Technical science and innovation

Volume 2022 | Issue 2 Article 5

7-15-2022

UZBEK COMMANDS RECOGNITION BY PROCESSING THE SPECTROGRAM IMAGE

M M. Musayev
Tashkent University of Information Technologies

I Sh Khujayorov
Samarkand branch of Tashkent University of Information Technologies

M I. Abdullaeva
Tashkent University of Information Technologies

M M. Ochilov
Tashkent University of Information Technologies

UZBEK COMMANDS RECOGNITION BY PROCESSING THE SPECTROGRAM IMAGE

M.M. Musayev[1], I.Sh. Khujayorov[2], M.I. Abdullaeva[1], M.M. Ochilov[1]
[1]Tashkent University of Information Technologies,
Amir Temur St, 108, 100200 Tashkent, Uzbekistan
[2]Samarkand branch of Tashkent University of Information Technologies
Shokhruh St., 47[a], 140100, Samarkand, Uzbekistan

Abstract*: This paper describes the most common algorithms with image approach convolutional neural network and two-dimensional DCT with machine learning classification KNN, SVM and RF. These algorithms are evaluated for applicability to the Uzbek language and a comparative analysis on the accuracy and recognition rate. The command words of the Uzbek language were chosen for the experiments. According to the results, it was found that both methods give high rates of recognition accuracy and are 92% (CNN) and 90% (2D-DCT+Zigzag+SVM). Also the combinations of 2D-DCT+Zigzag+ KNN and 2D-DCT+Zigzag+ RF with average recognition accuracy of 86% and 85%, respectively, were considered in the paper.*

Keywords*: Spectrogram image, feature extraction, speech classification, speech recognition.*

INTRODUCTION. Since the advent of machine learning algorithms and neural networks, there has been interest in automatic speech recognition for voice control of technical devices [1-8]. Today, automatic speech recognition has found its application in various fields such as information and reference systems, smart home, etc. At the same time, speech recognition is rapidly being introduced and developed in the areas of creating convenience for people when driving a car, by providing the ability to control the functions of the car using the voice. It is known that the recognition of most of the world's languages is deeply studied and developed to this day, while the Uzbek language in this vein is very superficially considered. Among the explosive and fricative consonants of the Uzbek language, there are sounds that have no analogues in other languages (for example, the Uzbek sounds "q, g', h"), which makes them especially difficult to recognize and requires a special approach

The main goal of speech recognition is to efficiently and accurately convert speech into individual words, which is complex due to the speaker's speech dynamics, accents, pronunciation, speech rate and performance. When studying the research on speech signal recognition and classification have shown that a set of features can be computed using a variety of algorithms and methods and the most widely used traditional methods are Hidden Markov Model based statistical analysis (HMM) [1, 3,5,6], signal element processing in the spectral domain (Fourier analysis, Wavelet analysis) [8,9], dynamic time wrapping (DTW), methods based on neural networks that allow the system to self-learn and self-improve [2,4,10-22].

This article is a review and comparison between the two main algorithms of Uzbek speech recognition. The methods differ from each other both in the method of feature extraction and in the method of speech classification. The paper consists of an introduction, the main part, where both methods are described, and an experimental part, where the results of the comparison of the methods are given. The article concludes with a list of conclusions.

MATERIAL AND METHODS. Methods of speech signal recognition based on images (spectrograms) have been investigated in [1, 5, 9, 10, 12, 15, 16]. Obviously, the spectrogram concentrates more information about the audio signal than most of the manual functions traditionally used for the analysis of speech signals [6, 7, 12, 15, 23], which justifies the great attention of world scientists to the two-dimensional transformations. For example, in the article [7, 13] the results of a study based on the calculation of a feature vector by the two-dimensional discrete-cosine method for phoneme recognition are given.

Moreover, two types of neural networks, semi-dynamic (Time Delay Neural Network) and static (Multilayer Preceptors) networks, where recognition accuracies of 72.4%-77.5% were obtained by the obtained feature vector [7]. The works [8], [11] and [17] show the use of spectrograms for speech phoneme recognition based on CNN. The authors of [2] describes emotion recognition methods and It can be seen the use of image processing methods to identify speech signals based on spectral imaging [4].

Description of speech recognition systems. Recognition of speech signals is achieved through the processes of pre-processing, plotting spectrogram image, extraction of the features from spectrogram image and speech classification by machine learning and neural network algorithms. The recognition process can be represented as a sequence (Fig.1).

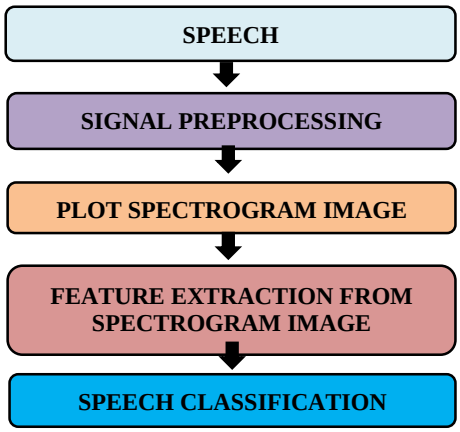

Fig.1. Sequence of speech recognition

The first stage gets speech signals with the given basic characteristics (discrete frequency, bit depth, number of channels). The second stage includes an initial preprocessing of the speech signal where the speech noise and the silence/pause zones are removed by filtering, normalization, averaging. The third stage generates an image of spectrograms obtained as a result of speech signal processing by transform methods. The fourth stage performs the process of features extraction from the spectrograms images.

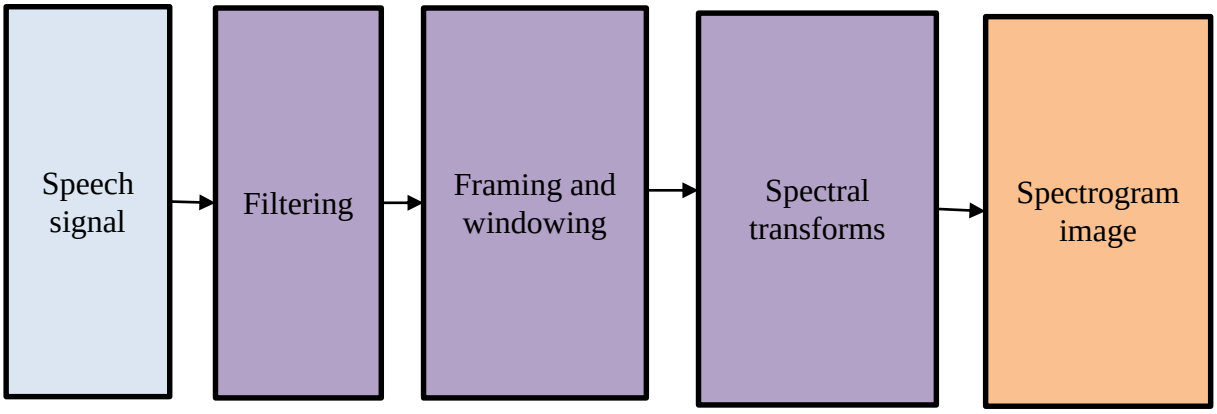

Fig.2. Spectrogram acquisition process

Fig.2 shows an expanded and detailed sequence for generating the spectrogram image of the speech signal. Processes in Fig.2 are parts of the speech signal preprocessing stage of Fig.1. Filtering, removes low frequencies (<100 Hz) and high frequencies (>4000 Hz) from the speech signal. The spectral entropy algorithm is used in the silence zone as removal stage for signal. In the framing step the speech signal is divided into frames with 256 lengths. The windowing (Hamming, Hann, Hamming, Bartlett) is used to reduce distortions in the segments as well as to smooth them out.

After spectral transformations (Discrete Cosine Transform, Fast Fourier transform, Wavelet) a spectrogram is generated. Spectrograms are processed by a special 2D algorithms (Convolution Neural Network, 2D-Discrete Cosine Transform) to extract the most important features from images and further processed by the machine learning (ML) algorithms and neural networks (NN). Training and testing of neural networks and machine learning classifiers is carried out according to Fig. 3.

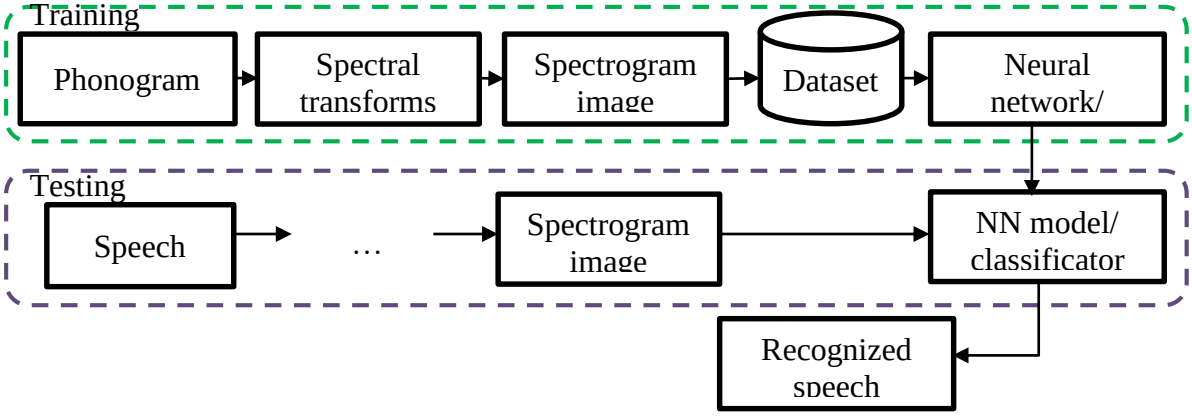

Fig.3. Spectrogram-based recognition process

Depending on the task it is necessary to choose the type of neural network and properly build its architecture. Next, considered two approaches that are sharply different from each other in their method and accuracy of speech recognition. They are speech recognition by using Convolution Neural Network and 2D-DCT.

Speech recognition by using convolution neural network**.** Convolution Neural Network (CNN) is an algorithm in the field of highly-performance deep learning in image classification and consists of several successive layers (Fig.4). Different CNNs differ in the choice of parameters.

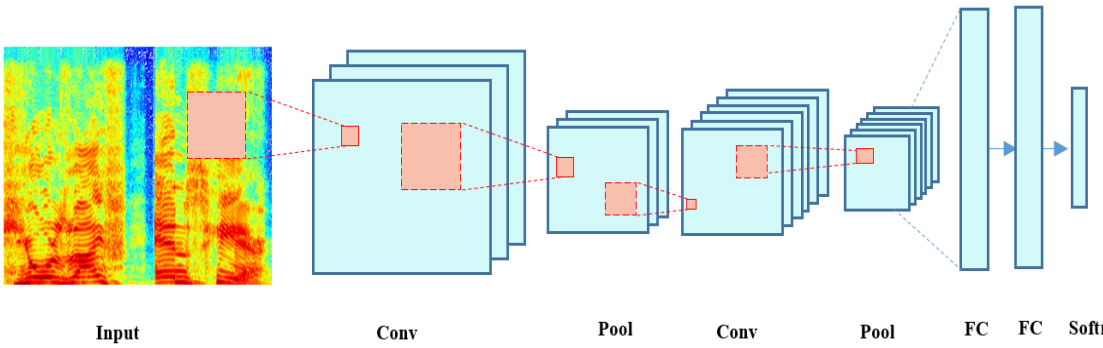

Fig.4. Structure diagram of a typical CNN

Common CNN neural network models include the following types of layers:

- Convolutional (CONV). The convolution layer includes a filter for each channel, the convolution core of which processes the previous layer in fragments (by summing the results

of the matrix product for each fragment)

- Activation (ACT or RELU, where we use the same or the actual activation function)
- Pooling (POOL). Pooling layer (in other words, subsampling, subsampling) isanon-linear compression of the attribute map, with a groupofpixels (usually 2×2 in size) compressed to one pixel, passingthrough a non-linear transformation.
- F

ully-connected (FC). A fully connected layer combines each neuron in one layer withneurons in other layers

- Batch normalization (BN). Batch normalization is a technique for improving the performanceand stability of artificial neural networks.
- Dropout (DO). Exception or dropout is a method of regularization of artificial neural networks, designed to prevent network retraining.

CNN can be generated based on the placement of the above layers sequentially. Normally, CNN can be represented as a text diagram as following:

INPUT(IMAGE) => CONV => RELU => FC => SOFTMAX 1

According to (1), the input data will be received and sent to the convolutional layer, then to the activation phase, after which it will be sent to completion, then to the full-link layer, and finally it will be classified based on the softmax classifier [44].

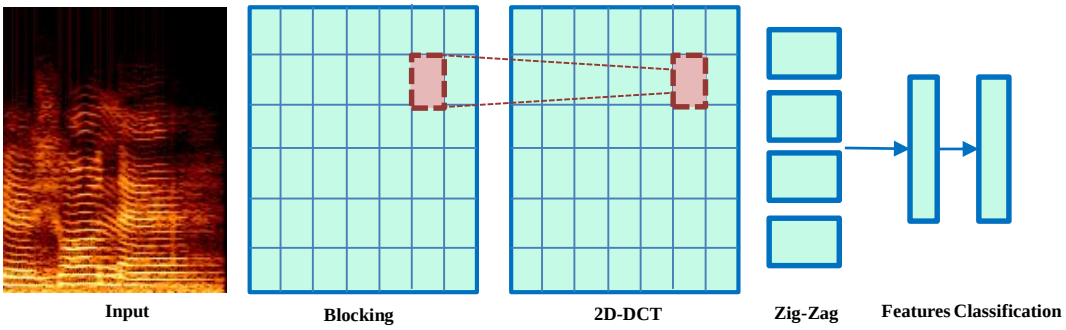

Fig.5. Structure diagram of speech recognition by 2D-DCT.

Recognition by two-dimensional DCT is performed by means of the following layers:

- Blocking. Two-dimensional DCT initially divides a spectrogram of size (MxN) into non-overlapping blocks of size PxQ. The default block size is square, i.e., P=Q=8;
- 2D-DCT. Equation (2) describes the DCT formula, where the coefficients $F(u,v)$for the block from the spectrogram $f(x,y)$ are calculated:

$$F(u,v) = \frac{2}{P}C(u)C(v)\sum_{x=0}^{P-1}\sum_{y=o}^{P-1} f(x,y) \times \left[cos\left(\frac{\pi u(2x+1)}{2P}\right) cos\left(\frac{\pi v(2y+1)}{2P}\right)\right] \quad 2$$

- Zig-zag. The reordering of the coefficients into a one-dimensional array is possible in a zig-zag manner according to the scheme in Fig. 6.
- Features. The generated array is the main features of the speech signal.
- Classification. The features are used to classify the speech signal into one or another class. At the stage of classification any classification algorithms can be used.

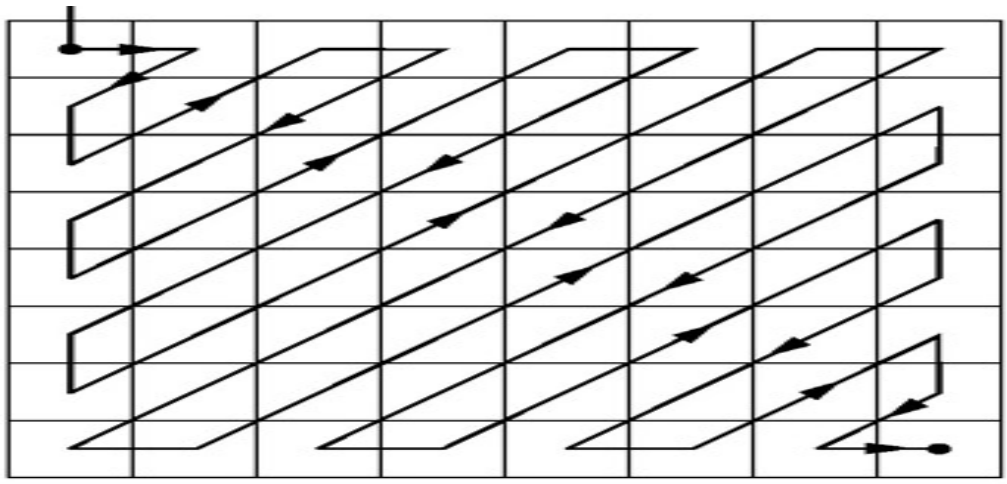

Fig.6. The decomposition scheme using the Zig-zag scan

RESULTS AND DISCUSSION. The purpose of this experiment is to conduct a comparative analysis between the above two methods in terms of accuracy and recognition rate of 6 selected individual words in the Uzbek language – chap (left), o'ng (right), chiroq (light), yaqin (close), uzoq (far away), yuv (clean). The selected words are the commands that can be used when driving a car.

Data collection. Men and women were involved for the experiment. The total number of speakers was 150. 16-bit mono audio files (benchmarks) in .wav format with a sampling rate of 22050 Hz were recorded. Each command word has an average of 150 benchmarks, which were used to train and test the model. 90% of the benchmarks were used for training and 10% for testing.

Table 1

Number of audio data used for training and testing

Methods	Train	Testing	Total number of benchmarks
CNN	135	15	150
2D-DCT+Zig-Zag	126	14	140

The results of the experiments showed that both image approaches are effective, as evidenced by the accuracy of recognition of each command word. The average accuracy of recognition of all words using the method 2D-DCT + Zigzag with classification algorithms with a teacher KNN, SVM, RF respectively were 86%, 90% and 85%. CNN method has an average accuracy of 92%.

Table 2

Experimental results

Words	CNN	2D-DCT + Zigzag		
		KNN	SVM	RF
Left	0,92	0,84	0,89	0,84
Right	0,94	0,85	0,90	0,86
Light	0,91	0,86	0,89	0,85
Far	0,89	0,87	0,91	0,85
Close	0,93	0,85	0,89	0,86
Clean	0,94	0,88	0,90	0,84

The results of 2D-DCT + Zigzag can be evaluated depending on the classification method used with it, taking into account the additional indicators shown in Table 3.

Table 3

Additional indicators of the models

	2D-DCT+ Zigzag + KNN	2D-DCT+ Zigzag + SVM	2D-DCT+ Zigzag + RF
Average recognition accuracy (%)	86	90	85
Recognition time	0,04	0,7	0,46
Optimal data quantity for training	>100	>120	>120

CONCLUSION. Recognition of speech signals by creating spectrogram images has several advantages and is that the spectrogram contains a wide set of characteristics in comparison with other speech recognition methods. In this paper, two 2D methods of speech signal recognition by means of their spectrograms are considered. Both methods are a striking confirmation of the fact that the spectrogram concentrates the most informative characteristics. It was found that each method has a distinctive amount of required training data. Experimental results show that based on the above approaches described in the article, high results can be achieved in the classification of the Uzbek word.

References:

1. P. Ibrahim., Y.R Srinivas. Speech recognition using HMM with MFCC-an analysis using frequency Spectral decomposing technique. "*Signal Image Processing an International Journal (SIPIJ)"*, **2010**.
2. A.M. Badshah., J. Ahmad., N. Rahim., S.W. Baik. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. 2017 "*International conference on platform technology and service"*, **2017**.
3. M. Gales., S. Young. The application of hidden Markov models in speech recognition. *"Foundations and Trends in Signal Processing",* **2007**. 195.
4. N.G. Andrew., Y. Zhang. "*Speech recognition using deep learning algorithms. published"*, **2013**.
5. M.B. Gulmezoglu. A novel approach to isolated word recognition. "*IEEE transactions on speech and audio processing",* **1999**. 620.
6. N.E. Sukmawati., A. Satriyo., R.A. Sutikno. Automatic Speech Recognition for Indonesian using Linear Predictive Coding (LPC) and Hidden Markov Model (HMM). *"Proceeding of 5th international seminar on new paradigm and innovation on natural science and its application",* **2015**.
7. M. Ahmadi., N.J Bailey., B.S. Hoyle. Phoneme recognition using speech image (spectrogram). "*Published in IEEE proceedings of third international conference on signal processing",* **1996**.
8. J. Zhang., S. Xiao., H. Zhang., L. Jiang. Isolated word recognition with audio derivation and CNN. *"Proceedings international conference on tools with artificial intelligence"*, **2018**. 336.
9. D. Polap., M. Woźniak. "*Image approach to voice recognition. 2017 IEEE symposium series on computational intelligence"*, **2018**. 1.
10. J.M. Padmanabhan., J.J. Premkumar. Machine learning in automatic speech recognition. "*A survey. IETE Technical review institution of electronics and telecommunication engineers",* **2015**. 240.
11. C. Glackin., J. Wall., G. Chollet., N. Dugan., N. Cannings. Convolutional neural networks for phoneme recognition. *"Proceedings of the 7th international conference on pattern*

recognition applications and methods", **2018**. 190.
12. L. Yingying., P. Siyuan., X. Nanfeng. Speech Recognition Method Based on Spectrogram. "*Proceedings of the international conference on mechatronics and intelligent robotics (ICMIR)"*, **2017**.
13. A.H. Waibel., T. Hanazawa., G. Hinton., K. Shikano., K. Lang. *"Phoneme recognition using time-delay neural networks",* **1989**.
14. W. Fisher., M. Doddington., R. George., M. Goudie., M. Kathleen M. "*The DARPA Speech recognition research database: specifications and status"*, **1986**. 93.
15. Q.T. Nguyen. Speech classification using sift features on spectrogram images. *"Vietnam journal of computer science",* **2016**. 247.
16. M. Al-Darkazali. *"Image processing methods to segment speech spectrograms for word level recognition"*, **2017**.
17. Y. Longhao., C. Jianting. "*Patients' EEG Data analysis via spectrogram image with a convolution neural network. conference: international conference on intelligent decision technologies",* **2016**.
18. B. Venkatesh., P. Andrej., J. Rasmusson., L. Lundberg. Classifying environmental sounds using image recognition networks. "*Procedia computer science"*, **2017**. 2048.
19. J. Dennis., H.D. Tran., H. Li. "*Spectrogram image feature for sound event classification in mismatched conditions. IEEE signal processing letters"*, **2010**. 130.
20. E. Geoffrey., N.S. Hinton., A. Krizhevskiy., I.R Sutskever., R. Salakhutdinov. "*Dropout a simple way to prevent neural networks from overfitting. journal of machine learning research",* **2014**. 1929.
21. A. Rosebrock. "*Deep learning for computer vision with python starter bundle",* **2017**.
22. S. Rekik., D. Guerchi., S.A. Selouani. "*Speech steganography using wavelet and fourier transforms"*, **2012**.
23. http://cs231n.github.io/convolutional-networks

OUTLOOK FOR GLOBAL ROAD SAFETY

K.Sh. Chariev, A.R. Yusupov
Tashkent State Transport University
Adilkhodjaev St 2, 100067 Tashkent, Uzbekistan

Abstract: *The World Health Organization estimates that 1.25 million people are killed in road accidents each year—almost 3,400 a day-and up to 50 million are injured. However, road traffic injuries are not equally common around the world; some countries suffer more than others, and the likelihood of being killed in a road accident depends on where the person lives. Almost 90% of all road accidents occur in low-and middle-income countries. Worldwide, the number of deaths per 100,000 populations (the death rate) ranges from less than 3 to almost 40. This figure is less than 9 in high-income countries, but averages about 20 in LMIC (low-and middle-income countries), with the Asian and African region showing the highest rate (46.5%) [1,2,3]. The article developed recommendations and provided proposals for improving the transport system, improving the economic, environmental situation, as well as road safety.*