International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Algorithm For Identification Based On Voice

Narzillo Mamatov, Abdurashid Samijonov, Nilufar Niyozmatova, Yusuf Yuldoshev , Naibakhon Mamadalieva

***Abstract*: This article proposes an algorithm for automating the process of personality recognition based on voice, provides an analysis of existing methods used to solve the problem that needs to be solved. A method was implemented based on the Gaussian mixture model, which distinguishes a person's voice with high accuracy. The components of this model allow you to simulate sound characteristics that are unique to each person. The results of the proposed algorithm and the use of voice recognition based on the results of the proposed algorithm are presented.**

***Keywords*: Algorithm, Identification, Voice, Model, Gaussian Mixture.**

I. INTRODUCTION

Currently, biometric technologies are widely used in various fields. In particular, internal affairs, access control, banking, etc. Identity recognition is a complex problem based on biometric symbols. Biometric symbols are unique to each person. This feature is also present in speech signals. Each person has important vocal characteristics that are determined based on the individual structure of the vocal apparatus. A person can easily recognize another person by voice, but creating an automated speech recognition system requires many complex tasks [4].Voice-based personality recognition consists of speech extraction from an audio stream, classification and recognition. As a rule, the speaker's recognition and verification tasks are primarily solved. The recognition and verification problem are solved by calculating the degree of similarity of the sample with the basic signals. The degree of similarity between the base and test samples is determined by the criteria of distance or probability [6]. Speaker recognition can be based on text or independent of text. When recognizing speech based on text, recognition is carried out by text prepared in advance or generated by the system, and when recognizing independently of text, by arbitrary text [1].

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Mamatov Narzillo*, Tashkent University Information Technologies named after Al-Kharezmi, Tashkent,
Niyozmatova Nilufar, Tashkent University Information Technologies named after Al-Kharezmi, Tashkent,
Samijonov Abdurashid, Bauman Moscow State Technical University, Moscow, Russia
Yuldoshev Yusuf, Tashkent University Information Technologies named after Al-Kharezmi, Tashkent,
Naibakhon Mamadalieva, Uzbekistan National University of **Uzbekistan** named after Mirzo Ulugbek, Tashkent,

This article discusses the task of automatic personality recognition based on their vocal characteristics, an algorithm for solving recognition problems, regardless of the text, is proposed.

Speech modeling techniques have come a long way from moderating symbol vectors to complex and discriminatory models [3]. The concept of effective models includes modeling data used in training. For example, by estimating the probability density function. Discriminatory models restrict certain classes [8].

Currently, when modeling an announcer for text-dependent systems, approaches such as Dynamic Time Warping (DTW) and the Hidden Markov Model (HMM) are common, and for text-independent systems Vector Quantification-VQ, model Gaussian mixture (GMM) and the support vector method (reference vector machine -SVM) [9].

Dynamic Time Warping (DTW) is a dynamic programming algorithm that allows you to determine the distance between two time series. Typically, such sequences have different lengths, so measurements are required to be carried out at different speeds. This algorithm is simple to implement and is effectively used in many applications, but in some cases, it does not give the expected result.

This algorithm eliminates the mismatch between the x and y axes based on the transformation, this can be done when one point in a given time series depends on several points in another time series [10]. Another difficulty in implementing this algorithm is the difficulty of aligning these two lines when the corresponding points of the series are located above or below each other [11].

The support vector method is used when recognition is implemented on the basis of several classes. In this case, the strategy "against each other" is often used. This requires the construction of q -classifiers. Where each classifier learns the difference from another class. When solving the identification problem, the object class is determined based on the classifier, which gives the separator function the maximum value.

The support vector method is theoretically justified and has the ability to perform classification with high accuracy, this allows the use of various classification approaches in accordance with the selected main function. The main disadvantage of this method is that it takes a long time to learn the choice of a kernel and solve a multiclass recognition problem [2].

The Gaussian mixture model can be used not only for modeling the characteristics of the speaker's sound, but also for recording environmental signals.

Each component of the model represents some common characteristics of sound; however, the speech of each speaker is individual. This model has proven its effectiveness with high recognition accuracy. Therefore, this approach can be effectively used in solving text-independent recognition problems. [12].

Finding the weighted sum of components M, which represents the model of a Gaussian mixture, according to the following formula [5].

$$P\left(\bar{x}\middle|\lambda\right)=\sum_{i=1}^{M}p_i b_i\left(\bar{x}\right), \quad (1)$$

where $\bar{x}$ – D-dimensional vector of random variables, p_i - weight of model components, $b_i\left(\bar{x}\right)$ - density distribution function of component models $\left(1\le i\le M\right)$.

$b_i\left(\bar{x}\right)$ - the density distribution function is calculated by the following formula:

$$b_i\left(\bar{x}\right)=\frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\bar{x}-\overline{\mu_i}\right)^T\Sigma_i^{-1}\left(\bar{x}-\overline{\mu_i}\right)\right\}, \quad (2)$$

where $\overline{\mu_i}$ - vector mathematical expectation, Σ_i - covariance matrix. In this case, the weight of the mixture must satisfy the following condition:

$$\sum_{i=1}^{M}p_i=1 \quad (3)$$

The model of a Gaussian mixture is formed on the basis of the mathematical vector of expectations, the covariance matrix and the weights of the mixtures and is expressed as follows:

$$\lambda=\left\{p_i,\overline{\mu_i},\Sigma_i\right\},\ i=\overline{1,M} \quad (4)$$

In this method, each speaker presents himself as his model of a Gaussian mixture.

II. STATEMENT AND SOLUTION OF THE PROBLEM

Building an automated voice recognition system based on a Gaussian mixture model requires the isolation and processing of incoming speech signals, the development of algorithms for generating model parameters and estimates, as well as solving problems of determining the number of components. In this case, an analog-to-digital conversion of the speech signal is carried out. In the case of sampling, the signal is divided into individual values of the quantum amplitude for a given period of time. In this work, discretization is performed by converting a given signal frequency to a signal of 1000 Hz based on the gradient sampling method. Signal values for 1 millisecond are converted to a single value based on the gradient. A complete recording of the signal obtained on the basis of gradient sampling is scanned in windows of a predetermined length. Typically, the window length is taken at intervals of 20-30 ms. To reduce calculations and ease of application of signal processing methods, each window is taken with a length of 25 ms. Then, the numbered signal is processed in small parts (frames), characteristic of the speech signal for specific voice components. It is assumed that it preserves the signal properties for a given time interval, and a window function is selected. The time window function accepts a non-zero value within a given time interval and sets the value to zero outside this time interval.

After that, the window function is tuned to a sequential signal frame and information is extracted from the speech frame. The information is then obtained by multiplying the value of the signal $x[t]$ at time t by the value of the window function $w[t]$, i.e.

$$y[t]=w[t]x[t] \quad (5)$$

Width (milliseconds), movement (milliseconds between consecutive window borders) and window shape are parameters of the window function.

In this study, a Hamming window was used with a width of L = 25 ms and a displacement of 10 ms:

$$w(t)=\begin{cases}0,54-0,46\cos\left(\dfrac{2\pi t}{L}\right), & 0\le t\le L-1\\ 0, \textit{ else}\end{cases} \quad (6)$$

After filtering each segment, a complete signal is extracted without noise and interference that interfere with speaker recognition.

It is necessary to extract information about the spectral components from the signal obtained in the previous stages of the algorithm. Where the discrete Fourier transform is applied. In the form of an input signal, a signal divided into frames is transmitted to the counter; at the output of the counter, a complex number $X[k]$ with the amplitude and phase of the incoming signal for each interval T. $X[k]$ is calculated by the following formula:

$$X_k=\sum_{n=0}^{N-1}x_n\exp\left(-\frac{2\pi i}{N}kn\right) \quad (7)$$

where $k=0,1,...,\left(N-1\right)$.

Then the transition from the magnitude of the speech frequency f to the value of the height (mel). Initially, the resulting spectrum is placed on the mel scale, this is done using the following formula

$$B\left(f_{hz}\right)=1127,01048*\ln\left(1+\frac{f_{hz}}{700}\right) \quad (8)$$

This operation simulates the existence of different sensitivities of a person's auditory ability at different frequency intervals.

Triangular filters are formed in each frequency range to collect energy values, and for each mel value obtained, logarithms are calculated. In order that various methods of transmitting incoming signals have a lesser effect on an individual speech assessment, logarithms are used.

The resulting values are converted to a frequency scale. The next step of the algorithm is the cepstrum signal. This conversion allows you to separate the sound source from the filter. The conversion properties allow you to generate a sound corresponding to the waveform with the main sound frequency of the audio channel. The filter contains most of the useful information.

Each signal segment is represented with 12 frequency cepstral coefficients, which are determined by the following formula:

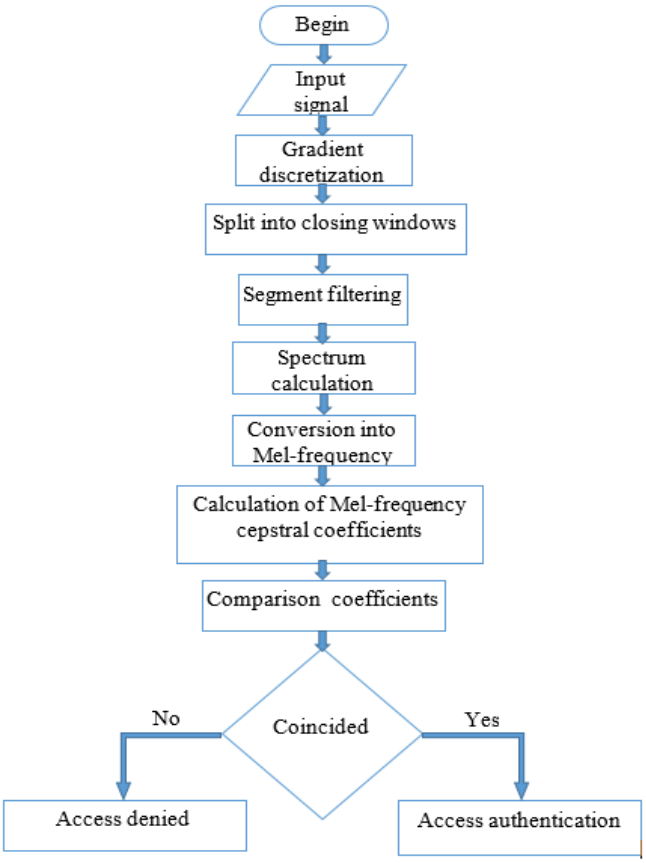
$$c(n)=\sum_{m=0}^{M-1}S(n)\cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M}\right) \quad (9)$$

where $0\le n< M$.

The time-dependent mel-frequency cepstral coefficient of the same phrases of two speakers is different, and the mel-frequency cepstral coefficient of different phrases of the same speaker is not very different [13].

After determining all the coefficients, the signal for identification is compared with all the reference signals stored in the database. As a comparison criterion, the Euclidean metric is used. A block diagram of a personality recognition algorithm based on vocal data is shown in Figure 1. Based on this block diagram, software was developed.

To form the initial parameters of the model for the speech signal vector, the cluster analysis algorithm is used. As a clustering algorithm, the modification algorithm (K-means ++) of the *K*-nearest neighbors' algorithm was chosen. Since in the present algorithm the final error is much smaller than in the *K*-nearest neighbors' algorithm.



Pic. 1. Flowchart of voice recognition automatic speaker identification algorithm.

III. RESULTS

Algorithm testing software was developed in the C ++ programming language. In total, 3 different phrases were taken from 45 people, of which 29 were male voice signals. The speech signal was recorded on a laptop in the form of a 16-24-bit audio file with a frequency of 16-44 kHz in stereo. The time of pronounced sentences varies from 40 to 90 seconds, and the duration of the control signal is 20-30 seconds. The algorithm was tested for operability with a different number of components of the Gaussian mixture model. The following table shows the results.

Based on existing features	**Based on the features of gradient sampling**
Number of features in 135 files:	
977 271 pcs	61 007 pcs
File size 135 files	
1914 Kb	124 Kb
Spent time	
11,7 sek	2,6 sek
Accuracy	
87,9%	97,3%

IV. CONCLUSION

From the results obtained on the basis of the algorithm for automatic personality recognition by voice, we can draw the following conclusions:

the method of "gradient discretization" was proposed, which reduces the file size and the time required for recognition, and also provides high recognition accuracy. This method can be used to compress audio files and speed up recognition;

it was determined that the model of a Gaussian mixture is more effective than other models in modeling the sound properties of an image, which allows speaker recognition with high accuracy.

determination of the parameters of the base model based on the *K*-means ++ algorithm allowed to increase the learning speed and identification accuracy.

it was determined that the optimal number of components for the effective operation of the system is 5. The speaker identification accuracy was 97.3%. This shows that the proposed algorithm can be used in access control systems.

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

1. Ahmad X. M. Introduction to the digital processing of speech signals: textbook. allowance / X. M. Akhmad, V.F. Zhirkov; Vladim. state Vladimir: Publishing house Vladim. state Univ., 2008. ¬ 192 p. ¬ISBN 5¬89368¬751¬5.
2. Konstantin Vorontsov Lecture on the method of support vectors http://www.ccas.ru/voron/download/SVM.pdf
3. Pervushin E. A. Review of the main methods of speaker recognition / E. A. Pervushin // Mathematical structures and modeling. 2011. - Vol. 24. - S. 41-54
4. Rybin S. V. Speech synthesis. Textbook on the discipline "Synthesis of speech" / S. V. Rybin. - St. Petersburg: ITMO University, 2014. - 92 p.
5. Sadykhov R. Kh. Models of Gaussian mixtures for speaker verification by arbitrary speech / R. Kh. Sadykhov, VV Rakush // Reports of BSUIR. - 2003. - No. 4. - S.98 - 103
6. Ramishvili G.S. (1981). Automatic recognition of the speaker by voice. M .: Radio and communication, 221 s
7. Shokina M. O. Application of the k-means ++ algorithm for clustering sequences with an unknown number of clusters // New Information Technologies in Automated Systems. - 2017. - No. 20.

8. Campbell J. P., Speaker Recognition: A Tutorial / J. P. // Proceedings of the IEEE. 1997. V. 85, N 9. P. 1437-1462.
9. Martin A., Przybocki M. The NIST 1999 Speaker Recognition Evaluation - An Overview // Digital Signal Processing. 2000. V. 10
10. Kim S. H. Pattern Matching Trading System Based on the Dynamic Time Warping Algorithm. Sustainability / S. H. Kim, H. S. Lee, H .J. Ko and others. 2018, 10, 4641.
11. Thi-Thu-Hong Phan Dynamic time warping based imputation for univariate time series data. Pattern Recognition Letters / Phan Thi-Thu-Hong, Emilie Poisson Caillault, Alain Lefebvre, André Bigand., Elsevier, 2017, <10.1016 / j.patrec.2017.08.08.019>. <hal-01609256>
12. Chow D. Speaker Identification Based on Perceptual Log Area Ratio and Gaussian Mixture Models / D. Chow, H. Waleed, A. Robust. - Auckland, New Zealand: 2002. - 65 p.
13. Chernetsova EA, Shishkin AD, Voice Identification Algorithm for Authorizing Access to Information // International Research Journal, No. 2, 2019