Ensuring the best quality and performance of modern speech technologies, today, is possible based on the widespread use of machine learning methods. The idea of this project is to study and implement an end-to-end system of automatic speech recognition using machine learning methods, as well as to develop new mathematical models and algorithms for solving the problem of automatic speech recognition for agglutinative (Turkic) languages.

Many research papers have shown that deep learning methods make it easier to train automatic speech recognition systems that use an end-to-end approach. This method can also train an automatic speech recognition system directly, that is, without manual work with raw signals. Despite the good recognition quality, this model has some drawbacks. These disadvantages are based on the need for a large amount of data for training. This is a serious problem for low-data languages, especially Turkic languages such as Kazakh and Azerbaijani. To solve this problem, various methods are needed to apply. Some methods are used for end-to-end speech recognition of languages belonging to the group of languages of the same family (agglutinative languages). Method for low-resource languages is transfer learning, and for large resources – multi-task learning. To increase efficiency and quickly solve the problem associated with a limited resource, transfer learning was used for the end-to-end model. The transfer learning method helped to fit a model trained on the Kazakh dataset to the Azerbaijani dataset. Thereby, two language corpora were trained simultaneously. Conducted experiments with two corpora show that transfer learning can reduce the symbol error rate, phoneme error rate (PER), by 14.23 % compared to baseline models (DNN+HMM, WaveNet, and CNC+LM). Therefore, the realized model with the transfer method can be used to recognize other low-resource languages

Keywords: ASR, transfer learning, end-to-end, low-resource language, connectionist temporal classification, attention

UDC 004.934
DOI: 10.15587/1729-4061.2022.252801

IDENTIFYING THE INFLUENCE OF TRANSFER LEARNING METHOD IN DEVELOPING AN END-TO-END AUTOMATIC SPEECH RECOGNITION SYSTEM WITH A LOW DATA LEVEL

Orken Mamyrbayev
PhD, Associate Professor,
Deputy General Director in Science*
Keylan Alimhan
Doctor of Science Degree
in Mathematical Sciences, Professor
Department of Mathematical and Computer Modeling
L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010008
Dina Oralbekova
Corresponding author
Researcher**
E-mail: dinaoral@mail.ru
Akbayan Bekarystankyzy
Researcher**
Bagashar Zhumazhanov
Software Engineering*
*Laboratory of Computer Engineering of Intelligent Systems
Institute of Information and Computational Technologies
Shevchenko str., 28, Almaty, Republic of Kazakhstan, 050010
**Department of Cybersecurity,
Information Processing and Storage
Satbayev University
Satpaev str., 22a, Almaty, Republic of Kazakhstan, 050013

Received date 12.01.2022
Accepted date 23.02.2022
Published date 28.02.2022

How to Cite: *Mamyrbayev, O., Alimhan, K., Oralbekova, D., Bekarystankyzy, A., Zhumazhanov, B. (2022). Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. Eastern-European Journal of Enterprise Technologies, 1 (9 (115)), 84–92. doi: https://doi.org/10.15587/1729-4061.2022.252801*

1. Introduction

Automatic speech recognition systems began to develop dynamically with the rapid development of computing technologies. Accurate results have been achieved in the field of speech recognition, with many models and methods used in commercial applications, justifying their use in these directions. Among the commercial applications for speech recognition, first is the introduction of call centers or interactive voice response (IVR) systems for automatic access to information, speech chatbots, etc. Call centers have implemented intelligent voice assistants that generate user questions in natural language, and the response is synthesized by the system in the user's language.

Primary automatic speech recognition systems consist of three modules: decoding, acoustic, and language models. The modular subsystem for speech recognition mainly consists of independent modules, and even the acoustic model depends on the HMM model as well as GMM models, which, in many cases, correspond to the pronunciation unit [1].

Breakthroughs in artificial intelligence have made it possible to improve the quality of speech recognition and

obtain accurate speech recognition results. In many research works, advanced recognition systems have been developed using deep learning methods. Machine learning is a popular method of speech recognition, where the HMM-DNN architecture and hybrid architecture of GMM-DNN were used to model the dynamic characteristics of speech signals. The end-to-end approach has a theoretical advantage using a hybrid DNN approach: they are trained to minimize cross-entropy between the predicted and actual HMM states so that optimization is performed in one step in order to improve speech recognition accuracy. Some studies used deep neural networks [2] to build acoustic models, and others used RNNs [3] to build language models with excellent results. Analyzing the research work of scientists, it was determined that to solve the speech recognition problem, neural networks can be used at all stages of recognition.

With the development of software and computing technologies, it is possible to successfully implement deep learning technology for speech recognition using end-to-end methods, as shown in many research papers. To develop an end-to-end model, it is necessary to eliminate all potential assumptions from the entire speech recognition system and build a single model optimized at the sequence level [4]. In the end-to-end method, all modules are trained simultaneously. This method can also be used to directly train the automatic speech recognition system, that is, without manual work with raw signals. The method under consideration in this study was developed using several recurrent and convolutional layers that function as an acoustic and language model, displaying speech inputs in transcriptions. Thus, the end-to-end method can receive the raw speech signal as the input and generate conditional phoneme class probabilities as the output. Currently, there are two types of architecture of the end-to-end method, namely, the connectionist temporal classification (CTC) and attention-based encoder-decoder models.

In the development of the automatic speech recognition system, as before, attention is paid to end-to-end methods; many studies have proven that performance and accuracy increase with an increase in the amount of data used for training. For example, in published studies, the best results in training big data were obtained by end-to-end systems based on CTC [5, 6] and attention-based encoder-decoder models. In end-to-end models, all parameters are calculated by the gradient descent method, which is easily influenced by the structure of the neural network. The models in question require less memory, which makes it possible to use them locally on mobile devices. End-to-end systems are trained with large corpora, but still do not reach the current level of performance. We assumed the following in this study: language uniqueness and multitasking for recognition; end-to-end models, in most cases, are not sufficiently trained; and models need a large amount of training data to be properly trained.

From the above, you can see the main problem, it concerns the recognition of low-resource languages, such as Kazakh, Kyrgyz, Azerbaijani, Tatar, Turkish, etc. The listed languages are included in the group of agglutinative languages or the Turkic group of languages. There are no large training data corpora for agglutinative (Turkic) languages, and other languages have corpora such as TIMIT, WSJ, LibriSpeech, AMI, and Switchboard that have over a thousand hours of training data.

To solve the above problems in this area, researchers proposed joint architectures, such as recurrent neural networks (RNNs) combined with a conditional random field (CRF) [7] and joint CTC-attention systems [8]. The architectures under consideration have the advantages of each submodel and model, and they introduce more explicit and strict restrictions, but we improved the quality and performance of end-to-end systems in this study. We think that for an end-to-end system, introducing complex computational layers into a model can better use correlations in both the time and frequency domains; a model with a large number of parameters is more difficult to train, and a data-based learning method without involving expert knowledge can be a vulnerability.

To date, the following methods have been used for end-to-end speech recognition of languages belonging to a group of languages of the same family (Turkic-speaking): transfer learning for low-resource languages and multitasking learning (MTL) for large resources [9, 10].

In this work, we constructed an end-to-end model with transfer learning aimed at recognizing Kazakh and Azerbaijani languages and solving problems related to limited speech resources. In our previous studies, we proved that for an end-to-end model without integrating language models, good results can be obtained [11].

Transfer learning is a method to adapt models trained on one dataset to another dataset. We thought that this approach would lead to the following improvements. Firstly, the use of the approach obtained from the Kazakh representation model would lead to a reduction in training time in comparison with training from scratch. Secondly, an end-to-end model trained using transfer learning requires less data for equivalent evaluation than models for the Azerbaijani language. Thus, we expected a decrease in the use of GPU memory, since we did not need to support gradients for all layers.

2. Literature review and problem statement

For Turkic languages, large and qualitatively annotated speech data are not available for teaching an end-to-end speech recognition system. There is a large demand for high-quality end-to-end speech recognition systems for these languages, which requires special methods to solve this issue.

In [9], two main end-to-end models were developed: connectionist temporal classification (CTC) and attention-based encoder-decoder models for recognizing Mandarin speech. During the study, the Chinese character was found to be a suitable unit for recognizing Mandarin speech. As a result of recognition, the attention model attained a CER of 35.2 % and the CTC model reached a CER of 35.7 %. Results show that encoder-decoder models based on the attention mechanism performed better than the CTC model in recognizing Mandarin speech, however, the process of extracting features from the incoming signal is not provided.

[10] presented an open-source platform for end-to-end speech processing called ESPnet. ESPnet mainly focuses on end-to-end automatic speech recognition and uses widely used dynamic neural network tools, Chainer and Py-Torch, as the main deep learning mechanism. Thus, it can be concluded that a number of experiments and comparisons with other works show that ESPnet achieves acceptable ASR performance.

One of the studies [12] presented the effects of transfer learning of end-to-end speech recognition systems based on deep neural networks. The original acoustic model was trained on a large corpus of call center phone records, and experiments showed that for all target training sizes, the

transfer models outperformed models trained only on the target data. The model that was trained using 20 hours of target data achieved a higher recognition accuracy than the original model by 7.8 %. The data obtained from experiments show that the transfer learning method had a good effect on the results of the Turkish language recognition system.

In [13], experimental results were reported for cross-lingual and multilingual network learning of eleven Romance languages for a total of 10.000 hours of data. The average relative increase over the baseline for monolingual learning was 4 % and 2 % for data-deficient and data-rich languages for cross-learning, respectively; and 7 % and 4 % for multilingual training, respectively. Here, the additional benefit of collaborative language learning on all data was achieved with doubling the training time, leading to good results.

Other researchers [14] discussed language-adversarial transfer learning. Adversarial learning was used to enable the common layers of the SHL model to learn more language-invariant features. Experiments were conducted on IARPA Babel datasets. The results showed that the target model trained using knowledge transferred from the adversarial SHL model achieves a 10.1 % relative reduction in word error rate compared to the target model trained using knowledge transferred from the SHL model. However, it can be seen that the model achieves good performance only by adding additional components, which makes the system heavier.

Another work [15] considered end-to-end acoustic modeling using convolutional neural networks (CNNs), where the CNN receives raw speech signals as input and estimates the conditional probabilities of the HMM state classes at the output. During the research and analysis of ASR in several languages and a variety of tasks, the following were proved:

1) the proposed approach consistently produces a better system with fewer parameters compared to the traditional approach of cepstral feature extraction followed by ANN training;

2) the difference from the conventional speech processing method in the proposed approach, the corresponding feature representations are studied by preprocessing the input source speech at the subsegment level (≈2 ms).

In particular, it was proven through analysis that the output layer is more legible than standard cepstral functions and can be transferred between languages and domains.

Many research works [16] used a variant of gradient descent optimization and mini-batch gradient descent. In this paper, four mini-batch selection strategies were constructed to represent variants of each function in the dataset for speech recognition tasks to improve the performance of a deep-learning-based speech recognition model. For this, gender-adjusted strategies and emphasis on the selection of mini-batches were proposed. Experiments showed that the proposed strategies perform better than the standard mini-batch sampling strategy.

In [17], a 335 h corpus for the Kazakh language was presented. As a result of the experiment, they showed that a sufficiently large training data set significantly improves the performance of a speech recognition system based on an end-to-end model compared to hybrid ones. Kazakh-speaking people usually include Russian words in their speech, since the Russian language is at the level of the state language. In this case, the recognition system may not work correctly in recognizing mixed speech.

The study [18] proposed a new method that takes a pre-trained model on the VoxForge dataset with 100 h of Russian speech and applies its knowledge as a basis for building its neural network within the framework of the transfer learning method. A 20 h corpus of Kazakh speech was also assembled to train the neural network. The trained model used two neural networks: LSTM and biLSTM. The results showed that the biLSTM model with an external Russian language model improved system performance, reducing LER to 32 %. The results obtained are not so good. This is most likely due to the fact that the Russian language is not agglutinative, word formation is very different from the Kazakh language.

In [19], an attention-based model is proposed for the automatic recognition of continuous Russian speech. The experiment used a small set of Russian speech, the total duration of which is more than 60 hours; using the proposed methods, the recognition accuracy was increased and the best performance was obtained in terms of speech decoding speed using the beam search optimization method. This work also presents an augmentation method that helps to increase the volume of the corpus, which will be useful for low-resource languages.

[20] discussed an insertion-based model (NAT). Insertion-based models solve the above mask prediction problems and can generate an arbitrary order in which the output sequence is generated. This model enhances the CTC by making it dependent on insert-based token generation without autoregression. They experimented with three publicly available benchmarks and the results were competitive with a strong autoregressive transformer with similar decoding conditions.

[21] constructed a hybrid architecture based on transformer-LSTM. The results showed that, in general, the model surpasses the conventional transformer ASR by 11.9 % in terms of WER, and this hybrid architecture offers much faster output compared to the LSTM and transformer architectures. The use of Transformer gives good recognition results, but it is very difficult to set up.

[7] provided a method of adding location to the attention mechanism to alleviate the problem of recognizing speech of different lengths. This method changes the attention mechanism, which prevents excessive concentration of attention on individual frames, which further reduced the PER to 17.6 %.

Other researchers [22] applied the deep neural network approach and used a corpus of a different type for linguistic enrichment of the language model. As a result of the experiment, the WER was 3.61 % and the optimal combination of architecture was obtained: this is a deep LSTMP with L2 regularization. Combining end-to-end methods than using them alone brings more improvements.

In [23], studies were conducted to improve the quality of universal automatic speech recognition (ASR) and pronunciation assessment. The idea was proposed based on research on computer resources for training pronunciation. The results of the work show that the problem of the approach to training pronunciation did not involve the entire phonological system.

The study [24] presented the development of a special ASR system, which specializes in the recognition of individual words of Spanish minimal pairs. The results show that the performance of this new ASR system is comparable to that of the conventional Google ASR system. The advantage of the new ASR is that it will allow future applications to obtain information about the quality of pronunciation at the phoneme level.

Based on the foregoing, it can be concluded that the combined use of end-to-end methods, as well as the use of transfer learning, improves the performance of the speech recognition system for low-resource languages with a limited data sample.

3. The aim and objectives of the study

The aim of the study is to develop an automatic speech recognition system for agglutinative languages such as Kazakh and Azerbaijani using the transfer learning method.

To achieve the aim of the study, the following objectives were set:

– to collect and develop the speech corpora with their transcriptions for the Kazakh and Azerbaijani languages for system training, create a corpus that includes all types of speech – prepared and spontaneous, and it must be taken into account;

– to implement the transfer learning method for training the system for speech recognition languages with limited data resources, evaluate the effectiveness of training, comparing the results obtained with the transfer model and basic models.

4. Materials and methods

4. 1. Bilingual training

To increase efficiency and quickly solve the problems associated with limited resources, we applied transfer learning in an end-to-end model. Although transfer learning is trained with multilingual data, it is more useful to use acoustic similarities from common layers. In this study, we thought that feature extraction from multilingual speech data would be an effective method to embed general acoustic knowledge into end-to-end models.In the first stage of the experiment, we trained several independent RNNs with common hidden layers using two language resources (Fig. 1). This figure depicts the input with two language data for transfer learning; the architecture of the automatic recognition system has common hidden layers with the output.

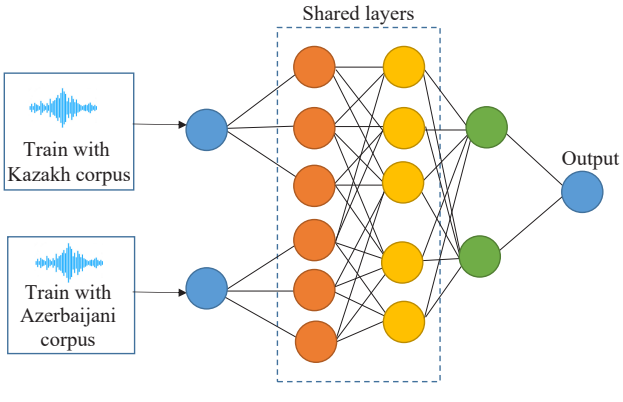

Fig. 1. Recurrent neural network with shared hidden layers

In our experiment, we used the maxout activation function with dropout training to avoid the problem of overfitting and identify the best common features.

During training, we trained two similar languages simultaneously. In the RNN, for each hidden layer, the output is described as follows:

$$x_n = y_n \times M_y,\ 1 \le n \le N, \tag{1}$$

where y_n is the output of each layer l for the nth frame, M_n is a vector filled with binary elements, and × denotes the dot product operation. Each element of the vector indicates whether the corresponding unit remains unchanged or not. The activation function for each hidden layer can be described as follows:

$$\max\left(w_1^T x + b_1, w_2^T x + b_2, \ldots\right), \tag{2}$$

where $w_1^T x$ is the activation output of layer l for the t-th frame and b_t is a vector of the same size.

Next, we performed a simple max pooling operation to average or calculate the maximum. The actual maximum output is calculated as follows:

$$y_n(i) = \max\left(s_n(k*i-2), s_n(k*i-1), s_n(k*i)\right),$$

$$1 \le n \le I, \tag{3}$$

where I is the number of output ones in each hidden layer, s_n is a vector consisting of these units, and k is pooling size. The RNN is linked with phonemes units generated by the GMM, while the inputs are classic low-level acoustic characteristics.

To extract low-dimensional features from the RNN, we first moved all parameters under the last hidden layer and added a new layer SoftMax with random parameters, and then adjusted the entire target RNN. Such adaptation of training, without destroying the structure of the neural network during training, allowed us to maintain maximum nonlinearity for subsequent processing.

4. 2. Feature extraction using the nonnegative matrix factorization method

Many studies have used MFCC, PLP, LPC, SVD, and PCA methods for feature extraction, and these methods produced good results. In this study, the application of such methods produced many redundant calculations and values. We used the NMF method to extract high-level features. Machine learning research related to the application of non-negative matrix factorization (NMF) has been proven to focus on non-negative elements in datasets that consist of text and speech, so they are easy to interpret. NMF is a dimensionality reduction method based on the low-rank approximation of the feature space. In addition to providing a reduction in the number of functions, NMF ensures that the functions are non-negative, producing additive models [25]. A detailed description of the method is given in [26].

Given the target weight matrix X with size $n \times m$, and all elements are positive integers $k < \min(m, n)$, NMF finds non-negative matrices W and H that minimize the norm of difference $X - WH$. W is a matrix whose columns contain the spectrum of voiced sounds, the matrix H can be interpreted as the time-dependent value of various signal characteristics that can be used as input. W and H are thus approximated non-negative factors of X. Fig. 2 shows the extraction of high-level features using NMF.

The features obtained using the NMF method are displayed in red, and P1, P2, P3, P4 are high-level features that are the minimums of high weight in Fig. 3. In this figure, we show the process of feature extraction. The resulting weights were chosen randomly and the weights changed with each training epoch. The weight matrix of a particular hidden layer was decomposed into two matrices according to the above-described process. We kept W and set H as the weight matrix of the new feature extraction layer. During the experiment, we determined that the features may have multilingual acoustic similarities caused by limited resources.

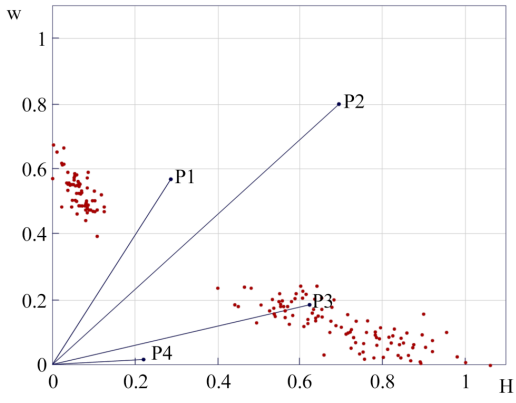

Fig. 2. High-level feature extraction using the negative matrix factorization method

4. 3. Joint application of connectionist temporal classification and attention

In this section, we present an end-to-end model for transfer learning based on the obtained high-level features. On the basis of previous works [27, 28], we built a joint end-to-end model based on two architectures, CTC and attention, in which the model has a single encoder and a combined decoder. The task was to transfer the monotonic constraint from the CTC to a decoder with an attention mechanism to improve the system performance after modeling.

As an intermediate process for our joint model, a network was considered to extract the features from the incoming signal. Thus, our extracted features were already high-level, and there was no need to map these original data to phonemes. In this work, we implemented our model using shallow bidirectional LSTMs [29].

We built an end-to-end model, as shown in Fig. 3, based on the hybrid CTC model and attention through the obtained high-level feature vectors.

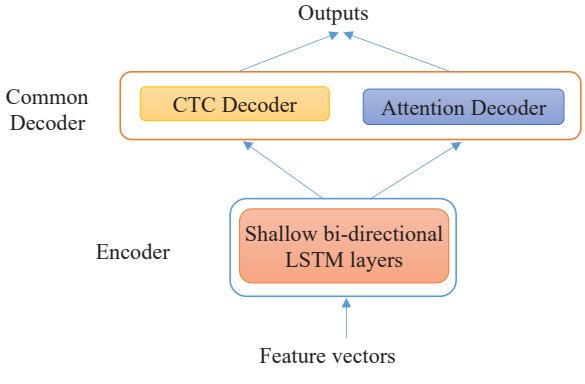

Fig. 3. The structure of the joint connectionist temporal classification and attention model

In the encoder component, the initial data of the symbol d_t at time t are determined at all inputs X:

$$P(d_t \mid X) = Soft\max(BiLSTM(X)). \quad (4)$$

After, we can form a probability distribution $P(S|X)$ over the input audio signal S under relatively independent conditions.

$$P_{CTC}(S \mid X) = \sum_{d \in N(S')} P(d \mid X) \approx \sum_{d \in N(S')} \prod_{t=1}^{T} P(d_t \mid X). \quad (5)$$

The model has three elements in the attention mechanism component. Bidirectional LSTM was used as the encoder. Location attention was used. Let $g_{k,t}$ be the attention weights combining the outputs of the kth encoder with the input of the t-th decoder. To calculate $g_{k,t}$, the previous weights g_{k-l}, I hidden outputs for the decoder l_{k-l}, and the outputs of the encoder h_t are calculated as:

$$f_n = F * g_{n-1}, \quad (6)$$

$$e_{k,l} = w^T \tanh\left(V^S s_n + V^H h_n + V^F f_{k,1} + b\right), \quad (7)$$

$$r_k = \sum_{t=1}^{T} g_{k,t} h_t, \quad (8)$$

$$P(s_k, \mid s_k, ..., s_{k-1}, X) = Decoder(r_k, l_{k-1}, h_t), \quad (9)$$

where F is a convolutional filter; g_n, T is the dimensional attention weight vector; w, V^S, V^H, and V^F are the adjustable weight parameters of multilayer perceptron; and r_k is a context vector needed to combine all encoder output sequences based on attention weights.

In our case, the posterior probability $p(S|X)$ of a model with an attention mechanism is formed without any conditional assumptions:

$$P(S \mid X) \approx \prod_k P(s_k \mid s_1, ..., s_{k-1}). \quad (10)$$

The attention mechanism model and the CTC loss function are defined as follows:

$$\begin{cases} J_{CTC} = -\ln P_{CTC}(S \mid X), \\ J_{attention} = -\ln P_{attention}(S \mid X). \end{cases} \quad (11)$$

To calculate the total loss function, it is necessary to take a combination of the CTC logarithmic linear function and attention:

$$J_t = \gamma J_{CTC} + (1-\gamma) J_{attention}, \ \gamma \in [0,1], \quad (12)$$

where γ is the CTC loss weight.

4. 4. Common decoder

To improve the quality of the recognition indicators, the CTC was included in a model with the attention mechanism. In this section, we more closely examine the joint decoder of the proposed model.

In the decoder, the attention mechanism determines the estimate of the assumption in the beam search:

$$g_{att}(b_l) = g_{att}(b_{l-1}) + \log P(z \mid b_{l-1}, X), \quad (13)$$

where b_l is a hypothesis with length l and z is the finite symbol of b_l.

Typically, the CTC decodes the output sequence framewise, while the attention model performs this process simultaneously. We decided to apply the CTC prefix probability and find an estimate of the CTC assumption:

$$g_{CTC}(b_l) = \log P(b_l, ..., \mid X). \quad (14)$$

After, we applied the single-pass decoding method [24] to connect the attention estimates and CTC; then, using λ, we

combined $g_{ctc}(b_l)$ and $g_{att}(b_l)$. This joint decoding produced the most possible audio sequence $\hat{S}$:

$$\hat{S} = \arg\max_{S} \left\{ \lambda g_{CTC}(b_l) + (1-\lambda) g_{att}(b_l) \right\}. \quad (15)$$

As can be seen, one part of our attention model is applied as a target model in speech recognition; in the other part of the model, the CTC loss function contributes to the target model in the decoding step.

Although many experiments were conducted with joint models of CTC+attention using deep structures of neural networks, their effectiveness in limited conditions was not observed. In addition, we used an RNN with the minimum possible number of layers in the encoder.

4. 5. Main restrictions and assumptions adopted during the study

During the study, it was found that:

– for languages that have a limited training set, but have a different system, like Russian, which is inflectional, the use of transfer learning did not effectively affect the results of the study;

– the proposed model is difficult to train on long input data with a duration of 20 sec or more, and it was decided to split the data into shorter sentences (up to 10 sec) from the training corpus.

5. Results of research of using the transfer learning method in the recognition of Turkic languages

5. 1. Corpus description and presettings

For the Kazakh corpus, to conduct the experiment, we developed a corpus of 400 hours of speech, with the corpus consisting of two parts: 200 h of pure speech and 200 h of spontaneous telephone speech. This corpus was assembled in the «Computer Engineering of Intelligent Systems» laboratory of the Institute of Information and Computational Technologies CS MES RK [11]. When creating the corpus, two types of speech were considered: prepared (reading) and spontaneous. In the corpus, sound files were divided into training and test parts at 90 % and 10 %, respectively.

The pure speech database consists of recordings of 380 native Kazakh speakers of different ages and sexes, as well as speech data from fiction audiobooks and audio data of news broadcasts.

The audio data were in .wav format. All audio data were converted to a single channel state. The PCM method was used to convert the data into digital form. The discrete frequency is 44.1 kHz and the bit depth is 16 bits.

The Azerbaijani language corpus was developed with a volume of 70 h of speech. The corpus contains speech data and was developed for conducting experiments in the field of automatic recognition of the Azerbaijani language. A total of 101 speakers participated in the recording (of which 55 % were men and 45 % were women). The corpus mainly includes young and middle-aged people. Thus, the group of speakers has a relatively small difference in age, profession, and education. Most speakers were recorded within one month.

The recording was captured in an office environment. The windows and doors were closed to avoid any external noise. Headphones with a noise-canceling microphone were used for recording. For increased efficiency, we chose phonetically rich words in which consonants dominate vowels. The database includes the read text consisting of 94.267 words in 1200 sentences (https://www.sketchengine.eu/corpora-and-languages/azerbaijani-text-corpora/).

All speech files were named with a unique identification code, as well as in Kazakh files.

For transfer learning, the Keras toolkit was used. The experiments were carried out on the AMD Ryzen9 server with a GeForce RTX3090 GPU. The datasets were stored on 1000 GB SSD memory to allow faster data flow during training.

5. 2. Experiments based on end-to-end models with transfer learning

In this section, we describe two experiments that were conducted to evaluate performance. First, we built a transfer learning model and evaluated the learning efficiency obtained from transfer learning. Secondly, we compared the resulting transfer model with the main baseline methods.

Our developed end-to-end models were trained using 32 phonemes of the Azerbaijani language and 28 phonemes of the Kazakh language. Only 60 phonemes were selected for evaluation. For training, we had 470 hours of data. For test and development (dev) data, we used 20 % and 80 %, respectively (Table 1).

Table 1

Data set

Setup	Duration/Proportion
Training	470 hours
Dev	80 %
Test	20 %

In the first stage of the experiment, we trained the CTC model, and then we trained the attention-based model separately. The basic CTC model consists of a directional six-layer BLSTM with 256 cells in each layer. For the attention-based model, the encoder is a directional three-layer BLSTM with 256 cells in each layer. The attention layer is location-based and has 120 cells. The decoder is a single-layer 256-cell LSTM. The dropout rates for inputs were –0.2, –0.5, and –0.1 for the encoder, attention, and decoder, respectively. We used the Adam algorithm to optimize the models. The decoding weight for CTC was 0.3. The beam research width at the decoding stage was 15. The learning process up to the 45^{th} epoch could not distinguish acoustically similar words between the Kazakh and Azerbaijani languages (Table 2). There are 71.649 similar acoustic words in this corpus.

Table 2

Acoustically similar words

Kazakh words	Azerbaijani words
алма[alma]	alma[almaq]
күн[kuun]	gün[guun]
ашық[ashyq]	açıq[achyykh]
қасық[qasyq]	qaşıq[qashyykh]
кітап[ki'taap]	torpaq[torpaaq]
алмас қылыш[almass qylysh]	almaz qılınc[almas qylynch]
қара топырақ[qara topyraq]	qara torpaq[qara torpaaq]
терең көл[tereng kyol]	dərin göl[daerin qyol]

The solution to the above problems was the model that had the best accuracy, chosen as the final model after 45 training epochs.

All RNNs were trained with a dropout rate of 0.2 for other hidden layers. The initial learning rate remained at the level of 0.3 for the first 26 epochs and then tripled after that. The accuracy of the learning process and verification by epochs are shown in Fig. 4.

To evaluate the phoneme recognition system, the phoneme error rate (PER) is a commonly used measure and it is calculated using the Levenshtein distance, where phonemes are taken instead of words.

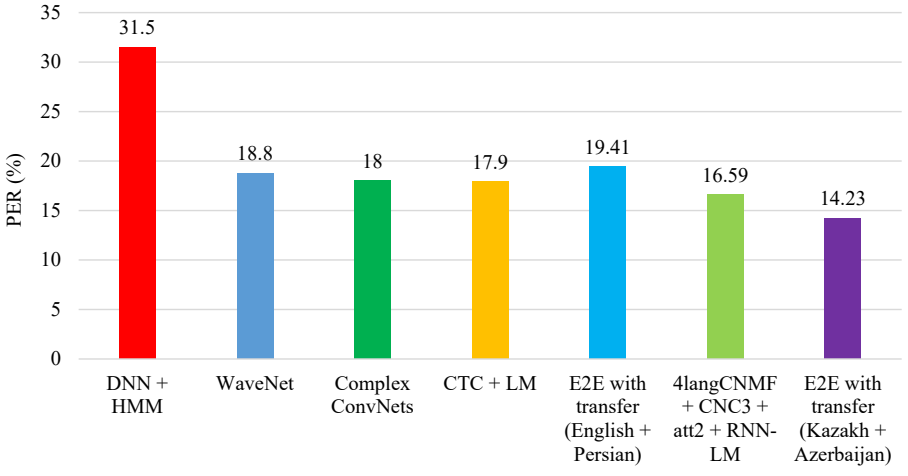

Fig. 5. Comparison of transfer learning results with some basic models

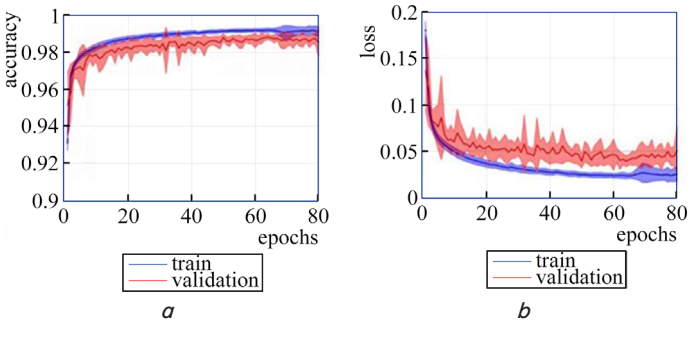

Fig. 4. The process of learning the system: *a* — by the accuracy of learning; *b* — by loss of learning

In this case, it is necessary to compare the recognized and reference sequence of the phoneme labels. All experimental data are shown in Table 3. After determining the error rate of the symbols, our model showed a PER of 14.23 %.

Table 3

PER for different speech recognition systems

Model	PER (%)
DNN+HMM [26]	31.5
WaveNet [27]	18.8
Complex ConvNets [28]	18.0
CTC+LM [11]	17.9
End-to-end with transferring (English+Persian language) [29]	19.41
4langAdaptCNMF+CTC3+att2+RNN-LM [30]	16.59
End-to-end with transferring (Kazakh+Azerbaijan language)	14.23

Fig. 5 shows a comparative diagram of all the results of the obtained transfer model with other models obtained without using the transfer method.

The results obtained will help us to draw further conclusions about the effectiveness of using methods in speech recognition.

6. Discussion of experimental results of the obtained model

In the second stage of the experiment, we compared the obtained transfer model with other models constructed without transfer learning (Table 3). We compared our model to baseline models, such as deep neural network with the hidden Markov model (DNN-HMM), and CTC+LM, which are detailed in our previous work [11, 26].

Another known end-to-end method for automatic recognition systems involves constructing a model display from raw input data in a sequence of audio/character sequences or introducing complex encoders [30, 31].

Studies [32, 33] used joint CTC-attention consisting of shallow recurrent neural networks (RNNs). The experimental results showed that the proposed transfer learning approach achieved the best performance amongst all considered end-to-end methods and is comparable to the current speech recognition system for TIMIT.

The results of our work are summarized in the last line in Table 3. Based on the research results, we conclude that the end-to-end with transferring (Kazakh+Azerbaijan language) joint model performs better than other models. When multilingual pre-trained features are introduced, our end-to-end model with transferring (Kazakh+Azerbaijani language) reaches a PER of 14.23 %, which is the best compared to those of other methods (Fig. 5).

These results confirm the effectiveness of our approach based on transfer learning in end-to-end models. Although our approach to transfer learning requires additional training procedures for feature extraction, it works better with fewer RNN layers for the end-to-end portion. We also list the PERs of some typical traditional methods and note that the proposed end-to-end models cannot outperform traditional speech recognition systems (Table 3). Note that our end-to-end models are trained without any regularization, except for dropout on BLSTM layers. The results show that the end-to-end system without transfer produced the worst results, despite performing at the current level. The result shows a significant improvement in the results produced by the system trained with transfer learning. The developed system of automatic speech recognition using transfer learning can

be applied to other low-resource languages such as Kyrgyz, Tatar, Uighur, Uzbek, etc. In this study, the most difficult aspect was data preparation and comparative analysis. Because we considered low-resource languages, the data were different and different methods and approaches were used. We conducted a large-scale analysis of the published works and selected approximately similar works for comparative analysis.

The results obtained are still inferior to the indicators of the human level of speech recognition, which requires additional experiments using other models or methods of speech recognition with limited training data. In addition, the proposed models and methods are not suitable for real-time speech recognition, meaning that after recording speech, you need to wait a few seconds to decode the audio into text, which is not very convenient to use.

In this way, we plan to conduct experiments with other types of end-to-end models with transformer for recognizing continuous Kazakh speech.

7. Conclusions

1. The quality indicator of an automatic speech recognition system is highly dependent on the training database. On the basis of this, speech corpora for two agglutinative languages were collected. To collect information in the Kazakh and Azerbaijani languages, audio files from open sources were selected, as well as clear and phone spontaneous speech were prepared. Thus, a speech corpus was assembled for the Kazakh language in the amount of 400 hours of speech, and for the Azerbaijani language, a speech corpus was formed amounting to 70 hours of speech.

2. For end-to-end speech recognition, we proposed a new approach based on transfer learning. In the first stage, the NMF algorithm is used to extract features. In the second stage, the joint CTC attention models are trained based on the extracted features through the NMF algorithm. Transfer learning is applied through bilingual learning and multitasking at two levels. The experimental results showed that our model works better than all end-to-end models and achieves high performance compared to a modern speech recognition system. Although our transfer learning approach improves the performance of end-to-end speech recognition models, it is necessary to check whether this approach also works for end-to-end learning with a relatively large resource. The result shows high-quality speech recognition. The experiments performed with two corpora prove that the transfer learning method can reduce the PER indicator by 14.23 % compared to the base models.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09259309).

References

1. Perera, F. P., Tang, D., Rauh, V., Lester, K., Tsai, W. Y., Tu, Y. H. et. al. (2005). Relationships among Polycyclic Aromatic Hydrocarbon-DNA Adducts, Proximity to the World Trade Center, and Effects on Fetal Growth. Environmental Health Perspectives, 113 (8), 1062–1067. doi: https://doi.org/10.1289/ehp.10144
2. Jaitly, N., Hinton, G. (2011). Learning a better representation of speech soundwaves using restricted boltzmann machines. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: https://doi.org/10.1109/icassp.2011.5947700
3. Rustamov, S., Gasimov, E., Hasanov, R., Jahangirli, S., Mustafayev, E., Usikov, D. (2018). Speech recognition in flight simulator. IOP Conference Series: Materials Science and Engineering, 459, 012005. doi: https://doi.org/10.1088/1757-899x/459/1/012005
4. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A. (2016). Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. Interspeech 2016. doi: https://doi.org/10.21437/interspeech.2016-1446
5. Rao, K., Peng, F., Sak, H., Beaufays, F. (2015). Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: https://doi.org/10.1109/icassp.2015.7178767
6. Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Fayed, Z. T. (2021). Arabic speech recognition using end-to-end deep learning. IET Signal Processing, 15 (8), 521–534. doi: https://doi.org/10.1049/sil2.12057
7. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y. (2015). Attention-based models for speech recognition. Advances in Neural Information Processing Systems, 577–585.
8. Putri, F. Y., Puji Lestari, D., Widyantoro, D. H. (2018). Long Short-Term Memory Based Language Model for Indonesian Spontaneous Speech Recognition. 2018 International Conference on Computer, Control, Informatics and Its Applications (IC3INA). doi: https://doi.org/10.1109/ic3ina.2018.8629500
9. Zou, W., Jiang, D., Zhao, S., Yang, G., Li, X. (2018). Comparable Study Of Modeling Units For End-To-End Mandarin Speech Recognition. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). doi: https://doi.org/10.1109/iscslp.2018.8706661
10. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y. et. al. (2018). ESPnet: End-to-End Speech Processing Toolkit. Interspeech 2018. doi: https://doi.org/10.21437/interspeech.2018-1456
11. Mamyrbayev, O., Alimhan, K., Zhumazhanov, B., Turdalykyzy, T., Gusmanova, F. (2020). End-to-End Speech Recognition in Agglutinative Languages. Lecture Notes in Computer Science, 391–401. doi: https://doi.org/10.1007/978-3-030-42058-1_33
12. Asefisaray, B., Haznedaroglu, A., Erden, M., Arslan, L. M. (2018). Transfer learning for automatic speech recognition systems. 2018 26th Signal Processing and Communications Applications Conference (SIU). doi: https://doi.org/10.1109/siu.2018.8404628

13. Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J. (2013). Multilingual acoustic models using distributed deep neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. doi: https://doi.org/10.1109/icassp.2013.6639348
14. Yi, J., Tao, J., Wen, Z., Bai, Y. (2019). Language-Adversarial Transfer Learning for Low-Resource Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (3), 621–630. doi: https://doi.org/10.1109/taslp.2018.2889606
15. Palaz, D., Magimai-Doss, M., Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. Speech Communication, 108, 15–32. doi: https://doi.org/10.1016/j.specom.2019.01.004
16. Dokuz, Y., Tufekci, Z. (2021). Mini-batch sample selection strategies for deep learning based speech recognition. Applied Acoustics, 171, 107573. doi: https://doi.org/10.1016/j.apacoust.2020.107573
17. Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M., Varol, H. A. (2021). A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. doi: https://doi.org/10.18653/v1/2021.eacl-main.58
18. Beibut, A. (2020). Development of Automatic Speech Recognition for Kazakh Language using Transfer Learning. International Journal of Advanced Trends in Computer Science and Engineering, 9 (4), 5880–5886. doi: https://doi.org/10.30534/ijatcse/2020/249942020
19. Markovnikov, N., Kipyatkova, I. (2019). Investigating Joint CTC-Attention Models for End-to-End Russian Speech Recognition. Lecture Notes in Computer Science, 337–347. doi: https://doi.org/10.1007/978-3-030-26061-3_35
20. Fujita, Y., Watanabe, S., Omachi, M., Chang, X. (2020). Insertion-Based Modeling for End-to-End Automatic Speech Recognition. Interspeech 2020. doi: https://doi.org/10.21437/interspeech.2020-1619
21. Zeng, Z., Pham, V. T., Xu, H., Khassanov, Y., Chng, E. S., Ni, C., Ma, B. (2021). Leveraging Text Data Using Hybrid Transformer-LSTM Based End-to-End ASR in Transfer Learning. 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). doi: https://doi.org/10.1109/iscslp49672.2021.9362086
22. Qin, C.-X., Zhang, W.-L., Qu, D. (2019). A new joint CTC-attention-based speech recognition model with multi-level multi-head attention. EURASIP Journal on Audio, Speech, and Music Processing, 2019 (1). doi: https://doi.org/10.1186/s13636-019-0161-0
23. O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I.et. al. (2018). Directions for the future of technology in pronunciation research and teaching. Journal of Second Language Pronunciation, 4 (2), 182–207. doi: https://doi.org/10.1075/jslp.17001.obr
24. Tejedor-García, C., Cardeñoso-Payo, V., Escudero-Mancebo, D. (2021). Automatic Speech Recognition (ASR) Systems Applied to Pronunciation Assessment of L2 Spanish for Japanese Speakers. Applied Sciences, 11 (15), 6695. doi: https://doi.org/10.3390/app11156695
25. Ding, C. H. Q., Tao Li, Jordan, M. I. (2010). Convex and Semi-Nonnegative Matrix Factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (1), 45–55. doi: https://doi.org/10.1109/tpami.2008.277
26. Schuller, B., Weninger, F., Wollmer, M., Sun, Y., Rigoll, G. (2010). Non-negative matrix factorization as noise-robust feature extractor for speech recognition. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. doi: https://doi.org/10.1109/icassp.2010.5495567
27. Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A., Turdalykyzy, T. (2019). Automatic Recognition of Kazakh Speech Using Deep Neural Networks. Lecture Notes in Computer Science, 465–474. doi: https://doi.org/10.1007/978-3-030-14802-7_40
28. Mamyrbayev, O., Oralbekova, D., Kydyrbekova, A., Turdalykyzy, T., Bekarystankyzy, A. (2021). End-to-End Model Based on RNN-T for Kazakh Speech Recognition. 2021 3rd International Conference on Computer Communication and the Internet (ICCCI). doi: https://doi.org/10.1109/iccci51764.2021.9486811
29. Mamyrbayev, O., Kydyrbekova, A., Alimhan, K., Oralbekova, D., Zhumazhanov, B., Nuranbayeva, B. (2021). Development of security systems using DNN and i & x-vector classifiers. Eastern-European Journal of Enterprise Technologies, 4 (9(112)), 32–45. doi: https://doi.org/10.15587/1729-4061.2021.239186
30. Van Den Oord, A., Dieleman, S., Zen, H. et. al. (2016). Wavenet: A generative model for raw audio. arXiv. Available at: https://arxiv.org/pdf/1609.03499.pdf
31. Zeghidour, N., Usunier, N., Kokkinos, I., Schaiz, T., Synnaeve, G., Dupoux, E. (2018). Learning Filterbanks from Raw Speech for Phone Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: https://doi.org/10.1109/icassp.2018.8462015
32. Kermanshahi, M. A., Akbari, A., Nasersharif, B. (2021). Transfer Learning for End-to-End ASR to Deal with Low-Resource Problem in Persian Language. 2021 26th International Computer Conference, Computer Society of Iran (CSICC). doi: https://doi.org/10.1109/csicc52343.2021.9420540
33. Qin, C.-X., Qu, D., Zhang, L.-H. (2018). Towards end-to-end speech recognition with transfer learning. EURASIP Journal on Audio, Speech, and Music Processing, 2018 (1). doi: https://doi.org/10.1186/s13636-018-0141-9