

Stages of Development of Speech Signal Processing, Problems and Algorithms

**Berdanov Ulugbek Abdumurodovich¹, Azamov Temur Narzullayevich¹,
Makhmudova Shakhzoda Yorkinovna²**

¹ Department of introduction ICT and digitization, Innovative Development Agency, Tashkent, Uzbekistan

² Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

*Corresponding author: u.berdanov@ilmiy.uz

Abstract: This article develops mathematical methods of learning and decision-making in speech recognition, when extracting informative properties from speech signals, intelligent algorithms based on a data set consisting of properties. Effective ways to isolate features from speech signals are explained in sequential steps. These sequences are reflected in the database of several solutions (HMM, CNN and DTV) based on their comparative analysis, and the results are presented in the form of a table. In the experiments, sounds, syllables and words of speech were evaluated based on familiar intelligent algorithms, and a high result was achieved using the CNN mode.

Keywords: Hidden Markov models, convolutional neural networks, Dynamic Time Wrapping, artificial intelligence algorithms, informative properties.

Received 11 September 2024

Accepted 17 October 2024

Published 4 December 2024

1. Introduction

The importance of processing and analyzing speech signals in the Uzbek language is growing in today's world. Speech recognition technologies are commonly utilized in different fields, especially in domains like virtual assistants, voice control systems, text-to-voice conversion, etc. To excel in these domains, comprehending the intricate structure of the speech signal and devising methods for its analysis is crucial.

The focus of this study is on utilizing artificial intelligence algorithms to process and analyze speech signals in the Uzbek language. The primary aim of the research is to identify essential parameters in speech signals and then utilize intelligent algorithms to enhance the decision-making process [1].

The research examines commonly utilized algorithms in artificial intelligence like dynamic programming, implicit Markov models, and neural networks. A study is conducted on their abilities and restrictions in handling speech signals, and a broad model is suggested. Experiments are carried out using the speech signal of the Uzbek language to assess the performance of this model. The examination of the findings provides new opportunities for structuring speech signal parameters, enhancing the effectiveness of smart algorithms, and

advancing Uzbek language speech recognition technologies.

Contemporary systems employ three primary categories of techniques to compute speech cognition. Hidden Markov models are part of the initial group. They analyze if an opening speech can be seen as a series of specific phonemes and the likelihood of transitioning between them. Recognition involves searching for the most probable sequence of phonemes for a given input signal [3]. The next method relies on comparing to the standard. Each word in the dictionary is compiled in a certain manner as per the standard. In the process of recognition, the closest word to the standard input signal is chosen. The artificial neural networks form the foundation of the third set of techniques. The core of the techniques is discovering a key function from the input signal that can identify if it fits into a specific class. Artificial neural networks are constructed based on the organization of biological neural networks and effectively handle various tasks [2].

Methods utilizing neural network technology play a unique role in the field of speech recognition. The outcome of recognition in these techniques is influenced by how a specific neural network type is structured and operates. Neural networks consist of connected basic processors that carry out simple tasks.

2. Stages of development of speech signal processing, problems and algorithms

The main goal of the research work is to develop speech recognition algorithms based on a limited



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

vocabulary, the results of which are carried out to this day, processing speech signals. This, algorithms are selected that process speech signals based on algorithms and models developed by scientists. The results of the analysis show that there are several algorithms that process speech signals. Based on the results of the study, the process of processing speech signals can be carried out in three stages [4]. The general guide structure of the algorithms of these steps is presented Figure1.

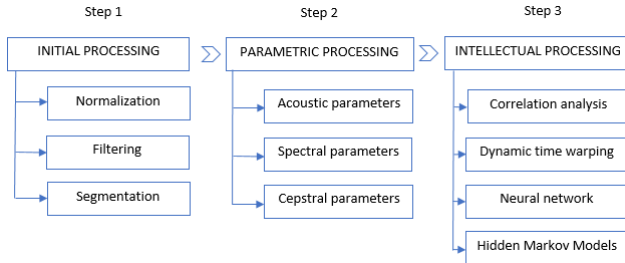


Fig. 1. The stages of processing speech signals.

The possibility of these stages can be used to get acquainted with the sounds, syllables and words of speech. Based on the task set based on the algorithms specified at these stages, the primary processing of the speech signal takes place at the first stage [5]. After that, the speech signal is divided into parameters and information about the speech signal is extracted using intelligent processing algorithms. At each stage, there are several algorithms for processing speech signals. Based on the problem statement, it is necessary to select algorithms. That is, based on a perfect analysis of the task, parameters are selected and the potential of intelligent algorithms is used. At these stages, there are several types of algorithms, of which one or more types can be used simultaneously [6].

The formation of speech signals is described as depending on the characteristics of the speaker. The algorithms of these stages can also be used in the process of processing speech signals. These stages form a set of several algorithms in their working mode. Each stage is organized based on the sequence of algorithm changes depending on the task statement. These steps are described below.

2.1. The stage of primary processing.

During the primary processing of speech signals, the speech signal is converted from analog to digital form. This task is performed by analog to digital converters. At the same time, discrete quantization and encoding are performed. This is often a speech signal converted to WAV file format, the sampling rate is quantized in the range of 8 kHz – 19.2 kHz, and the recording channel is selected mono. In real-time speech signal processing tasks, the speech signal is quantized at a sampling frequency of 22 kHz. These parameters are standard when selecting a speech signal. The reason for this is the voice signal [7]. The WAV file format is an expression of a speech signal in an uncompressed state, and the sampling rate of 22 kHz is the standard sampling rate chosen by scientists when processing speech signals. After the speech signal is presented in digital form, it is processed by the following initial processing algorithms.

Normalization. The digital expression of the speech signal is adjusted to the standard margin at different intervals. Such normalization intervals for speech signals [0:1], [-1:1], [-3:3], [-128:128], [-256:256], [-512:512] there are intervals. These intervals ensure that the values of the speech signal fall within one standard interval. When processing speech signals, the interval [-1:1] is often chosen.

Filtering. Speech signal filtering is used to clear the speech signal from interference and eliminate external influences. To do this, you can use filters with finite pulse characteristics and filtering algorithms with infinite pulse characteristics. After the filtering process, the silence states in the speech signal are lost [8].

Segmentation. Segmentation is the process of dividing the values of a speech signal into frames of a given size. Currently, in modern speech signal processing systems, speech signal frames include 64, 128, 256, 512, 1024, 2048 accepts the coefficients of the speech signal. When processing speech signals, one of these segmentation intervals is selected, depending on the formulation of the question. As a result, the speech signal is divided into frames.

The process of mirroring a segmented speech signal performs the task of approximating the initial and final values of the frame being determined to zero. The main reason for this is that with parametric processing, spectral values do not tend to infinity [9]. There are such types of mirroring algorithms as rectangular mirror, Hann mirror, Hemming mirror, Blackman mirror, Kaiser mirror. The Hemming mirror for speech signals is one of the most commonly used types of mirrors by scientists.

2.2. The stage of parametric processing.

At the stage of parametric processing, acoustic, spectral, formant frequency, and cepstral parameters of the speech signal are determined. In the general account, the choice of a specific parameter is selected depending on the formulation of the question. Each parameter has its own mathematical algorithm. At the same time, the completeness of the parameters is checked.

The spectral density of speech signals is calculated using mirror functions. Here the mirror size changes without stripes. This approach characterizes the activity of the human auditory system. The first MFCC algorithm was S.B. Davis and P. Mermelstein proved that speech can be used to solve the dating problem. The MFCC algorithm is calculated in the following sequence.

A discrete cosine transforms (DCT) and discrete fourier transform (DFT) more information about these converters is described in this source [22].

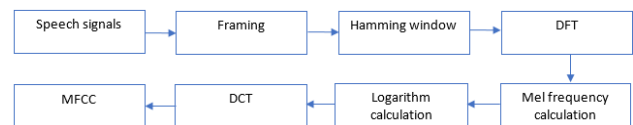


Fig. 2. Algorithm for obtaining MFCC values

The algorithm for obtaining MFCC values consists of several successive stages, the algorithm for calculating the Mel-frequency is shown above.

2.3. The stage of intelligent processing.

The intelligent processing stage is an important stage. At this stage, the grouped set of parameters is processed based on decision-making algorithms. Dynamic programming algorithms, Markov chains, and neural network algorithms can be attributed to this group of stage algorithms. The results of the intelligent processing stage are summarized [10]. That is, the results of the processed frames are analyzed and a final conclusion is made about the state of the speech signal.

Parametric characteristics of speech are important when processing speech signals. That is, as a result of processing the parameters of human speech based on intelligent algorithms, a conclusion is made about the speech signal. The state of the conclusion depends on the formulation of the question. We will fully familiarize ourselves with the general theory of intelligent algorithms below.

3. Application of artificial intelligence algorithms in speech recognition

Uzbek language based on the analysis of the above algorithms for processing speech signals, speech signals were divided into parameters. Based on these parameters, it is necessary to determine the speech signal. The decision-making process is carried out by intelligent algorithms. Artificial intelligence algorithms come in several types, depending on the tasks they are used for. Their algorithms, which are currently widely used in speech signals, include dynamic programming algorithms, implicit Markov models, and neural network algorithms. These artificial intelligence algorithms are used to make decisions based on the parameters of speech signals. The main purpose of the research work is to organize a parametric set of speech signals and propose a generalized model that allows making decisions based on these parameters as a result of the effective use of intelligent algorithms and evaluating its effectiveness. As a result of the experiments, a set of parameters was formed on which spectral analysis processing speech signals was based [11]. As a result of selecting the necessary parameters from this set of parameters, a general structure based on dynamic programming algorithms, implicit Markov models and neural networks was created. This structure is shown in Figure 2. Based on this question, the use of all parameters and decision-making algorithms in the structure is not required. That is, based on the task, parameters are selected and the possibility of decision-making algorithms is used.

More about fast-Fourier transform (FFT) is described in this source [22].

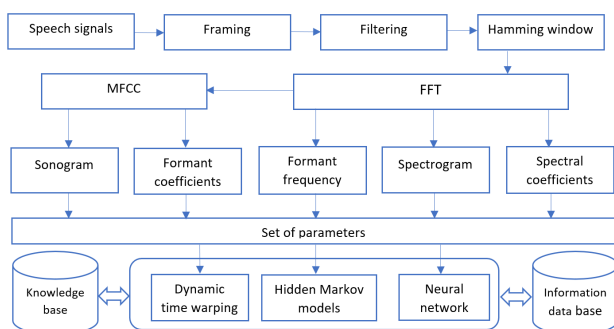


Fig. 3. Speech processing as a general structure.

A user familiar with the speech of the Uzbek language, this general structure was the first to apply the process of shredding a recorded speech signal. Using the above algorithm of the grinding filter [12]. After that, the smoothed signal is segmented. Segmentation is used to divide the speech signal into frames. After that, the task of extracting parameters from each frame is displayed. A discrete Fourier transform is performed from spectral algorithms with parametric decomposition. Using spectral algorithms, spectral distortions, formant frequencies and spectrograms of the speech signal can be returned as a signal parameter. In addition, spectral algorithms can help to obtain MFCC parameters and return as signal parameters in the MFCC algorithm utility. These parameters are used for general parameters.

The solution of the problem is determined from the selected parameters with the addition of the basics of the artificial intelligence algorithm. Artificial intelligence algorithms dynamic programming algorithm, hidden Markov models and neural network algorithm were analyzed during the study. This method is able to solve a number of practical issues that call into question the processing of human speech. An example is the definition of personality based on human speech, the content of human speech and use in the management system in order.

3.1. Speech recognition solutions based on Dynamic Programming

DTW (Dynamic Time Warping) - this algorithm is considered a kind of dynamic programming algorithms. It is used to compare the algorithm with the parameters of speech signals. There are several types of algorithms for comparing the parameters of speech signals, each of which requires that the parameters be of the same length during the comparison process. Unlike other comparison algorithms, it allows you to compare two vectors of different lengths. An effective aspect of the DTW algorithm is that this algorithm itself is able to determine the difference between vectors of different lengths. Therefore, this algorithm is widely used in the processing of speech signals.

When using this algorithm to recognize speech sounds, syllables and words, MFCC signal parameters were obtained. It is known that the MFCC values in one frame of a given speech signal are 12. When organizing an ad database for a given speech M frame, a sum of MFCC winds of $12 \times m$ is generated. As a result, a reference database is formed for a given speech signal [13]. The main problem is comparing the incoming anonymous voice signal with the database of ads. The number of MFCC values of the incoming speech signal frames is 12, and the number of frames is N. The result is an unknown speech signal, which is $12 \times n$, and the signal, which is $12 \times m$, requires comparison with some data. The DTW algorithm is effective when comparing signals with different resolutions. The DTW algorithm builds the path of the least action from the sequence when solving such problems. There are several types of computing speech signal information using the DTW algorithm:

- symmetrical, classic (Figure 4, a);

- asymmetric (Figure 4, b);
- fast (Figure 4, b).

Each variant of this type of calculation of the DTW algorithm has a different speed and different effect, depending on the parameters of the speech signals.

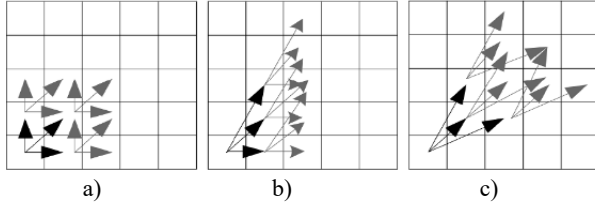


Fig. 4. Types of DTW algorithm.

The DTW algorithm, based on MFCC parameters, compares an unknown speech signal of $12 \times N$ with a standard data of $12 \times m$ with some of the unknown speech signal $X(x_1, x_2, \dots, x_n)$ and standard data some are absorbed into $E(e_1, e_2, \dots, e_m)$. As a result, vectors of length of two different sequences are formed. This is calculated using the following formula.

This is followed by the calculation of local deviations between the two sequences. Local deviation is the distance between the i - frame of an unknown speech signal and the j - frames in the standard database. The most common method of calculating local deviations is the method of calculating the absolute deviation between the values of the two elements, that is, the Euclidean distance. This is calculated using the following formula [14].

$$D_l(x_i, e_j) = \|x_i - e_j\| = \sqrt{\sum_{k=1}^N (x_{ik} - e_{jk})^2} \quad (1)$$

In addition, there are several types of calculation of local deviations. They are calculated using the formula [14]:

- Manhattan distance;

$$D_l(x_i, e_j) = \|x_i - e_j\| = \sum_{k=1}^N |x_{ik} - e_{jk}| \quad (2)$$

- Minkowski distance;

$$D_l(x_i, e_j) = \|x_i - e_j\| = \sqrt[m]{\sum_{k=1}^N (x_{ik} - e_{jk})^m} \quad (3)$$

- Chebyshev distance;

$$D_l(x_i, e_j) = \|x_i - e_j\| = \max_{k \in 1, N} |x_{ik} - e_{jk}| \quad (4)$$

As a result, a deviation matrix Ω_{nm} is constructed, consisting of n rows and m columns. Using the deviation matrix Ω_{nm} , a path W of length K is constructed. This path indicates the correspondence between the vectors X and E . The element k of the path W is defined as $w_k = (i, j)$, $D(w_k)$ with $D_l(x_i, e_j)$ is considered as an equivalent formula. The length of the K -path is determined by the following condition.

$$\min(m, n) \leq K < m + n - 1 \quad (5)$$

The W path is determined as a result of the following three conditions. There are many ways to fulfill these three conditions, but the optimal definition of this path is carried out using the following formula [15]:

$$DTW(X, E) = \min \left\{ \frac{1}{K} \sqrt{\sum_{k=1}^K D(w_k)} \right\} \quad (6)$$

The algorithm used to recognize sounds, syllables and words of the Uzbek language by the DTW method is shown in Figure 5 below.

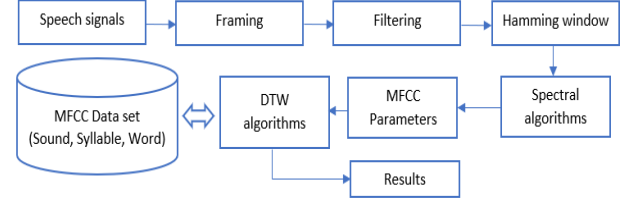


Fig. 5. Speech recognition algorithm based on the DTW algorithm DTW.

In this structure of the speech signal processor, the MFCC parameters of sounds, syllables, and words are stored in the announcement database. During processing of the incoming speech signal, the MFCC parameters of the signal are compared with the MFCC parameters in the ad database using the DTW algorithm. As a result, the process of recognizing the incoming speech signal takes place. When creating an ad database using the DTW algorithm, it is recommended to group the database. Therefore, the algorithm creates a separate database of sounds, syllables and words. This is due to a reduction in the calculation time during the comparison process. In the issue of speech recognition based on the DTW algorithm-based speech recognition method, the results of the ban for a single speaker gave a minimum accuracy of 70% to 99%. These results are presented in Table 1 below in the context of speech sounds, grouped into groups.

Table 1. Results of sound recognition using the DTW algorithm.

Classification of speech sounds (Uzbek language)		Result based on DTW algorithm (%)	
		Min	Max
Vowel sounds	A, O, O', U E, I	92	99
Crawling noises	S, Z, SH, J, CH	76	82
Sonar sounds	R, L, M, N	83	92
Tongue back sounds	K, G, X	97	96
Deep tongue sounds	Q, G', H	73	86
Sonorous sound	B, V, Z, D, J, G, G', Y, P, F	82	94
Muted sounds	S, T, SH, CH, K, Q, X, H	77	95

Like any decision-making algorithm, DTW has a number of disadvantages. In this algorithm, the number of simple calculations increases depending on the size of the computational matrix [16]. This creates computational complexity. The large number of columns I and rows J of the deviation matrix generates the following disadvantages:

- the complexity of the process of studying large digital matrices and searching in them;
- calculating a large number of deviations takes a long time;
- to increase small calculations in the matrix.

Due to the above disadvantages, the DTW algorithm has generally proven itself well in determining sounds and syllables. In addition to the slippery and deep language, the error in the background sounds was found unsatisfactory. Thus, the DTW algorithm can serve well mainly for correcting learning and speech disorders.

3.2. The Markov model speech recognition process

Based on the implicit Markov model, NFCC parameters are taken as an input parameter when solving the problem of familiarity of sounds, syllables and words of speech. To build an implicit Markov model using these parameters, it is necessary to select a sequence of states.

Experiments show that using the hidden Markov model, it is necessary to determine the number of hidden cases when recognizing sounds, syllables and words of speech. The number of hidden states depends on mass spectrometry. The hidden Markov model is the most effective solution for recognizing speech sounds, syllables and words, it is necessary to select the number of hidden cases. There are 29 sounds in the Uzbek language, each of which allows you to express one hidden state in a hidden Markov model.

The solution to the problem is to determine the highest probability of the existence of each frame of a given speech signal (MFCC frame parameters) in these 29 hidden states. This decision determines which sound the given frame of the speech signal belongs to. As a result of sequential frame processing, the sequence of sounds in the speech signal is determined. Based on the identified sequence of sounds, it is possible to artificially determine syllables and words in a speech signal [17]. To do this, an implicit Markov model of a sequence consisting of n -states is first constructed. In solving the problem, $N=29$ is considered.

The implicit Markov model is based on modeling processes consisting of implicit N -states. The transition between states is carried out in the time interval t according to Probst's law, and the transition probability matrix is described by ANN. A schematic representation of the transition scheme between the hidden states of the hidden Markov model is shown in Figure 6.

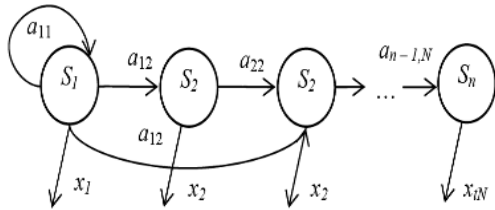


Fig. 6. Hidden cases of the Hidden Markov model.

The presence of the model in a given state i corresponds to a certain stationarity of the observed signal in a given time interval. A simple physical interpretation of the hidden Markov model arises: sometimes it is considered a process that gradually changes its properties.

Multidimensional probability distribution function $f_j(x)$ at the next transition to a new state i , at time t , an output vector is formed, called the parametric vector x_t . As a

result of the hidden Markov model, the length the parameter is a sequence of vectors $[x_1, x_2, \dots, x_t]$ [16].

The advantage of the Hidden Markov model over other intellectual algorithms is that it has different lengths and the ability to work with sequences and parameters.

$$f_j(x) = \sum_{i=1}^M w_i p_i(x) \quad (7)$$

where M is the number of components in the model; w_i is the weight of the component; $p_i(x)$ is the normal distribution of the event measure D . The $p_i(x)$ function is defined by the expression:

$$p_i(x) = \frac{1}{2\pi^{D/2} |\sigma_i|^2} + \exp \left\{ -\frac{1}{2(x-\mu_i)^T \sigma_i^{-1} (x-\mu_i)} \right\} \quad (8)$$

where D is the size of the vector; μ_i is the expectation vector; σ_i is the covariance matrix.

Hidden Markov models the analysis process, as in any other expert system, when processing speech information is carried out in two stages.

1) Training - concealment the Markov model of the Baum-Welch algorithm is used in the training of speech signals. Scientists have proved that this algorithm is one of the optimal algorithms for learning speech signals

2) Definition - a sequence of vectors for determining the probabilities observed in $S_i [x_1, x_2, \dots, x_T]$ probability maximization is determined using the Viterbi algorithm. A certain probability determines the outcome. The following is a brief description of the above algorithms [19].

The process of learning a hidden Markov model consists in determining the following parameters using a set of training samples:

- A matrix of probabilities of transition between A_{NN} states is constructed;
- For each case, using the Gauss algorithm, the parameters in the A_{NN} matrix are determined: mathematical expectation, covariance matrix and weight.

When constructing a probability matrix and determining parameters using the Gauss algorithm, the forward-backward algorithm is used to solve these problems using two iterative forward-backward algorithms and the Baum-Welch algorithm. In the forward-backward algorithm, $a_j(t)$ is the correct probability distribution, and $\beta_j(t)$ is the inverse probability distribution. The solution $a_j(t)$ with certain values consists in finding the state j of the hidden Markov model at time $t [x_1, x_2, \dots, x_T]$ gives a probability distribution [17].

$$a_j(t) = P(x_1, x_2, \dots, x_t | \text{state}_t = j) \quad (9)$$

where $a_j(t)$ i $a_j(t-1)$ the quantities are related to an iterative expression:

$$a_j(t) = \left[\sum_{i=2}^{N-1} a_i(t-1) A_{i,j} \right] f_j(x_t) \quad (10)$$

where A_{ij} is the probability of transition from state i to state j , $f_j(x_t)$ is the probability of observing vector x in state j .

In the implicit Markov model, the inverse function $\beta_j(t)$ is the time t of the state j in $[x_{t+1}, x_{t+2}, \dots, x_T]$ allows us to determine convergence in the sequence of observations and is determined by the formula [17]:

$$\beta_j(t) = P(x_1, x_2 \dots x_t | \text{state}_t = j) \quad (11)$$

where $\beta_j(t)$ va $\beta_j(t+1)$ the amounts are related in a similar way:

The values of $a_j(t)$ i $\beta_j(t)$ are equal to $[x_1, x_2, \dots, x_t]$ makes it possible to determine the state of the hidden Markov model at time t when observing the sequence and is calculated using the following formula.

$$L_j(t) = \frac{1}{p} a_j(t) \beta_j(t) \quad (12)$$

At the next stage of training, the Baum-Welch recovery algorithm allows you to overestimate the model parameters using the above expressions. Let R be a training sample, then the probability of transition from state i to state j is defined as [18]:

$$\tilde{A}_{i,j} = \frac{\sum_{r=1}^R \frac{1}{p_r} \sum_{t=1}^{T_r-1} a_i^r(t) A_{ij} f_j(x_{t+1}^r) \beta_j^r(t+1)}{\sum_{r=1}^R \frac{1}{p_r} \sum_{t=1}^{T_r} a_i^r(t) \beta_j^r(t)} \quad (13)$$

Each state J and each component M of the Gaussian mixture, the mathematical expectation, the covariance matrix and the weight are determined by the following expressions:

$$\hat{\mu}_{j,m} = \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t) x_t^r}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t)} \quad (14)$$

$$\hat{\sigma}_{j,m} = \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t) x_t^r x_t^r - \hat{\mu}_{jm}^r (\sum_{t=1}^{T_r} L_{jm}^r(t) x_t^r - \hat{\mu}_{jm}^r)^T}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t)}$$

$$\hat{w}_{j,m} = \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t)}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jm}^r(t)}$$

For high-quality training of the hidden Markov model, many samples of speech signal frames (MFCC parameters) in the amount of several tens or even hundreds of copies are required. It is also necessary to monitor the state of linear independence of the training samples, otherwise the covariance matrix will be violated in the learning process, which will lead to a complete failure of the model. To avoid this condition, it is advisable to constantly monitor the learning process[19]. When processing speech signals with hidden Markov models developed on the basis of these mathematical algorithms, the following requirements must be met:

- The choice of kepsstral coefficients (MFCC) as parametric vectors in speech signal recognition.
- Development of an algorithm for normalization of parametric vectors;
- Choosing the number of m states of the N model and the components of the Gaussian algorithm.

It should be noted that the implicit Markov model does not have a universal algorithm for determining the above parameters, and each case has a high recognition accuracy depending on the problem being solved, many experiments may be required to achieve the result.

The MFCC parameters of each detected frame are set to ready for processing in an implicit Markov model, and processing is performed based on the underlying structure (Figure 7).

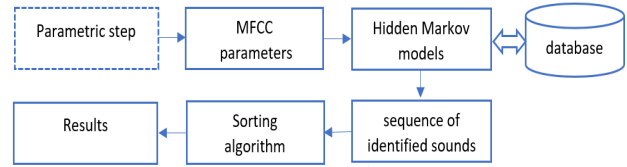


Fig. 7. An algorithm for recognizing speech information based on the Hidden Markov model.

In the structure, the MFCC parameters of each frame of the speech signal are entered as parameters into the Markov model. To increase the accuracy in this structure, the frames were worn 60% apart. This, of course, leads to an increase in the number of processing frames in the speech signal and an increase in processing time. It is proved that such sorting of frames gives high accuracy of experimental results [20]. The Markov algorithm performs the task of determining the correspondence of the obtained parameters to the parameters in the database and determines in which of the 29 specified states the MFCC parameters of the frame have the highest probability. As a result of algorithmic processing of the parameters of the MFCC, a certain sequence of sounds is formed (in the form of text).

These sound sequences are summarized using sorting algorithms. There are certain types of sounds in this that are discarded if their sequence is less than 15. If the number of frames exceeds 15, this set of sound sequences is returned as a result.

Of the 20 two-syllable words selected for the study, 16 were found clearly, and 4 were found indeterminately. There is one mistake in each of the vaguely defined words.

The MFCC parameters of each detected frame are set to ready for processing in an implicit Markov model, and processing is performed based on the underlying structure (Fig.7).

4. Speech recognition based on CNN

Convolutional Neural Networks (CNN) is also used in speech recognition systems. Convolutional neural networks are mainly designed for speech recognition and processing, and are also widely used in natural language processing and other fields. CNN networks allow you to maintain and effectively analyze the spatial structure of the input data.

- Convolutional layer: performs convolutional operations with input data. This layer analyzes the input data using small filters (cores).
- Activation Function: After each convolution, an activation function (for example, Relu) is applied, which introduces zero linearity.
- Pool layer: After convolution, layers of the maximum pool or average pool are usually applied, which reduces the spatial resolution and reduces the computational complexity of the model.
- Fully Connected (FC) layer: One or more fully connected layers are added after convolution, these layers are used to classify the results.
- Dropout: The method of accidental disconnection (dropout) of some neurons of the network is used to prevent overvoltage.

The principle of operation of a convolutional neural network is implemented in the following sequence.

1. Input data: CNN is usually designed for image processing, but can be processed as speech signal spectrograms or in other forms.
2. Convolutional layers: Each convolutional layer defines local properties by maintaining small filters for the input data. Filters slide over the image, creating dot multiplication and sums at each location.
3. Activation functions: After each convolution, an activation function (for example, ReLU) is applied, which adds zero linearity to the property map.
4. Pool Layers: Pool layers reduce spatial resolution and extract the most important features. For example, the max-pooling layer takes the highest value in a small area.
5. Fully connected layers: One or more fully connected layers are applied at the end, and these layers act as classifications.
6. Output layer: The last layer usually supports the SoftMax function, providing the classification result.

Convolutional neural networks (CNNs) are one of the main components of modern artificial intelligence and machine learning. They are highly accurate in performing many complex tasks such as image processing and speech recognition. CNN's work has a wide range of applications due to the effective analysis of spatial objects [21].

A neural network with two architectures is built in a speech recognition algorithm based on a convolutional neural network. The first architecture is based on the MFCC parameter of a speech signal with 12 input vectors and the FFT parameters of a speech signal with 128 input vectors. The neural network principle of the algorithm checks the result based on the MFCC parameter of the first speech signal. If the result meets the specified requirement i.e., above 90%, the result will be printed. Otherwise, it performs the computational process of a neural network based on the parameters of the FFT of the speech signal. Here the result is printed if it is 35% higher. Otherwise, the information in the form of "file error" will be printed.

These algorithms can be used not only to recognize children's speech pronunciation, but also to solve problems of recognizing Uzbek speech sounds, syllables and word content. The wide range of tasks solved by neural networks today limits the possibility of developing a functional, powerful, universal network with various algorithms.

Because each neural network uses different resources that do not match in the problems being solved. Therefore, when designing neural networks, the structure is developed depending on the type of problem being solved, as well as the scope of application. When developing this structure, the following approach was created.

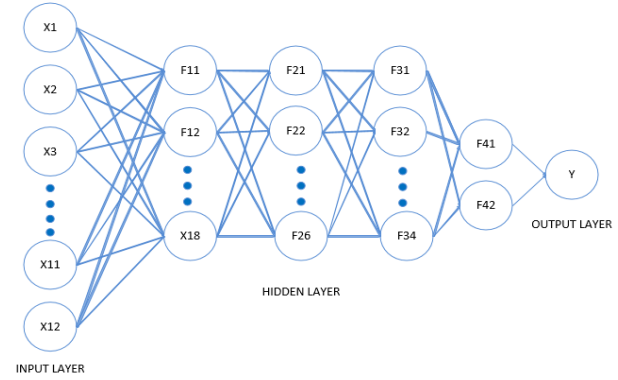


Fig. 8. The general structure of the neural network.

Speech word recognition used a multilayer perceptron neural network with a classical multilayer structure with the above-mentioned sequential connections and activation functions of a sigmoid neuron. The input layer of the network consists of 12 elements (MFCC parameters) $X_1, X_2, \dots, X_{12}$ has four hidden layers and consists of one output layer.

The first hidden layer is $F_{1n}=8$ elements, the second hidden layer is $F_{2n}=6$ elements, the third hidden layer is $F_{3n}=4$ elements, the fourth hidden layer is $F_{4n}=2$ elements. These dimensions are determined based on experiments, and the size Y of the output layer always consists of 1 element. The output layer, equal to 1, is based on the fact that the child pronounces one word during the speech.

In the structure of a neural network, each synapse has its own synaptic connection or its weight. The current state of a neuron is determined by a randomly found sum of values at its input [19].

$$F_{i,N} = \sum_{i=1}^n x_i w_i \quad (15)$$

At the output of the neuron there is a function for its state:

$$Y = f(F_{i,N}) \quad (16)$$

One of the most common of these functions is a nonlinear function, that is, a logistic function in the form of $F(X)$ or a sigmoid function

$$f(x) = \frac{1}{1+e^{-\alpha x}} \quad (17)$$

Reduction in the indicator α , the sigmoid moves to a position as if it is stretched, at $\alpha = 0$ a horizontal line is formed, all points of which are equal to 0.5. and with an increase in the value of the pointer α , the function $X = 0$ at point T makes it possible to convert one into a function of a jump variable.

This activation function has the following advantages:

- this function is bipolar, and, accordingly, all signals in the neural network will also be bipolar, which will reduce the number of repetitions of the learning algorithm of this network;
- the hyperbolic tangent function is quickly calculated on a computer, which allows you to speed up the work of neural networks both in speech recognition mode and in learning mode.

The mathematical model of the neural network used in this paragraph is presented as follows [20]:

$$Y = f\left[\sum_{N=1}^2 F_{4,N} f\left(\sum_{N=1}^4 F_{3,N} f\left(\sum_{N=1}^6 F_{2,N} f\left(\sum_{N=1}^8 F_{1,N} X_1^{12}\right)\right)\right)\right] \quad (18)$$

where $F_{1,N}$, N are the neural cells of the hidden layer, X_1^{12} - are the MFCC coefficients, f is the activation function, y is the outgoing neural cell.

In the general structure of recognition of children's speech pronunciation, we can see that the spectra obtained on the basis of the Fourier transform algorithm are sent directly to the neural network. To do this, the structure of the neural network is based on the model below.

$$Y = f\left[\sum_{N=1}^2 F_{4,N} f\left(\sum_{N=1}^4 F_{3,N} f\left(\sum_{N=1}^6 F_{2,N} f\left(\sum_{N=1}^8 F_{1,N} X_1^{128}\right)\right)\right)\right] \quad (19)$$

where $F_{1,N}$ are the neural cells of the hidden layer, X_1^{128} - are the spectral coefficients, f is the activation function, Y is the outgoing neural cell.

In this formula, we see that the number of spectral coefficients is 128. This is due to the fact that the speech signal contained 256 values when divided into frames. When spectral values are obtained by the Fourier transform, 128 real frequencies and 128 abstract frequencies are obtained from each frame. When the spectral coefficients were aimed at the neural network, only the actual frequencies were obtained [21].

We can see the decision-making process in their neural network by taking the MFCC coefficients from each frame of the speech signal. Spectral values can also be determined by neural networks on this basis.

Based on the sequence of these algorithms, it is possible to determine how correctly the speech signal is recognized in the speech development of children (Table 2)

Table 2. the result of comparing artificial intelligence algorithms.

№	Speech signal type	DTW	HMM	NN
1.	Sound	73	77	87
2.	Syllable	74	81	95
3.	Word	82	85	91

The sequence of algorithms consists in making decisions on neural networks based on the distribution of the speech signal by parameters (MFCC and spectral parameters). At the same time, the probability that the speech signal will be correctly pronounced is 90% higher. This means that the sequence of algorithms for processing Uzbek speech characterizes the correct organization.

5. The effectiveness of speech recognition algorithms based on intelligent algorithms

It can be seen from the results of the analysis that the use of the Markov model is considered productive when recognizing words in Uzbek speech. The principle of word recognition based on parameterization of two- and three-syllable words of the Uzbek language is implemented. The work on the parameters of the speech signal was carried out in the Markov model as destructive values. When determining two-syllable words, out of 20 selected words,

4 were found vague, and the remaining 16 were clear. There is one mistake in each of the four words found ambiguously. No error was found in the three-syllable words. The degree of determination in processing Uzbek words based on the hidden Markov model gave an accuracy of 80-85%. Thus, the proposed model of processing Uzbek words can be assessed as effective. But the calculation takes a long time, and constant monitoring is necessary when learning speech information. Otherwise, this will lead to incorrect formation of the covariance matrix. This characterizes a high error rate.

6. Conclusion

According to the results of the analysis, it can be seen that from the artificial intelligence algorithms used to recognize Uzbek speech, neural networks turned out to be effective based on experimental results. It is also efficient with computational time from other artificial intelligence algorithms with the simplicity of the computational process in neural networks.

The accuracy of comparative indicators of artificial intelligence algorithms conducted in the research work was estimated as a percentage. This is due to the fact that the main goal was to determine the type of analyzed artificial intelligence algorithms that provide high accuracy. The main purpose of comparing the Markov chain, the DTW algorithm and neural networks is to check the correspondence of pointers to complexity:

- accuracy (must be above 90%);
- real-time compatibility;
- occupied memory capacity;
- simplicity of the algorithm.

Experiments have shown that there are certain requirements for the learning and testing process when working with neural networks. Therefore, this algorithm is more efficient than other artificial intelligence algorithms. Speech signals it is advisable to use neural network algorithms in solving field problems.

Acknowledgment

This paper was carried out at the Tashkent University of Information Technologies named after Muhammad al-Khorazmi in BV-Atex-2018-134—"Analysis and creation of software for speech signal recognition algorithms in digital signal processors" and 24/19-F "In the rehabilitation of speech defects in preschool and school-aged children" thanks for using the results of scientific research projects "development of computer programs".

References

- [1] Abdullaeva, M., I. Khujayorov and M. Ochilov (2021). Formant set as a main parameter for recognizing vowels of the Uzbek language, International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 1-5.
- [2] M. M. Mahmudovich, A. M. Ilkhamovna and T. B. Shukhrat Ogli, (2022). Image Approach to Uzbek Speech Recognition, IEEE 22nd International Conference on Communication Technology (ICCT), Nanjing, China, 2022, pp. 1201-1206.
- [3] Musaev M., Abdullaeva M., Ochilov M. (2022). Advanced Feature Extraction "Method for Speaker Identification

- Using a Classification Algorithm". AIP Conference Proceedings, 2022, pp 256-265.
- [4] W. Chan (2016). Speech recognition with attention-based recurrent neural networks, US Patent 9,990,918, 2018.
- [5] W. Chan (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).— IEEE. 2016, pp. 4960—4964.
- [6] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig (2016) Achieving human parity in conversational speech recognition, Technical Report MSR-TR-2016 pp.71.
- [7] O. Abdel-Hamid (2014) Convolutional neural networks for speech recognition, IEEE/ACM Transactions on audio, speech, and language processing. Vol. 22, 2014, pp. 1533—1545.
- [8] Lama, P. (2010) Speech recognition with dynamic time warping using MATLAB, Project Report, CS 525, Spring 2010.
- [9] B. W. Gawali. (2010) Marathi isolated word recognition system using MFCC and DTW features, Proc. of Int. Conf. on Advances in Computer Science. Vol. 1, Citeseer. 2010, pp. 21—24.
- [10] Y. LeCun, Y. Bengio (1995). Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks Vol. 3361, no. 10, 1995.
- [11] P. Ahmadi and M. Joneidi (2014). A new method for voice activity detection based on sparse representation, 2014 7th International Congress on Image and Signal Processing, Dalian, 2014, pp. 878-882.
- [12] V. A. Volchenkov and V. V. Vityazev (2016) Development and testing of the voice activity detector based on use of special pilot signal, 5th Mediterranean Conference on Embedded Computing (MECO), Bar, 2016, pp. 108-111.
- [13] Z. Fan, Z. Bai, X. Zhang, S. Rahardja and J. Chen (2019). AUC Optimization for Deep Learning Based Voice Activity Detection, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp.
- [14] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio, (2018) Light gated recurrent units for speech recognition, IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, pp. 92–102.
- [15] D. Palaz, M. Magimai-Doss, and R. Collobert (2015). Convolutional neural networks-based continuous speech recognition using raw speech signal, in Proc. of ICASSP, April 2015.
- [16] S. M. Mon and H. M. Tun (2015). Speech-to-text conversion (STT) system using hidden Markov model (HMM), International Journal of Scientific & Technology Research, vol. 4, no. 6, Jun. 2015 pp.349-352.
- [17] Savchenko V. V. (2017). Study of the stationarity of random time series using the principle of the information-divergence minimum, Radiophysics and Quantum Electronics, vol. 60, no. 1, 2017, pp. 81—87.
- [18] Pahini A. Trivedi (2014). Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks, International Journal of Engineering Development and Research, Volume 2, Issue 4, 2014.
- [19] G. Harsha Vardhan, G. Hari Charan (2014). Artificial Intelligence & its Applications for Speech Recognition, International Journal of Science and Research (IJSR) Volume 3 Issue 8, August 2014.
- [20] Weng C., Yu D., Seltzer M. L., Droppo J. (2014). Single-channel mixed speech recognition using deep neural networks, IEEE ICASSP. 2014, pp. 5632-5636.
- [21] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition" - IEEE, Signal Processing Magazine, 2012.
- [22] Shaw Talebi. (2020). The Fast Fourier Transform (FFT)", Published in The Startup 7 min read Dec 4, 2020.