









Automatic Recognition of Kazakh Speech Using Deep Neural Networks

Orken Mamyrbayev¹ , Mussa Turdalyuly¹ ,
Nurbapa Mekebayev² , Keylan Alimhan¹ ,
Aizat Kydyrbekova² , and Tolganay Turdalykyzy¹ 

¹ Institute of Information and Computational Technology,
050010 Almaty, Kazakhstan

morkenj@mail.ru, mkt_001@mail.ru

² al-Farabi Kazakh National University, 050040 Almaty, Kazakhstan

Abstract. This article presents a deep neural network (DNN) system based on automatic speech recognition for Kazakh language, developed using the Kaldi speech recognition tool. DNNs are initialized using the restricted Boltzmann machines (RBM) and are trained using cross-entropy as the objective function and the standard back propagation of error. In order to achieve optimal results, the training has been modified based on peculiarities of Kazakh language. A 76 hours-corpus has been used in training. Results are compared for two different sets of values between classical models and various DNN settings.

Keywords: DNN · ASR · Kazakh speech recognition · LM

1 Introduction

The creation of natural-language man-machine interfaces and, in particular, automatic speech recognition systems has recently become one of the main areas and tasks in the field of artificial intelligence. Speech technologies provide a more natural user interaction with computing and telecommunications complexes compared to the standard graphical interface.

With the development of personal computers and a wide range of public information and entertainment services, speech and then multimodal interfaces are now more focused on application in social intelligent services, which imposes its conditions on speech processing systems. In particular, the vocabulary of lexical units increases, the variability of speech increases, and processing should be carried out in real time to maintain a natural dialogue with the user. The development of a compact way of presenting the dictionary is especially relevant for agglutinative languages with a relatively rich morphology. To take into account the variability and learning models of phonemes and words require huge text and speech materials, the preparation of which requires meticulous expert work.

In [1], three types of speech disruptions that are most characteristic of spontaneous speech were analyzed: (1) voiced pause, (2) repetition of words, (3) modification of the sentence from the very beginning. As a material, speech corpus Spoken Dutch Corpus (CGN) and Switchboard-1 were used. The number of voiced pauses made up 3% of all

lexical units in these corpora. Most often these were interjections, and they were located in all parts of sentences. The relative number of repetitions was approximately 1%. And the twenty most frequent repetitions are short words consisting of one syllable.

In [2], an audiovisual detector of voiced pauses was used to filter unwanted speech failures in multimedia recordings of lectures. The recorded multimedia corpus of lectures lasting about 7 h contained an image of the tablet computer screen, on which the lecturer made handwritten notes displayed to the audience on a multimedia projector, as well as a sound stream with the lecturer's speech and background noise. The analysis of the body showed that the vast majority of hesitations occurs when the lecturer does not use a tablet (or pad, data tablet), so a two-stage algorithm was used to filter the pauses. First of all, the moments of time were determined when the image on the monitor screen did not change, and then only during these periods of time the search for filled pauses in the audio stream was carried out. In the analysis, voiced pauses with a duration of more than 120 ms, pronounced in isolation (i.e., those containing segments with silence before and after hesitation), as well as within a word, were considered. The use of preliminary segmentation of the audio segments and the video analysis from a tablet helped to increase the recognition accuracy of the hesitations up to 85%.

In the rapid development of speech technology is associated with the development of artificial neural networks and is now becoming increasingly popular research on the use of DNN for recognition of Kazakh speech. At the same time, there are no effective systems of automatic recognition of Kazakh speech at the moment and the development of ASR is relevant.

In this article, we consider the method of creating an automatic speech recognition system using DNN using Kaldi tools. In this study, the existing speech corpus was expanded, the speech and text corpus for the Kazakh language was assembled, and acoustic and language models were created on the basis of neural network (NN), which allowed to increase the accuracy of recognition of Kazakh speech.

For speech preprocessing, we used the following algorithms: Mel-cepstral coefficients (MFCC) and Perceptual Linear Prediction Coefficients (PLP). For acoustic modeling, it uses the hidden Markov model (HMM), the Gaussian distribution mixture model (GMM), Subspace Gaussian mixture model (SGMM) and deep neural networks (DNN). Language modeling is performed using finite transducers (FSTs) with linear algebra support — the BLAS and LAPACK libraries.

The paper is organized as follows. Section 2 describes the work on the relevant scientific research area. Section 3 discusses data preprocessing methods Sect. 4 describes the methodology for automatic speech recognition. Section 5 describes the DNN architecture and Sects. 6, 7 discuss the results of the experiment and the conclusion.

2 Related Works

Currently DNN is often used in speech research for speech recognition and the results of the research show good results. For example, studies [3] present a system of recognition of spontaneous Czech, Slovak and Russian speech for processing interviews of Holocaust witnesses. In this paper, basic transcriptions were created automatically using a specific set of rules, and for many words several transcription variants were generated to

take into account the phonetic phenomena of continuous speech (for example, assimilation of consonants at the word boundary). Then transcriptions describing spoken pronunciation variants were created, as well as for the Russian language and accent, as interviews were taken not only from the residents of Russia, but also from Russians living in Ukraine, Israel, and the USA. In addition, non-speech phenomena were modeled. The size of the corpus used to create acoustic models for the Russian language was 100 h and DNN was used. The language model was a bigram model using the return method (Katz's backing-off scheme). With a dictionary size of 79 thousand transcriptions, the percentage of incorrectly recognized words was 38.57%.

Another class of speech recognition applications is shorthand.

Most often, with such a task, some monologue is carried out, recorded in fairly good acoustic conditions using a headset microphone. Therefore, in contrast to systems of mass service, where speech comes through telephone channels and/or is recorded on the street, automatic transcribing systems receive a speech signal with a much better recording quality. Since there are softer requirements for recognition speed, the system can process the speech signal in several passes, using the methods of adaptation to the voice of the speaker and the applied problem [4].

Scientists from Russia conducted a study on the recognition of continuous Russian speech using DNN confidence), described in [5]. A method using finite state machine-based converters was used for speech recognition. It was shown that the proposed method allows to increase the accuracy of speech recognition in comparison with hidden Markov models.

The study [6] compares the language models constructed using a feedforward neural network and a recurrent neural network. Three different implementations of the language model on neural networks were used: (1) LIMSI software tools for creating a feedforward neural network in which the output layer is limited to the most frequent words; (2) feedforward neural network with clustering (the entire dictionary is used); (3) recurrent neural network with clustering. Experimental results showed that language models built using a feedforward neural network, work worse than recurrent neural networks. On the test data, the recurrent network showed an improvement of 0.4% compared to the use of a feedforward neural network.

3 Speech Preprocessing

The conversion of input data into a set of features is called feature extraction. Speech recognition efficiency deteriorates dramatically in the presence of noise due to spectral mismatch between training and testing data. With conventional MFCC feature extraction, the logarithm function is used for the energy of the Mel filter Bank to reduce their dynamic range. Root cepstral analysis replaces the logarithmic function with a constant root function and gives the RCC coefficients. The coefficients of the RCC showed the best resistance to noise. In the RCC method, the compressed speech spectrum is calculated as shown in (1):

$$L_d(n) = L(n)^\tau, 0 \leq \tau \leq 1 \quad (1)$$

where $L_d(n)$ - compressed spectrum, $L(n)$ is the original spectrum, τ is the compression ratio, and m - filter bank index. Feature extraction involves simplifying the amount of

resources required to accurately describe a large dataset. The feature extraction was performed using 13 MFCC coefficients [7].

Therefore, relation (1) expands, as shown in (2):

$$L_d(n) = L(n)^{\tau(m)}, 0 \leq \tau(m) \leq 1 \quad (2)$$

where the compression ratio depends on the frequency band and is called non-uniform spectral compression. We show that by incorporating a speech recognition system into the process of adjusting the compression ratio, the recognition rate is further improved.

4 The Proposed Automatic Speech Recognition System

The methodology of our work is as follows:

4.1 Construction of Experimental Speech Corpus

Over the past ten years, a number of speech corpuses have been created in the world, containing up to a thousand speakers recorded in various environmental conditions. The recording of acoustic data for the creation of an acoustic corpus of the language was carried out at the Institute of Information and Computational Technologies of the Scientific Committee of the Ministry of Education and Science of the Republic of Kazakhstan in Almaty. For this, a sound-proofing, professional recording studio from Vocalbooth.com was used (Fig. 1). The cabin for recording audio data consists of two noise insulation layers, with the same hermetic door. The interior design consists of a pyramid-shaped sound-absorbing acoustic material of red color and the cabin is equipped with a silent air exchange system. The studio is designed to record high quality audio data.



Fig. 1. Soundproof professional recording studio of the company Vocalbooth.com (Color figure online)

The recorded audio materials have been preserved with expansion .wav. Each sentence was saved as a separate file, and the name consisted of the following identifiers:

<Region_code> + <gender> + <birth_year> + <initials_name> + <education_code> + <text_number> + <sentence_number_in_text>

For example, speaker from the Almaty region, named Mamyrbayev Orken, male, born in 1979, with a higher education, voiced the text number 5 and sentence 82, will be identified as 05M79OM3_T005_S082.

All audio materials have the same characteristics:

- file extension: .wav;
- method of converting to digital form: PCM;
- discrete frequency: 8 kHz;
- digit capacity: 16 bits;
- number of audio channels: one (mono).

As speakers, people were selected without any problems with the pronunciation of speech. For research purposes and further use of the data, the speakers were surveyed according to a previously created template (Fig. 1). 200 speakers of different ages (age from 18 to 50 years) and genders were used for recording. It took an average of 40–50 min to sound and record one speaker. For each speaker, a text consisting of 100 sentences was prepared. The sentences were recorded in separate files. Each sentence consists of an average of 6–8 words. Sentences are selected with the most rich phoneme of words. Text data were collected from news sites in the Kazakh language, and other materials were used in electronic form. In total 76 h of audio data were recorded. During recording, transcriptions were created – a description of each audio file in a text file. The corpus created gives us, firstly, work with large volumes of databases, checking the proposed system characteristics and, secondly, studying the effect of database expansion on the recognition rate.

4.2 Acoustic Model

The acoustic model $p(x|w)$ provides conditional probability of a sequence of feature vectors x , given a sequence of words w has occurred. This can be thought of as a measure of acoustic similarity of the input features to a sequence of words, regardless of the grammatical correctness of that word sequence. For an ASR system each word may be represented by a sequence of sub-word units called acoustic states. During acoustic model training, the statistics of each state are calculated from the occurrences of feature vectors corresponding to that state. For ASR with very large vocabulary sizes of thousands of words, due to data sparsity it is not feasible to accumulate sufficient statistics for each word separately. We would like to recognize even those words which may have few or no occurrences in the training data. To alleviate this problem, the words are defined as sequences of phonetic units called phonemes, just like the word pronunciations are represented in language dictionaries. Such a representation based on sub-word units is called a pronunciation lexicon. Each word in the lexicon may have one or more pronunciations.

4.3 Kazakh Language Model

Language model - allows determining the most likely word sequences. The complexity of constructing a language model depends largely on the specific language. So, for the English language, it is enough to use statistical models (the so-called N-grams). For agglutinative languages with a relatively rich morphology, statistical models are not suitable and hybrid models are used.

The language model $p(w)$ gives the prior probability of the word sequence w . Basically it shows how likely a word sequence is to be uttered, based on grammatical rules of a language. Since this model only depends on the text and is independent of acoustic data, therefore large amounts of text available on the books, journal, articles etc., can be used as input source. Additionally, we want the language model to capture the topic-specific information for special ASR systems. To capture certain characteristics associated with human speech e.g. certain grammatical errors common in speaking, repetitions, hesitations etc., transcriptions of spoken text are also a useful input data source. Since the total number of possible word sequences is unlimited, simplifying assumptions have to be made to have reliable non-sparse estimates. The standard way to calculate the language model probabilities is through accumulation of counts of neighbouring words. It assumes that the probability of current word w_n depends only on the previous $m-1$ words $w_{n-1} \dots w_{n-m+1}$.

A speech recognition involves a number of different components such as feature extraction, acoustic modeling, language modeling and DNN, as shown in the Fig. 2.

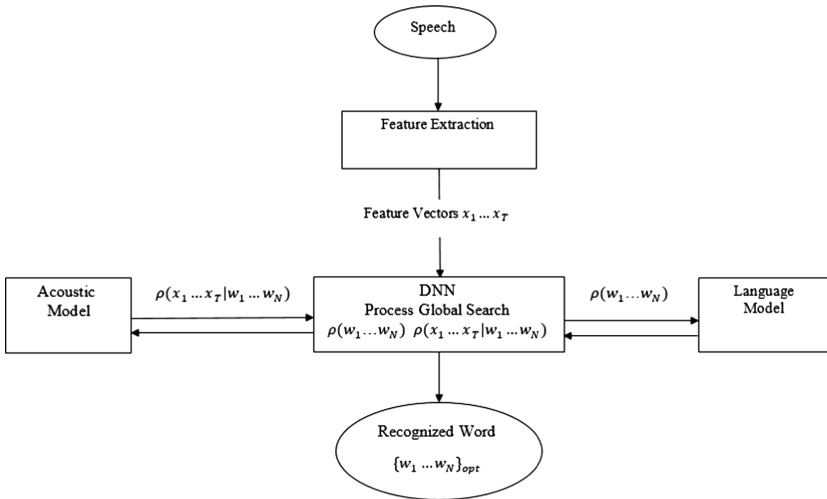


Fig. 2. Overview of an ASR system

5 The DNN Architecture and Training

For the development of ASR, we used the Kaldi tool and the DNN library in it, the modified Karel Vesely setting on the CUDA graphics processor was used for training.

We consider the DNN model:

First layers L

$$v^l = f(z^l) = f(W^l v^{l-1} + b^l), \text{ for } 0 < l < L,$$

where $z^l = W^l v^{l-1} + b^l \in R^{N_l \times 1}$, $v^l \in R^{N_l \times 1}$, $W^l \in R^{N_l \times N_{l-1}}$, $b^l \in R^{N_l \times 1}$, and $N_l \in R$ respectively, the excitation vector, the activation vector, weight matrix, displacement vectors and the number of neurons in layer l. $v^0 = 0 \in R^{N_0 \times 1}$ - observation vector, $N_0 = D$ that is the element size and $f(\cdot) : R^{N_l \times 1} \rightarrow R^{N_l \times 1}$ activation function in relation to the excitation vector elementwise. In most applications, the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

or hyperbolic tangent function

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

is used as an activation function. Next, we consider the algorithm for this model [8].

Algorithm Direct DNN calculation.

```

1: procedure ForwardComputation(O)
    > Each column O is an observation vector.
2:    $V^0 \leftarrow O$ 
3:   for l ← 1; l < L; l ← l + 1 do    > L total number of
    layers
4:      $Z^l \leftarrow W^l V^{l-1} + B^l$     > Each column  $B^l$  is  $b^l$ 
5:      $V^l \leftarrow f(Z^l)$     > f(.) may be sigmoidal tanh, ReLU,
    other functions
6:   end for
7:    $Z^L \leftarrow W^L V^{L-1} + B^L$ 
8:   if regression then    > regression task
9:      $V^L \leftarrow Z^L$ 
10:  else
11:     $V^L \leftarrow softmax(Z^L)$ 
12:  end if
13:  Return  $V^L$ 
14: end procedure

```

During training, we use the algorithm of single-stage selection by the Monte-Carlo method in the Markov chain. RBM have Gauss-Bernoulli units and is trained at an initial learning rate of 0.01 and other RBM have Bernoulli-Bernoulli units. Training was not controlled, the number of iterations was 4, the number of hidden layers was up to 6 and the number of units per layer was up to 2048.

6 Experimental Results

In the course of this work, feature extraction methods such as MFCC and acoustic, language model, DNN were investigated. The results were evaluated by the word error rate (WER) for classical models. The results indicating the vertical axis are percentages, and the horizontal axis: training monophonic models (Mono), the passage of the first (Tri1) and second (Tri2) and third (Tri3) thyryphon (Fig. 3). The best result is 36.76% WER for SAT Training.

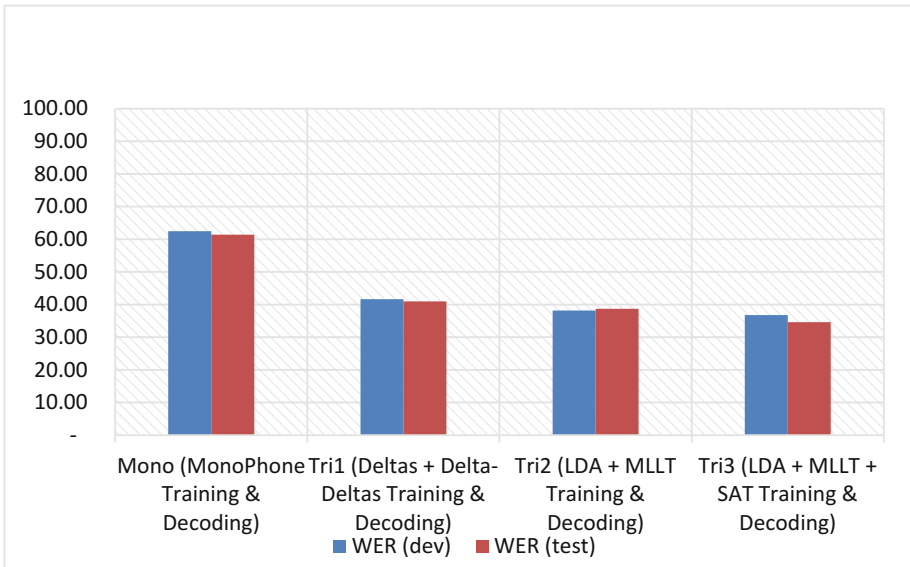


Fig. 3. The rule set of the classic model.

The results obtained with the application of DNN using from 0 to 6 hidden layers. The optimal result of 32.72% WER was obtained for 6 hidden layers, and this was an improvement over the classical models (Fig. 4).

It is important to note that performance is improved when the volume of the corpus for training is large. The best results have been obtained using DNN and SMBR algorithm.

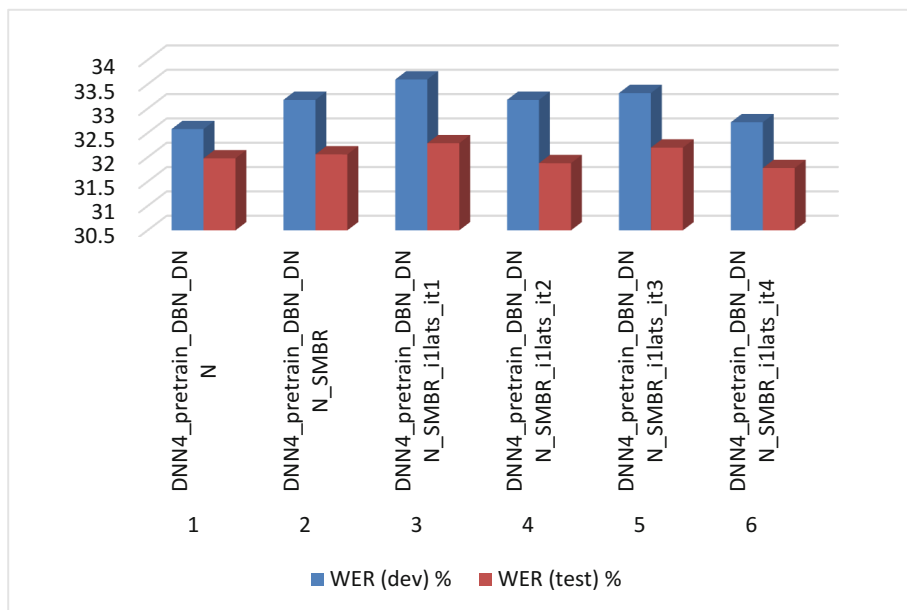


Fig. 4. Results of different DNNs.

7 Conclusion and Future Work

In this article we have developed and implemented a system of automatic speech recognition of Kazakh speech, which works on the basis of DNN. According to the results of the study we can see that it is better to use DNN for automatic speech recognition than classical algorithms. The paper analyzed the existing models and methods and considered a speech compression algorithm using the MFCC algorithm and gave an example of the ASR architecture. In this regard, it was stated that the MFCC and DNN methods provided the best results. The test error rate reached 0.56% for the corpus with 76 h of speech.

Future work will focus on improving the training corpus and exploring different optimization approaches for designing and implementing ASR for real-time applications such as voice-controlled robots.

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.

References

1. Stouten, F., Duchateau, J., Martens, J.-P., Wambacq, P.: Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation. *Speech Commun.* **48**, 1590–1606 (2006)
2. Tsiaras, V., Panagiotakis, C., Stylianou, Y.: Video and audio based detection of filled hesitation pauses in classroom lectures. In: *Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, 24–28 August 2009, pp. 834–838 (2009)
3. Psutka, J., Ircing, P., Psutka, J.V., Hajič, J., Byrne, W.J., Mirovsky, J.: Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In: *Proceedings of Eurospeech, Portugal, Lisboa, 4–8 September 2005*, pp. 1349–1352 (2005)
4. Young, S., et al.: *The HTK Book (for HTK Version 3.4)*, Cambridge, UK, 375 p. (2009)
5. Karpov, A., Kipyatkova, I., Ronzhin, A.: Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: *Proceedings INTERSPEECH-2011*, Florence, Italy, pp. 3161–3164 (2011)
6. 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*, pp. 135–140 (2014)
7. Behbahani, Y.M., Babaali, B., Turdalyuly, M.: Persian sentences to phoneme sequences conversion based on recurrent neural networks. *Open Comput. Sci.* **6**, 219–225 (2016)
8. Yu, D., Deng, L.: *Automatic Speech Recognition*, p. 315. Springer, London (2014). <https://doi.org/10.1007/978-1-4471-5779-3>