Journal of Computer Science Engineering and Information Technology Research (JCSEITR)
ISSN (Print): 2250-2416; ISSN (Online): Applied
Vol. 14, Issue 1; Jun 30 2024, 13-20
© TJPRC Pvt. Ltd.

TRANS STELLAR
•Journal Publications •Research Consultancy

AUTOMATIC SPEAKER RECOGNITION THROUGH VOICE USING DEEP NEURAL NETWORKS

MAMATOV NARZILLO[1], BABOMURODOV OZOD[2] & DUSANOV KHURSHID[3]

[1]*Head of The Department, Doctor of Technical Sciences, Professor, "TIIAME" National Research University*

[2]*Executive Director, Doctor of Technical Sciences, Professor, Kazan Federal University Branch in Jizzakh*

[3]*Phd Student, Jizzakh Branch of the National University of Uzbekistan Named After Mirzo Ulugbek*

ABSTRACT

In this paper, text-independent speech recognition systems were considered. Character extraction was performed using Mel Scaled Filter Banks (MSFB). A deep neural network method was studied for automatic speaker identification by voice. A model is built by clustering the feature vectors for each speaker. Voices are modeled using a deep neural network (DNN). In the database, voice samples of all speakers are collected in the form of a file. From the results, it can be said that deep neural network using cepstral features gives good results for speaker recognition system.

Keywords: Cepstral Coefficient, Feature, Identification, Verification, Text-Dependent, Text-Independent, Framing, Mel-Scaled Filter Banks, Hamming, Probability, Signal, Speech.

Received: Dec 04, 2023; **Accepted:** Dec 19, 2023; **Published**: Jan 04, 2024; **Paper Id**: JCSEITRJUN20242

Original Article

INTRODUCTION

Speech is the main means of communication between people, through which, it is possible to determine various information, including the meaning of words, the speaker's emotional state, gender, and personality. It is also possible to convey information such as thoughts and ideas through speech. Speech production includes articulation, speech, and fluency. It is a set of natural abilities of a person, which is usually formed by the coordinated movement of about 100 muscles connected to the nerves of the person [1]. According to its physical properties, it is an acoustic signal that is continuously variable in time.

In recent years, the rapid development of voice recognition systems has helped to overcome many of the main challenges for artificial intelligence systems. The task of voice recognition is to distinguish, classify, and respond accordingly to human speech from an input audio signal. In this case, two sub-tasks are usually separated: identification and verification [1].

Identification: a closed set of individuals is entered into the system along with some test data. This involves the system classifying an unknown voice as belonging to one of N reference individuals.

Verification: Two parts of speech are required in the system. This requires the system to determine whether the two segments are spoken by the same person or by a different person, and this is called overt identification.

PROBLEM DEVELOPMENT

Voice recognition is a routine that checks whether an unknown voice belongs to a specific voice in a set of voices that have been previously identified by voice. According to the set property, it is divided into open or closed set definitions. In the open-set condition, an observed speech sample may not belong to a predefined set of speakers. In contrast, closed

set identification assumes that the test sample belongs to one of the samples in the set [2]. Solving the problem of identification in an open set is more complicated, because it is necessary to establish a clear mechanism to determine whether the test sample belongs to one of the existing samples. The term "audio identification" is widely used as a general term for audio identification. This applies to any mode of operation that includes user identification (Figure 1).

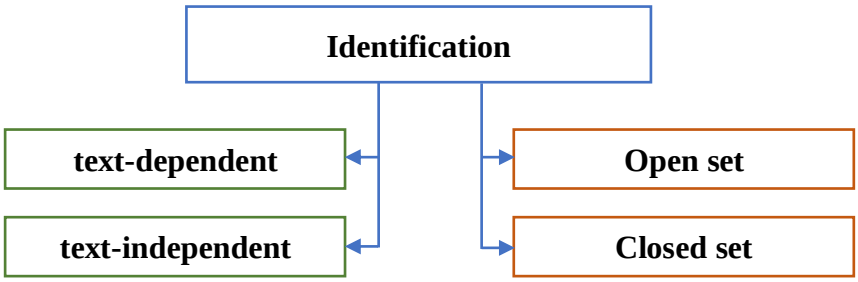


Figure 1: Identification of a person based on his voice

Identifying a speaker by comparing a given voice sample with voice samples stored in an existing database is called speaker identification. The result of identification is a list of candidates, and this list is numbered, it is formed based on a certain threshold.

There is a symbol that divides identification systems into two classes, which is a grammatical representation of the speech signal used:

– text-dependent systems grammatically use exactly one speech signal. In other words, training and authentication require the speaker to say exactly the same word.

– text-independent systems do not depend on the grammatical structure of the used speech signal, that is, the speaker is not required to say exactly the same word during training and authentication, and he can say arbitrary text.

Nowadays, the number of unauthorized access to various objects and resources by unauthorized users is increasing. This requires the development of user identification systems and applications with high efficiency and recognition speed. The effectiveness of recognition systems depends on the model, method and algorithms used in it. Biometric identification methods and algorithms, which have a number of advantages, and advantages have a special place in this. Biometric indicators based on unique individual characteristics of a person allow reliable description of an individual user. In addition, variability and insufficient data should be listed as one of the problems where background noise can also affect speaker recognition. Problems can be related to user or technical errors. These are problems of variability, problems of insufficient data, problems of background noise.

These problems indicate the need for scientific and experimental research on the creation of new approaches, mathematical methods and techniques that ensure high accuracy of identification and reliability of its results in conditions of noise and various noises added to processed sound signals.

FEATURE EXTRACTION

The effectiveness of recognition systems depends on how the characters are selected. The better the initial character space is chosen, the higher recognition quality. The problem of speech recognition begins with the formation of a character vector from the speech signal.

For a more accurate description of the signal, the speech segments are taken overlapping. The process of creating

speech segments is carried out using the window method, by multiplying the signal with some window function so that the spaces at the window boundaries are relaxed.

The feature extraction process is shown in Figure 2.

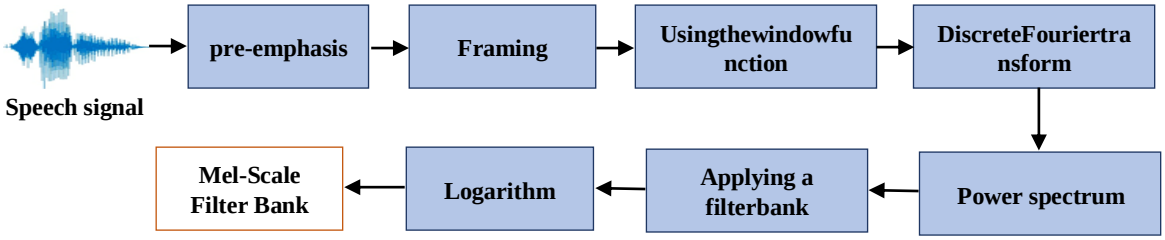


Figure 2: Mel-Scaled Filter Banks

An audio signal that can be used to divide a speech signal into frames is considered as an example of 1 second recorded at 16 kHz. A typical 25 ms (millisecond) audio signal consists of 400 samples. [7]

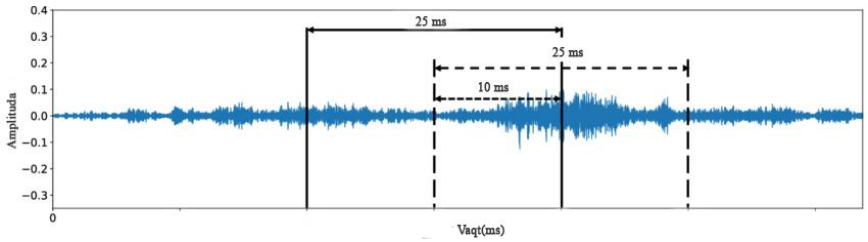


Figure 3: Dividing the Speech Signal into Frames

In order to obtain the accuracy in time, it is necessary to divide the signal into overlapping frames. The signal is separated using a window function, usually the Hamming window function is used.

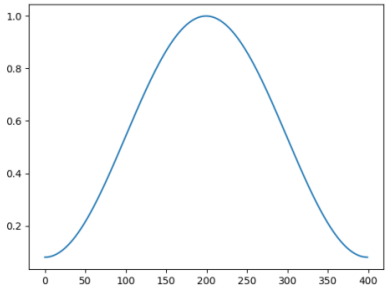


Figure 4: Hamming Window

The Hamming window is used to solve speech related problems. Using the Hamming window, the signal is replaced by the following formula:

$$w[t] = 0.54 - 0.46\cos\left(\frac{2\pi t}{N-1}\right), \quad 0 \le t \le N-1 \qquad (1)$$

Discrete Fourier transform (DFA) describes what frequencies are present in the signal, but does not localize them in time. To account for this, DFA is applied to the signal frames to get an idea of what frequencies are present in each

frame. Creates a time-frequency characteristic for the audio in the frame. This method has an accuracy trade-off for the frequency-time axis. The windows must be small enough for the desired timing accuracy. A short-term Fourier transform for a speech signal [7] (2) is performed by the expression:

$$H(n,k)=\sum_{n=1}^{N}x(n)w(n)e^{\frac{-2\pi ikn}{N}}\quad 0\le k\le K \tag{2}$$

Here: $x(n)$- the signal in the time domain, N - the length of the window consisting of n samples, K - the length of the DFA. The formula for the power spectrum sample energy is obtained as follows:

$$S(n,k)=\left|H(n,k)\right|^2 \tag{3}$$

Triangular filter sets can be used from 20 to 40 (26 standard). In this case, to calculate the energies of the filter bank, each filter bank is multiplied by the power spectrum and coefficients are generated. Once this is done, 40 numbers are obtained that tell you how much energy is in each filter bank. Mel-filter bank is calculated by the following formula(4):

$$f_{mel}=2595\cdot\log_{10}\left(1+\frac{f}{700}\right) \tag{4}$$

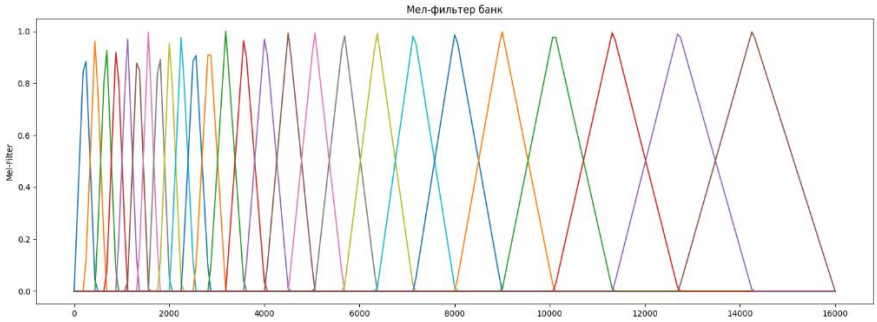


Figure 5: View of a triangular filter in a Mel-filter bank

Humans perceive low-frequency changes more clearly than high-frequency changes. Logarithmization has a similar property. At low values of the input x, the gradient of the log function is high, but at high values of the input, the gradient value is smaller. This allows us to apply logarithmization to the output of a Mel-filter suitable for the human auditory system:

$$MSFB=20log_{10}\left(S(n,k)\cdot F_m(k)\right) \tag{5}$$

DEEP NEURAL NETWORKS (DNN)

Backpropagation and stochastic gradient descent are used to train deep neural networks. One of the latest trends is the development of deep neural networks that allow to solve the problem of automatic identification of a speaker based on its sound without using additional methods of character set formation or classification methods. In this research work, a deep neural network was used to model it based on the sound of a speaker. A deep neural network is based on convolutional neural networks (CNN), which uses a d-vector approach [3]. Despite the variation of layers and capacities, all systems include the same important elements, that is, a part working at the frame level (convolution), pooling (polling), and a fully

connected (fully connected) layer for calculating the location of models in hyperspace. The main idea of the aggregation layer is to aggregate information on all input frames by calculating the average (average polling) or average standard deviation (statistical polling).

We propose a new architecture based on a neural network for speech recognition through SpeakerNet. SpeakerNet consists of three main parts: Encoder, Pooling layer and Decoder. The encoder is based on the QuartzNet architecture developed for ASR [4]. It consists of 4 blocks, where each block consists of 1D depth-separated convolutions, batch norm, ReLU, and full-partition layers. The encoder converts variable-length audio into a sequence of acoustic features that can be used to extract higher-order features. The pooling layer maps the temporal sequence of acoustic features into a vector of fixed length by calculating statistics on the acoustic features. A decoder consisting of a series of fully connected layers compares fixed-length vectors of size D with an array of N votes to calculate the probability that the current segment belongs to a voice in the training set. The deep neural network architecture is presented in Figure 6.

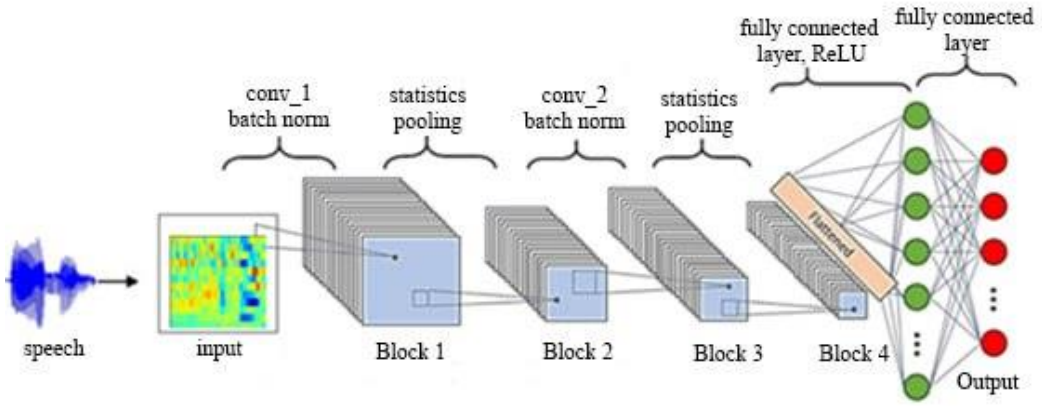


Figure 6: Deep Neural Network Architecture

Studies have shown that the specific type of activation function does not significantly affect the final quality of the neural network. Therefore, in practice, easily calculated functions (but conditionally non-linear) are taken as activation functions. As an activation function, the ReLU activation function, which is currently most widely used, is shown in Fig 6.

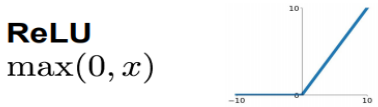


Figure 6: ReLU Activation Function

The number of neurons in the last layer corresponds to the number of classes, each neuron gives an estimate of class belonging to the object, and the softmax function was used as the activation function of the last layer to estimate the class membership (that is, to have a certain probability distribution):

$$\text{soft}\max(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad (6)$$

As an error function, the cross-entropy loss [5] error function given in the following expression was used:

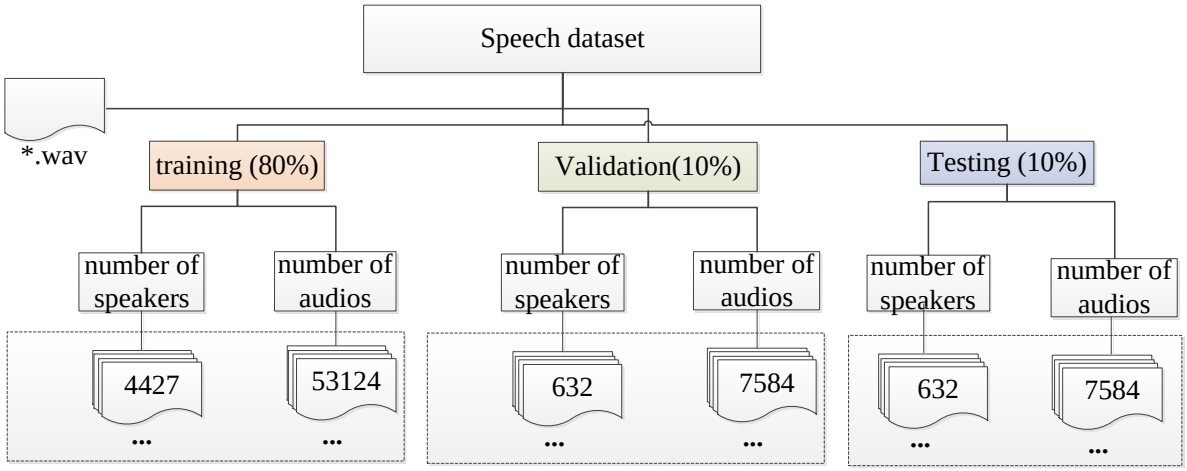
$$L = -\sum_{i}^{C} y_i \log\left(Y(s)_i\right) \quad (7)$$

RESULTS AND DISCUSSIONS

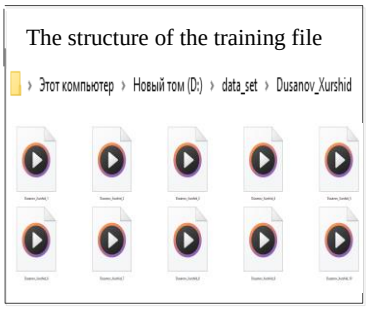
Software design was carried out in Python programming language based on existing and proposed models, methods and algorithms. The PyQt5 library was used to develop the user interface of the software. A database was created in the software to store the data of registered users, and the SQLite database management system was used for this. In addition, the PyTorch library was used to work with deep neural networks. System registration and recording interfaces are shown in Figure 8. The result of the study is presented in Table 1. The speech samples used in this work were recorded using the Audacity audio editor. Sampling rate 16000 Hz (16 bit, mono). Figure 9 provides a description of the database[6].

Table 1: The Degree of Accuracy of the Mixed Base is Given in Percent

Algorithm	VCTK + TIMIT	VoxCeleb1+ VoxCeleb2	TIMIT + SITW	NutqUzData
PLP+VQ	82,3	84,3	80,4	87,7
MFCC+VQ	83,2	86,2	83,1	88,8
PLP+GMM	84,8	85,8	82,7	90,1
MFCC+GMM	85,1	87,5	82,3	93,4
GRU	94,5	95,2	90,6	96,3
SpeakerNet	97,6	98,5	98,1	99,1



Enrollment Requirements:
1. format - *.wav
2. The number of channels is mono
3. Sampling rate – 16kHz
4. space for each sample – 16 bits
Recording conditions - recorded in normal (home, street, morning, afternoon, evening) conditions



Speech dataset:
Training-5060 speakers, 60710 audios.
Validation-number of 632 speakers, number of audios 7584.
The number of test-632 speakers, the number of 7584 audios.
Total: 6324 speakers, 75888 audios.

Figure 7: NutqUzData is an Overview of the Speech Dataset

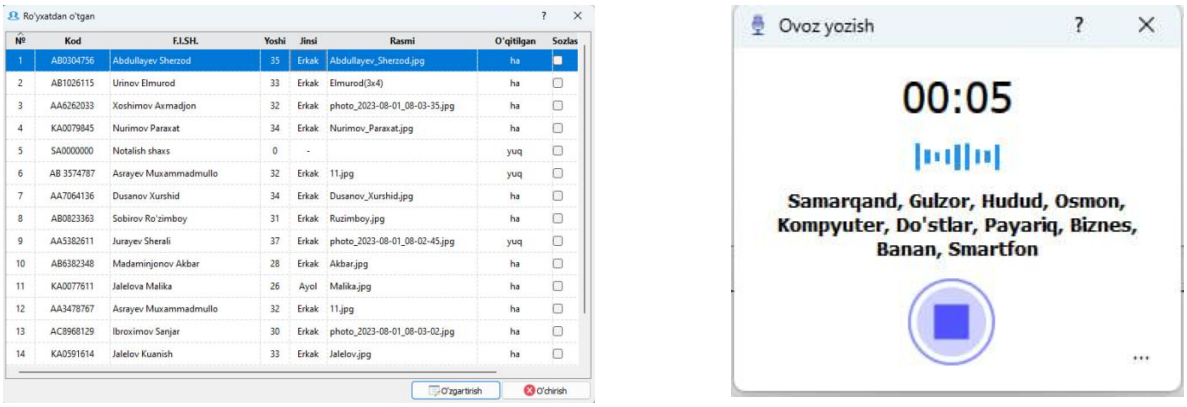


Figure 8: System Registration and Recording Interfaces

CONCLUSIONS

In this article, text-free speech recognition systems were considered. Character extraction was performed using mel-scaled filter banks (MSFB). Voices are modeled using a deep neural network (DNN). Using the obtained features, a model was built by clustering the character vectors from each speaker. At least 10 samples of all speakers are collected in the database.

Experimental analyses have shown that a high result of recognizing individuals with features based on MSFB can be obtained. From the results, it can be said that applying a deep neural network using filter banks of characters gives good results for creating a voice recognition system.

The developed voice control system can be used to perform a large number of tasks: controlling computer applications, mobile platforms or robotic devices, for example, loading robots. The presented approach allows creating speech recognition systems based on open technologies and a personal computer equipped with a microphone.

REFERENCES

1. *Zhongxin Bai, Xiao-Lei Zhang. Speaker recognition based on deep learning: An overview. Center of Intelligent Acoustics and Immersive Communications (CIAIC) and the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China*
2. *Mamatov, N.S., Niyozmatova, N.A., Yuldoshev, Y.S., Abdullaev, S.S., Samijonov, A.N. Automatic Speech Recognition on the Neutral Network Based on Attention Mechanism Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)this link is disabled, 2023, 13741 LNCS, страницы 100–108*
3. *Cutajar M, Gatt E, Grech I, Casha O, Micallef J. Comparative study of automatic speech recognition techniques. IET Signal Proc 2013;7(1):25–46.*
4. *Narzillo, M., Abdurashid, S., Parakhat, N., Nilufar, N.Automatic speaker identification by voice based on vector quantization method. International Journal of Innovative Technology and Exploring Engineering.2019.*
5. *Dr.H.B.Kekre, Ms.Vaishali Kulkarni, Speaker Identification by Vector Quantization, International Journal of Engineering Science and Technology Vol. 2 (5), 2010, 1325-1331*
6. *Niyozmatova, N.A., Mamatov, N.S., Tulyaganova, Sh.A., Samijonov, A.N., Samijonov, B.N.Methods for determining speech activity of uzbek speech in recognition systems* AIP Conference Proceedings, *2023, 2789, 050019*
7. *Mamatov N.S., Dusanov X.T. Nutq signali belgilar to'plamini shakllantirishning MFCC usuli. Zamonaviy innovatsion tadqiqotlarning dolzarb muammolari va rivojlanish tendensiyalari: Yechimlar va istiqbollari. Respublika miqyosidagi ilmiy-texnik anjuman 2022 yil 13-14 may 179-182 bet.*