IRSTI 49.40.01
UDC 004.92

DEVELOPMENT OF A PERSONALITY RECOGNITION SYSTEM BY VOICE

O.Zh.Mamyrbayev[1,2], A.S.Kydyrbekova[1,2], D.O.Oralbekova[1,3], B.Zh.Zhumazhanov[1], O.Mohamed[4]
[1]Institute of information and computational technologies, 050010 Almaty
[2] Al-Farabi Kazakh National University, 050040 Almaty, Kazakhstan
[3]Satbayev University, 050040 Almaty, Kazakhstan
[4]Putra University, Malaysia
(e-mail: kas.aizat@mail.ru https:// https://orcid.org/0000-0001-5740-4100?)
(e-mail: morkenj@mail.ru https://orcid.org/0000-0001-8318-3794)

Abstract. the paper discusses the development of a voice recognition system based on cepstral coefficients on the chalk scale (Mel-frequency Cepstral Coefficients, MFCC) and mixed Gaussian models (Gaussian Mixture Models, GMM). Within the framework of this work, the existing methods for solving the problem of automatic speaker identification by voice were investigated. As a result, a complete review of the subject area was carried out and one of the advanced algorithms for solving the problem was implemented, based on the application of the Gaussian mixture model. The components of the Gaussian mixtures simulate the individual characteristics of the voice, which allows highly accurate distinguishing of human voices. It has been experimentally proven that traditional MFCCs using DNN and i-vector classifiers can achieve good results. The aim of the research is to create a simple and convenient automatic speaker recognition system. The description of the created speech database for the Kazakh language is given. Experiments have shown quite good results of the system FAR 8.05%.

Keywords: voice recognition system; MFCC; GMM; speech database.

Introduction

In practice, voice recognition systems are widely used [1]: access to databases, bank accounts, forensics. Automatic voice recognition is a computational task of verifying a user's claimed compliance using characteristics derived from the user's voice. In the case of automatic voice recognition, the speech signal is processed to obtain information specific to the owner of the voice [2,3]. This information is used to create a non-playable speaker ID that is different from the original one. This makes the voice recognition process a secure method of authenticating the user, as opposed to passwords or tokens, which prevent theft, duplication, or forgetting of the voice. Unlike other biometric technologies, which are primarily image-based and require expensive hardware such as a fingerprint sensor or an eye scanner, voice recognition systems are designed for use on any standard public telephone or telephone network. The ability to work with standard telephony equipment allows you to support a wide range of biometric voice applications in a variety of environments.

The Institute of Information and Computer Technologies has been conducting research in the field of voice recognition for several years. As a result of the experience gained, a prototype of a voice recognition system was developed and a speech base for the Kazakh language was compiled. The aim of the study was to create a simple and convenient automatic voice recognition system.

The general scheme of the created system is shown in Fig. 1. The application is built in the MatLab programming language. The system consists of several main subsystems: verification and identification.

User credentials and voice models are stored in a SQLite database. Speech patterns are recorded in the registration subsystem, and after processing these patterns, a speech model is created. New user compatibility is added to the database along with the voice model. The engine also creates a Universal Background Model (UBM) attack model. In the identification subsystem based on the speech model, the search for the nearest users is carried out according to a certain indicator. The result can be a list of users or an empty list. The list is sorted in ascending order of distance from the user being sorted.

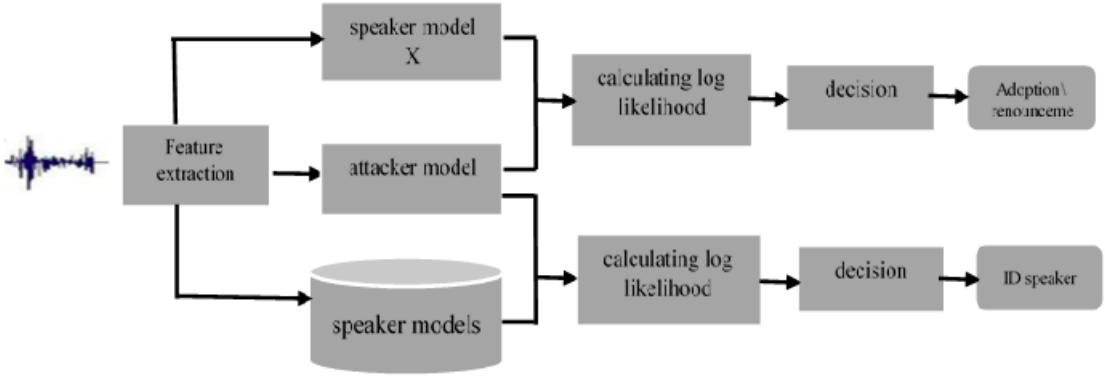

Figure 1- General architecture of the speaker verification and identification system

The verification subsystem verifies the identity of the owner of the vote. To do this, the owner of the voice must enter their identification number in the text field and speak the keyword. If authentication is successful, the user's credentials will be displayed and access granted. You can also update users' voice models based on new speech pattern recordings in this system.

Creation of a speech database for the Kazakh language

The performance of an automated test system is highly dependent on the speech database. The operation of the automatic recognition system is influenced by many factors. These include recording conditions, environment, recording devices, duration, speaker gender, age group, and more. It makes no sense to expect a good result from an automatic checking system without knowing the writing conditions. Here are some important reasons why you need a speech impediment:

a) the corpus can display the language of the content appropriate to the geographic environment using audio and voice data;

b) the practical characteristics of harmonized speech, which are not reflected in the textual database, reflect well the individual characteristics of users;

c) unlike text corporations, the speech corpus reflects prosodic information, as well as the chosen pronunciation style of the chosen sociocultural model.

Database description

The speech corpus for the Kazakh language was created by the Institute of Information and Computer Technologies. There are records of 86 speakers (21 men and 65 women). The recording was made in the office using the Cool Edit Pro software. All information includes: single numbers, single words, combinations of numbers and fragments of text.

The database was recorded in .wav format using a microphone with a sampling rate of 11025 Hz and a resolution of 16 bits per session. All native speakers. The total amount of data for each speaker is about 2500KB. The average speaking time for each speaker is about 100 seconds. There are 86 speakers with IDs from 1001 to 1086.

Speech signals output

Research shows that the perception of human speech sounds does not have a linear scale. Therefore, using the MFCC coefficients [4,5], it is possible to more accurately determine the human auditory system. This allows for better data processing.

The speech signal consists of sounds of different frequencies. For each actual audio frequency f, measured in Hz, the subjective pitch is measured on a scale called "chalk". It is a linear frequency below 1000 Hz and logarithmically above 1000 Hz. The following approximation formula is used to calculate the frequency f

$$mel(f)=2595*\log_{10}(1+\frac{f}{700}). \quad (1)$$

To extract the MFCC coefficients, the signal is divided into frames, to which a window is

applied to reduce spectral distortion. As a result, we get a signal of the form

$$y(n)=x(n)w(n),\ 0\le n\le N-1. \tag{2}$$

Applying the Hamming window

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),0\le n\le N-1. \tag{3}$$

finally, using the Discrete Cosine Transform (DCT), the logarithmic spectrum is converted back to the time domain, and the MFCC coefficients are calculated

$$C_i=\sqrt{\frac{2}{N}}\sum_{j=1}^{L}\log(m_j)\cos\left(\frac{nj}{L}(i-0.5)\right). \tag{4}$$

Gaussian mixed models

Mixed Gaussian models are used to model the distribution of feature vectors received from each user (Mixture Models, GMM) [6,7]. The GMM can be thought of as a nonparametric multivariate Probability Density Function (PDF) that is capable of simulating arbitrary distributions and is the preferred method of speaker modeling.

One of the main advantages of GMM is the ability to create smooth approximations of arbitrary shapes of distributions. GMM has a fast learning phase compared to other approaches, the models can be easily scaled and updated when new speakers are added.

GMM distribution of feature vectors for speaker S is a weighted linear combination of M unimodal densities of Gaussians $b_i^S(x)$, each of which is parameterized by the expectation vector μ_i^s and the covariance matrix ∑Si. These parameters are collectively represented by the following entry:

$$\lambda_S=\{p_i^S,\mu_i^S,\Sigma_i^S\},\quad i=1,...,M \tag{5}$$

where p_i^s – are mixed weights satisfying the condition

$\sum_{i=1}^{M}p_i^S=1.$

each speaker has its own model λ_S .

For the feature vector x, the mixed density for the speaker S is calculated as

$$p(x|\lambda_S)=\sum_{i=1}^{M}p_i^S b_i^S(x), \tag{6}$$

where

$$b_i^S(x)=\frac{1}{(2\pi)^{D/2}\left|\Sigma_i^S\right|^{1/2}}\exp\left\{\frac{1}{2}(x-\mu_i^S)'\Sigma_i^S(x-\mu_i^S)\right\} \tag{7}$$

For a given sequence of feature vectors X{x1,x2,…,xT}, which are assumed to be independent, the logarithmic likelihood [8] of the speaker model λ_S is presented in the form

$$L_S(X)=\log p(X|\lambda_S)=\frac{1}{T}\sum_{t=1}^{T}\log p(x_t|\lambda_S). \tag{8}$$

To identify a speaker, the last equation is calculated for each speaker registered in the system. This paper uses a GMM with 32 blends for each model.

The speaker's identity is determined by the model with the highest value. Various algorithms are used to find the maximum likelihood of models. One of them is the EM (Expectation-Maximization) algorithm [11].

In this case, for each speaker S, we find the following values:

mixed weights:

$$p_i = \frac{1}{T}\sum_{t=1}^{T} pr(i|x_t, \lambda), \tag{9}$$

mathematical expectation:

$$\mu_i = \frac{\sum_{t=1}^{T} pr(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} pr(i|x_t, \lambda)}, \tag{10}$$

covariance matrix:

$$\Sigma_i = \frac{\sum_{t=1}^{T} pr(i|x_t, \lambda)x_i^2}{\sum_{t=1}^{T} pr(i|x_t, \lambda)} - \mu_i^2, \tag{11}$$

where the posterior probability for component i has the form

$$pr(i|x_t, \lambda) = \frac{p_i b_i(x)}{\sum_{k=1}^{M} p_k b_k(x)}. \tag{12}$$

In modern voice recognition systems, UBM is used to simulate an alternative hypothesis. To simulate the possible space of acoustic possibilities, the standard TV-JFA method was used, which is one of the most effective methods in the field of voice testing [9,10]. The voice model in this approach is as follows:

$$M = m + Ux + Vy + Dz, \tag{13}$$

where M is the supervisor of the mixture of Gaussian distributions (GMM-models) of the speaker's voice, m is the supervisor of the parameters of the universal background model (UBM), U, V, D are their own matrix channels (Eigen Channel), respectively, their own voices and residual variability.

In the space of total variability, the i-vector is obtained using factor analysis, analyzed in the average supervectors of the UBM model and in the T-matrix of total variability. In this case, the voice model is described in the following relationship:

$$M = m + Tw, \tag{14}$$

where w is a low-dimensional vector in the space of possibilities.

In our systems, UBM is a mixture of Gaussian models of the described characteristics. To train the T-matrix and UBM, the features obtained from the training base of the competition were used. The UBM diagonal covariance matrix was studied using the EM-algorithm (Expectation - Maximization) [11]. It has voiceover characteristics and channel independent characteristics. UBM is a set of GMMs trained to distribute functions independently of the speaker.

Experimental results

Speech samples specially written for the experiments were taken from the speech database in question. Initially, each speech signal was preprocessed. The Hamming window length was 25 msec and the overlap was 12.5 msec. The following 20 MFCCs were obtained. The training was

carried out in 5 samples for each speaker. The duration of the teaching presentation was about 10 seconds. Testing was conducted on 3-second samples.

Table 1. Experiment Results

GMM с 32 смесями					
Duration of training	Duration of testing	Accuracy of recognition	FAR	FRR	EER
10 сек	3 сек	93,7 %	8 ,05%	7 ,72%	7 ,89%

The experimental results are presented in Table 1. Recognition accuracy, pseudo-acceptance rate (FAR) and pseudo-rejection rate (FRR), as well as the equal error of type I and II errors (error rate, EER) are given.

Conclusion

The work describes the developed prototype of the voice recognition system. The results of testing the system in the Kazakh language on the basis of the collected speech are considered. Experiments have shown high results of his work. In the future, research will continue to develop an automatic voice recognition system to further enhance its security and recognition performance.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP08855743).

References

[1] J.P. Campbell. Speaker recognition: A tutorial. Proc. IEEE, (1997). 9(85). 1437–1462.

[2] J. Benesty M. Sondhi, Y. Huang, Springer handbook of speech processing, Springer, 2007. 1176 p.

[3] L. Rabiner B.H. Juang, Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993. 277 p.

[4] S. Furui, Digital speech processing, synthesis, and recognition. MarcelDekker, 2000. 452 p.

[5] S. Furui. Cepstral analysis techniques for automatic speaker verification. IEEE Tran. acoust., speech, signal processing, 1981. 27. 254-277.

[6] D.A. Reynolds. Speaker identification and verification using. Gaussian mixture speaker models. Speech Communication. 1995. 17. 91-108.

[7] D.A. Reynolds. A Gaussian mixture modeling approach to textindependent speaker identification. Ph.D. Thesis. – Georgia Institute of Technology, September, 1992.

[8] M.N.Kalimoldayev, O.Zh.Mamyrbayev, A.S.Kydyrbekova, N.O.Mekebayev, Algorithms for Detection Gender Using Neural Networks// International journal of circuits, systems and signal processing, ISSN, 2020. 14. 1998-4464.

[9] Mamyrbayev O., M.Turdalyuly, N.Mekebayev, K.Alimhan, T.Turdalykyz., A.Kydyrbekova A. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. Intelligent Information and Database Systems Proceedings, Part II, Indonesia, 2019.

[10] Kydyrbekova A., Othman M., Mamyrbayev O., Akhmediyarova A., B.Zhumazhanov Identification and authentication of user voice using DNN features and i-vector. Cogent Engineering. 2020. 7 (1751557). 1 - 21.

[11] A.P. Dempster, N.M. Laird, D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977. 39. 1-38.