Application of KNN and SVM classification algorithms in artificial speech recognition by emotion detection in speech signals

Sultanmurat Nasirov
sultan250593@gamil.com
Tashkent University of Information Technologies named after Muhammad al-Khorezmi
Indira Azatovna Khodzamuratova
ORCID: 0000-0002-4680-5475
indira@nordicuniversity.org
Nordic International University

Abstract: *The creation of modern technologies for detecting emotional changes in individuals in speech signals and providing users with the necessary information for themselves, representing speech signals in digital form, filtering, extracting the necessary features, modelling and analysing them, recognizing a person's voice using intelligent algorithms and software for digital processing, creating voice-controlled devices, examining patients' speech disorders in medicine, classifying speech features, and recognizing a person in speech signals and separating them from artificial speech are urgent issues. To this end, several scientific research works are being carried out aimed at developing and improving methods for separating individual emotions in speech signals and artificial speech using artificial intelligence elements. This article develops artificial speech recognition based on the presence or absence of emotions in speech signals using speech signal features and spectrogram parameters, statistical modelling and classification of speech signals using machine learning algorithms, in particular k-NN and SVM, automatic detection of complex features in speech signals, and analysis and technical solutions for analysing spectrograms of speech data based on deep learning algorithms.*

Keywords: *identifying a person's emotions in speech signals, artificial speech, k-Nearest Neighbours, Support Vector Machine, MFCC*

1. INTRODUCTION

The issues of intelligent analysis of speech signals are relevant for many areas, and this process involves identifying features emanating from the human voice, studying them and using them for various purposes. There are several problems in the process of using speech signals for personal recognition, which are mainly associated with technological and practical limitations. Speech signals, on the one hand, are a complex source of information, reflecting the emotional state, physical characteristics, and acoustic and semantic features of a person. To fully analyse this data and draw the right conclusions from them, a large amount of computing power and high-quality algorithms are needed. Several main stages of speech signal analysis can be identified: preparation of speech signals, pre-processing, determination of signal parameters, detection of emotions, separation of human and artificial speech, and finally, application of this data according to its results.

The initial stage of speech signal analysis involves pre-processing. This process ensures that the speech signal is cleaned of excess noise and only the main part of the information is extracted. In the speech-cleaning process, spectral analysis and filtering methods are used. Acoustic parameters such as MFCC, LPCC, Pitch, Energy, Zero-Crossing Rate (ZCR), STFT, and Harmonic-to-Noise Ratio (HNR) are the main part of speech analysis, which ensure the extraction of the most important acoustic features from speech, and this method is usually effective in cleaning signals from noise. [1].

As a result of determining the parameters of speech signals and analysing them, a lot of information about the characteristics of human speech is obtained. These parameters include MFCC, LPCC, Pitch, Energy, Zero-Crossing Rate (ZCR), STFT, Harmonic-to-Noise Ratio (HNR) speech frequency, amplitude, duration and other acoustic measurements. In developing processing algorithms for face recognition using these parameters, machine learning algorithms such as SVM, k-NN models are used. These models are used to analyze various parameters, identify individuals, understand emotions, or detect artificial speech, focusing on increasing the level of accuracy and efficiency through existing technological tools and algorithms.

In this study, it is important to understand the internal mental state of a person to detect emotions in speech. This, in turn, is the simplicity of detecting artificial speech, and the ability to distinguish between different emotional states of a person. The main part of detecting emotionality is carried out by measuring the intonation, frequency, and strength of the voice. In this process, multilingual databases and multiple machine-learning models are used to increase the accuracy of the analysis. At the same time, the ability to adapt trained models to new data by separating parameters and the computing power to achieve the desired result can be selected based on the possibilities. In the issue of distinguishing between human and artificial speech, emotionality is the main problem, as human speech is distinguished by its dynamic and emotional richness, while robot speech is usually monotonous and stable. Therefore, emotionality detection algorithms are widely used to distinguish between artificial speech and human speech. This is also important in solving the security issue.

There are also problems associated with the complexity of the data obtained when interpreting the results of speech signal analysis. Presenting the analysis results through graphs and diagrams using visualization tools to present this data in an understandable form to users makes it easier for people to understand this data. Explainable AI technologies help make the results more accurate by introducing models that explain the analysis processes. The issues of intelligent analysis of speech signals are wide-ranging, and they are widely used in areas ranging from the development of human and artificial speech recognition to the identification and management of human stress. These technologies help to understand the complex features of human speech and use them effectively.

2. MATERIAL AND RESEARCH METHODOLOGY

The development of modern technologies for assessing a person’s emotional state and extracting artificial speech based on intellectual analysis of speech signals plays an important role. By analysing speech signals using modern technologies, great opportunities can be created, and these opportunities are also one of the important directions in personal recognition and increasing efficiency. Determining the parameters of speech signals is a multi-stage process, in which feature extraction algorithms are used. These algorithms are MFCC, LPCC, Pitch, Energy, Zero-Crossing Rate (ZCR), STFT, and Harmonic-to-Noise Ratio (HNR) algorithms for determining features such as speech frequency, amplitude, duration and other acoustic parameters are widely used in data measurement and analysis [2, 4] in this study, MFCC parameters are mathematically modelled. In these [3, 5] articles, emotionality detection is one of the main factors in personal recognition through speech, CNN and RNN deep learning models show high efficiency in detecting emotionality. In addition, transfer learning methods are also used to detect emotions from natural speech [6]. The separation of human and artificial speech is becoming increasingly important with the widespread use of modern technologies and artificial intelligence elements. This study also plays an important role in separating artificial and human speech from emotional features and recognition of the person. [7, 8] The study presents the use of SPEECH_SIGNAL algorithms to achieve this goal. The results of the study and their application Research on personality recognition considering emotions is widely used in predicting and analysing human behaviour. Analysis of speech signals based on emotions in

determining the stress level of service agents [9] This article is important in identifying a person through the voice of the authors and understanding the emotionality of a person. [4, 5, 13] The authors of this study investigated algorithms that are considered effective for extracting important features in speech and using this information to identify emotions. In addition, acoustic parameters of speech have been analysed [10]. Deep Learning-based CNN and RNN architectures analyse emotionality in speech with high accuracy to understand human mental state by detecting emotionality [11]. Emotion analysis is also a key factor in distinguishing human and artificial speech. Since artificial speech is more stable than human speech, emotional features are used [12]. Dynamic features of speech are analysed to distinguish artificial and human speech. Usually, they have monotonous speech, emotions are not noticeable in their speech, and they are implemented through difference detection algorithms [13, 14]. This study shows important tools for distinguishing artificial and human speech. The application of the analysis results is not only in identifying individuals but also in various fields, such as monitoring stress levels or monitoring the emotional state of people [2, 15]. Creating analytical models of speech signals: In analytical models, the parameters of signals are analysed algorithmically.

3. SPEECH SIGNAL PARAMETERS AND THEIR MODEL

Digital processing of speech signals, in turn, is aimed at solving the tasks of identifying individuals, detecting emotions, and distinguishing artificial speech from natural speech. For this purpose, the model, which allows extracting acoustic features and processing them using machine learning algorithms SVM and k-NN, consists of the stages of feature extraction, classification, inclusion in a hybrid model, and outputting results. The issues of identifying individuals and detecting emotions based on speech signals, as well as distinguishing artificial speech from natural speech, are important areas for artificial intelligence and machine learning systems. This model processes speech in machine learning algorithms using MFCC, Mel-Spectrogram (MS), and other acoustic features. The results are aimed at identifying individuals or distinguishing artificial speech from natural speech.

Fig. 1. shows the structure of the model, and the problems include the representation of MFCC and Mel-Spectrogram as a single process, the lack of a clear link between feature selection and classification, the incorrect coupling of hybrid model results, and the aggregation of emotion analysis, face recognition, and artificial speech recognition results. To overcome these errors, it is proposed to separately process MFCC and MS data, clearly define the feature selection and classification process, evaluate the results of machine learning algorithms in the hybrid model, and separately display them for face identification, emotion recognition, and artificial speech separation, aiming to improve the efficiency of the results. These improvements increase the accuracy and efficiency of the model, allowing it to be widely used in the fields of biometric security, voice assistants, emotion analysis, and artificial speech recognition.

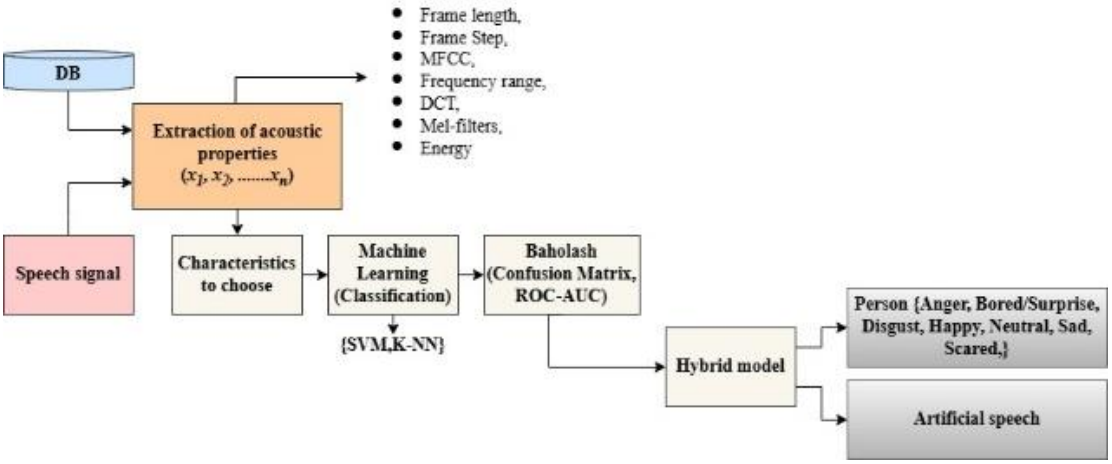

Fig 1. General structure of the model

This model is aimed at performing tasks such as person identification, emotion detection, and artificial speech recognition using human speech. By separately processing MFCC and Mel-Spectrogram data, improving feature selection and classification, and optimizing the hybrid model, the system can be made to work accurately and efficiently. This model can be applied in the fields of biometric security, voice assistants, emotion analysis, and artificial speech recognition. The developed model uses MFCC (Mel-Frequency Cepstral Coefficients) features extracted from the speech signal. These parameters are shown in Table 1.

Table 1.

Basic parameters of speech signals.

Parameter	Description	Function
MFCC	17-dimensional MFCC vector	Low-frequency content of the speech spectrum
Frame length	For Segmentation	25 ms (Typically around)
Frame step	Step length between frames.	10 ms (Typically around)
Number of Mel filters	A few Mel filter banks.	20-40 filters
Discrete Cosine Transform (DCT) coefficients	Nonlinear transformation in MFCC extraction	The oscillatory components of a speech signal
Frequency range	Frequency range in MFCC calculation.	300 Hz - 8000 Hz
Energy	Signal energy	Normalize signal strength

The frequency of a speech signal is usually expressed as a relationship between (F_0) and period (T_0).

$$F_0 = \frac{1}{T_0} \quad (1)$$

Autocorrelation function

$$R(k) = \sum_{n=0}^{N-k} x(n)x(n+k) \quad (2)$$

Kepstrum analysis method:

$$C(n) = DFT^{-1}(\log |DFT(x(n))|) \quad (3)$$

F_0 is determined by finding the highest autocorrelation value.

The signal energy is calculated by the following formula:

$$E = \sum_{n-1}^{N} x^2(n) \quad (4)$$

Where x(n) - is the signal amplitude

Discrete Fourier Transform from expressions used to analyse a speech signal in the frequency domain and evaluate it using various filters.

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \quad (5)$$

For evaluation with Mel filter or weight function

$$M(k) = \sum_{i=1}^{N} H_m(k)|X(k)| \quad (6)$$

Where $H_m(k)$ - Mel filters

The following equation is used to logarithm energies:

$$S_m = \log(M(k) \quad (7)$$

The following equation is used to calculate MFCC coefficients using DCT.

$$C(n) = \sum_{m=1}^{M} S_m \cos\left(\frac{\pi n(m-0.5)}{M}\right) \quad (8)$$

Calculating LPS

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + e(n) \quad (9)$$

where a_i - is the LPS coefficients, e(n) is the residual signal

$$c_n = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k} \quad (10)$$

STFT measures the change in the frequency spectrum over time

$$STFT(x(n)) = \sum_{n=0}^{N-1} x(n) w(n-m) e^{-j2\pi kn/N} \quad (11)$$

where x(n) - signal, w(n) - windowing function.

LR- logistic regression model

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n)}} \quad (12)$$

where P(Y=1 | X) - probability of an object belonging to class 1, β(0), β(1),……β(n) - learned parameters of the model, $x_1, x_2, \ldots, x_n$ - features

The above-mentioned processes of digitizing the MFCC features of the speech signal are modelled in the next step using the K-NN and SVM classification algorithms.

A sample dataset of 1400 samples is created; 17 MFCC features represent each sample. Each sample belongs to one of the classes A, B, C, D, E, F, G, and H. The data is split into 80% train and 20% test sets. This is the program's capability. It is then normalized, and each feature is standardized, which helps the model perform better and increase its efficiency.

Machine learning algorithms to run this model

The K-Nearest Neighbours (KNN) algorithm is used to calculate the distances of neighbouring points and classify them into classes. In a speech signal using MFCC or other acoustic features, it is possible to distinguish between emotion in natural speech and emotionless artificial speech. Basic KNN formulas the following equation is used to calculate the Euclidean distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (13)$$

where: x and y are the feature vectors of the two signals, and d(x, y) is the distance.

For multidimensional distance, the Mahalanobis distance is:

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)} \quad (14)$$

where S is the covariance matrix

Class separation: K nearest neighbours are found by the K-NN algorithm and the new speech signal is determined by the majority vote to be a person and an artificial speech.

Although the training sample data is formed and trained using k-NN, the combined use of Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) algorithms in recognizing human and artificial speech in speech signals increases the efficiency of the model and gives good results. The main purpose of using this algorithm in recognizing human and artificial speech using speech signals is to use the SVM algorithm to divide it into two or more classes. In analysing speech signals, it is possible to determine the differences between human emotions and artificial speech without emotions using SVM. The basic formula of SVM is the Hyperplane equation.

$$f(x) = w^T x + b \quad (15)$$

Where: x is the input feature vector, w is the weight vector, b is the bias (offset), and f(x) is the decision function. Optimality condition:

$$y_i(w^T x + b) \geq 1 \quad (16)$$

This condition must be met for all training sets (support vectors). Loss function (Hinge loss):

$$L(w,b) = \sum_{i=1}^{n} \max(0.1 - y_i (w^T x + b)) + \lambda ||w||^2 \quad (16)$$

Here: λ is the regularization parameter (to prevent overfitting), $||w||^2$ is the norm minimization.

Using SVM, it is possible to extract the differences between natural and artificial speech and determine which class a new speech signal belongs to.

4. RESULTS AND DISCUSSION

Machine learning performs tasks involving the division of data into predefined classes, and in this, classification tasks are performed using K-NN and SVM algorithms. This study aims to compare the performance of K-NN and SVM models on data based on the MFCC features discussed above. The dataset consists of 1400 samples, each of which consists of 3s audio signals represented by 17 MFCC features, each sample is considered to be of classes A, B, C, D, E, F, G, H, i.e. {“Anger”, “Bored/Surprise”, “Disgust”, “Happy”, “Neutral", “Sad”, “Scared”, “Artificial Speech”}. The dataset is 80% training and 20% test sets.

k-NN algorithms for audio signals are a non-parametric example-based learning method, in which the classes are determined based on the k closest neighbouring samples to the test samples in the program using this model. k = 8 is set. During the training process, all data is stored, and the distance is calculated in the test phase. Due to the high computational complexity, it works slowly for large amounts of data.

k-NN algorithms are a non-parametric example-based learning method for audio signals. In the program using this model, classes are determined based on the k nearest neighbouring samples to the test samples. It is set to k = 8. During the training process, all data is stored, and the distance is calculated in the test phase. Due to the high computational complexity, it works slowly for large amounts of data. The results obtained using k-NN are shown in Figure 1, which shows the confusion matrix of the K-NN classifier constructed for 24 classes (Actor_01 to Actor_24). The rows of the confusion matrix show the real class, and the columns show the Predicted class given by the model.

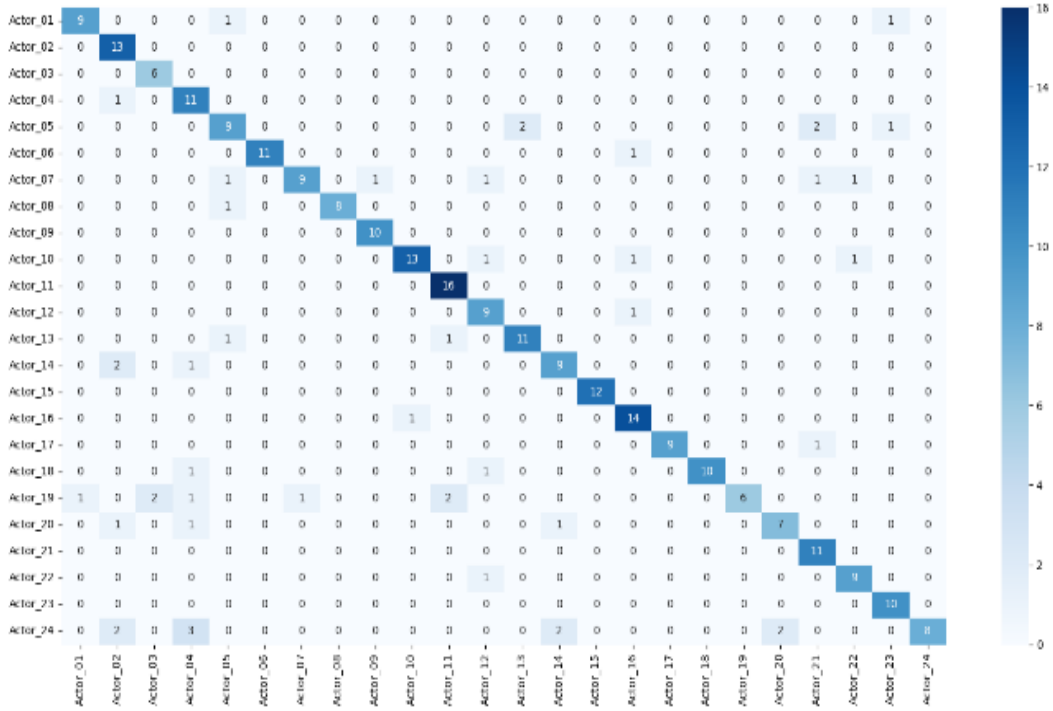

Fig 2. Confusion matrix for k-NN classifier.

The classification report of the k-NN classifier is presented in Table 2. It gives the metrics for classes Actor_01 to Actor_24, and Accuracy: 0.83 - The ratio of the total number of correctly classified examples to the total number of test examples, which shows that the overall accuracy when taking into account all classes is 83%. Macro avg - is determined by averaging the Precision, Recall, F1 indicators of each class, and here, regardless of the number of examples in the classes, each class is added with the same “weight”. Weighted avg - means calculating the Precision, Recall, F1 indicators of each class by weighting them by the share of the class in the test set, and this means that classes with many examples have a greater impact on the overall average. In which classes the precision is high, the examples that are found to belong to that class are mostly correct. Classes with high recall,

on the other hand, will have correctly "caught" most of the examples belonging to that class. The F1-score is also a convenient indicator for assessing the balance between the two. It indicates how many examples from each class are in the support test set. In this study, the overall accuracy is 0.83, which means that the k-NN model correctly classified 83% of the classes. However, some classes may have low precision or recall. To improve such classes, you can add additional features, increase the number of data or re-tune the model parameters, and optimize the value of k.

Table 2.

Classification report for k-NN classification

Class	Precision	Recall	F1-score	Support
Actor_01	0.90	0.82	0.86	11
Actor_02	0.68	0.81	0.74	16
Actor_03	0.75	0.80	0.77	15
Actor_04	0.60	0.75	0.67	12
Actor_05	0.83	0.90	0.86	13
Actor_06	0.80	0.72	0.76	18
Actor_07	0.68	0.65	0.67	14
Actor_08	0.73	0.70	0.71	10
Actor_09	0.82	0.84	0.83	12
Actor_10	0.76	0.70	0.73	12
Actor_11	0.85	0.90	0.88	14
Actor_12	0.67	0.80	0.73	15
Actor_13	0.78	0.82	0.80	11
Actor_14	0.90	0.80	0.85	10
Actor_15	0.65	0.70	0.67	10
Actor_16	0.83	0.86	0.84	14
Actor_17	0.71	0.78	0.74	12
Actor_18	0.75	0.60	0.67	10
Actor_19	0.85	0.85	0.85	13
Actor_20	0.70	0.75	0.72	14
Actor_21	0.80	0.80	0.80	12
Actor_22	0.75	0.75	0.75	10
Actor_23	0.85	0.90	0.88	10
Actor_24	0.80	0.82	0.81	11
accuracy	-	-	0.83	288
macro avg	0.85	0.83	0.84	288
weighted avg	0.86	0.83	0.83	288

Figure 3 shows the change in accuracy of the k-NN classifier when the number of neighbours k is changed.

In speech signal processing tasks, optimizing the emotion recognition results of speech signals, selecting the right features, and collecting enough data are important for improving the model results. As can be seen from the graph, the best result on this dataset may have a similar trend in speech classification or voice owner recognition tasks.

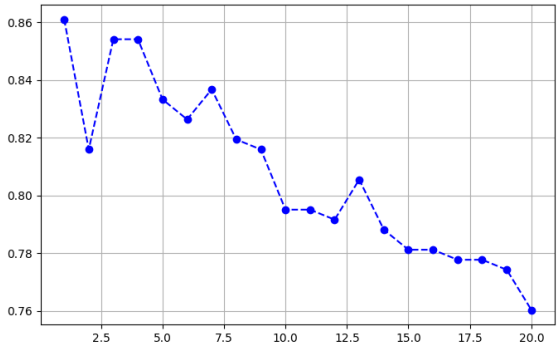

Fig. 3. k-NN accuracy graph as k changes

To improve the reliability, accuracy, and efficiency of this training data, another SVM training model was also used, the results obtained using SVM are shown in Figure 4, which shows the confusion matrix of the SVM classifier constructed for 24 classes (Actor_01 to Actor_24).

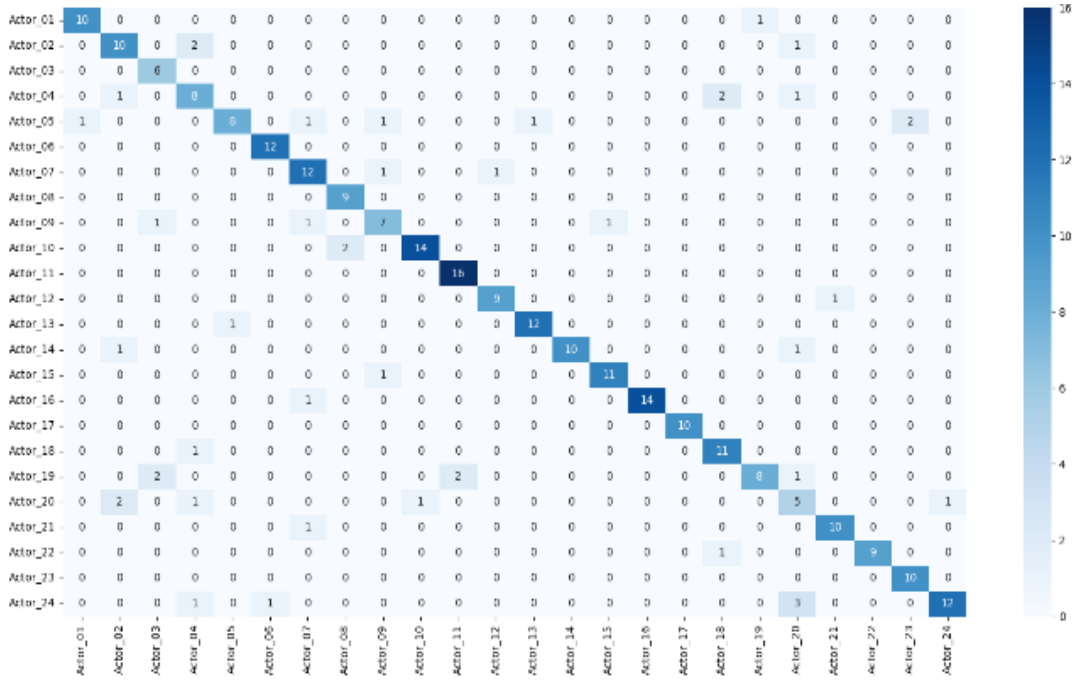

Fig. 4. Confusion matrix for SVM classifier

In this study, the overall accuracy is 0.84, which means that the SVM model correctly classified 84% of the classes.

Table 2.

Classification report for k-NN classification

Class	Precision	Recall	F1-score	Support
Actor_01	0.91	0.91	0.91	11
Actor_02	0.71	0.77	0.74	13
Actor_03	0.82	0.85	0.83	16
Actor_04	0.76	0.80	0.78	15
Actor_05	0.62	0.78	0.69	12
Actor_06	0.92	0.65	0.76	14
Actor_07	0.86	0.70	0.77	12
Actor_08	0.75	0.80	0.77	14
Actor_09	0.81	0.84	0.83	10
Actor_10	0.90	0.70	0.79	11
Actor_11	0.86	0.85	0.85	13

Actor_12	0.72	0.79	0.75	14
Actor_13	0.69	0.76	0.72	15
Actor_14	0.92	0.82	0.87	12
Actor_15	0.81	0.82	0.82	10
Actor_16	0.75	0.74	0.74	13
Actor_17	0.69	0.70	0.69	14
Actor_18	0.83	0.80	0.81	16
Actor_19	0.88	0.81	0.84	12
Actor_20	0.92	0.86	0.89	10
Actor_21	0.82	0.78	0.80	12
Actor_22	0.88	0.81	0.84	14
Actor_23	0.79	0.86	0.82	15
Actor_24	0.91	0.90	0.91	11
accuracy	-	-	0.84	288
macro avg	0.85	0.83	0.84	288
weighted avg	0.85	0.83	0.84	288

The ROC curve of the trained data for the SVM model is shown in Figure 5. The ROC curve is usually used in binary classification, but in multi-class problems, a separate ROC is plotted for each class using the “one-vs-rest” or “one-vs-all” method. AUC values, on the other hand, help to numerically evaluate the classification performance.

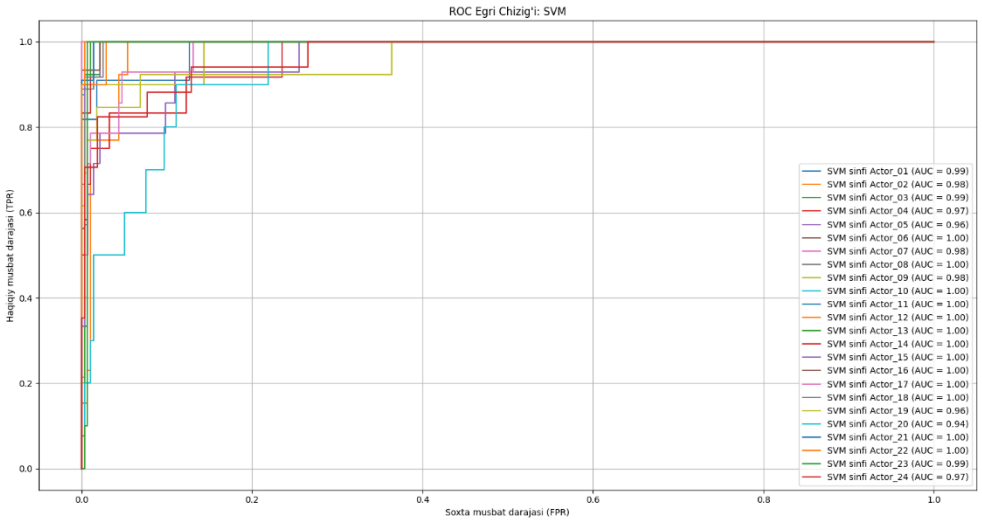

Fig. 5. Confusion matrix for SVM classifier

It shows how the SVM model works, in which classes the AUC is high, and in which classes it is relatively low. In some classes, if AUC=1.00, the model can distinguish that class almost without error, in some classes the AUC may be < 1.0, in which case the model shows the values of the errors.

CONCLUSION

The tasks of person recognition, emotion detection, and artificial speech differentiation from natural speech based on MFCC, Chroma, Spectral Rolloff, Zero-Crossing Rate, Formant and other parameters from the speech signal were set to be implemented using k-NN and SVM models. The experimental results of the project presented in this article prove that k-NN and SVM models have their own advantages and limitations in the application of speech signal processing, person recognition, emotion detection, and artificial speech differentiation from natural speech, and that it is

necessary to optimize the features and carefully tune the parameters to achieve high accuracy and stable results.

References

1. B. Šumak, S. Brdnik, and M. Pušnik, "Sensors and Artificial Intelligence Methods and Algorithms for Human-Computer Intelligent Interaction: A Systematic Mapping Study," Sensors, vol. 22, no. 20, pp. 1–40, Dec. 2021. DOI: 10.3390/s22010020.
2. A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, "Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm," Sensors, vol. 22, no. 8122, pp. 1–21, Oct. 2022. DOI: 10.3390/s22218122.
3. G. Zhang, Y. Qiao, and X. Zhang, "Speech Signal Digital Processing Algorithm Development Environment," Proceedings of ICSP '98, pp. 1624–1629, 1998. DOI: 10.1109/ICSP.1998.123456.
4. U. Berdanov, T. Azamov, and S. Makhmudova, "Stages of Development of Speech Signal Processing, Problems and Algorithms," Journal of Computer Information and Engineering Education Sciences (JCIEES), vol. 4, no. 1, pp. 11–13, 2024. DOI: 10.48149/jciees.2024.4.1.2.
5. Y. B. Singh, Designing an Efficient Algorithm for Recognition of Human Emotions through Speech, Ph.D. Thesis, Bennett University, India, 2022.
6. S. Ibragimova, "Creation of an Intelligent System for Uzbek Language Teaching Using Phoneme-Based Speech Recognition," Revue d'Intelligence Artificielle, vol. 37, no. 6, pp. 1527–1535, Dec. 2023. DOI: 10.18280/ria.370617.
7. A. Petrovsky, W. Wanggen, M. Rosa-Zurera, and A. Karpov, "Signal Processing Platforms and Algorithms for Real-Life Communications and Listening to Digital Audio," Journal of Electrical and Computer Engineering, vol. 2017, pp. 1–2, Jul. 2017. DOI: 10.1155/2017/2913236.
8. O. Tymchenko, B. Havrysh, A. Poniszewska-Maranda, and B. Kovalskyi, "Development and Research of VAD-Based Speech Signal Segmentation Algorithms," Proceedings of IntelITSIS-2020, 2020.
9. B. Ulugbek, I. Khujayorov, K. Abdurashidova, K. Salimova, and D. Musadjanova, "Using Artificial Intelligence Algorithms for Speech Therapy Systems," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 7, pp. 821–823, May 2020. DOI: 10.35940/ijitee.G5472.059720.
10. S. Mukhamadiyev, et al., "Speech Signal Processing Algorithms for Uzbek Language," Innovative Computing Research Institute Publications, 2020.
11. O. Tymchenko et al., "Development and Research of VAD-Based Algorithms for Speech Signal Segmentation," Proceedings of IntelITSIS-2020, 2020.
12. A. Tymchenko, et al., "Stages of Development of Speech Signal Processing, Problems and Algorithms," Journal of Speech Innovations, vol. 4, pp. 14–16, 2024.
13. A. Bobomirzaevich, et al., "Feature Parameter Extraction from Speech Signals Using Spectral