



THE ISSUE CONTAINS:

Proceedings of the 4th
International Scientific
and Practical Conference


**INNOVATIVE DEVELOPMENT
IN THE GLOBAL SCIENCE**

Boston, USA
26-28.08.2024


SCIENTIFIC COLLECTION
INTERCONF

No 214
August, 2024


CHEMISTRY AND MATERIALS SCIENCE

	Shumykin S.O. Shevchenko V.G. Ryagin S.L. Onyshchenko R.V.	IMPROVING THE MECHANICAL PROPERTIES OF THE WORKING SURFACE OF STEEL PARTS OF MACHINES OPERATING IN SPECIFIC CONDITIONS	98
---	---	---	----




AGROTECHNOLOGIES AND AGRICULTURAL INDUSTRY

	Талыбов Т.Г. Наджафов Дж.С.	БИОМОРФОЛОГИЧЕСКИЕ И АГРОТЕХНИЧЕСКИЕ ОСОБЕННОСТИ НЕКОТОРЫХ АБОРИГЕННЫХ СОРТОВ ВИНОГРАДА НАХЧЫВАНСКОЙ АВТОНОМНОЙ РЕСПУБЛИКИ АЗЕРБАЙДЖАНА	101
---	--------------------------------	--	-----



MODELING AND NANOTECHNOLOGY

	Игамбердиев Х.Э. Фозилова М.М.	ХАРАКТЕРНЫЕ ОСОБЕННОСТИ И СВОЙСТВА ПРОИЗВОДСТВЕННЫХ ОБЪЕКТОВ С РЕЦИКЛОМ	110
---	-----------------------------------	--	-----

INFORMATION AND WEB TECHNOLOGIES

	Kozhukhivskyi A.D. Kozhukhivska O.A.	FUZZY SET AND FUZZY LOGIC ANALYSIS METHODS	116
	Mamatov N.S. Babomurodov O.J. Dusanov K.T.	STAGES OF PERSONAL RECOGNITION BASED ON VOICE	122
	Корчинський В.В. Сокур Д.І. Белова Ю.В.	АНАЛІЗ ВРАЗЛИВОСТЕЙ ТА СТРАТЕГІЇ ЗАХИСТУ ВЕБ-ДОДАТКІВ ВІД CSRF-АТАК	129

MILITARY AFFAIRS AND NATIONAL SECURITY

	Мельник А.П. Балковий А.В.	ПІДХІД ДО ВИБОРУ ОСНОВНИХ ПОКАЗНИКІВ ДЛЯ ПРОГНОЗУВАННЯ СТАНУ АРТИЛЕРІЙСЬКИХ ВОЄПРИПАСІВ	138
	Ярошенко О.В.	ПОРЯДОК ОБЛІКУ ВИРОБНИЧОЇ ДІЯЛЬНОСТІ РЕМОНТНО-ВІДНОВЛЮВАЛЬНОГО ОРГАНУ	152

INFORMATION AND WEB TECHNOLOGIES

Stages of personal recognition based on voice

**Mamatov Narzillo Solidjonovich¹, Babomurodov Ozod Jo'rayevich²,
Dusanov Khurshid Toshpulotovich³**

¹ Professor, Doctor of Technical Sciences;
National Research University «Tashkent Institute of Irrigation and Agricultural
Mechanization Engineers»; Republic of Uzbekistan

² Doctor of Technical Sciences, professor, Executive director of the branch;
Kazan Federal University in the city of Jizzakh; Republic of Uzbekistan

³ PhD student;
Mirzo Ulugbek National University of Uzbekistan Jizzakh Branch; Republic of Uzbekistan

Abstract. In this article, the stages of speech recognition and recognition of a person through human voice were considered. The procedure of speech data set formation, pre-processing of the speech signal was carried out using voice activity detection (VAD) algorithm, mel-frequency cepstral coefficients (MFCC) method of character set formation, and recognition Gaussian mixture model (GMM) and deep neural network (DNN) method. A Gaussian mixture model was studied for automatic person identification by voice. From the results, it can be said that the deep neural network method using cepstral features gives good results for creating a voice recognition system.

Keywords: GMM, dataset, VAD, scalar threshold, frame, Hamming window, mathematical expectation.

The issues of introducing digital-logic devices, programming, commanding and controlling them with various forms and tools are considered as the most urgent problems of our day. One of the important aspects of human-machine communication is the principle of commanding the machine and mutual data sharing [1]. There are a number of approaches, including the use of speech signals. According to this principle, a command is given by a person to a machine in the form of speech, and automatic recognition from the speech by the machine is important. The development and implementation of speech recognition systems provides convenience and ease. We can see such systems as a means of directly commanding devices in security and military, from industrial production to our daily lives [2].

INFORMATION AND WEB TECHNOLOGIES

Speech recognition and identity recognition are different concepts, and identity recognition is based on human voice. Recognition of a person based on human voice is carried out in several stages:

1. Formation of speech database;
2. Preprocessing of the speech signal;
3. Forming a set of characters from the speech signal;
4. It consists of obtaining results through recognition methods (GMM and neural network).

1. Formation of speech database. Creating a dataset in voice recognition systems is the main part of the overall work. According to scientists working in this field, the reliability and accuracy of intelligent systems depends more on the size and quality of the generated dataset than on the architecture of the built model.

A number of factors and tools are an important aspects of recording speech datasets. For example, some equipment used as technical means, Audio data recording tool, Recording format, Speech dataset size and recording technique can be included in this category.

Each person's speech fragments were collected to form this speech dataset. The following requirements were imposed on him:

1. file format - *.wav;
2. number of channels - 1 (mono);
3. sampling rate - 16 kHz;
4. memory space for each sample - 16 bits;
5. recording conditions - recorded in normal (home, street and other places) conditions.

2. Preprocessing of the speech signal. Voice Activity Detection (VAD) is used in audio signal processing and speech recognition systems to determine the presence or absence of human speech. VAD is usually the first step in cleaning a speech stream and usually used to remove any non-speech segments (including silence, cross-talk, laughter, ambient noise, etc.) and should only be submitted to subsequent processing steps. saves the speech. VAD plays a crucial role in various applications including voice assistants, automatic speech recognition (ASR), etc.[4]

x for the input signal, it is necessary to determine whether it is speech or non-speech. In this case, the VAD algorithm can be expressed as a function:

$$\bar{y} = \begin{cases} 0, & x - \text{нүтк эмас} \\ 1, & x - \text{нүтк} \end{cases}$$

INFORMATION AND WEB TECHNOLOGIES

The probability that speech is present is the probability that x is speech. A possible definition of VAD is:

$$VAD(x) = \begin{cases} 0, & S(x) < \theta \\ 1, & S(x) \geq \theta \end{cases}$$

here θ - scalar threshold.

Energy-based VAD [2]. This method assumes that frames without speech have less energy than frames with speech. A speech signal is not a stationary signal. Most notably, a person sometimes speaks with energy and sometimes does not. It is then clear that signal energy can be used as an indicator of speech presence.

3. Forming a set of characters from the speech signal. The effectiveness of recognition systems depends on how the character set is chosen. The better the initial character phase is chosen, the higher the recognition quality. The problem of recognizing a person based on human voice also begins with the formation of a set of symbols from the speech signal. Nowadays, there are many ways of forming speech signal symbols. For example, LPCC, PLP, MSFB, MFCC, etc[3].

The MFCC method of forming a set of characters, which is widely used in voice recognition, was used. MFCC includes:

3.1. An audio signal that can be used to divide a speech signal into frames is considered as an example of 1 second recorded at a frequency of 16 kHz. A typical 25 ms (millisecond) audio signal consists of 400 samples.[8]

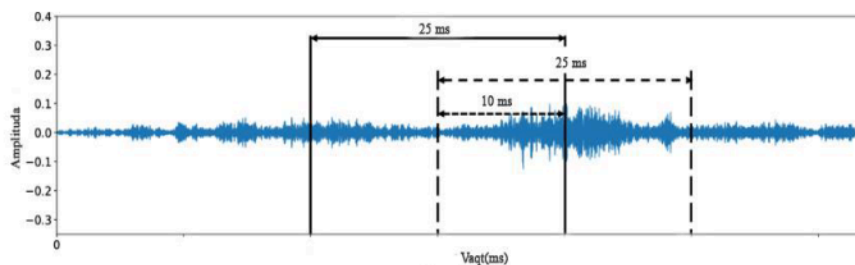


Figure 1

Dividing the speech signal into frames

3.2 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.

INFORMATION AND WEB TECHNOLOGIES

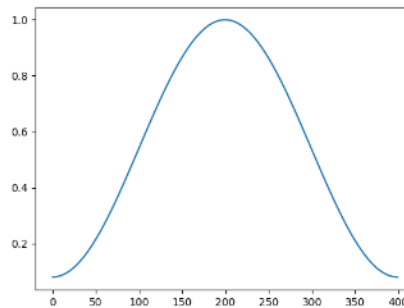


Figure 2
Hamming window

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

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

3.3 Discrete Fourier transform (DFA) describes what frequencies are present in the signal, but does not localize them at that 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. The short-term Fourier transform for a speech signal is performed by the expression (2):

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

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

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

3.4 Triangular filter sets can be used from 20 to 40 (26 in standard). In this case, to calculate the energies of the filter bank, each filter bank is multiplied by the power

INFORMATION AND WEB TECHNOLOGIES

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:

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

3.5 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:

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

3.6 In this step, the reverse transformation is done for the output from the previous step. After the MFCC method applies the discrete cosine transform, 13 coefficients of the signal are generated.

$$C_k = \alpha_k \sum_{i=1}^M X_i \cos \frac{\pi k(2i+1)}{2M}, k = 0, 1, 2, \dots, K \quad (6)$$

4. Gaussian Mixture Model (GMM - Gaussian Mixture Model) is a statistical model widely used in pattern recognition. This is an approach to identify a person based on their voice that provides a probabilistic model of the speaker's voice. An important aspect of any accent modeling is to collect and find the weights of each accent and mixture mean vector from the training speech. The parameters of the Gaussian mixture model are estimated using the maximum likelihood - Gaussian mixture model. The probability density function of the shapes is approximate. The GMM probability density function can be represented by the covariance matrix (Σ_i), the mathematical expectation (μ_i) and a set of mixture weight parameters (K_i). The multivariate Gaussian mixture model is given by equation (7) [7].

INFORMATION AND WEB TECHNOLOGIES

$$K_i = \frac{1}{\sqrt{(2\pi)^D/2|\Sigma_i|^{1/2}}} \exp\left[-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right] \quad (7)$$

The proposed algorithm consists of classification methods after feature extraction. The analog signal is first converted into digital form. Features of Uzbek speech signals are calculated using MFCC. Then the resulting feature is classified using a Gaussian mixture model. The final result is calculated using the maximum logarithmic likelihood function. The speaker detection algorithm using the MFCC-GMM method is presented below.[5]

The MFCC-GMM method for recognizing a person based on human voice consists of the following steps:

Formation set of characters: the speech signal is pre-processed and divided into frames. Each frame is then converted to MFCC coefficients. The number of coefficients is usually from 12 to 20.

Feature Normalization: MFCC coefficients are normalized to have zero mean and unity variance. This step helps to improve the accuracy of the voice recognition system.

MFCC training: The GMM is trained on the normalized MFCC coefficients for each individual. The number of components in a GMM can vary depending on the size of the training dataset.

Speaker Modeling: 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.[6]

Voice recognition: to recognize a person from a test speech signal, the MFCC coefficients of the test signal are first extracted and normalized. Then, the probability of the test signal belonging to the GMM of each candidate is calculated. The speaker with the highest probability is defined as the speaker of the test signal.

Conclusion. This article provides information on the stages of identity recognition, starting with the formation of a speech dataset, pre-processing, formation set characters, and the use of a Gaussian mixture model. Speech signal processing technology is explained, and the problems of variability, insufficient information, and background noise are identified as difficulties in acquiring identity recognition systems.

References:

- [1] Маматов Н.С., Ю.Ш.Юлдошев, Ш.Ш.Абдуллаев, А.Н.Самижонов, Х.Т.Дусанов
Нутқ сигнали белгиларини шакллантириш "Ёнфин-портлаш хавфсизлиги"

INFORMATION AND WEB TECHNOLOGIES

- илмий-амалий электрон журнал, № 1(8) 2022
- [2] Маматов Н.С., Нуримов П.Б., Самижонов А.Н. Нутқ сигналларида овоз фаоллигини аниқлаш алгоритмлари «Ахборот коммуникация технологиялари ва дастурий таъминот яратишда инновацион ғоялар» Республика илмий-техник конференцияси 17-18 май 2021 йил
 - [3] S. Dhingra, G. Nijhawan and P. Pandit, Isolated Speech Recognition using MFCC and DTW, International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(2), 2013.
 - [4] 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.
 - [5] A. Larcher, K. A. Lee, B. Ma, and H. Li. Text-dependent speaker verification: Classifiers, databases and rsr2015. Speech Communication, 60:56-77, 2014.
 - [6] 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
 - [7] Suchitha TR, Bindu AT. Feature extraction using MFCC and classification using GMM. Int J Sci Res Dev (IJSRD) 2015;3(5): 1278-83.
 - [8] 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.