig. 1) [1].

This processing could be done with analog or digital bandpass filters. Modern speech processing systems use digital frequency filters software implemented. The voice command signal is passed through a correcting filter with a transfer function:

$$W(z) = \sum_{k=0}^{m} a_k z^{-k}, \quad (1)$$

where a_k are constant coefficients, m is an integer ($m>0$), k is the coefficient number. Most often, $m=1$, and the transfer function has the form:

$$W(z) = a_0 - a_1 z^{-1}. \quad (2)$$

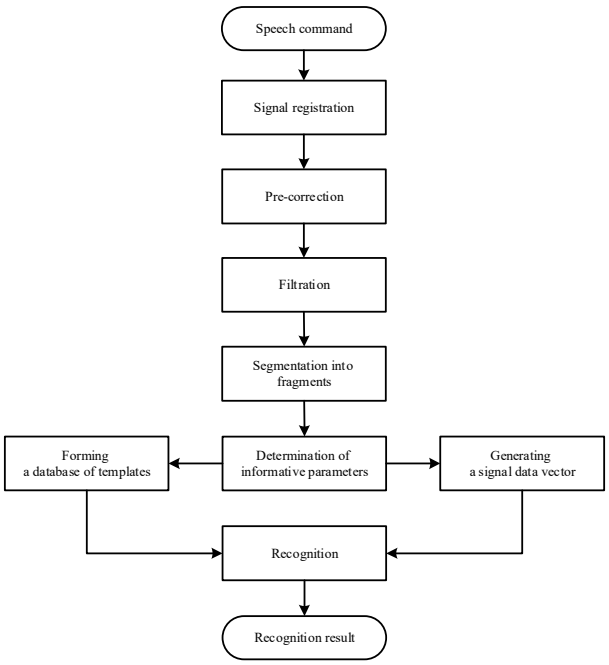

Fig. 1. System operation algorithm of the speech recognition commands

An important stage of input signal preprocessing is signal level normalization. This allows you to reduce recognition errors associated with the fact that the speaker can pronounce words at different volume levels [3].

The extraction of unique signal parameters is the compilation of a chain of vectors of speech signal features. Speech is a non-stationary signal, but due to the inertness of the vocal tract within a short enough period of time (from 10 to 40 ms), its characteristics do not change, that is, it can be considered stationary [2].

The feature extractor works with small fragments of the speech signal, which should overlap each other to improve the recognition accuracy.

The decomposition of the signal into components can be represented as:

$$s(t) = \sum_{i=1}^{I-1} imf_i(t) + r_I(t), \qquad (3)$$

where $imf_i(t)$ is the empirical mode (EM) (Intrinsic Mode Functions, IMF), $r_I(t)$ is the decomposition residue, $i = 1, 2,\ldots, I$ is the EM number.

Formation of the obtained EM spectrum of Hilbert:

$$HHT(t) = \sum_{i=1}^{T} a_i^2(t) \cdot e^{q \int \omega_k(t)dt}, \qquad (4)$$

where $a_i^2(t) = \sqrt{imf_i(t)^2 + IMF_i(t)^2}$ is the module of the instantaneous value of the signal amplitude of each EM $imf_i(t)$ – EM signal, $IMF_i(t) = \frac{1}{\pi}\int \frac{imf_i(\tau)}{t-\tau} d\tau$ – conjugate Hilbert signal EM, τ – time shift proportional to the phase of the signal, $\omega(t) = 2\pi fj$ – cyclic frequency of each EM, j - imaginary unit.

The values $a(t)$ and $\omega(t)$ are determined from the analytical signal $Z_i(t)=imf_i(t)+jIMF_i(t)$ of each EM.

It has been experimentally established that the optimal length of such fragments should correspond to an interval of 10 ms, the «overlap» from 50 %. Studies [3] have shown that speech is best represented by features obtained in the frequency domain. These features include linear prediction coefficients (LPC) [1, 3, 5], perceptual coefficients of linear prediction (Perceptual Linear Prediction – PLP) [3], Mel-Frequency Cepstral Coefficients (MFCC) [3].

The methods using LPC are to compose a feature vector from linear prediction coefficients. These methods are based on the possibility of approximating the current sample of a speech signal using a linear combination of previous samples.

In the method using LPC, there is a drawback associated with the peculiarity of the perception of various frequencies by a person. It has been established [2] that the resolution of the human ear is unevenly distributed over the spectrum: in the low frequency region it is higher than in the high frequency region. This effect is described using a psychoacoustic bark scale [2].

To take into account these features, while retaining the essence of the method for finding LPC, PLP is used. The method using PLP is that, before finding the linear prediction coefficients, the speech signal is passed through filters, the bandwidths of which change in accordance with the barc-scale. The disadvantage of this method in comparison with the method using LPC is a much more complicated calculation of the coefficients due to the imposition of the barc scale on the input signal.

Most frequently used method in modern speech recognition systems is method with MFCC. This is due to the simplicity of its implementation with similar indicators of recognition quality compared to the LPC and PLP methods.

Methods for processing signal parameters (recognition) in modern literature are divided into the following main groups [3]:

– methods based on comparison with a standard;

– methods based on the construction of decision functions;

– methods based on hidden Markov's models;

– hybrid methods.

Thus, the analysis of the methods of speech recognition systems showed that it is more expedient to choose the MFCC method to highlight the characteristic features of a speech signal, and to recognize voice commands, use a method based on the construction of decision functions and implemented using artificial neural networks (ANN).

III. EXPERIMENTAL STUDY OF A SPEECH COMMAND USING AN ARTIFICIAL NEURAL NETWORK

The choice of a suitable neural network structure directly depends on the speed of command execution and on the resources consumed. An analysis of ANN topologies presented in [9] showed that recurrent neural networks, in particular, a bidirectional neural network of long short-term memory (Bidirectional Long Short-Term Memory – BiLSTM) [10]. The choice of this neural network is due to the fact that the network uses both a priori and a posteriori information, i.e. the network will also receive the next letter during the return pass, thus opening access to future information. Therefore, a neural network can be trained not only to supplement information, but also to fill in the gaps, so,

for example, instead of expanding the picture along the edges, it can complete the missing fragments in the middle.

The study used the ready-made open database of speech commands Google Speech Commands Dataset. This base is used to train a neural network.

Thus, the main task of this work is to recognize a keyword in a noisy signal using a neural network. As a type of neural network architecture, a recurrent bidirectional deep learning neural network Bidirectional Long Short-Term Memory and a method for obtaining mel-frequency cepstral coefficients were used (MFCC).

In the work, the key word «выключить» (translation from Russian means «off») is recognized. Recognition takes place in several stages:

1) Checking the baseline of the keyword definition;

2) Creation of a training sample of a speech command signal without interference;

3) Training the BiLSTM neural network for the keyword using MFCC sequences extracted from the speech command signal;

4) Verification of the training accuracy of the neural network by comparing the validation baseline with the network output;

5) Conducting recognition of the speech command.

The validation signal consists of a 34 second speech containing the «выключить» keyword that appears randomly in the signal.

Loading and visualizing the validation signal (Fig. 2).

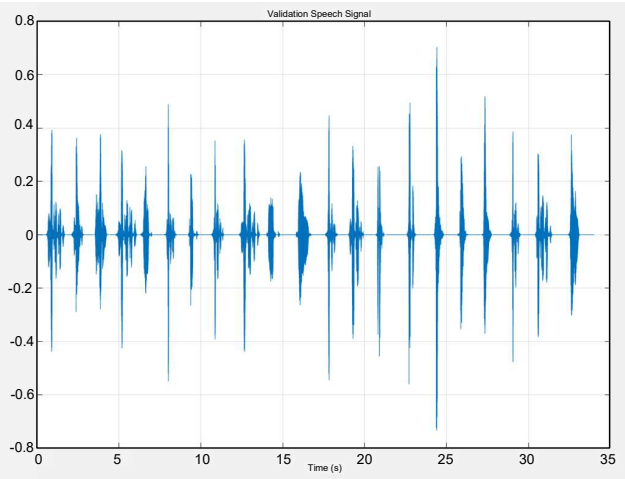

Fig. 2. Visualized validation signal. The speech signal consists of the following words: «выключить, выключить, right, выключить, выключить, left, выключить, happy, house, two, выключить, выключить, marwin, bed, seven, выключить, cat, one, выключить, выключить, one»

Using the additional software package Matlab \ AudioLabeler [6, 7], create a mask of the recorded speech command, which will also be a keyword. First, let's load the recorded speech signal into AudioLabeler. After launching the application, the intervals of the speech signal with words are automatically determined (selected), and masks of all words spoken in the audio recording are created.

The resulting labels of the masks of the words of the recorded audio recording are automatically exported to the Workspace Matlab.

In this work, a six-layer neural network BiLSTM is used, consisting of:

– input;

– two hidden bi-directional layers BiLSTM;

– one fully connected layer;

– the layer containing the logistic activation function;

– classification layer.

A schematic representation of an artificial neural network BiLSTM is presented in the form of a graph (Fig. 3).

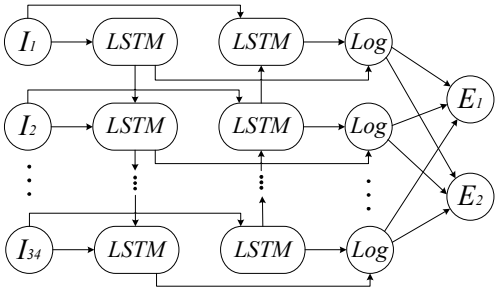

Fig. 3. Schematic representation of an artificial neural network BiLSTM: I – input layer, the number of neurons corresponds to the number of MFCC equal to 34, Log – layer containing logistic activation function, E – output layer, corresponds to the number of classes

Based on the data obtained, we built a baseline for the search for keywords. The result of constructing the baseline for the search for keywords is shown in Fig. 4.

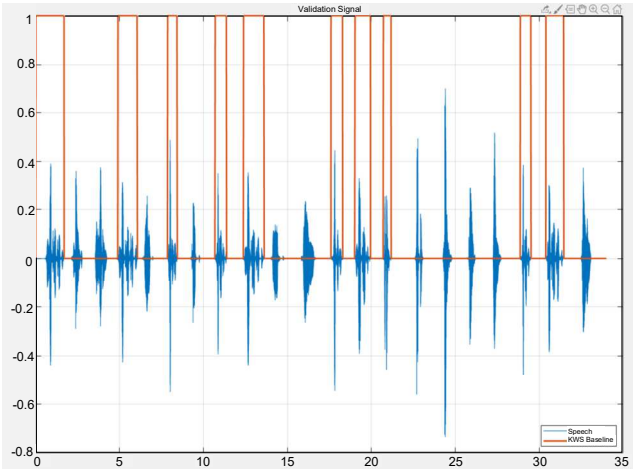

Fig. 4. Speech signal with a plotted baseline of keyword search (KWS)

The training dataset contains approximately 65,000 statements of 30 seconds duration, including the keyword «выключить». The received data was divided into two classes: contain a keyword and no keyword. This data partitioning will allow us to generate training samples containing a maximum of 10 words, while choosing a random location for the keyword.

The deep learning procedure for the neural network is performed by extracting the MFCC coefficients of the sample signal. In this case, the calculation of the MFCC coefficients is carried out by sliding the window on the input signal, and thus, therefore, the feature matrix is shorter than the input

speech signal. Each row in the feature matrix corresponds to 128 samples from the speech signal.

Description of the architecture of the neural network BiLSTM is as follows:

```
layers = [sequenceInputLayer(numFeatures)
bilstmLayer(150,"OutputMode","sequence")
bilstmLayer(150,"OutputMode","sequence")
fullyConnectedLayer(2)
softmaxLayer
classificationLayer];
```

The optimal number of neurons n = 150 of the hidden layer was determined empirically.

The training time for the neural network took 18 minutes 11 seconds. The learning result is shown in Fig. 5.

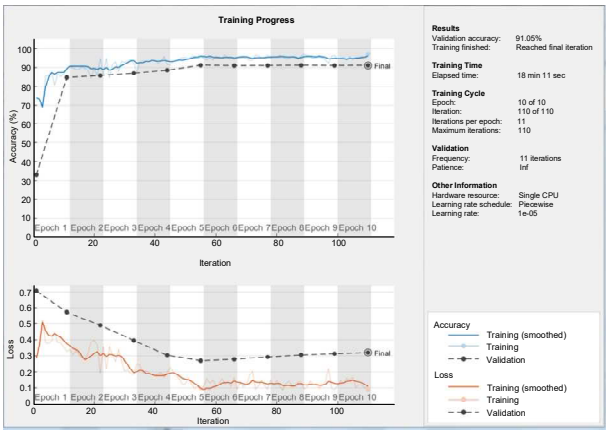

Fig. 5. Results of training the neural network BiLSTM

Analysis of the obtained learning results (Fig.5) shows that the recognition accuracy is 91.05%.

The result of the work of the BiLSTM neural network for keyword recognition is shown in Fig. 6.

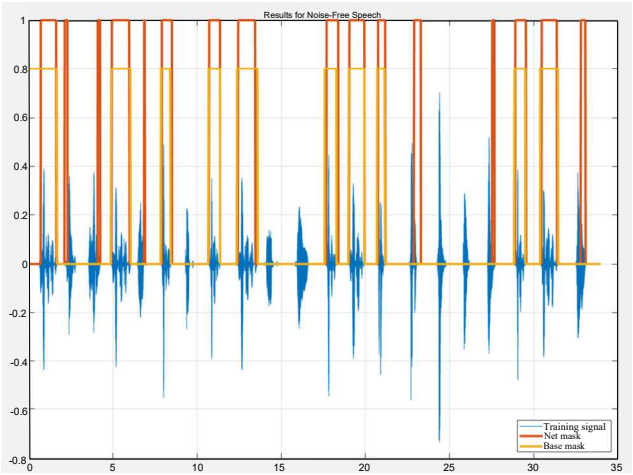

Fig. 6. Visualization of the BiLSTM neural network by keyword recognition

When comparing the obtained recognition results on the graph (Fig.6, red graph), it can be seen that the neural network correctly determines the boundaries of the desired command "turn off" (Fig. 6, orange graph), but at the same time captures some other areas, this is due to the fact that the keyword has similar morphemes from other words. The problem of identifying non-target signal sections can be solved by replenishing the training base of audio recordings.

IV. Conclusion

Thus, the recognition of voice commands using artificial neural networks is less affected by noise and does not depend on the individual characteristics of the operator's voice. The conducted research of the speech command «выключить» using the artificial neural network BiLSTM, which is a recurrent 6-layer neural network, showed the effectiveness of recognition of the speech command. It was determined that the selected artificial neural network adequately defines the boundaries of the required command «выключить» (off). The recognition error does not exceed 10%. It is possible to increase the recognition accuracy by replenishing the training base of audio recordings. The results of this work may represent interest to developers of equipment that is controlled by voice commands.

References

[1] O.S. Agashin and O.N. Korelin, “Methods of digital processing of a speech signal in the problem of recognizing isolated words with the use of signal processes”. vol. 4, Proceedings of the Nizhny Novgorod State Technical University. R.E. Alekseeva, 2012, pp. 32–44.

[2] . L. Rabiner, R. Shafer, Digital processing of speech signals, Moscow: Radio and communication, 1981.

[3] K.A. Makovskin, “Hybrid models: hidden Markov models and neural networks, their application in speech recognition systems”. Computing Center.” A.A. Dorodnitsyn, Moscow, 2006, pp. 40–95.

[4] Z. Guohua, “Ant Colony Clustering Algorithm and Improved Markov Random Fusion Algorithm in Image Segmentation of Brain Images”, International Journal Bioautomation, 2016, vol. 20(4), pp. 505–514.

[5] Li Junbing, “Application of BP Neural Network Algorithm in Biomedical Diagnostic Analysis”, International Journal Bioautomation, 2016, vol. 20(3), pp. 417-426.

[6] M. Schultz, MATLAB 14. Programming, numerical methods. - Moscow: BHV-Petersburg, 2016, p. 928.

[7] L.F. Chaparro, Signals and Systems Using MATLAB. Moscow, 2011, p. 768.

[8] D. Chikhachev (2019, Aug.) World market forecasts and research [Online]. Available: http://www.techportal.ru/marketsandmarkets/

[9] I. Biryukov. (2016, Oct.) The neural network zoo [Online]. Available: https://tproger.ru/translations/neural-network-zoo-2

[10] S. Hochreiter, J. Schmidhuber, “Long short-term memory.” Neural Comput. 1997, vol. 9, pp. 1735–1780.