Neural architectures for gender detection and speaker identification

Orken Mamyrbayev\textsuperscript{1,2*}, Alymzhan Toleu\textsuperscript{1}, Gulmira Tolegen\textsuperscript{1} and Nurbapa Mekebayev\textsuperscript{1,2}

Abstract: In this paper, we investigate two neural architecture for gender detection and speaker identification 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/speaker-specific features automatically. The experimental results reveal that the models using z-score and Gramian matrix transformation obtain better results than the models only use max-min normalization of MFCC. In terms of training time, MLP requires large training epochs to converge than CNN. Other experimental results show that MLPs outperform CNNs for both tasks in terms of generalization errors.

Subjects: Computer Science; Artificial Intelligence; Computer Science; General

Keywords: MLP; CNN; gender detection; speaker identification

ABOUT THE AUTHOR

Orken Mamyrbayev, PhD, Associate Professor, head of the Laboratory of computer engineering of intelligent systems at the Institute of Information and Computational Technologies. In 2002 he received Master's degree at the Kazakh National Pedagogical University named after Abay in specialty Computer Science. In November 2014, he received the PhD degree in Information systems at the Kazakh National Technical University named after K. I. Satpayev. He is a member of the dissertation council “Information Systems” at L.N. Gumilyov Eurasian National University in the specialties Computer Sciences and Information systems. His research team received 1 patent for an invention and 5 copyright certificates for an intellectual property object in software. Currently, he manages 2 scientific projects, such as the Development of multilingual automatic speech recognition technology using deep neural networks and Methods and models for searching and analyzing criminally significant information in unstructured and poorly structured text arrays.

PUBLIC INTEREST STATEMENT

Voice gender and speaker identification are interesting and challenging tasks. In this paper, we present two approaches for both tasks by using the various architecture of the neural networks: multi-layer perceptron (MLP) and convolutional neural network (CNN).

These models are compared with each other in terms of their training time, model performance, model size, and generalization errors. Generally, people would think(expect) that CNN should be better than MLP because CNN is invented later than MLP. Indeed, from a single run experiment, we see that CNN can outperform the MLP in terms of their prediction accuracy. But through the 500 times experiments, we notice that MLP is much better than CNN in terms of generalization error. In this work, we also applied different techniques for audio signal transformation and tested them for both two tasks.
1. Introduction

Automatically detecting gender and identifying speakers through a speaker’s voice is an important task in the audio signal processing area. Gender detection deals with finding out whether a speech spoken by a male or a female. This task is very crucial for gender-dependent automatic speech recognition (ASR), which let the ASR system be more accurate than gender-independent systems. Speaker recognition is the process of automatically recognizing the speakers on the basis of individual information carried in the speech wave, which can be categorized into speaker identification and speaker verification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker identification, on the other hand, is the process of determining which registered speakers provide the input speech.

In general, gender detection and speaker identification can be viewed as classification problems, in which the former classifies the input audio into two categories and the latter classifies the input audio into the number of registered speakers. Many approaches have been proposed for gender classification, the most commonly used methods are decision tree (Naeem et al., 2013), support vector machine (SVM) (Lian & Lu, 2006), Bayesian network (Darwiche, 2010), K-Nearest neighbor (Cunningham & Delany, 2007) and Random forest (Ho, 19).

As known that there are many features can be used for audio signal processing, namely by MFCC, intermediate vectors (i-vectors) (Lee, Tsao, Wang, & Jeng, 2014), energy entropy, short-time energy and spectral centroid, etc. Gaussian Mixture Models (GMMs) with MFCC (Reynolds, Quatieri, & Dunn, 2000) is the traditional way of solving the task of speaker identification. This framework has been extended to use i-vector (Lee et al., 2014) and joint factor analysis (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010) to form compact representation for an utterance which has specific voice-related characteristics. In a recent decade, deep learning starts to dominate the different areas of artificial intelligence, such as machine learning (Du, Lee, Li, Wang, & Zhai, 2018; Qu, Bengio, & Tang, 2019), natural language processing (Mamyrbayev et al., 2019; Schmidhuber, 2015; Toleu, Tolegen, & Makazhanov, 2017a, 2017b), computer vision (Girdhar, Carreira, Doersch, & Zisserman, 2019; Pang, Zha, Cao, Shi, & Lu, 2019) and audio signal processing (Li et al., 2017; Villalba, Brummer, & Dehak, 2017), etc.

In this paper, we use different neural architectures for both gender detection and speaker identification. We apply multi-layers perceptron (MLP) (Li, Chen, Shi, Tang, & Wang, 2017; Youse, Youse, Fathi, & Fogliatto, 2019) and convolutional neural networks (CNNs) (Liew, Khalil-Hani, Radzi, & Bakhteri, 2016; Mamyrbayev et al., 2019) in order to able to learn gender/speaker-specific features from original MFCC vectors which do not cover specific voice-related characteristics of the speech signal. Another aim of this work is to compare two neural architectures’ performance for both tasks with different settings/ways: 1) different feature transformation, 2) different model size, 3) adding the noise signal to test set to measure the model’s generalization error. We evaluate our methods on Kazakh Speech corpus (Mamyrbayev et al., 2019). The experimental results show that MLP outperform CNN in terms of generalization error and MLP requires large training epochs to converge than CNN for both tasks.

The rest of the paper is organized as follows: Section 2 introduces the task description with notations. Section 3 illustrates the signal feature extraction process. Section 4 describes our approach for both tasks. Section 5 reports the experimental outcomes. Section 6 summarizes the results of this work as a conclusion.

2. Task description: gender detection and speaker identification

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 $X = x_1, x_2, \ldots, x_n$ denote 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 signals $X$. 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 defined in the following: i) $(X, G) = (x_1, g_1), \ldots, (x_n, g_n)$ for gender detection; ii) $(X, S) = (x_1, s_1), \ldots, (x_n, s_n)$ for speaker identification. For those two 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.

3. 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) (Sahidullah & Saha, 2012) 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.

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, 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.

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:

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

where $\mu$ is the mean and $\text{std}(M)$ standard deviation.

---

**Figure 1.** An original audio signal.

**Figure 2.** MFCC feature of the audio signal.
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 ways to solve variable length issue by following transformation:

\[ M^g = M^z \times M^z \]  

Where \( M^g \in \mathbb{R}^{40\times40} \), 40 is the number of MFCC features. Then we could apply the atten operation to \( M^g \) or use its 2-D form.

### 4. Models

Neural networks (NN) can be considered as a classifier function with parameters, and a neural network with several layers can be seen as a composition of functions defined as follows:

\[ f_\theta(\cdot) = f_\theta^{(1)} \left( f_\theta^{(1)}(\cdot) \right) \]  

where \( \theta \) denotes parameters of NN and \( l \) is the number of the layers. In the following, we describe our two neural network architectures for gender and speaker detection tasks: i) feed-forward NN, it is a multi-layer perceptron and refers to MLP; ii) convolutional NN, it refers to CNN.

#### 4.1. Feed-forward neural networks

To better describe the model, let us start by a simple neural network. As known, a single-layer perceptron (Auer, Burgsteiner, & Maass, 2008; Freund & Schapire, 1999) 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.

#### 4.2. Convolutional neural networks

Convolutional neural networks (CNN) (Collobert & Weston, 2008; Krizhevsky, Sutskever, & Hinton, 2012; Van den Oord, Dieleman, & Schrauwen, 2013) 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 fully-connected 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, 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.

### 5. Experiments

We conducted a series of experiments to evaluate the MLP and CNN models for gender and speaker detection tasks:

- the first experiment is designed to analyze the effectiveness of extracted features from MFCC, which are tested for both models. As described in section 3, we use two types of features and compare them, i) the normalized-flattened original MFCC as a long feature vector, and ii) using z-score on MFCC, then turn it into the Gram matrix (equation 2).
in order to compare two models effectively, we test the trained two types of models 500 times by adding different noise (normal distribution with zero mean and one variance) to the test set.

- visualization of the audios after the model training for both tasks.

For model evaluation, we report precision, recall, F1-score results under the different model setup configuration.

### 5.1. Data sets

We use the data-set from the study (Mamyrbayev et al., 2019). Tables 1 and 2 show the statistics of the data sets for gender and speaker recognition. It can be seen that there are total 855 and 570 audios in training and test set for gender detection task. The number of audios for Male and Female are 448 and 407 in training set. The remaining 570 audios for Male and Female as test set.

Table 2 shows the training and test sets for speaker identification. It can be seen that there are 19 speakers in total, and each of them has around 60 audios for training. It is a quite small training set and not able to train any data-hungry deep learning models well. But in this work, for speaker identification, we use this data set to train MLP and CNN models and do the evaluations.

### 5.2. Model setup

The neural architectures of MLP and CNN with small and large models are tested in the experiments. We train two versions of our models to assess the trade-off between performance and size.

| Table 1. Data sets for gender detection |
|----------------------------------------|
| Data sets | #Male | # Female | #Total |
| Train | 488 | 407 | 855 |
| Test | 302 | 268 | 570 |

| Table 2. Data sets for speaker identification. The number of each speaker’s audios are shown below at corresponding each Speaker ID |
|-------------------------------------------------------------------------------------------------------------------------------------|
| Data sets | Speaker ID | #Total |
| Train | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 | 1140 |
| Test | 12 12 19 18 17 15 12 9 16 16 12 14 10 14 15 18 21 16 19 | 285 |

| Table 3. Hyper-parameters of MLP used for both tasks |
|-----------------------------------------------------------------|
| MLP | Small | Large |
| Layers | Layer-1 | Layer-2 | Layer-3 | Layer-1 | Layer-2 | Layer-3 |
| Hidden units | 128 | 64 | 32 | 512 | 256 | 128 |
| Dropout | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| Activation function | Relu | Relu | Relu | Relu | Relu | Relu |
The architectures of small and large models are summarized in Tables 3 and 4. The main hyperparameter is the number of hidden units which we set to 128, 64, 32 and 512, 256, 128 for small and large models, respectively. We use Relu as the activation function for all layers and the dropout is set to 0.15.

### 5.3. Results

Figure 3 displays the distribution of audios with MFCC features and the features after z-score normalization and Gramian matrix transformation. Figure 3(a,c and e) are the distribution of audio with original MFCC, after z-score normalization and Gramian matrix transformation for gender’s data-set respectively. Similarly, Figure 3(b,d and f) are the audio’s distribution for the speaker’s data-set. It can be seen that the distribution of original MFCC for gender and speaker’s audios is not separated by gender or speakers or we can say that they are mixed together and hard to distinguish the gender or speaker audios/After the z-score transformation, the distribution of audios becomes denser than the original ones. More unexpectedly, the audios after z-score and Gramian matrix transformations, the distribution of audios for gender and speakers, are changed. From Figure 3(e,f), we can observe that the audio signal for gender and speaker are more distinguishable, and a few audios are mixed. It turns out that If we train/test models on this data-set with such distribution, it is easy to reach 100% accuracy.

Table 5 lists the results for gender and speaker recognition with different features (L and G) and model (small and large) settings. It can be observed that the model trained/tested after Gramian transformation (denoted as G in the Table) gave 100% F1-score, which confirms the results mentioned above. Comparing the results of models that use long-flattened MFCC vector with max-min normalization (denoted as L in the Table 5), we can see that fully-connected MLP outperforms CNN for gender detection. Figure 4 shows the training curve of those models for gender detection, and it can be seen that MLP tends to need more training epochs to converge than CNN (Figure 4(a,b)). Here, we only show the small-sized models, and the situation for the large one is similar. From Figure 4, it can be seen that MLP and CNN with Gramian matrix (G) takes relatively small training epochs than L.

Let us move on to the results of speaker identification in which we also used different feature settings and the model size. Table 6 shows the results, and we can see that the model using the feature L obtained relatively worse outcomes no matter which models we use. F-score of the models using L in the range of 19% to 36%. The Gramian matrix transformation seems to raised up model performance distinctly. F-score of both models almost reach 99%, which are higher than the model trained with L.

Figure 5 shows the training curve for speaker identification. As we can see that the model with L takes large training epochs to converge; in contrast, the model with G takes fewer training steps. Another issue that can be found in this figure a, b is that the training accuracy of the model with L reaches 99% gradually, the validation results remain at 20% level.
5.4. Generalization ability

In order to compare the models’ generalization error, we conducted another experiment: added the noise signal to test set, each of them was normally distributed with zero mean and one variance, and the already trained models with G were tested 500 times. Figures 6 and 7 depict how the F-score of the modes are distributed. It can be seen that for gender and speaker recognition, the CNNs cannot outperform MLPs when adding different normal distributed noisy signals to the test set. One of the possible reasons for that each layer of MLP is fully connected, and the input interacts with each other in more dimensions. In contrast to this, the CNNs have a convolution layer which commonly treated as a region feature extractor that uses a slide 2-D window with a specified step and a shared weight on an input to extract region features, and the inputs interact only in the regions. The integration between one input region
Table 5. Results for gender recognition with different features: L denotes for using the normalized fattened log mfcc vector; G denotes for using the z-score and Gramian matrix transformation. P is precision, R is recall, and F1 is F1-score.

| Models   | Male | Female | Macro, avg |
|----------|------|--------|------------|
|          | P    | R      | F1         | P    | R      | F1         |
| L MLP-small | 79.56 | 85.09 | 82.24 | 75.37 | 80.67 | 80.23 |
| L MLP-large | 77.22 | 92.05 | 83.98 | 69.40 | 77.82 | 80.72 |
| G MLP-small | 100  | 100   | 100    | 100  | 100   | 100    |
| G MLP-large | 100  | 100   | 100    | 100  | 100   | 100    |
| L CNN-small | 79.68 | 82.16 | 81.36 | 69.40 | 76.11 | 79.64 |
| L CNN-large | 76.57 | 88.74 | 81.20 | 84.54 | 78.01 | 79.68 |
| G CNN-small | 100  | 100   | 100    | 100  | 100   | 100    |
| G CNN-large | 100  | 100   | 100    | 100  | 100   | 100    |
Figure 4. Training curve of MLP and CNN for gender detection.

![Training curve of MLP and CNN for gender detection.](image)

Table 6. Results for speaker identification

| Models | Macro, avg. |
|--------|-------------|
|        | P    | R    | F1   |
| L      |      |      |      |
| MLP-small | 20.76 | 20.36 | 19.94 |
| MLP-large | 21.76 | 21.92 | 21.50 |
| G      |      |      |      |
| MLP-small | 98.66 | 98.57 | 98.58 |
| MLP-large | 99.36 | 99.28 | 99.30 |
| L      |      |      |      |
| CNN-small | 36.78 | 35.68 | 35.25 |
| CNN-large | 36.77 | 36.51 | 36.16 |
| G      |      |      |      |
| CNN-small | 99.17 | 99.09 | 99.12 |
| CNN-large | 99.17 | 99.47 | 99.47 |

Figure 5. Training curve of MLP and CNN for speaker identification.

![Training curve of MLP and CNN for speaker identification.](image)
with another one is not captured, except when we choose the step size small enough. Another possible reason is that MLPs have more trainable parameters than CNNs since MLPs have fully connected layers and CNNs have shared weights for convolution layers. As a result, it can be seen that CNNs gives a large generalization error than MLPs for gender and speaker identification.

5.5. Visualization

Figure 8 shows the visualization of the test set after model training for gender and speaker identification of both models with different feature forms (L and G). We use the trained models on the test set, and obtain the output of the intermediate layer as representations of the audios then use t-SNE (Tur, Hakani-tur, & Oazer, 2003) to visualize. Figure 8(a–c) and d show the visualization results for gender detection, it can be seen that the MLP seems better classifying the audios into two classes: male and female. One of the criteria to evaluate the clustered results is to see two distances: intra-class and inter-class distance. The former is to measure the distance between elements in a class and the smaller distance indicates the better results. The latter is to measure the distance between different class and the large distance indicates the better results.

Comparison of Figure 8(c,d) show that MLP can classify the audios into a more dense class than CNN, it can be seen that the intra-class distance is smaller than CNN's. Figure 8(e–g) and h show the visualization results for speaker identification. Figure 8(g,h) show the results for model with G, and as we can see that MLP for speaker identification have large inter-class distance than CNN's.
6. Conclusion

In this paper, we applied different neural network architectures for gender detection and speaker identification tasks. The comparisons of two neural networks were made in different ways: 1) various features transformation (L and G), 2) different model sizes (small and large), and 3) adding noise signal to test set to measure the model’s generalization errors (tested 500 times). The results show that for both tasks, the two types of neural networks obtain relatively better results after applying the z-score and the Gramian matrix transformation. In terms of training time, the MLP requires more training epochs to converge when only use the max-min normalization to MFCC features. The model size does not have a significant effect on the model’s performances, and the large models only give a minor improvement. Another comparison result shows that the MLPs outperform CNNs in these experiments in terms of generalization error. The visualization results show that MLPs can classify the audios into more dense classes than CNNs for both tasks.
Acknowledgements
This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan. IRN APOS131207 Development of technologies for multilingual automatic speech recognition using deep neural networks.

Funding
The authors received no direct funding for this research.

Author details
Orken Mamyrbayev1,2
E-mail: morkenj@mail.ru
ORCID ID: http://orcid.org/0000-0001-8318-3794

Alymzhan Toleu1
E-mail: alymzhan.toleu@gmail.com
ORCID ID: http://orcid.org/0000-0001-9246-319X

Gulmira Tolegen1
E-mail: gulmira.tolegen.cs@gmail.com

Nurbapa Mekebayev1,2
E-mail: nurbapo@mail.ru

1 Institute of Information and Computational Technologies, Almaty, Kazakhstan.
2 al-Bari Kazakh National University, Almaty, Kazakhstan.

Citation information
Cite this article as: Neural architectures for gender detection and speaker identification, Orken Mamyrbayev, Alymzhan Toleu, Gulmira Tolegen & Nurbapa Mekebayev, Cogent Engineering (2020), 7: 1727168.

References
Auer, P., Burgsteiner, H., & Maass, W. (2008, Jun). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Networks: the Official Journal of the International Neural Network Society, 21(5), 786–795. doi:10.1016/j.neunet.2007.12.036

Collobert, R., & Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 160–167. ICML’08, ACM, New York, NY (2008).

Cunningham, P., & Delany, S. (2007, April). k-nearest neighbour classifiers. Multiple Classifier System, 1–17.

Darwiche, A. (2010, December). Bayesian networks. Communications of the ACM, 53(12), 80–90. doi:10.1145/1859204.

Dekok, N., Penny, P. J., Dekok, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing, 19(4), 1–11.

Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2018). Gradient descent nds global minima of deep neural networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, 1–45. CoRR abs/1811.03804.

Freund, Y., & Schapire, R. E. (1999, Dec). Large margin classification using the perceptron algorithm. Machine Learning, 37(3), 277–296. doi:10.1023/A:1007662407062

Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A.: Video action transformer network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: California (June 2019).

Ho, T. K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition. 1, 278–283. ICDAAR’95, Montreal, Quebec, Canada, IEEE Computer Society, Washington, DC(1995).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in neural information processing systems, Massachusetts Institute of Technology Press, 25, 1097–1105. Curran Associates, Inc.

Lee, H. S., Tsoa, Y., Wang, H. M., & Jeng, S. K.: Clustering-based i-vector formulation for speaker recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Singapore, 1101–1105 (January 2014).

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., & Zhu, Z. (2017). Deep speaker: An end-to-end neural speaker embedding published. In ArXiv 2017, CoRR abs/1705.02304.

Li, L., Chen, Y., Shi, Y., Tang, Z., & Wang, D. (2017). Deep speaker feature learning for text-independent speaker verification. Published in ArXiv 2017, CoRR abs/1705.03670.

Lian, H. C., & Lu, B. L. (2006). Multi-view gender classification using local binary patterns and support vector machines. In J. Wang, Z. Yi, J. M. Zurada, B. L. Lu, & H. Yin (Eds.), Advances in neural networks - ISNN 2006 (pp. 202–209). Heidelberg: Springer Berlin Heidelberg, Berlin.

Liew, S. S., Khalil-Hani, M., Radzi, F., & Bakhteri, R. (2016, March). Gender classification: A convolutional neural network approach. Turkish Journal of Electrical Engineering and Computer Sciences, 24, 1248–1264. doi:10.3906/elk-1311-58

Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekov, A., & Turdalykyz, T. (2019). Automatic recognition of kazakh speech using deep neural net-works. In N. T. Nguyen, F. L. Gao, T. P. Hong, & B. Trowinski (Eds.), Intelligent information and database systems (pp. 465–474). Cham: Springer International Publishing.

Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Mukhsina, K., Alimhan, K., BabaAli, B., ... Akhmetov, B. (2019, January). Continuous speech recognition of kazakh language. ITM Web of Conferences, 24, 01012. doi:10.1051/tmconf/20192401012.

Naem, M., Khan, A., Qureshi, S. A., Riaz, N., Zul Kar, S., & Bhutto, A. R. (2013). Gender classification with decision trees. International Journal of Signal Processing, Image Processing and Pattern Recognition 6(2), 165–176. February, 2013.

Pang, B., Zhao, K., Cao, H., Shi, C., & Lu, C.: Deep rnn framework for visual sequential applications. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: California, 423–432. (June 2019).

Qu, M., Bengio, Y., & Tong, J. (2019). GMNN: Graph markov neural networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. CoRR abs/1905.06214.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. In Digital signal processing, Academic Press, London, UK, 19–41 (2000) doi:10.1006/dsp.1999.0361

Sahidullah, M., & Saha, G. (2012, May). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Communication, 54, 543–565. doi:10.1016/j.specom.2011.11.004

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, Manno-Lugano, Switzerland,
Toleu, A., Tolegen, G., & Makazhanov, A.: Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2, 666–671. Association for Computational Linguistics, Vancouver, Canada (July 2017a).

Toleu, A., Tolegen, G., & Makazhanov, A.: Character-based deep learning models for token and sentence segmentation. In: Conference: 5th International Conference on Turkic Languages Processing (TurkLang 2017). Kazan, Tatarstan, Russian Federation (October 2017b).

Tur, G., Hakkani-tur, D., & Oazer, K. (2003, June). A statistical information extraction system for Turkish. Natural Language Engineering, 9(2), 181–210. doi:10.1017/S135132490200284X

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), Advances in neural information processing systems. 26, 2643–2651. Curran Associates, Inc.

Villalba, J., Brummer, N., & Dehak, N. (2017, August). Tied variational autoencoder backends for i-vector speaker recognition, INTERSPEECH 2017, August 20–24, 2017, Stockholm, Sweden, 1004–1008.

Youse, M., Youse, M., Fathi, M., & Fogliatto, F. (2019, October). Patient visit forecasting in an emergency department using a deep neural network approach. Kybernetes Ahead-of-print, 46, 643–651.