IRSTI 50.01.85

D. Almukhametova[1]*, D. Kuanyshbay*[2]*, N. Askar*[3]
[1,2,3]Suleyman Demirel University, Kaskelen, Kazakhstan

SPEECH RECOGNITION BASED ON CONVENTIONAL NEURAL NETWORKS

Abstract. In this research work, the problem of speech recognitionis considered in the form of an analysis of the numbers from 1 to 10 recorded by the speakeron the dictaphone. The paper uses the method of recognizing the spectrogram of an audiosignal using convolutional neural networks. Also written and implemented an algorithm for processinginput data, and an algorithm for recognizing spoken words. In this work, the qualityof recognition was assessed for a different number of convolutional layers. A comparison of the recognition quality is made in cases when the input data for the network are the spectrogramof the audio signal or the first two formants extracted from it. The recognition algorithm was tested using examples of male and female voices with different pronunciation lengths.

Keywords: spectrogram, formant, algorithm for learningneural networks.

***

Аннотация. В данной исследовательской работе рассматриваетсяспособы распознавания речи в виде анализа записанных диктором на диктофон цифр от 1до 10. В работе используется метод распознавания спектрограммы звукового сигнала с помощьюсверточных нейронных сетей. Так же написан и реализован алгоритм для обработкивходных данных, и алгоритм распознавания произнесенных слов. В работе была данаоценка качество распознавания для разного количества сверточных слоев. Произведеносравнение качества распознавания в случаях, когда входными данными для сети являютсяспектрограмма звукового сигнала или выделенные из нее первые две форманты. Тестированиеалгоритма распознавания произведено на примерах мужского и женского голосовс разной длительностью произношения.

Ключевые слова: спектограмма, формант, алгоритм обучениянейронной сети.

***

Аңдатпа. Бұл зерттеу жұмысында сөйлеуді тану мәселесі диктофонға дикторға жазып алған 1-ден 10-ға дейінгі сандарды талдау түрінде қарастырылады.Қағаздаконволюциялық жүйке желілері арқылы дыбыстық сигнал спектрограммасын тануәдісі қолданылады. . Сондай-ақ, енгізілген деректерді өңдеу алгоритмі жәнеайтылған сөздерді тану алгоритмі жазылған және енгізілген. Бұл жұмыста конволюциялыққабаттардың басқа саны үшін тану сапасы бағаланды. Тану сапасын салыстыру желіүшін кіріс деректері аудио сигналдың спектрограммасы немесе одан алынған алғашқыеекі формант болған жағдайда жасалады. Тану алгоритмі әр түрлі айтылу ұзындығыбар ерлер мен әйелдер дауыстарының мысалдары арқылы тексерілді.

Түйін сөздер: спектрограмма, формант, нейрондық желініоқыту алгоритмі.

Introduction

In this work, a convolutional neural network (CNN) is used, since such networks currently show one of the best results in the field of image recognition [8]. This feature of the CNN allows us to consider not the temporal realization of an audiosignal (speech), but it's spectrogram, which is recognized as an image. An important feature of the CNN is its resistance to changes in scale and image displacements, which, obviously, takes placein the case of spectrograms of speech signals [9, 10]. Supervised neural networks require data onwhich the network will train. In our case, the input data library will contain the spoken words recordedon the dictaphone. To improve the quality of recognition, it is advisable to processthese signals - to highlight the most important characteristics and present them in a form suitablefor feeding to the CNN input. Thus, the problem can be formulated as follows: the implementationof a human word recognition algorithm using a convolutional neural network, aswell as a data preprocessing algorithm for recognition.

Aim and objectives of research

The aim of this research work was to study a synthesizedspoken number recognition algorithm based on convolutional neural networks. Recognizeand test the realizations of several numbers in a female voice, as well as those pronounced bythe main speaker quietly and loudly, at a slow and fast pace.

Background of Literature Review

Research in the field of automatic speech recognitionand speech synthesis has received a lot of attention over the past fifty years, given that theinterest in automation in general began even earlier - more than sixty years ago. The idea of creating artificial intelligence (AI) appeared in the 50s of the twentieth century, when computer scientistsasked the question: "Can a computer think?" It began to be actively investigated in the1980s [1]. One of the proposed methods for implementing AI was artificial neural networks (ANNs), the use of which was focused, among other things, on the task of speech recognition, includingrecognition of phonemes or several words [2]. Research in the field of speech recognitioncontinues to this day [3–6], and the creation of AI in the human sense is still a lot of science fiction literature [7].

Discussion

Input processing.

Processing of input data Digits from 1 to 10 recordedon a dictaphone were used as examples of speech. The neural network was trained not on the time realization of signals, but on the images of their spectrograms. Speech signals were recordedusing a dictaphone with a sampling rate of 44.1 kHz. To construct spectrograms on the time signal realization, segments with a length of 1024 samples were selected. From these segments, thecalculation of the fast Fourier transform was made. In order to reduce the recognition error,preliminary processing was carried out (Fig. 1), including the following stages:

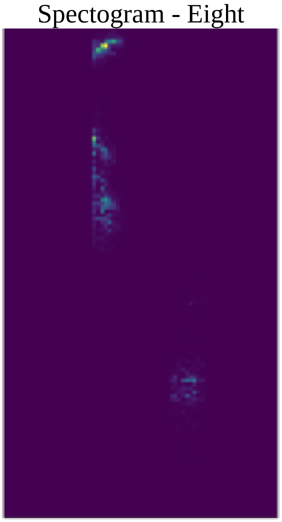


Figure 1. (Spectrogram of word "Eight")

- limiting the signal in the time domain to highlightthe informative part of the signal;
- limiting the frequency range of the signal. In [11], it is shown that the frequency range up to 3.4 kHz (the frequency range for the transmission of telephoneconversations) is not enough for high-quality recognition, since two similar wordscan have the same spectral components in this area and differ at higher frequencies. Therefore, the frequency range of spectrograms was used up to 8 kHz;
- normalization of the spectrum intensity. To work with signals of different loudness, it is advisable to normalize the brightness of the spectrograms;
- decreasing the image resolution. The quality of recognition strongly depends on the amount of data on which the network will be trained. If thereis a lot of data (the images will have a high resolution), then a lot of computing power will berequired to train the neural network. To increase the speed of performing operations in theexperiments carried out, the images of the spectrograms were reduced to a size of 50 × 50 pixels;

• increasing the contrast in order to highlight theinformational components of the signal against the background of noise (the speech was recorded notin a studio environment).

Convolutional neural network synthesis. In order for the implemented CNN to produce the correctresults, you need to create three directories in which data about audio signals willbe stored:

• directory of training data (train), which will beused to train the neural network;

• directory of verification data, which will checkthe quality of training during the training itself;

• directory of test data (test), which can be used to test the trained network. In order to minimize the recognition error, networktraining must be carried out on a large data array - from 1000 to 10000 [1]. But since it takesa lot of time and a lot of computing power to prepare such a volume of data and training, a muchsmaller sample of data was used in this work: 100 images of spectrograms (10 realizations for 10 digits) for training and 50 images each for checking and testing (5 implementations for 10 digits). The network has been trained over 30, 40 and 50 eras. Sigmoid and ReLU were used as an activationfunction in different layers (Fig. 2). It is also known that the complexity and quality of aneural network depends on the number of layers [7]. Therefore, to optimize the algorithm, experiments were carried out with sequential build-up of layers up to four. The initial structure of the SNA (Fig. 3) contained the following elements:

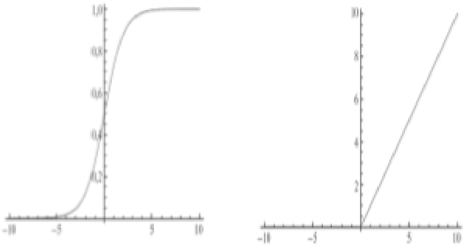


Figure 2. Activation functions of the “sigmoid” and ReLU

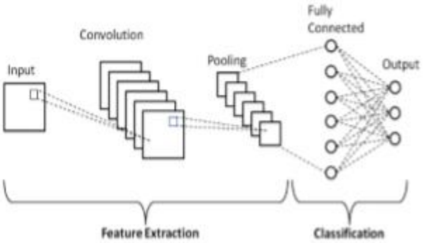


Figure 3. CNN structure

• a convolutional layer that highlights 32 features;core 3 × 3; ReLU activation function; • subsampling layer (MaxPooling) with a 2 × 2 core;

• a layer for transforming a two-dimensional arrayinto a vector;

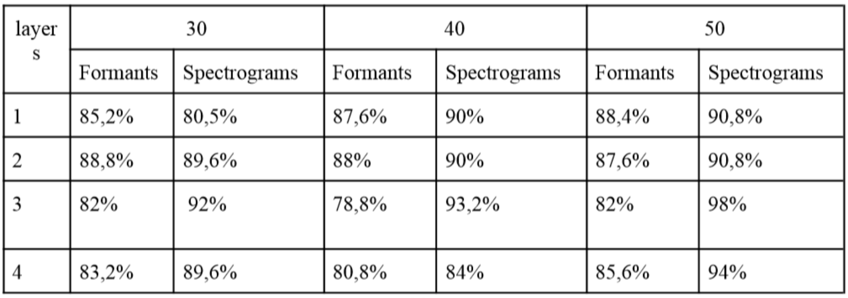
• fully connected layer of 32 neurons; ReLU activationfunction;

• Dropout layer against overfitting;

• fully connected layer of 10 neurons; activationfunction "sigmoid".

Recognition results

To obtain statistical data, experiments were carriedout in which 5 neural networks were created for 30, 40, 50 epochs with the number of convolutionallayers from 1 to 4. Then the recognition quality was assessed. A different approach to recognition was also implemented. There is a known method for recognizing spoken numbers by teaching a neural networknot on spectrograms, but on images of formants (narrow sections of the highest spectrumintensity) of speech signals [1]. Often, the first two formants are clearly visible on spectrograms,and their shape can be used as a feature for recognition. The implemented formant extraction algorithm made it possible, after additional processing of the input data, to obtain binary images.Testing the CNN trained on formant images (Fig. 4) showed that the recognition quality decreased in comparison with the recognition of spectrogram images directly (Table 1.).



layers	30		40		50	
	Formants	Spectrograms	Formants	Spectrograms	Formants	Spectrograms
1	85,2%	80,5%	87,6%	90%	88,4%	90,8%
2	88,8%	89,6%	88%	90%	87,6%	90,8%
3	82%	92%	78,8%	93,2%	82%	98%
4	83,2%	89,6%	80,8%	84%	85,6%	94%

Table 1.

```
Audio: three
Text: three
```

Figure 4. (Result)

Thus, of the two considered applications of the CNN,it is preferable to use the recognition of spectrograms or set more stringent requirements forthe formant extraction algorithm (up to the extraction of 4 or 5 formants). Although, having analyzed the appearance of the spectrograms, it is pertinent to note that not only the formants determinethe general picture of the spectrogram, but also extended sections that occupy a large area.

Conclusion

The synthesized algorithm for the recognition of spokennumbers based on convolutional neural networks showed an accuracy of about 80–98% of therecognized signals in the test sample. Realizations of several numbers in a female voicewere successfully recognized, as well as those pronounced by the main speaker quietly and loudly,at a slow and fast pace. The following facts about the work of neural networks have been investigatedand confirmed:

- an increase in the number of learning epochs increasesthe quality of recognition;
- the complication of the network structure, thatis, the introduction of additional layers for the selection of features, leads to an increase in the quality of recognition.

On the other hand, if the structure is too complex, the recognition worsens, as the neural network is retrained. The advantages of the chosen algorithm can be attributedto its relative simplicity, since the synthesis of the CNN structure was carried out in Python usingthe Keras library for working with neural networks. The advantages of the method used are due to the properties of the CNN: the stability of the algorithm to signal stretching in time andthe automatic extraction of characteristic features of the image, which are then used for recognition.

References

1 Tebelskis, J. *Speech recognition using neural networks.* Pittsburgh, Carnegie Mellon University, 1995, 180 p.

2 Juang, B.H. Automatic speech recognition. Atlanta, Georgia Institute of Technology, (2000): pp. 1–24.

3 Hazrati, O., Ghaffarzadegan, S., Hansen, J.H.L. Leveraging automatic speech recognition in cochlear implants for improved speech intelligibility under reverberation. *Proceedingsof the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane* (2015): pp. 5093–5097.

4 Suh, Y., et al. Development of distant multi-channel speech and noise databases for speech recognition by in-door conversational robots. Proceedings of the 20th Conference of theOriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, 2017, pp. 1–4.

5 Meltzner, G.S., Heaton, J.T., Deng, Y., et al. Silent speech recognition as an alternative communication device for persons with laryngectomy. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25 (12), (2017): pp. 2386–2398.

6 Dominguez-Morales, J.P., et al. Deep spiking neuralnetwork model for time-variant signals classification: a real-time speech recognition approach. Proceedings of the 2018 InternationalJoint Conference on Neural Networks (IJCNN). Rio de Janeiro, 2018, pp. 1–8.

7 Chollet F. Deep lerning with Python. Shelter Island, Manning Publication, 2018, 384 p.

8 Stanford Vision Lab. ImageNet Large Scale Visual Recognition

Challenge (ILSVRC). URL: http://image-net.org/ challenges/LSVRC (accessed 13.12.2018).

9 Guo, T., Dong, J., Li, H., Gao, Y. Simple convolutional neural network on image classification. *In 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (2017): pp. 721-724.

10 Albawi, S., Mohammed, T.A., Al-Zawi S. Understanding of a convolutional neural network. International Conference on Engineering and Technology (ICET). Antalya, 2017. P. 1–6.

11 Pieraccini, R. The voice in the machine. Building computers that understand speech. Cambridge, Massachusetts: MIT Press, 2012. – 360 p.