Algorithms for Detection Gender Using Neural Networks

Maksat KALIMOLDAYEV[1], Orken MAMYRBAYEV[1],
Nurbapa MEKEBAYEV[1,2], Aizat KYDYRBEKOVA[1,2]
[1]Institute of Information and Computational Technologies, Kazakhstan
[2]al-Farabi Kazakh National University, 050040 Almaty, Kazakhstan

Abstract- In this paper, we investigate two neural architecture for gender detection tasks by utilizing Mel-frequency cepstral coefficients (MFCC) features which do not cover the voice related characteristics. One of our goals is to compare different neural architectures, multi-layers perceptron (MLP) and, convolutional neural networks (CNNs) for both tasks with various settings and learn the gender -specific features automatically.

Keywords- MFCC, CNN, MLP,NN, speech recognition, audio signal, neural network,.

I. INTRODUCTION

Automatic speech recognition is a dynamically developing area in the field of artificial intelligence.

Speech recognition is the process of recognizing words spoken by a person based on an automatic speech signal. The length of a word can be different at different frequencies, and the same length depends on the same words, different parts of words are different from the apparent rate of difference in the environment. You need to align the time to get the distance between the speeches (represented as vector sequences). The comparison of the concept of dynamic alignment by spectral sequence of words has been used to solve the problems. The problem of speech recognition is a current problem today.

Most modern methods used to solve it require large computational resources, the amount of which is often limited. The impossibility of wide application of many algorithms today, for example, in mobile devices makes researchers look for more effective methods.

This article describes the algorithm and analysis of speech recognition methods, the identification of the shortcomings of each of them. Development of the program for speech recognition and experiment.

The speech recognition module includes two main digital signal processes: function extraction and function matching. The first one processes the word spoken by the user and generates its functions. The spoken speech is first converted into a digital domain, and the digital sampled speech is processed to extract functions using the MFCC approach, which evaluates the vocal track filter. In section I pre-processing of the speech signal is discussed, in section II we consider the selection of speech characteristics using the MFCC algorithm , the architectures of the automatic speech recognition system are considered in section III.

II. TASK DESCRIPTION: GENDER DETECTION

Voice gender detection aims to automatically detect the author's gender through audio signals. Similarly, speaker identification is to distinguish the author's identity (name or ID) by analysis his/her audios.

Let $= x_1, x_2, \ldots x_n$denotes a series of audio signals as input. $G = g_1, g_2, \ldots g_n$is a binary vector of 0/1 for gender categories corresponding to the audio sig-nals X. Here, we use 1 to denote Female, and 0 for Male. $S = s_1, s_2, \ldots s_n$ denotes speaker's ID, we use unique number to distinguish speakers. The training element pairs for the two tasks can be define in the following: (X,G)=(x_1, g_1),…,(x_n, g_n) for gender detection. For those tasks, we use X as input and extract corresponding signal features then use different neural networks to train models for the gender detection and speaker identification.

III. SIGNAL FEATURE EXTRACTION

Like many speech processing tasks (speech recognition, etc), the first step is to extract features which can be used for identifying linguistic content contained in the audio signals and for discarding the background noise information. Mel Frequency Cepstral Coefficients (MFCC) [23] are the state-of-the-art features widely used in many speech processing applications. Before describing the MFCC, let us show an original audio signal shown in Figure 1. An original signal consists of thousands or millions of numbers, It can be considered as a very long vector which contains the linguistic content and noise.

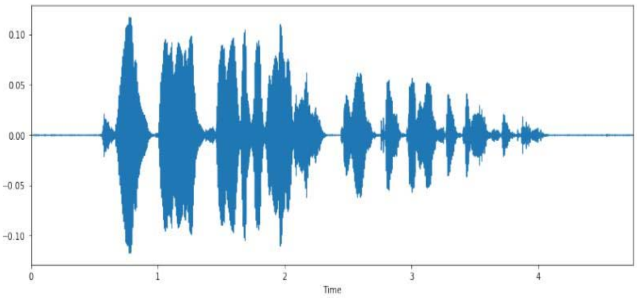


Fig. 1: An original audio signal.

In this work, we use MFCC to perform gender/speaker detection/identification, and the way of extracting MFCC feature is not the focus of this paper. In practice, we apply LibROSA[3], a python package for audio signal analysis. Its function of librosa.feature.mfcc was used for the extraction purpose of MFCC. The extracted features are shown in Figure 2.

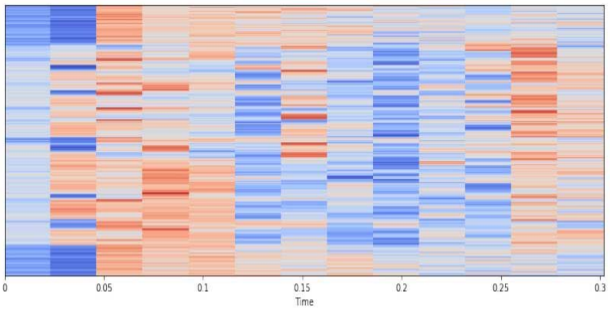


Fig. 2: MFCC feature of the audio signal.

In practice, we set the number of MFCC features to 40 then the dimension of an MFCC for audio is $M \in \mathbb{R}^{|40\times n|}$Max-min normalization is computed to each MFCC features and it refers to MFCC original in the following. We tried an alternative normalization, the z-scores, for MFCC features by the following calculation:

$$M^z = \frac{M-\mu}{std(M)} \qquad (1)$$

whereμ is the mean and $std(M)$ standard deviation.

One of the standard ways to handle the variable length of input is to find the max-length of the audios and padding its MFCC features with zero value if the length is less than max-length. One of the efficient way to solve variable length issue by following transformation:

$$M^g = M^z \times M^{z^T} \qquad (2)$$

where $M \in \mathbb{R}^{|40\times 40|}$, 40 is the number of MFCC features. Then we could apply the flatten operation to M^g or use its 2-D form.

IV. FEED-FORWARD NEURAL NETWORKS

To better describe the model, let us start by a simple neural network. As known, a single-layer perceptron [3, 1] is a NN with no hidden units, which only contains an input layer and an output layer. There is no non-linear feature extraction, which means the outputs are computed directly from the sum of the product of weights corresponding to the input. We use the MLP, and it is an NNs composed of many perceptions and MLP can learn or extract non-linear features. Generally speaking, MLP consists of an input layer, some number of hidden layers, and an output layer. Figure 3 shows the general architecture of the MLP. It can be seen that the NN has input, several hidden layers and the output layer.

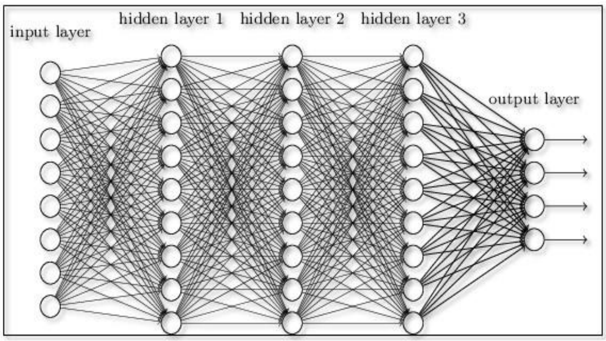


Fig. 3: Architecture of NN for the gender/speaker detection.

Convolutional neural networks (CNN) [4, 2, 6] are a specialized kind of neural network for processing the data with 2-D grid-like topology. CNN has been tremendously successful in practical applications. Unlike MLPs, which uses fullyconnected layers to extract features, CNN leverages two important ideas that can help improve the model: sparse interactions and parameter sharing. The former is a feature extraction process with a smaller kernel than the input. For example, Gender Detection and Speaker Identification with Neural Networks 5 when processing audio, the input signals might have thousands or millions of numbers, instead of feed such a long vector to NN, CNN can detect small and meaningful features by capturing local information. Parameter sharing refers to using the same parameter for the smaller kernel sliding on a 2-D input. A typical CNN consists of three stages: i) use convolution layers to perform a set of linear activation. ii) each linear activation run through a non-linear activation function. iii) use pooling function to modify the output of the layer further. Figure 4 shows the general architecture of CNN. It can be seen that this architecture composed of an input layer, two convolutional layers, two max-pooling layers and a fully-connected layer followed by an output layer. For gender detection and speaker identification, we train CNN models with such architecture.

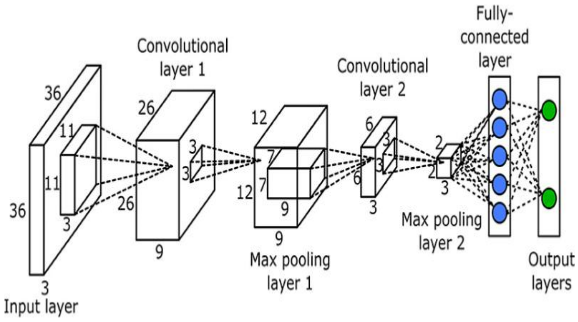


Fig.4: Architecture of CNN for the gender/speaker detection.

A. Models and Algorithms

Gender detection identification process

$$h_i^1 = \sigma^{(1)}\left(\sum_j w_{ij}^{(1)} x_j^{(1)} + b_i^{(1)}\right) \quad (1)$$

$$h_i^2 = \sigma^{(2)}\left(\sum_j w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}\right) \quad (2)$$

$$\hat{y}_i = \sigma^{(3)}\left(\sum_j w_{ij}^{(3)} h_j^{(2)} + b_i^{(3)} \qquad\right) \quad (3)$$

$\sigma(z)$- non-linear activation function;

h_i– hidden layer;

x_j – signal input;

$\hat{y}$ - model's output;

$\theta = \left\{w_{ij}^{(1)}, w_{ij}^{(2)}, w_{ij}^{(3)}, b_i^{(1)}, b_i^{(2)}, b_i^{(3)}\right\}$ – model's parameters;

$\sigma^{(1)}(x)$ және $\sigma^{(2)}(x)$ are recti_er activation functions:

$$\sigma^{(1)}(x) = \sigma^{(2)} = x^{+} = \max(0, x) \quad (4)$$

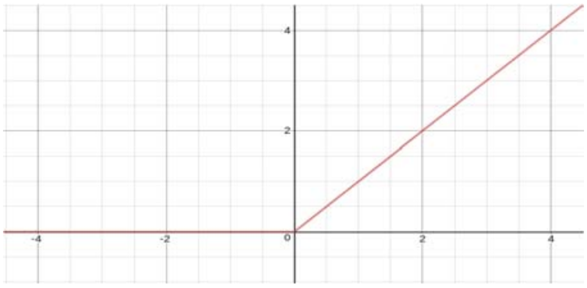


Figure 1: Plot of the Recti_er.

$\sigma^{(3)}(z)$ - is softmax functions,, $i = 1, \ldots, K$ and $z = (z_1, \ldots, z_k) \in R$;

$$\sigma^{(3)}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (5)$$

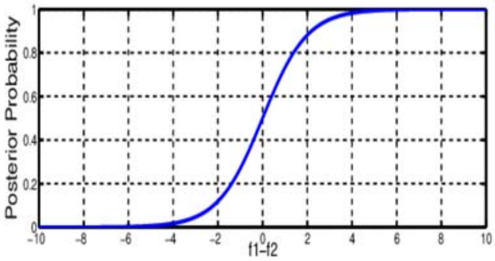


Figure 2: Plot of the softmax.

Back-Propagation:

$$E_i = \frac{1}{2}\sum_{j=1}^{K}\left(\hat{y}_{ij} - y_{ij}\right)^2 \quad (6)$$

$$\frac{\partial E_i}{\partial w_{ij}^3} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^{(3)}(z)} \cdot \frac{\partial z^{(3)}}{\partial w_{ij}^3} \quad (7)$$

$$\frac{\partial E_i}{\partial w_{ij}^2} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^3(z)} \cdot \frac{\partial z^3}{\partial h_j^{(2)}} \cdot \frac{\partial h_j^{(2)}}{\sigma^{(2)}(z)} \cdot \frac{\partial z^{(2)}}{\partial w_{ij}^{(2)}} \quad (8)$$

$$\frac{\partial E_i}{\partial w_{ij}^1} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^3(z)} \cdot \frac{\partial z^{(3)}}{\partial h_j^{(2)}} \cdot \frac{\partial h_j^{(2)}}{\sigma^{(2)}(z)} \cdot \frac{\partial z^{(2)}}{\partial h_j^{(1)}} \cdot \frac{\partial h_j^{(1)}}{\partial \sigma^{(1)}(z)} \cdot \frac{\partial z^{(1)}}{\partial w_{ij}^{(1)}}$$
(9)

Update

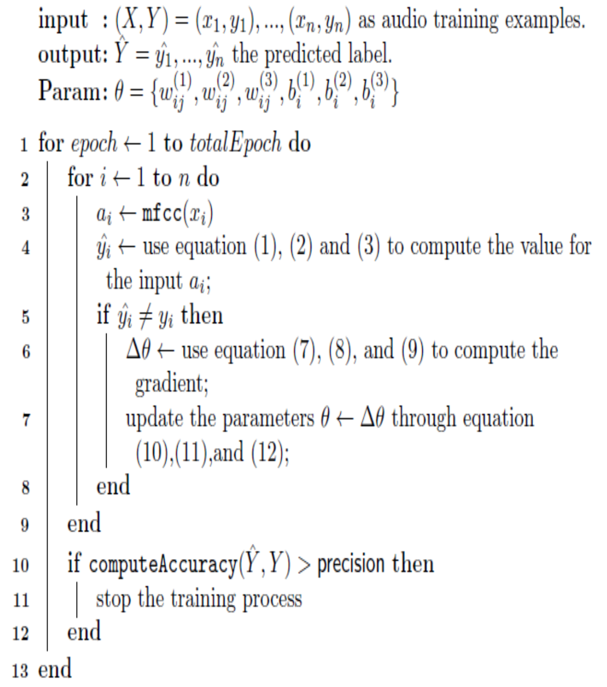
$$\Delta w_{ij}^{(3)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(3)}} \quad (10)$$

$$\Delta w_{ij}^{(2)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(2)}} \quad (11)$$

$$\Delta w_{ij}^{(1)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(1)}} \quad (12)$$

Algorithm

The algorithm for determining gender specificity is shown below



```
input  : $(X, Y) = (x_1, y_1), \ldots, (x_n, y_n)$ as audio training examples.
output: $\hat{Y} = \hat{y}_1, \ldots, \hat{y}_n$ the predicted label.
Param: $\theta = \{w_{ij}^{(1)}, w_{ij}^{(2)}, w_{ij}^{(3)}, b_i^{(1)}, b_i^{(2)}, b_i^{(3)}\}$

for $epoch \leftarrow 1$ to $totalEpoch$ do
    for $i \leftarrow 1$ to $n$ do
        $a_i \leftarrow$ mfcc($x_i$)
        $\hat{y}_i \leftarrow$ use equation (1), (2) and (3) to compute the value for
          the input $a_i$;
        if $\hat{y}_i \neq y_i$ then
            $\Delta\theta \leftarrow$ use equation (7), (8), and (9) to compute the
              gradient;
            update the parameters $\theta \leftarrow \Delta\theta$ through equation
              (10),(11),and (12);
        end
    end
    if computeAccuracy($\hat{Y}, Y$) > precision then
        stop the training process
    end
end
```

Algorithm gender identification

V. EXPERIMENTAL SETUP

We conducted a series of experiments to evaluate the MLP and CNN models for gender detection tasks. The experiments were designed to evaluate the MLP and CNN models training process with its final test results under the different model setup configuration. Precision, recall, F1-score and accuracy metrics are reported for model evaluation.

A. Data Sets

Table 1 show the statistics of the data sets for gender detection. It can bee seen that there are total 1125 and 300 audios in training and test set for gender detection task. The number of audios for Male and Female are 600 and 525 in training set. The remaining 150 audios for Male and Female test set.

Table 1: Data sets for gender detection.

Data sets	#Male	# Female	#Total
Train	600	525	1125
Test	150	150	300

VI.RESULTS

We run small and large MLP and CNN models for gender detection and speakers identification. Below we report the results of precision, recall and F1-score for both two tasks independently. Unless stated otherwise we refer to the F1-score when comparing model performances.

A. Gender Detection

Table 5 shows the results of gender detection corresponding to Male and Female categories, and the overall macro results are also shown. First of all, for gender detection, we can confirm the results of CNN-based models are better than MLP-based models. It can be seen, the CNN-small model beats the model of MLP-large and the MLP-large has a large number of trainable parameters than former's.

In order to know how the parameter setups affect the model performances, we compare the results of the same architectures but with different model parameter sizes: MLP-small vs. MLP-large and CNN-small vs. CNN-large. In MLP-based model's results, the MLP-large F1-score of Male has slight improvement (around 1%) over the small ones. This is not the case for Female category, the F1-score of MLP-large drops distinctly around 6% when the model size was increased. One possible explanation of the phenomenon is that the number of Female training audios is less than the Male ones, it can be seen from Table 1. As a result, the MLP-large cannot outperform the MLP-small models. From the figures 6 (top) and 7 (top), we can compare the training and testing process of both two MLP-large and MLP-small models. As we can see from the curves, those models were trained around 500 epochs on the same data sets, both models' training accuracy close to 1, which indicates the models were well trained for the given training examples, no significant signs of over-fitting and under-fitting be found. It can be seen from the curves of the testing process that the test results are fluctuating during the training process. Since we do not have the development set, for the final test results, instead of choosing the best test results among all test results obtained from the training process, we report the results of the last model to shows the models real generalization.

CNN-small vs. CNN-large comparison shows a significant improvement for both Male and Female gender detection. It is very clear from the over macro avg. column of Table 2. Compare to MLP-based models, CNN shows different results, MLP does not gain improvements with increased model size, but CNN does. From the neural architecture point of view, we can explain these results like the MLPs process an audio signal as a long MFCC real value vectors, but the CNN process more natural MFCC features which is a matrix with a shape of (length, 13), then uses several convolution/max-pooling layers to extract the higher abstract features. From the comparisons are presented in Figure 3, we could observe that CNN-based model takes fewer epochs to convergence for gender detection, MLP takes 500 epochs in general, and CNN takes 50 only.

Table 2. Recall and F1-score results for Gender detection with respecr to categories of Male and Female. Macro avg.denotes for the overall macro Precision, Recall and F1-score.

Models	Male			Female			Marco avg.		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
MLP-small	53.71	62.66	57.84	55.2	46	50.18	54.45	54.33	54.01
MLP-large	52.04	68	58.95	53.84	37.33	44.09	52.94	52.66	51.52
CNN-small	63.29	**79.33**	70.41	**72.32**	54	61.83	67.80	66.66	66.12
CNN-large	**73.57**	68.86	**71.03**	70.62	**75.33**	**72.09**	**72.09**	**72**	**71.96**

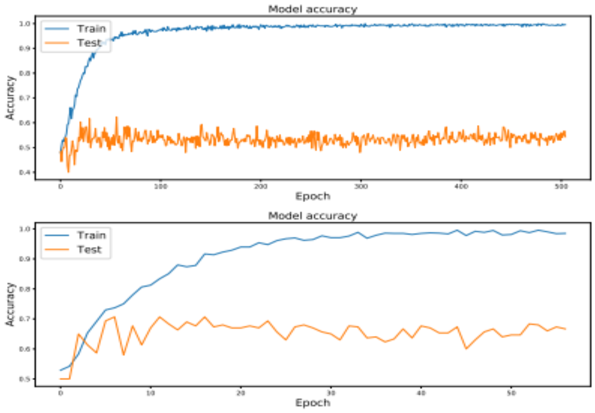
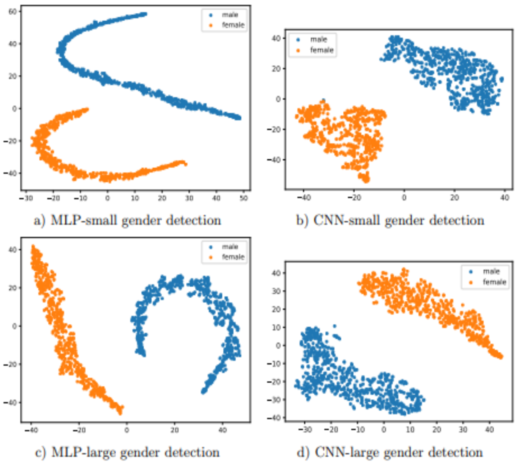


Fig. 3: Accuracy curve of training and test processes for Gender detection with small MLP (top) and CNN (bottom) models.

B.Visualization

Figure 4 show the visualization of the training audios after the model training for gender detection. We extract the feature vector from the layer before the output layer of each model as a representation of the input audio. It can be seen from Figure 10, the audios are classified into two classes: Male and Female. The visualization results of MLP and CNN are different, MLP's result shows obvious strong cohesiveness of the audios examples compared to CNN's. The CNN visualization results also shows the case and the audios are distributed over a wide area compared to MLP's. As experimental results show that CNN outperforms MLP for both gender detection. It can be seen, MLP locates the same speaker audios into a small area, but CNN's are located in a wide area, and these visualization results may indicate CNN may have better generalization than MLP.



a) MLP-small gender detection b) CNN-small gender detection

c) MLP-large gender detection d) CNN-large gender detection

Fig. 4: Visualization of the training audios after training the MLP and CNN models for Gender detection.

VI. CONCLUSION

In this paper, we demonstrated that convolutional neural networks can achieve consistent improvements over the feed-forward neural networks for both gender detection. The experiments include two neural architectures: MLP and CNN with different model setups. CNN outperforms MLP for both two tasks, the relative gains for gender detection range from around 10% to almost 20%. Several aspects are remarkable about this result.

First, the comparison results of the same architecture have shown that the large model setup for MLP of gender detection does not seem to help much. CNN-large seems to perform well or better than a small CNN.

Second, CNN uses convolution layers to extract features from signals, which makes the model able to have input with the original shape of MFCC features. MLP consists of several fully connected-layers which cannot process a feature with matrix, must turn the original MFCC features into a long feature vector. This could be an explanation of why CNN performs better than MLP for both tasks.

Third, the comparison of the training process of MLP and CNN shows that the former usually takes a large number of epochs, and the letter one takes remarkably fewer epochs. The visualization results show that MLP locates the same speaker audios into a small area, but CNNs are located in a relatively wide area, and these results may indicate CNN have better generalization than MLP.

Acknowledgements. IRN-AR05131207 "Development of the technology of multilingual automatic speech recognition using deep neural networks".

References

[1] Auer, P., Burgsteiner, H., Maass, W.: A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Netw. 21(5), 786-795 (Jun 2008)

[2] Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th Interna-tional Conference on Machine Learning. pp. 160-167. ICML'08, ACM, New York, NY, USA (2008)

[3] Cunningham, P., Delany, S.: k-nearest neighbour classifiers. Mult Classif Syst (04 2007)

[4] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques" Issue 3, March 2010.

[5] Watcher, M. D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., "Template Based Continuous Speech Recognition", IEEE Transaction on Audio, Speech, & Language Processing, 2007.

[6] Gupta, R., and Sivakumar, G., "Speech Recognition for Hindi Language", IIT BOMBAY, 2006.

[7] Ingle V., Proakis J. Digital Signal Processing Using Matlab V4 – Boston: ITP, 1997.

[8] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.

[9] Barsky A.B., Neural networks: recognition, management, decision-making.

[10] Cheong Soo Yee and abdul Manan ahmad, Malay Language Text Independent Speaker Verification using NN-MLP classifier with MFCC, 2008 international Conference on Electronic Design.

[11] Wu Junqin, Yu Junjun, "An Improved Arithmetic of MFCC in Speech Recognition System," IEEE Transaction on Audio Speech processing, and Language, pp.719-722, 2011.

[12] Gurpreet Kaur, Harjeet Kaur, "Multi Lingual Speaker Identification on Foreign Languages Using Artificial Neural Network with Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[13] L. Rabiner and G. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, 1993.

[14] H. Hasegawa, M. Inazumi, "Speech Recognition by Dynamic Recurrent Neural

Networks," Proceedings of 1993 International Joint Conference on Neural Networks.

[15] Furui, S., 50 years of progress in speech and speaker recognition. SPECOM 2005, Patras, 2005: pp. 1-9.

[16] BabaAli B., Wójcik W., Mamyrbayev O., Turdalyuly M. «Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction» SIGMA-NOT, Przeglad Elektrotechniczny, № 94(6), 2018

[17] Orken Mamyrbayev, Keylan Alimhan, Bagashar Zhumazhanov, Tolganay Turdalykyzy, Farida Gusmanova. End-to-End Speech Recognition in Agglutinative Languages. Asian Conference on Intelligent Information and Database Systems. ACIIDS 2020: Intelligent Information and Database Systems pp 391-401.

[18] Ashkan Tashk, Jürgen Herp, Esmaeil Nadimi, Automatic Segmentation of Colorectal Polyps based on a Novel and Innovative Convolutional Neural Network Approach, WSEAS Transactions on Systems and Control, Volume 14, 2019, pp. 384-391