Chapter

Voice Identification Using Classification Algorithms

Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova, Tolga Ihsan Medeni and Aigerim Yessentay

Abstract

This article discusses the classification algorithms for the problem of personality identification by voice using machine learning methods. We used the MFCC algorithm in the speech preprocessing process. To solve the problem, a comparative analysis of five classification algorithms was carried out. In the first experiment, the support vector method was determined—0.90 and multilayer perceptron—0.83, that showed the best results. In the second experiment, a multilayer perceptron with an accuracy of 0.93 was proposed using the Robust scaler method for personal identification. Therefore, to solve this problem, it is possible to use a multi-layer perceptron, taking into account the specifics of the speech signal.

Keywords: speaker identification, classification, speech recognition, MFCC

1. Introduction

In the era of informatization, many high-tech products gradually entered our daily life and significantly changed our life habits. On the other hand, information technologies continue to evolve towards a more human-centered approach. Biometric identification technology, which provides us with simpler and more convenient methods for identifying specific people, has gradually replaced some of the existing authentication methods that should be explored before people will be able to manage them properly Face recognition systems used in public places, law enforcement organizations [1], and Siri voice mobile assistant on iPhone, Bixby Voice on Galaxy [2], are examples of biometric identification results.

The recognition of a person by his voice is one of the forms of biometric authentication, which makes it possible to identify a person by a combination of unique voice characteristics and refers to dynamic methods of biometrics. Speaker recognition is a technology that can automatically identify the speaker based on the speech waveform, that reflects the physiological and behavioral characteristics of speech parameters from the speaker. Like traditional speaker recognition systems, there are two stages, namely, training and testing. These are the main stages of speaker recognition. Learning is the process of extracting phonetic characteristics from a speaker that has already been recorded or saved as a sample, storing them in a database, and familiarizing the system with the characteristics of the speaker's voice. Testing is the process of comparing questionable sound and phonetic characteristics from a speaker recognition database. Two popular sets of features, often used in the analysis of the speech signal are the Mel frequency cepstral coefficients (MFCC) and the linear prediction cepstral coefficients (LPCC). The most popular recognition models are vector quantization (VQ), dynamic time warping (DTW), and artificial neural network (ANN) [3].

The study of speech technologies for the Kazakh language is conducted in Kazakhstan. Kazakh language belongs to agglutinative languages. Agglutinative languages that have a system in which the dominant type of inflection is the "gluing" of various formants of suffixes or prefixes, each of which has only one meaning. The Turkic, Mongolian, Korean languages are agglutinative. In our country, the personal identification system in the Kazakh language has not been developed yet and research in this area is relevant.

This article deals with the problem of identification of a person using the classification model through the use of artificial neural networks. The paper is organized as follows. Section 2 describes the work on the relevant scientific research area. Section 3 discusses data preprocessing methods. Section 4 describes the methodology of automatic identification. Sections 5 and 6 discuss the results of the experiment and the conclusion.

2. Related works

A common feature of agglutinative languages, such as Finnish, Kazakh, and Turkish, is that, until now, attempts at personal identification and speech recognition have not led to comparable performance with English systems [4]. The reason for this is not only the difficulty of modeling the language, but also the lack of suitable resources for speech and text learning. In [5, 6] the systems aim to reduce active vocabulary and language models to a possible size by clustering and focusing.

Recently, neural networks have become dominant in various machine learning fields. One of them is natural language processing (the sequence of characters/words can be considered as another type of signals) in which the multilayer perceptron (MLP) and long-short-term memory (LSTM) network, two standard classifiers are widely used for the tasks of disambiguating morphological forms or identifying the boundary of sentence/tokens.

In many papers [7, 8], it was shown that the use of ANNs in conjunction with the HMM can improve the accuracy of speech recognition. Acoustic models are usually based on deep neural networks, which are artificial neural networks of direct propagation, containing more than one hidden layer between the input and output layers. For training, the backpropagation method is used.

The review article considers that feature extraction is one of the most important tasks in the identification system, which significantly affects the process and performance of the system. In the review analysis, the existing proposals and implementations of methods for identifying the features of the identification system were considered. Analysis of the results shows that MFCC-based approaches have been used more than any other approach and, moreover, it was revealed that the current trend of the identification system research is to solve important identification system problems, such as adaptability, complexity, multilingual recognition and noise resistance [9].

In one of the works [10], speech pre-processing method was considered using the VAD algorithm, which proves that this algorithm improves the performance of speech recognition. The study presents the principles of operation and the block diagram of the VAD algorithm in recognition of Kazakh speech.

Toleu et al. [11] proposed character-based MLP and LSTM models that can jointly identify the boundary of sentences and tokens. In order to extract the high-level abstract features, the proposed models project the characters embedding into low-dimensional space which could allow us to judge the different variety of signals. The models were tested for three languages: English, Italian and Kazakh. The experimental results show that character-based MLP and LSTM models for sentence and token segmentation have positive effects in terms of F-measure and the error rates compared to existing models.

Clustering algorithms are used to partition an existing set of speech segments into groups according to similarity of their attributes. Parametric algorithms for determining initial points (centroids) and subsequent cluster propagation are proposed in [12] and can be applied for speech classification task solving.

The work [13] considers the question of using a throat microphone (laryngophone) as an additional modality for phonetic segmentation of the speech signal into acoustic sub-word units. The new algorithm is proposed for the automatic speech signal segmentation based on the use of changing dynamics analysis of the throat-acoustic correlation (TAC) coefficients, which can be used for subsequent speech segment classification.

In the works of Russian scientists can be found a study on the recognition of continuous Russian speech, using deep belief networks (DBN), described in [14]. A method using finite state transducers was used for speech recognition and it was shown that the proposed method allows to increase the accuracy of speech recognition in comparison with hidden Markov models.

3. Feature extraction and configuration parameters

In identification tasks, the main process is speech pre-processing. In this study, we select MFCC [15] as a tool for extracting voice dynamics functions. The speech pre-processing process is described in **Figure 1**.

Voice signals in the time domain change very quickly and dramatically, but if we convert speech signals from the time domain to the frequency domain [16], then the corresponding spectrum can be clearly defined. Our system separates the signals into frames and calls the window function to increase the continuity of voice signals in the frame. DCT is being used for quantitative evaluation of spectral energy data into data units that can be analyzed by MFCC [17]. The MFCC parameters are in the range of analyzed frequencies 300–8000 Hz, as well as 16 cepstral.

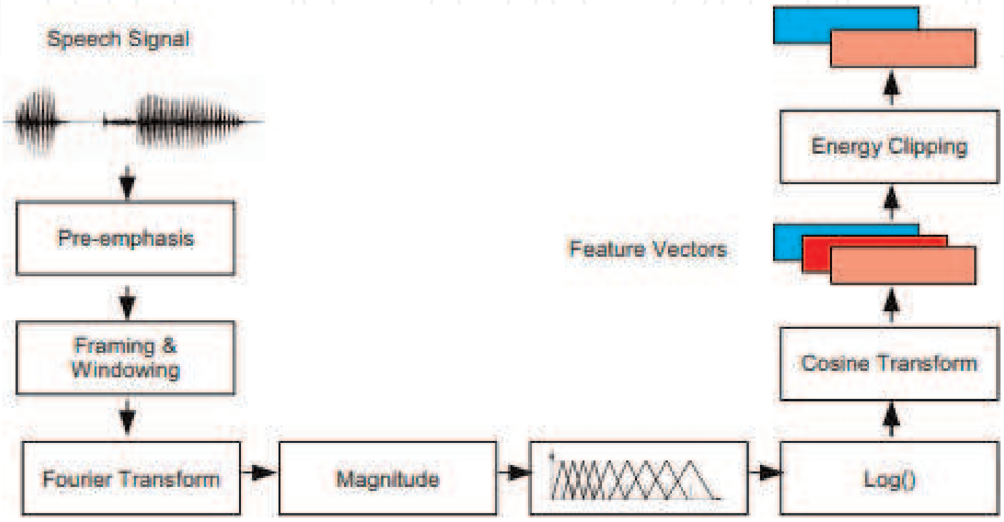


Figure 1.
Steps involved in extracting MFCC feature vectors.

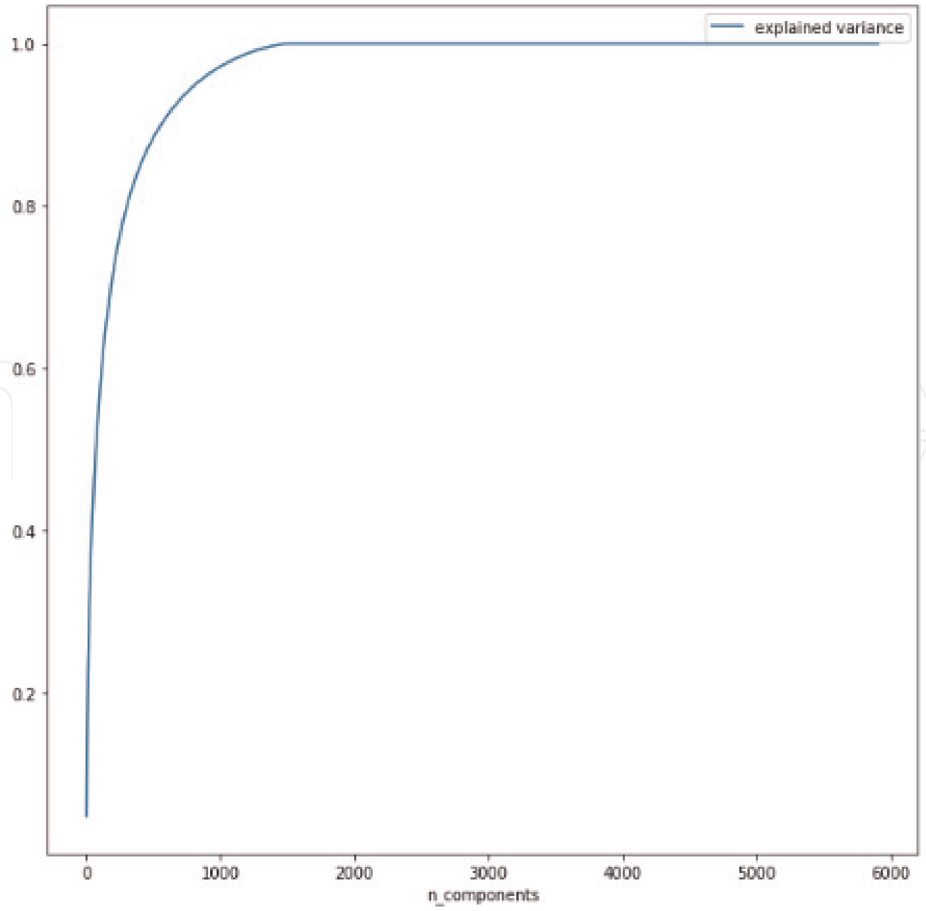


Figure 2.
Preservation of dispersion with decreasing dimension by the method of principal components.

As a result, 5904 features were obtained for each audio file. Each audio file was marked with the initials of the speaker whose voice was recorded in it. The resulting dataset had a dimension of 1480 × 5904.

To visualize the data, the principal components method was used to reduce the dimension of the vector space from 5904 features to two- and three-dimensional space [18]. Maintaining the dispersion in the reduction of the dimensionality by the principal component analysis shown in **Figure 2**.

As can be seen from the above graph, 100% of the variance is preserved when the data dimension is reduced to 1479 features. However, as experiments with classification models and data standardizers have shown, such a reduction in dimension critically affects the accuracy of the classification.

4. The proposed speech identification system

The methodology of our work is as follows:

4.1 The design of the experimental data

Data for analysis were provided by the laboratory of "computer engineering of intelligent systems." The data set consists of 1480 audio recordings from 20 speakers with 74–75 recordings. Each audio recording consists of phrases in Kazakh with an average length of 6 seconds. To identify the speaker, we collected the following data: name, gender, place of birth, year of birth (**Table 1**).

Label	Origin	Name	Middle name	Gender	Birthplace	Year of birth
MZA	Masimkanova	Zhazira	Auezbekkyzy	Female	Almaty	20.03.1982
IMT	Iskakova	Moldir	Tasbolatkyzy	Female	Almaty	01.01.1994
DAZ	Duisenbaeva	Aigerim	Zhanbolatovna	Female	Almaty	15.05.1995
ZEA	Zhetpisbaev	Erlan	Alibekovich	Man	Almaty	23.05.1995
SSM	Samrat	Sanjar	Muhametkaliuly	Man	Almaty	12.07.1996
...	...	...	...	...	...	...

Table 1.
Information about the speakers.

To increase the accuracy of recording audio materials, a soundproofing, professional recording studio from Vocalbooth.com was used.

All audio materials have the same characteristics:

- file extension: .wav;
- digital conversion method: PCM;
- discrete frequency: 44.1 kHz;
- digit capacity: 16 bits;
- number of audio channels: one (mono).

The sound and recording of one speaker took an average of 40–50 minutes of time including the time required to prepare the speaker, equipment and doubles, which corresponds to the 74–75 files received, a total length of 7–8 minutes for each speaker.

4.2 Classification algorithms

For the speaker identification problem we took the following known classification algorithms.

4.2.1 Extra-Trees algorithm

The Extra-Trees algorithm builds an ensemble of unpruned decision or regression trees. The algorithm's two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample to grow the trees. The Extra-Trees algorithm is given in **Table 2**.

It has two parameters: K, the number of attributes randomly selected at each node and n_{min}, the minimum sample size for splitting a node. It is used several times with the original learning sample to generate an ensemble model.

4.2.2 KNN algorithm

The K-nearest-neighbor (KNN) algorithm measures [19] the distance between a query scenario and a set of scenarios in the data set.

```
Trees_node(M)
Input signal: the local learning subset M corresponding to the node
Output signal: a tree [$a < a_c$] or nothing
    • If Tree(S) is TRUE then return nothing;
    • Otherwise select K attributes {$a_1$,...,$a_K$} among all non constant (in S) candidate attributes;
    • Draw K trees {$s_1$,...,$s_K$}, where $s_i$ = Random_split(S, $a_i$), ∀i = 1,..., K;
    • Return a tree $s_*$ such that Score($s_*$, S) = $\max_{i=1,...,K}$ Score($s_i$, S).
Random_split(S,a)
Inputs: a subset S and an attribute $a$
Output: a split
    • Let $a^S_{max}$ and $a^S_{min}$ denote the maximal and minimal value of a in S;
    • Draw a random cut-point $a_c$ uniformly in [$a^S_{min}$, $a^S_{max}$];
    • Return the tree [$a < a_c$].
Tree(S)
Input: a subset S
Output: a Boolean
    • If |S| < $n_{min}$, then return TRUE;
    • If all attributes are constant in S, then return TRUE;
    • If the output is constant in S, then return TRUE;
    • Otherwise, return FALSE.
```

Table 2.
Extra-Trees algorithm.

To classify each of the test sample objects, the following operations should be performed sequentially:

- To calculate the distance to each training sample feature.
- Select k objects of the training sample, the distance to which is minimal.
- The class of the object being classified is the class most often found among the k nearest neighbors.

4.2.3 SVC algorithm

In order to use an SVC to solve a linearly separable, binary classification problem we need to:

- create **H**, where $H_{ij} = y_i y_j x_i \cdot x_j$
- find α so that
- $\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha$
- is maximized, subject to the constraints
- $\alpha_i \geq 0 \ \forall_i \ and \ \sum_{i=1}^{L} \alpha_i y_i = 0$
- this is using a QP solver.
- calculate $w = \sum_{i=1}^{L} \alpha_i y_i x_i$

- determine the of Support Vectors S by finding the indices such that $\alpha_i > 0$
- calculate $b = \frac{1}{N_s} \sum_{s \in S} (\alpha_m y_m x_m \cdot x_s$
- each new point x' is classified by evaluating $y' = \text{sgn}\,(w \cdot x' + b)$.

4.2.4 MLPClassifier algorithm

Multi-layer perceptron is a supervised learning algorithm that learns [20] a function $f(X) = R_n : R_n \rightarrow R^0$ by training on a speech dataset, where n is the number of dimensions for input and 0 is the number of dimensions for output. Given a set of features $X = x_1, x_2, \ldots, x_n$, it can learn a non-linear function approximator for either classification (**Figure 3**).

The input layer consists of $x_1, x_2, \ldots, x_n$ representing the input features. The output layer receives the values from the last hidden layer and transforms them into output values.

4.2.5 Gaussian NB algorithm

Naive Bayes gives the probability of a data point $X = x_1, x_2, \ldots, x_n$ belonging to class C_k as proportional to a simple product of $n + 1$ factors. The class prior $p(C_k)$ plus n conditional feature probabilities $p(x_i|C_k)$. Specifically,

$$p(C_a) \prod_{i=1}^{n} p(x_i|C_a) > p(C_b) \prod_{i=1}^{n} p(x_i|C_b)$$

$$p(C_a|x_1, \ldots, x_n) > p(C_b|x_1, \ldots, x_n)$$

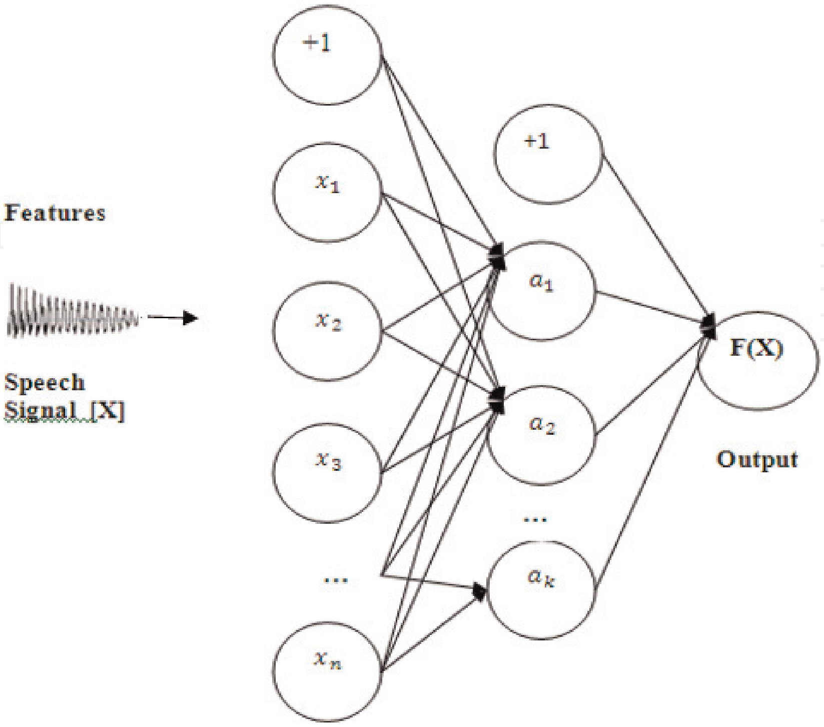
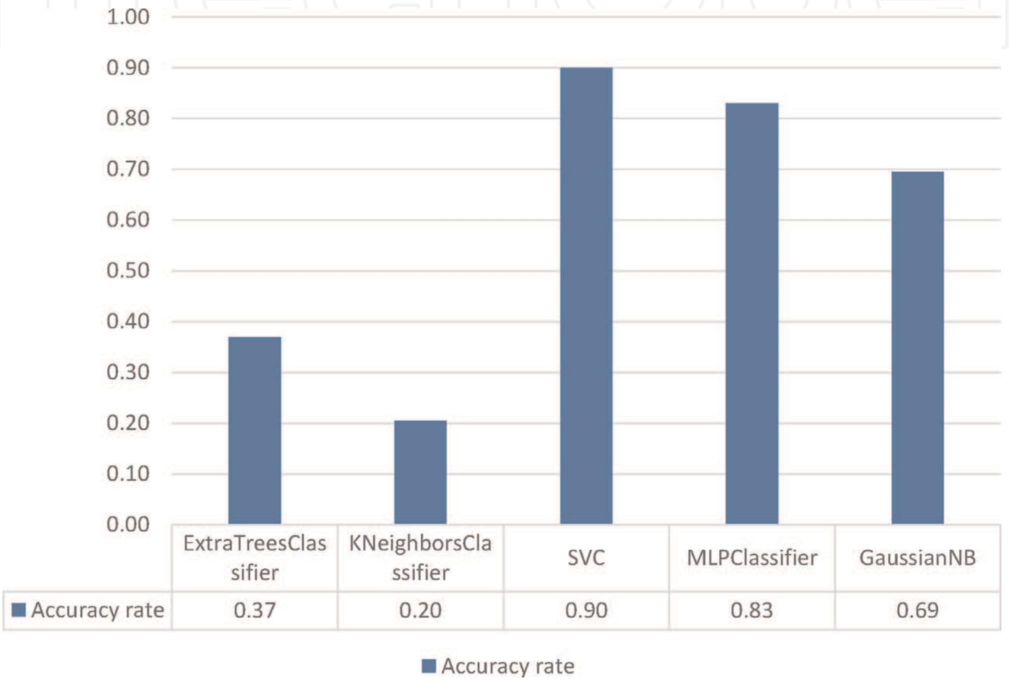


Figure 3.
MLP architecture.

Thus, the most likely class assignment for a data point $x_1, x_2, \ldots, x_n$ can be found by calculating $p(C_a) \prod_{i=1}^{n} p(x_i|C_k)$ for $k = 1, \ldots, K$ and assigning $x_1, x_2, \ldots, x_n$ the class C_k for which this value is largest.

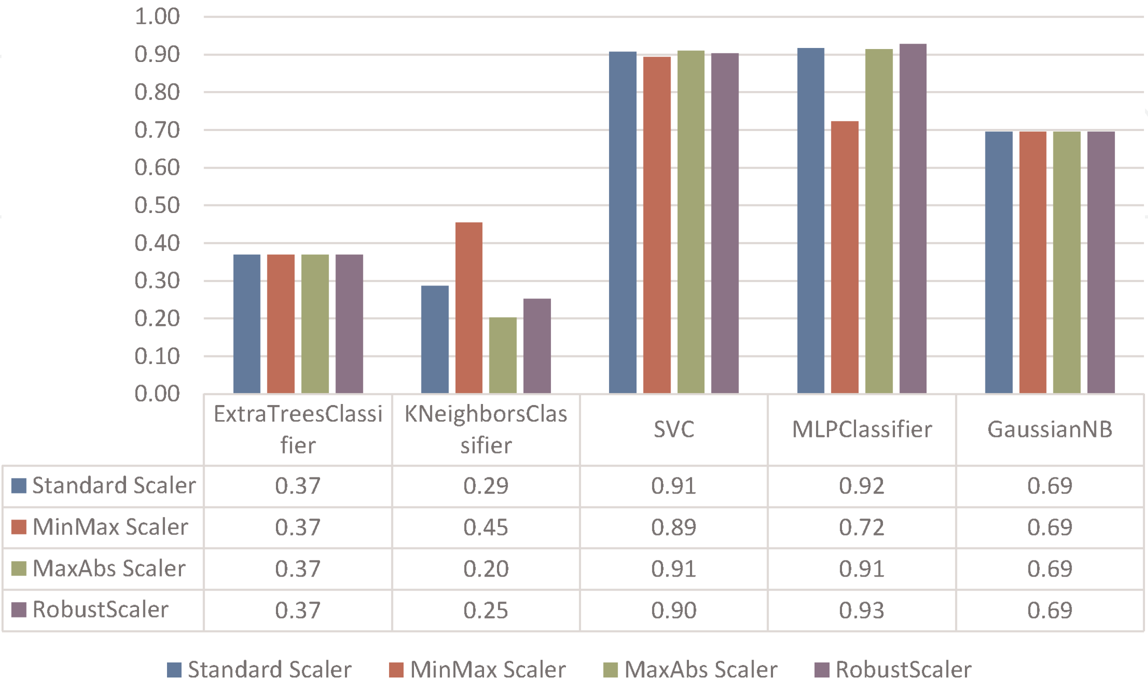
5. Results and discussion

In Section 4.2 we applied the algorithms considered for the problem of personality identification and made a comparative analysis. Comparative analysis and



	ExtraTreesClassifier	KNeighborsClassifier	SVC	MLPClassifier	GaussianNB
■ Accuracy rate	0.37	0.20	0.90	0.83	0.69

Figure 4.
Classification accuracy on a data set.



	ExtraTreesClassifier	KNeighborsClassifier	SVC	MLPClassifier	GaussianNB
■ Standard Scaler	0.37	0.29	0.91	0.92	0.69
■ MinMax Scaler	0.37	0.45	0.89	0.72	0.69
■ MaxAbs Scaler	0.37	0.20	0.91	0.91	0.69
■ RobustScaler	0.37	0.25	0.90	0.93	0.69

Figure 5.
Accuracy of classification when scaling data by various methods.

experiments have shown that the best results were obtained using the support vector machine and multilayer perceptron (**Figure 4**).

As can be seen from the diagram, the support vector machine and the multilayer perceptron showed the best results, 0.90 and 0.83, respectively.

To improve the results by different methods, the scaling was carried out and the results were slightly changed (**Figure 5**).

Now the highest accuracy was shown by the multilayer perceptron—0.93 when scaled by the Robust scaler method, and the support vector machine was relegated to the background, although it improved its result in accuracy from 0.90 to 0.91 when scaled by the Standard scaler and MaxAb scaler methods.

If we reduce the dimension of speech features to 1479 using the principal components method, the classification accuracy will change as in **Table 3**.

The purpose of the comparative analysis was to determine the degree of influence of classification algorithms for the problem of personality identification, as well as a comparative evaluation of the SVC and MLPClassifier Algorithms. Experiments conducted on the training set of speech data showed results that allow to

Algorithm name	Standard scaler	MinMaxScaler	MaxAbsScaler	RobustScaler
ExtraTreesClassifier	0.128125	0.128125	0.128125	0.128125
KNeighborsClassifier	0.043571	0.052143	0.050089	0.060982
SVC	0.051875	0.134732	0.097500	0.157679
MLPClassifier	0.002589	0.051875	0.082054	0.098393
GaussianNB	0.324286	0.324286	0.324286	0.324286

Table 3.
The accuracy of classification on the data with a decrease in dimension.

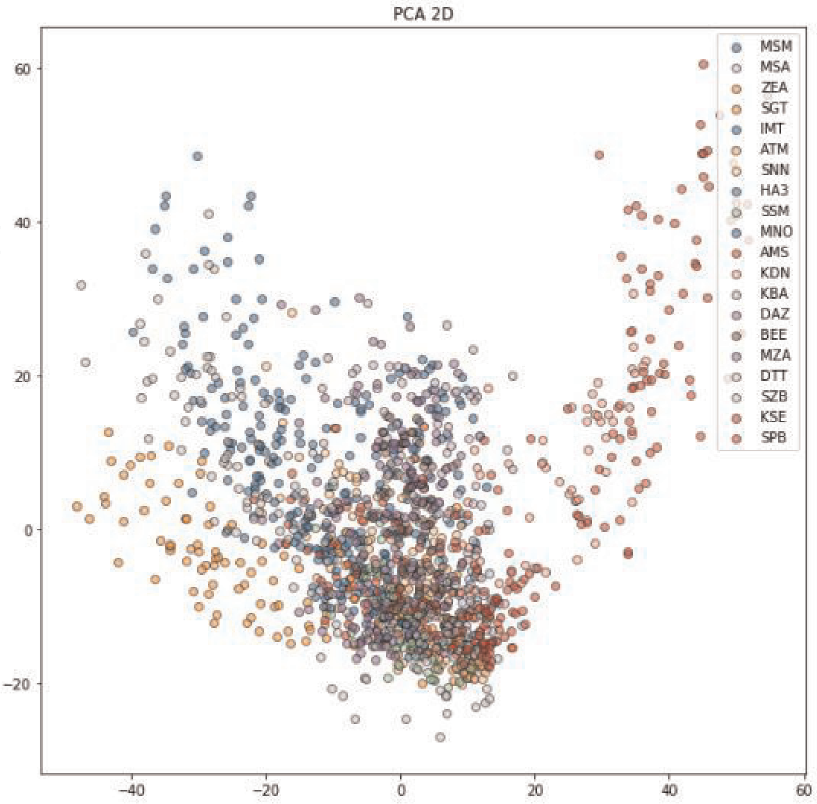


Figure 6.
Two-dimensional representation of speech data.

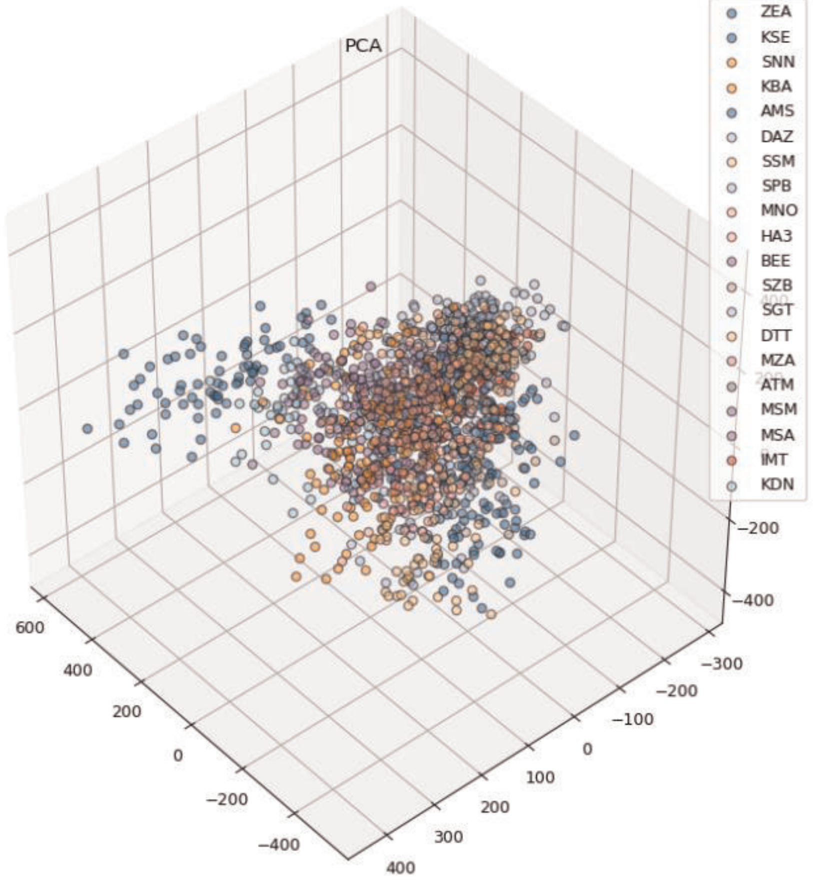


Figure 7.
Three-dimensional representation of speech data.

speak about the prospects of these algorithms. The data obtained were presented in **Figures 6** and 7.

The results of the classification when scaling data by various methods turned out to be significantly different from the results obtained during preliminary experiments.

6. Conclusions

In this paper, a number of classification algorithms and speech preprocessing issues were considered. Based on the analysis of the experimental results, a multilayer perceptron with an accuracy of 0.93 was proposed for scaling by the Robust scaler method and we will be able to classify the speech signal with the help of a multilayer perceptron. Further from the data we identified the personality.

In our further studies, we would like to solve the problem of verifying the identity of the data obtained.

Acknowledgements

This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan. IRN AP05131207 Development of technologies for multilingual automatic speech recognition using deep neural networks.

Author details

Orken Mamyrbayev*, Nurbapa Mekebayev, Mussa Turdalyuly,
Nurzhamal Oshanova, Tolga Ihsan Medeni and Aigerim Yessentay
Institute of Information and Computational Technologies, Almaty, Kazakhstan

*Address all correspondence to: morkenj@mail.ru

References

[1] Zhan C, Li W, Ogunbona P. Face recognition from single sample based on human face perception. In: International Conference Image and Vision Computing New Zealand; 2009. pp. 56-61

[2] Beigi H. Fundamentals of Speaker Recognition. Springer Science & Business Media; 2011

[3] Yella S, Gupta N, Dougherty M. Comparison of pattern recognition techniques for the classification of impact acoustic emissions. Transportation Research Part C: Emerging Technologies. 2007;**15**(6): 345-360

[4] Aida-zade K, Xocayev A, Rustamov S. Speech recognition using support vector machines. In: AICT'16. 10th IEEE International Conference on Application of Information and Communication Technologies; 2016

[5] Serizel R, Giuliani D. Vocal tract length normalization approaches to DNN-based children's and adults' speech recognition. In: IEEE Workshop on Spoken Language Technology; 2014. pp. 135-140

[6] Popović B, Ostrogonac S, Pakoci E, Jakovljević N, Delić V. Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. In: Ronzhin A, Potapova R, Fakotakis N, editors; SPECOM 2015; LNCS; Heidelberg: Springer; Vol. 9319; 2015. pp. 186-192

[7] Psutka J, Ircing P, Psutka JV, Hajič J, Byrne WJ, Mirovsky J. Automatic Transcription of Czech, Russian, and Slovak Spontaneous Speech in the MALACH Project. In: Proceedings of Eurospeech. Lisboa; Portugal; 4-8 September 2005. pp. 1349-1352

[8] Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine. 2012;**29**(6):82-97

[9] Tirumala SS, Shahamiri SR, Garhwal AS, Wang R. Speaker identification features extraction methods: A systematic review. Expert Systems with Applications. 2017;**90**: 250-271

[10] Kalimoldayev MN, Alimkhan K, Mamyrbayev OJ. Methods for applying VAD in Kazakh speech recognition systems. International Journal of Speech Technology. 2014b;**17**(2):199-204

[11] Toleu A, Tolegen G, Makazhanov A. Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; ACL; 2017. pp. 666-671

[12] Krassovitskiy A, Mussabayev R. Energy-based centroid identification and cluster propagation with noise detection. In: Nguyen N, Pimenidis E, Khan Z, Trawiński B, editors. Computational Collective Intelligence. Lecture Notes in Computer Science. Vol. 11055. Cham: Springer; 2018. pp. 523-533. DOI: 10.1007/978-3-319-98443-8_48. ICCCI 2018

[13] Mussabayev RR, Kalimoldayev MN, Amirgaliyev Ye N, Mussabayev TR. Automatic speech segmentation using throat-acoustic correlation coefficients. Open Engineering. 2016;**6**:335-346

[14] Karpov A, Kipyatkova I, Ronzhin A. Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: Proc. INTERSPEECH-2011; Florence, Italy; 2011. pp. 3161-3164

[15] Mamyrbayev O, Turdalyuly M, Mekebayev N, Alimhan K,

Kydyrbekova A, Turdalykyzy T. Automatic recognition of Kazakh speech using deep neural networks. In: Asian Conference on Intelligent Information and Database Systems; 07 March 2019. pp. 465-474

[16] Mamyrbayev OZ, Kunanbayeva MM, Sadybekov KS, Kalyzhanova AU, Mamyrbayeva AZ. One of the methods of segmentation of speech signal on syllables. Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. 2015:286-290

[17] Aida-Zade K, Ardil C, Rustamov S. Investigation of combined use of MFCC and LPC features in speech recognition systems. World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering. 2007: 2647-2653

[18] Voitko VV, Bevz SV, Burbelo SM, et al. Automated system of audio components analysis and synthesis In: Proceedings of SPIE; 2019. p. 110450V. DOI: 10.1117/12.2522313

[19] Hazmoune S, Bougamouza F, Mazouzi S. A new hybrid framework based on hidden Markov models and K-nearest neighbors for speech recognition. International Journal of Speech Technology. 2018;**21**(3):689-704

[20] Ribeiro FC, Santos CRT, Cortez PC, et al. Binary neural networks for classification of voice commands from throat microphone. IEEE Access. 2018; **6**:70130-70144