SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

SEMARAK ILMU
Journal of
ADVANCED RESEARCH IN APPLIED SCIENCES AND ENGINEERING TECHNOLOGY
ISSN 2462 - 1943

MFCC-GMM Method for Speaker Identification by Voice

N.A.Niyozmatova[1], N.S. Mamatov[2], X.T. Dusonov[3], B.N. Samijonov[4], A.N.Samijonov

[1] Digital Technologies and Artificial Intelligence, "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan
[2] Digital Technologies and Artificial Intelligence, "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan
[3] Department of Computer Science and Programming, Jizzak branch of Mirzo Ulugbek National University of Uzbekistan, Jizzak, Uzbekistan
[4] Student, Sejong University, South Korea, Seoul, Korea
[5] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

ARTICLE INFO

Article history:
Received
Received in revised form
Accepted
Available online

Keywords:

MFCC, VQ, GMM, Frame, Hamming, Fourier transform, Mel-filter, code word.

ABSTRACT

Voice, as a characteristic of humans, provides great opportunities for communication and identification. Today, voice recognition systems are widely used in many areas of human activity. However, the problem of developing perfect voice recognition systems is still considered an urgent task by researchers. Especially when the speech sample duration is relatively short, it is important to solve the problem of low recognition accuracy. Therefore, in this article, the Mel-frequency cepstral coefficients (MFCC) feature set extraction algorithm and Gaussian mixture model (GMM) were researched in the implementation of identification, based on which experimental researches were conducted on the recognition of a person based on his voice. In this, the speech samples of male and female speakers recorded in different environments were used, and the efficiency of the methods was compared by applying the MFCC-GMM and MFCC-VQ methods to them. As a result, the MFCC-GMM method is found to be more accurate than the MFCC-VQ method. That is, in the text-dependent condition, the accuracy of the speech recognition ranged from 82.8% to 94.5%, and in the text-independent condition, it showed an accuracy of 79.5% to 87.4%.

1. Introduction

Recognition of a speaker based on his voice is highly dependent on his anatomical features and habits. In addition, the voice is also related to diseases and emotional factors. Also, external factors can strongly influence the quality of recognition.

Voice recognition is the identification of a speaker by comparing the voice attributes with the attributes of existing speech samples [1]. Speech and speech recognition requires knowledge of acoustic, phonetic, and linguistic features [2-4]. The problem of identifying a speaker based on his voice belongs to the category of multidisciplinary problems such as face recognition, language recognition and speech recognition, most of the tasks of which are directly related to short speech or human instructions. Voice recognition can be text-dependent or text-independent, according on the nature of the problem to be solved. If the recognition text-dependent, recognition is based on pre-prepared or system-generated text. If the recognition is text-independent, user can use arbitrary text. There are several internal and external factors that affect the voice recognition system. The small number of words used also complicates the process of identifying a speaker [5].

Nowadays, there are solutions have been proposed for determining a speaker's gender, age, etc. based on voice, but there are many problems in creating systems for identifying the identity of the speaker. Fast and expensive hardware is required to operate a reliable recognition system in real time. Voice recognition is usually done through training and testing [6,7]. At the training stage, personality traits and their values are calculated from the speech stream. Features are used to create models of different individuals [8]. In the testing phase, speech samples of unknown individuals are compared using models and classification methods [9]. The efficiency and reliability of the experiments and the identification rate are evaluated in terms of accuracy and error rate [10].

2. Literature Review

The analysis of the feature extraction techniques used in voice recognition for Vector Quantization (VQ), Gaussian Mixture Model (GMM), and Neural Network (NN) is provided by Bimbot et al., [11], which states that Mel-frequency cepstral coefficients (MFCC), one of the feature extraction techniques, perform well with noise.

In a study on speech emotion detection, Todkar et al., [12] discovered that seven emotions performed well. Features were created by MFCC and LPC and modeled by the GMM model. Rohinikumar et al., [13] proposed that the MFCC and IHC feature extraction methods be compared with the GMM and UBM modeling approaches; the MFCC and GMM pair turned out to be the better option. Maximum log-likelihood (MLL) estimate was utilized for identification based on MFCC and VQ/GMM on three bases, and the results demonstrated a high degree of accuracy in identifying sounds in noisy situations.

The topic of automatic person recognition was studied by Sithara et al., Barai et al., (2017) [14, 15]. Barai et al., (2018) [16] proposed that the problem of automatic voice recognition for Arabic numerals in noisy situations be examined. The MFCC-GMM and MFCC-VQ algorithms are explored, and a combination of them called the MFCC-VQ-GMM is developed with excellent results.

The use of voice recognition technology in smart home control systems is demonstrated by Ouisaadane et al., [17], where the MFCC and GMM algorithms are the two key elements that are highlighted. In this instance, the accuracy of identifying the voice of an individual who is registered was high, but the accuracy of identifying the voice of an unregistered person was quite low.

3. Methodology

Below is a description of the algorithm used in the proposed approach. It also sheds light on the usefulness and reliability of voice recognition systems. MFCC [18,19], VQ [20] and GMM [21] algorithms are used in this work.

Mel-frequency cepstral coefficients (MFCC) is one of the effective feature set extraction methods used in speech signal description. MFCC consists of the following six steps:

1. Framing. This requires recording the audio signal at a frequency of 16 kHz, which can be used to divide the speech signal into frames. A typical 25 ms (millisecond) audio signal consists of 400 samples.

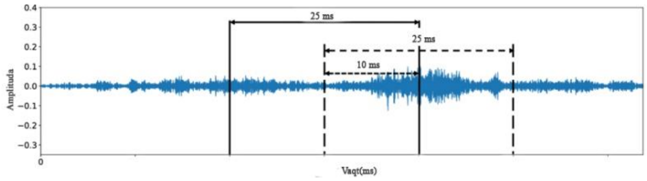

Fig. 1. Dividing the speech signal into frames

2. Hamming window. Since the first 400 samples start at 0, the next 400 samples start at 160, which means 240 samples intersect.

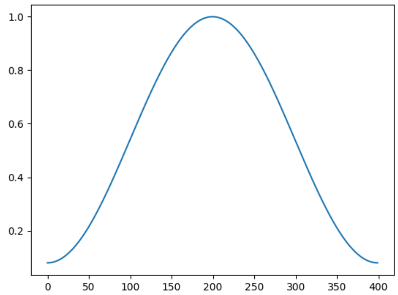

Fig. 2. Hamming window

The Hamming window is widely used to solve speech problems. The signal through the Hamming window is replaced by the following formula:

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

where t -is time, N -is the window length of n samples, and w[t]- is the Hemming window in the expression at time t.

3. Discrete Fourier transform (DFA): A signal is converted from the time domain to the frequency domain by applying the discrete Fourier transform. For speech signals, analysis in the frequency domain is more convenient than in the time domain. The Fourier substitution for a time signal is performed by the following formula:

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

where w(n)- is the Hamming window in expression (Eq.1), $x(n)$ -is the signal in the time domain, N - is the length of the window consisting of n samples, K - is the length of the DFA.

The formula for the power spectrum sample energy is obtained as follows:

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

4. Application of Mel-filter bank. Triangular filter sets can be obtained from 20 to 40 (26 standard). In this case, to calculate the energies of the filter banks, each filter bank is multiplied by the power spectrum and coefficients are generated. Once this is done, 40 values are generated that tell you how much energy is available in each filter bank. In this case, usually 1/2 or 1/4 of the elements are transferred to the mel space. For 512 Fure elements, this means filters with 256 or 128 mel steps. (Eq.4) converts the frequency to the Mel scale. (Eq.5) converts to the opposite frequency. Mel-filter bank is calculated by the following formula:

$$M(f) = 1125ln\left(1 + \frac{f}{700}\right) \quad (4)$$

$$M\left(f\right)=700\left(e^{\frac{m}{1125}}-1\right) \tag{5}$$

where f-is number of filter.

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

$$F_m\left(k\right)=\begin{cases}0, & k<f\left(m-1\right)\\ \dfrac{k-f\left(m-1\right)}{f\left(m\right)-f\left(m-1\right)}, & f\left(m-1\right)\le k<f\left(m\right)\\ 1, & k=f\left(m\right)\\ \dfrac{f\left(m+1\right)-k}{f\left(m+1\right)-f\left(m\right)}, & f\left(m\right)<k\le f\left(m+1\right)\\ 0, & k>f\left(m+1\right)\end{cases} \tag{6}$$

$$MSFB=20log_{10}\left(S(n,k)\cdot F_m(k)\right) \tag{7}$$

where m-is Mel's scale coefficient.

6. Discrete cosine substitution. In this step, the reverse substitution is performed for the output of the previous step. 13 coefficients of the signal are generated after the MFCC method applies the discrete cosine transform.

$$S_{MFCC_i}=\sum_{k=1}^{M}X_k\cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{M}\right],\ i=1,2,...,M \tag{8}$$

where X_k - is the number of Mel-scaled filter bank filters.

After extracting features from the speech signal, a voice model is built for each speaker. The following methods are used for modeling:

1. Vector Quantization (VQ) is a digital signal processing method that involves encoding digital data using a limited number of vectors [22]. This method is commonly used in image, speech compression, and data analysis and pattern recognition. The vector quantization process is usually performed in the following steps:

- divide data into smaller groups or clusters;
- determine a representative value for each cluster, known as a code word, that is closest to the values of the data points in the cluster;
- assign each data point to the closest cluster to its code word;
- encoding data by replacing each data point with a suitable code word;

Vector quantization can also cause data loss. Because the original data points are replaced by their codewords, they may not represent the original data correctly.

It helps to reduce the storage and transmission requirements while maintaining the original signal quality. By grouping similar data points and representing them by a single vector, vector quantization can effectively reduce the number of bits needed to represent a signal. It is used for data compression and pattern recognition.

The Gaussian Mixture Model (GMM) is a statistical model widely used in face recognition, and it is a method that provides a probability model for identifying a speaker based on his voice. An important aspect of any accent modeling is to collect and find a training sample of each accent and mixture weight mean vector. The parameters of the Gaussian mixture model are calculated by ML-maximum likelihood. The probability density function of the shapes is approximate. The GMM probability density function can be expressed by the covariance matrix ($\sum_i$), the mathematical expectation (μ_i) and a set of mixture weight parameters *(*K_i) [23]. A multivariate Gaussian mixture model is defined by the following expression.

$$K_i = \frac{1}{\sqrt{\left(2\pi\right)^{D/2}\left|\sum_i\right|^{1/2}}}\,\text{эх}n\left[-\frac{1}{2}\left(x-\mu\right)^T \sum\nolimits_i^{-1}\left(x-\mu_i\right)\right] \tag{9}$$

The proposed algorithm consists of classification methods after feature extraction. The characteristics of speech signals are calculated using MFCC. The resulting feature is then classified using a GMM. The final result is calculated using the maximum logorrheic likelihood function. Below is the algorithm for detecting speakers using the MFCC-GMM method.

The MFCC-GMM method for recognizing a speaker based on his voice is implemented in the following steps:

Feature extraction. The speech signal is preprocessed and divided into frames. Then each frame's MFCC coefficients are calculated. Usually, the number of coefficients is taken from 12 to 20.

Feature normalization. MFCC coefficients are normalized to have a mean of zero and a variance of unity. This step serves to increase the accuracy of the voice recognition system [24].

Training GMM. The GMM is trained on the normalized MFCC coefficients for each individual. The number of components in a GMM can vary according to the size of the training sample set.

Modeling the speaker. Once the GMM is trained, it can be used to model each speaker's speech. For each speaker, the GMM represents the distribution of their speech signal.

Voice recognition. To recognize a speaker from the test speech signal based on his voice, the MFCC coefficients of the test signal are first calculated and normalized. Then the probability of the test signal corresponding to the GMM for each candidate is calculated. The candidate with the highest probability is taken as the variable of the test signal.

The MFCC-GMM method was used to recognize the Uzbek speech signal based on the speaker’s voice with promising results. This method has been shown to be superior to other voice recognition methods in terms of accuracy and efficiency. The success of the MFCC-GMM method is due to the way MFCC extracts features from speech signals and the effectiveness of GMM in modeling the speech of different speakers [25].

The MFCC-GMM method has been shown to be an effective way to recognize the Uzbek speech signal based on the speaker’s voice. The effectiveness of this method is important in the separation and modeling of characters in voice recognition [26].

4. Results

A set of 5 male (M) and 5 female (W) voice samples was used in the experiments. For each speaker, recorded speech signals from 15 different environments are matched. The tables show the result of recognition using the MFCC-VQ method and the MFCC-GMM method for Uzbek speech. Tables 1 and 2 show the text-independent and text-dependent voice recognition results, respectively. Figure 3 shows graphs for text-independent voice recognition, and Figure 4 shows results for text-dependent recognition. The overall accuracy of both methods is shown in Figure 5.

Table 1
Text-independent voice recognition results

Algorithm	Umid	Akbar	Ilkhom	Aziz	Jamshid	Munisa	Dilbar	Imona	Samiya	Xusnora	Result
MFCC-VQ	79,6	80,3	77,6	81,7	79,3	78,5	79,6	80,6	78,7	79,3	79,5
MFCC-GMM	84,3	87,4	82,4	84,6	85,1	83,7	81,9	82,2	84,2	83,7	83,9

Table 2
The result of text-dependent speaker recognition based on voice

Algorithm	Umid	Akbar	Ilkhom	Aziz	Jamshid	Munisa	Dilbar	Imona	Samiya	Xusnora	Result
MFCC-VQ	85,7	87,4	84,2	80,5	81,6	81,4	82,2	80,7	83,5	82,1	**82,8**
MFCC-GMM	91,4	94,5	90,9	91,9	93,6	90,6	93,4	91,3	92,8	92,9	**91,3**

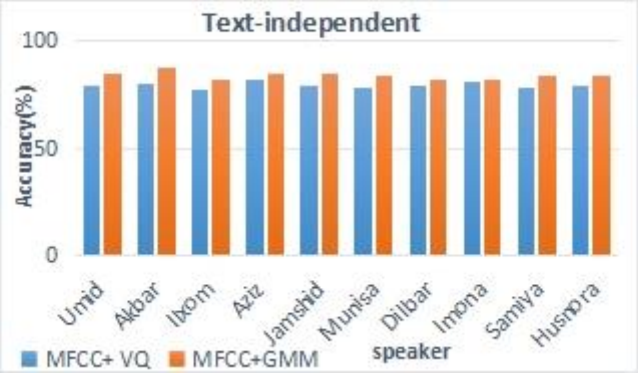

Fig. 3. Accuracy plot of text-independent speaker recognition based on voice

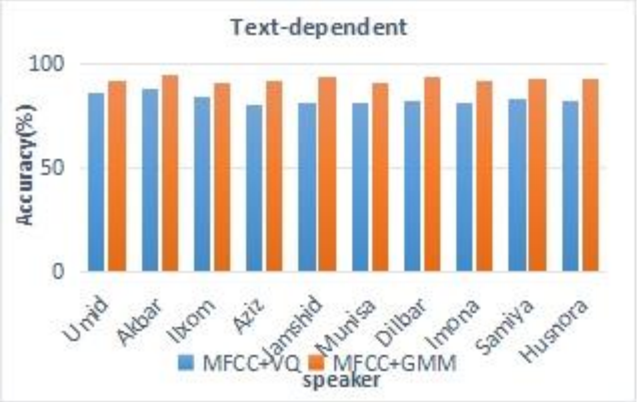

Fig. 4. Accuracy plot of text-dependent speaker recognition based on voice

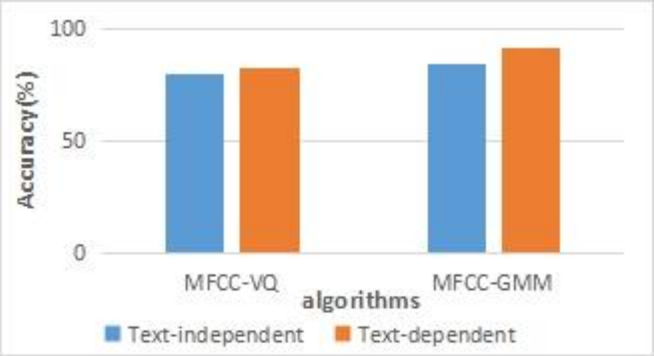

Fig. 5. Total accuracy plot of MFCC-VQ and MFCC-GMM algorithms

From the obtained results, it was found that there are some differences between MFCC-VQ and MFCC-GMM algorithms. It showed that the MFCC-GMM algorithm outperforms the MFCC-VQ algorithm in terms of accuracy in both text-independent and text-dependent voice recognition. Also, the comparison results of MFCC-VQ and MFCC-GMM algorithms through literature analysis and experimental research are presented in Table 3.

Table 3.
Comparison results of MFCC-VQ and MFCC-GMM algorithms

Aspect	**MFCC-VQ**	**MFCC-GMM**
Modeling approach	Uses VQ for modeling	Uses GMM for modeling
Expression of speech	Quantizes a character vector directly into codewords	Represents speech symbols as a mixture of Gaussians
Model complexity	The size of the fixed-size codebook is determined	It is determined by the number of Gaussian components and parameters
Flexibility / Generalization	Discrete representation is not flexible	Adaptability and better generalization ability
Recognition efficiency	Limitations may be encountered in accurately modeling variability	Shows high accuracy in recognition
Implementation complexity	Simple	Complicated

5. Conclusions

The results of the MFCC feature set separation algorithm and the GMM modeling algorithm are better than the results of the MFCC feature set separation algorithm and the vector quantization VQ modeling algorithm in solving the problem of person identification based on voice. use has shown to give high results.

In this research, the issue of recognizing a person based on his voice was investigated, and MFCC-VQ and MFCC-GMM algorithms were used to generate speech features, and then their comparison was carried out. MFCC-VQ and MFCC-GMM are both algorithms used in voice recognition systems. However, in the research work, they were found to have significant differences in modeling and recognition and were presented in tables.

As a result of the experimental studies, it was found that the MFCC-GMM algorithm performed about 4.4% better than the MFCC-VQ algorithm in text-independent voice recognition, and about 8.5% better in text-dependent voice recognition.

Using GMM in combination with MFCC features for speech modeling improves the accuracy of the voice recognition system compared to the traditional VQ approach.

References

[1] Kenny, P., Pierre, ., Najim, D., Vishwa, G., Pierre, D. 2008. "A Study of Interspeaker Variability in Speaker Verification." IEEE Transactions on Audio, Speech, and Language Processing 16 (5): 980–88. https://doi.org/10.1109/tasl.2008.925147.

[2] Aida-Zade, K. R., Cemal, A., Rustamov, S. 2017. "Investigation of Combined Use of MFCC and LPC Features in Speech Recognition Systems." International Journal of Computer and Information Engineering 1 (8): 2647–53. https://doi.org/10.5281/zenodo.1314791.

[3] Mamatov, N. S., Niyozmatova, N. A., Samijonov, A. N., Samijonov, B. N. 2022. "Construction of Language Models for Uzbek Language." 2022 International Conference on Information Science and Communications Technologies (ICISCT), September, 1–4. https://doi.org/10.1109/icisct55600.2022.10146788.

[4] Watanabe, Sh., Takaaki, H., Suyoun, K., John, R., and Tomoki, H. 2017. "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition." IEEE Journal of Selected Topics in Signal Processing 11 (8): 1240–53. https://doi.org/10.1109/jstsp.2017.2763455.

[5] Bao, L., Xueju, Sh. 2016. "Improved Gaussian mixture model and application in speaker recognition." In Control, Automation and Robotics (ICCAR), 2016 2nd International Conference, April, 387–90. https://doi.org/10.1109/iccar.2016.7486761.

[6] Mamatov, N., Niyozmatova, N. A., Abdullaev, Sh. Sh., Samijonov, A.N., Erejepov, K.K. 2021. "Speech Recognition Based on Transformer Neural Networks." 2021 International Conference on Information Science and Communications Technologies (ICISCT), November. https://doi.org/10.1109/icisct52966.2021.9670093.

[7] Mamatov, N., Niyozmatova, N. A., Yuldoshev, Yu. Sh., Abdullaev, Sh. Sh., Samijonov, A.. 2023. "Automatic Speech Recognition on the Neutral Network Based on Attention Mechanism." In Lecture Notes in Computer Science, 100–108. https://doi.org/10.1007/978-3-031-27199-1_11.

[8] Niyozmatova, N., Mamatov, N., Samijonov, A., Rahmonov, E., Juraev, Sh. 2020. "Method for Selecting Informative and Non-informative Features." IOP Conference Series: Materials Science and Engineering 919 (4): 042013. https://doi.org/10.1088/1757-899x/919/4/042013.

[9] Alimurodov, A.K., Churakov, R.R. 2015. "Rewiew and classification methods for processing speech signals in the speech recognation systems". Measurement. Monitoring. Control 2.

[10] Bimbot F., Bonastre, J. F., Fredouille C., Gravier G., Magrin-Chagnolleau, I., Meignier, S., Reynolds D. A. 2004. A tutorial on textindependent speaker verification. EURASIP journal on applied signal processing: 430-451.

[11] Todkar, Satyam P., Snehal S. Babar, Rudrendra U. Ambike, Prasad B. Suryakar, and J. Laxmi Prasad. 2018. "Speaker Recognition Techniques: A Review." 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, India, April, 1–5. https://doi.org/10.1109/i2ct.2018.8529519.

[12] Rohinikumar, Subhashree, and G. N. Rathna. 2016. "Speech Emotion Recognition: Performance Analysis Based on Fused Algorithms and GMM Modelling." Indian Journal of Science and Technology 9 (11). https://doi.org/10.17485/ijst/2016/v9i11/88460.

[13] Sithara, A., Thomas, A., Dominic, M. 2018. "Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications." Procedia Computer Science 143 (January): 267–76. https://doi.org/10.1016/j.procs.2018.10.395.

[14] Barai, B., Das, D., Das, N., Basu, S., Nasipuri, M. 2017. "An ASR System Using MFCC and VQ/GMM With Emphasis on Environmental Dependency." 2017 IEEE Calcutta Conference (CALCON), Kolkata, India, December, 362–66. https://doi.org/10.1109/calcon.2017.8280756.

[15] Barai, B., Das, D., Das, N., Basu, S., Nasipuri, M. 2018. “Closed-Set Text-Independent Automatic Speaker Recognition System Using VQ/GMM.” In Advances in Intelligent Systems and Computing, 337–46. https://doi.org/10.1007/978-981-10-7566-7_33.

[16] Ouisaadane, A., Saïd, S., Miloud, F. 2020. “Arabic Digits Speech Recognition and Speaker Identification in Noisy Environment Using a Hybrid Model of VQ and GMM.” TELKOMNIKA Telecommunication Computing Electronics and Control 18 (4): 2193. https://doi.org/10.12928/telkomnika.v18i4.14215.

[17] Malik, R. A., Setianingsih, C., Nasrun, M. 2020. “Speaker Recognition for Device Controlling using MFCC and GMM Algorithm.” 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, November, 1–6. https://doi.org/10.1109/icecie50279.2020.9309603.

[18] Niyozmatova, N. A., N. Mamatov, Sh. A. Tulyaganova, Abdurashid Samijonov, and Boymirzo Samijonov. 2023. “Methods for Determining Speech Activity of Uzbek Speech in Recognition Systems.” AIP Conference Proceedings, January. https://doi.org/10.1063/5.0145438.

[19] Xie, Xiaoping, Hao Cai, Can Li, Yu Wu, and Fei Ding. 2023. “A Voice Disease Detection Method Based on MFCCs and Shallow CNN.” Journal of Voice, October. https://doi.org/10.1016/j.jvoice.2023.09.024.

[20] Kekre, H., Kulkarni, V. 2010. “Speaker Identification by using Vector Quantization”. International Journal of Engineering Science and Technology. 2.

[21] Přibil, Jiří, Anna Přibilová, and Jindřich Matoušek. 2014. “GMM Classification of Text-to-Speech Synthesis: Identification of Original Speaker’s Voice.” In Lecture Notes in Computer Science, 365–73. https://doi.org/10.1007/978-3-319-10816-2_44.

[22] Gersho, A., and Robert M. Gray. 1992. Vector Quantization and Signal Compression. Springer eBooks. https://doi.org/10.1007/978-1-4615-3626-0.

[23] Mamatov, N., N. A. Niyozmatova, and Abdurashid Samijonov. 2021. “Software for preprocessing voice signals.” International Journal of Applied Science and Engineering 18 (1): 1–8. https://doi.org/10.6703/ijase.202103_18(1).006.

[24] Mamatov, N., Samijonov Abdurashid, Nurimov Parakhat, and Niyozmatova Nilufar. 2019. “Automatic Speaker Identification by Voice Based on Vector Quantization Method.” International Journal of Innovative Technology and Exploring Engineering 8 (10): 2443–45. https://doi.org/10.35940/ijitee.j9523.0881019.

[25] Wiedecke, Bernd, N. Mamatov, Mirabbos Payazov, and Samijonov Abdurashid. 2019. “Acoustic Signal Analysis and Identification.” International Journal of Innovative Technology and Exploring Engineering 8 (10): 2440–42. https://doi.org/10.35940/ijitee.j9522.0881019.

[26] Narzillo, M., Abdurashid, S., Parakhat, N., & Nilufar, N. 2019. “Karakalpak Speech Recognition With CMU Sphinx.” International Journal of Innovative Technology and Exploring Engineering 8 (10): 2446–48. https://doi.org/10.35940/ijitee.j9524.0881019.