International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Automatic Speaker Identification by Voice Based on Vector Quantization Method

Mamatov Narzillo, Samijonov Abdurashid, Nurimov Parakhat, Niyozmatova Nilufar

Abstract: ***In this paper, the systems of speaker identification of a text-dependent and independent nature were considered. Feature extraction was performed using chalk-frequency cepstral coefficients (MFCC). The vector quantization method for the automatic identification of a person by voice has been investigated. Using the extracted features, the code book from each speaker was built by clustering the feature vectors. Speakers were modeled using vector quantization (VQ). Using the extracted features, the code book from each speaker was built by clustering the feature vectors. Codebooks of all announcers were collected in the database. From the results, it can be said that vector quantization using cepstral features produces good results for creating a voice recognition system.***

Keywords: cepstral coefficient, criteria, feature, identification, method, model, phoneme, probability, signal, speech.

I. INTRODUCTION

Speech is one of the important elements of human activity, allowing a person to understand and transmit the knowledge of the world around to other people. Oral speech is manifested in a person in the form of statements in sound form.

Oral speech of each person has individual characteristics, which are determined by the characteristics of the structure of his vocal apparatus.

The task of recognizing the personality by voice is to isolate, classify and respond accordingly to human speech from the input audio signal. In this case, two subtasks are usually distinguished: identification and verification [1].

Identification is the process of determining a person by the voice pattern by comparing this pattern with the templates stored in the database. The result of identification is usually the name of the person registered in the system, the template of which most likely corresponds to the input voice pattern.

Verification is the process by which a presented voice signal sample is compared with a sample in a database. When verifying, along with the voice sample, the identity identifier is also transmitted in the database, the sample of which will be compared. As a result, either positive or negative will be displayed.

In addition, voice recognition systems can be divided into text-dependent and text-independent. In text-dependent recognition, fixed phrases and phrases generated by the system are used. Text-independent systems allow you to recognize arbitrary speech.

There are the following problems of the task of recognizing a person by voice, which should be considered when building a solution:

- emotional state of the announcer;
- acoustic environment (noise and interference);
- different channels of communication in learning and recognition (?);
- natural voice changes

II. FORMULATION OF THE PROBLEM

Speaker Identification System. The process of speaker identification is divided into two main stages: model training and testing.

At the first stage, speech samples are collected from the speakers, which are used to train the models for each speaker. The collection of registered models is also called the speakers database.

At the second stage, the test samples of unknown speakers are compared with the database. Both steps involve the extraction of features. This stage allows you to get from speech characteristics, dependent on the speaker. The goal of feature extraction is also to accelerate the subsequent comparison of characteristics. Then, at the registration stage, the resulting model is stored in the speakers database.

At the identification stage, the extracted features are compared with the models stored in the speakers database. Based on these comparisons, the final decision on the identity of the speaker is made. This process is presented in Figure 1.

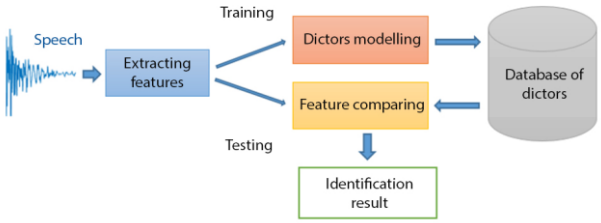

Fig. 1. The process of learning and speaker identification.

Manuscript published on 30 August 2019.
*Correspondence Author(s)
Mamatov Narzillo*, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan,
Samijonov Abdurashid, Bauman Moscow State Technical University, Russia Federation
Nurimov Parakhat, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan
Niyozmatova Nilufar, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan

III. FEATURE EXTRACTION

In recognition systems, the speech signal is divided into short segments, and each segment is converted into a feature vector, as a result, the input signal is represented by a sequence of feature vectors. Characteristic vectors are usually computed for short signal segments using the assumption that speech can be viewed as stationary at these short intervals. For a more accurate description of the signal speech segments are taken with overlap. The process of creating speech segments is performed using the window method, that is, by multiplying the signal with some window function so that the gaps at the window borders are relaxed.

The process of extracting features is presented in Figure 2.

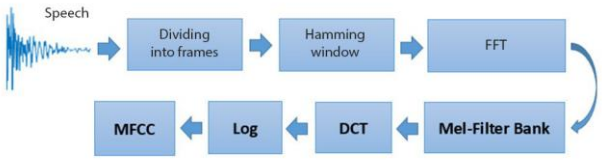

Fig. 2. The process of extracting features.

Usually for these purposes the Hamming window is used, in this case the window function takes the form:

$$w(n) = 0.53836 - 0.46164\cos\left(\frac{2\pi n}{N-1}\right)$$

where N is the width of the window. The next step of feature extraction is the transformation of each frame from the time domain into the frequency domain (calculation of the signal spectrum) using the discrete Fourier transform. This step is usually performed as a fast Fourier transform, which is an effective implementation of the discrete Fourier transform. Discrete Fourier transform:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, \ (k = 0,1,\ldots,N-1);$$

where N is the width of the window.

In order to more accurately simulate the perception of sound energy by the human auditory system, an analysis is performed for each frame using a comb of chalk filters. Chalk filters are based on a chalk scale, which is a logarithmic scale similar to that of a human being. The chalk scale is determined by the following formula relative to the frequency scale f, measured in hertz:

$$f_{mel} = 1125\ln\left(1+\frac{f}{700}\right)$$

Comb chalk filters performed with overlapping triangular weight functions.

A vector made up of the energy of all elements of the frequency resolution in each chalk filter is a chalk spectrum. Due to the overlap of the triangular weight functions of the chalk filters, the adjacent components of the chalk spectrum vector mutually correlate with each other.

With the use of a discrete cosine transform on a logarithmic chalk spectrum, the mutual correlation between adjacent components is greatly reduced. At the output of digital signal processing, so-called chalk-frequency cepstral coefficients are formed.

$$c[l] = \sum_{m=0}^{M-1} \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right) cos\left(\pi l\left(m+\frac{1}{2}\right)/M\right),$$

where $0 \le l \le M$.

Chalk-frequency cepstral coefficients (MFCC) mainly use a set of feature vectors in speech recognition, as they improve performance in relation to most other parameters.

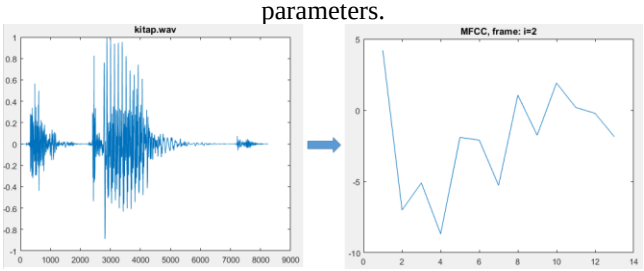

Fig.3. Getting stranded frequency coefficients

Feature matching. Comparison of features includes the actual procedure for identifying an unknown speaker by comparing the extracted features from his voice input with the features from a set of well-known speakers. Vector quantization (VQ) is used to match features in this article.

IV. VECTOR QUANTIZATION (VQ)

Vector quantization (VQ) [3] is an efficient data compression method and has been successfully used in various applications, including encoding and recognition based on vector quantization.

To generate code books, the LBG algorithm is used [2, 3]. The steps of the LBG algorithm are as follows [4]:

1. Develop a 1-vector codebook; it is the centroid of the entire set of training vectors.

2. Double the size of the codebook, dividing each current codebook *y*_n in accordance with the rule:

$$y_n^+ = y_n(1+\varepsilon)$$

$$y_n^- = y_n(1-\varepsilon)$$

where n varies from 1 to the current size of the code book, and ε is the splitting parameter.

3. Find the centroids for the split codebook. (i.e., the code book is twice as large)

4. Steps 2 and 3 are repeated until the M codebook is developed.

Euclidean distance measure

The Euclidean distance is used to measure the similarities or differences between two vectors. Comparison of the input sound fragment is performed by measuring the Euclidean distance between the feature vector of the source fragment and the models (code books) in the database. The smallest average minimum distance is selected by the formula

$$d(x,y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

where x_i is the i-th vector of input features, y_i is the i-th vector feature in the code book, d is the distance between x_i and y_i .

V. RESULT AND DISCUSSION

The system was implemented in MATLAB 2016 on the Windows 10 platform. The system is shown in Figure 4. The result of the study is presented in Table 2. Speech samples used in this work are recorded using the Audacity audio editor. The sampling rate is 16000 Hz (8 bits, mono). Table 1 shows the database description. Samples are collected from different announcers. Samples are taken from each announcer in two sessions so that you can create a training model and test data.

Table 1: Database Description

Language	Karakalpak
Number of speakers	15
Type of speech	Oral speech
Recording conditions	Silent room
Sampling frequency	16000 Гц
Resolution 8 bits / s	8 bits / s

Table 2: Speaker Identification Results

The number of words for learning	The number of words for testing	Result (Speaker Identification Rate)
5	15	98%

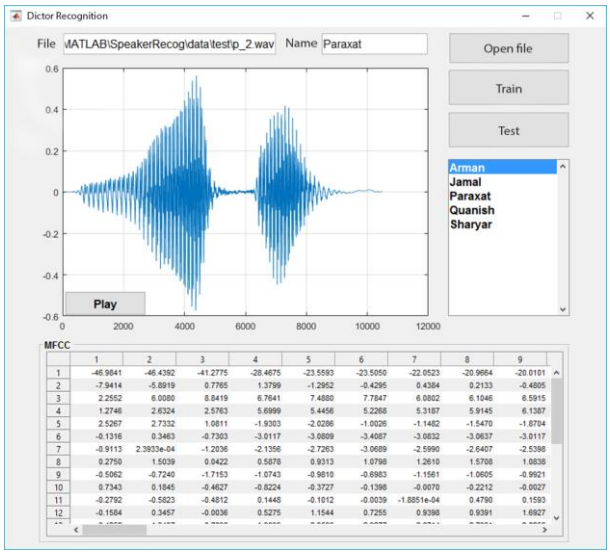

Fig.4. Speaker identification system

VI. CONCLUSION

In this paper, the systems of speaker identification of a text-dependent and independent nature were considered. Feature extraction was performed using chalk-frequency cepstral coefficients (MFCC). Speakers were modeled using vector quantization (VQ). Using the extracted features, the code book from each speaker was built by clustering the feature vectors. Codebooks of all announcers were collected in the database.

Experimental analyzes have shown that it is possible to obtain a high result of identification of individuals with features based on the MFCC. From the results, it can be said that vector quantization using cepstral features produces good results for creating a voice recognition system.

The developed voice control system can be used in a large number of tasks: control of computer applications, mobile platforms, or robotic devices, such as loader robots. The presented approach allows you to create speech recognition systems based on open technologies and a personal computer equipped with a microphone.

REFERENCES

1. Campbell J.P. Speaker Recognition: A Tutorial // Proceedings of the IEEE. 1997. Vol. 85, No. 9. P. 1437-1462.
2. Y.Linde, A.Buzo, and R.M.Gray .: 'An algorithm for vector quantizer design," IEEE Trans. Commun. ', Vol. COM-28, no. 1, pp. 84-95, 1980.
3. A.Gersho, R.M.Gray .: 'Vector Quantization and Signal Compression', Kluwer Academic Publishers, Boston, MA, 1991.
4. 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