









Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System

Dina Oralbekova^{1,2} , Orken Mamyrbayev² , Mohamed Othman³ , Keylan Alimhan⁴ , Bagashar Zhumazhanov² , and Bulbul Nuranbayeva⁵ 

¹ Satbayev University, Almaty, Kazakhstan
dinaoral@mail.ru

² Institute of Information and Computational Technologies, Almaty, Kazakhstan

³ Universiti Putra Malaysia, Kuala Lumpur, Malaysia

⁴ L.N. Gumilyov, Eurasian National University, Nur-Sultan, Kazakhstan

⁵ Caspian University, Almaty, Kazakhstan

Abstract. Architecture end-to-ends are commonly used methods in many areas of machine learning, namely speech recognition. The end-to-end structure represents the system as one whole element, in contrast to the traditional one, which has several independent elements. The end-to-end system provides a direct mapping of acoustic signals in a sequence of labels without intermediate states, without the need for post-processing at the output, making it easy to implement. Combining several end-to-end method types perform better results than applying them separately. Inspired by this issue, in this work we have realized a method for using CRF and CTC together to recognize a low-resource language like the Kazakh language. In this work, architectures of a recurrent neural network and a ResNet network were applied to build a model using language models. The results of experimental studies showed that the proposed approach based on the ResNet architecture with the RNN language model achieved the best CER result with a value of 9.86% compared to other network architectures for the Kazakh language.

Keywords: Automatic speech recognition · End-to-end · Connectionist temporal classification · Conditional random fields · ResNet

1 Introduction

It's no secret that some of the processes processed by humans are not so easy to implement at the machine level, this concerns speech recognition systems. Today there are many speech recognition applications and software. Companies like Google and Yandex have achieved a fairly good level of accurate word recognition in speech for languages with a large body of training data. However, the data quality of the systems requires improvements in decoding sequences from speech.

Conventional automatic speech recognition systems were built on the basis of independent components, this is an acoustic model, a language model and a vocabulary, which were tuned and trained separately. The context-dependent states of phonemes are

forecasted by the acoustic model, the language model and lexicon determine the most possible sequences of spoken phrases.

In traditional speech recognition systems, a Hidden Markov Model (HMM) model with a Gaussian Mixture Model (GMM) has been widely used. With the help of HMM, statistical models of words were built, and GMM represents the unit of pronunciation, i.e. distribution of signals within a certain period of time [1]. The development of deep learning technologies has contributed to the improvement of other scientific areas, which includes speech recognition. Deep neural networks began to be used for acoustic modeling instead of GMM, which led to improved results [2]. Thus, the HMM-DNN architecture, i.e. a combination of these methods is becoming popular in the field of continuous speech recognition.

In the future, the end-to-end model became widespread, which trains the components of the traditional model simultaneously without isolating individual elements, representing the system as a single neural network. In research papers [3–5] DNN was used to develop an acoustic model, language models and dictionaries were implemented by RNN. In addition, convolutional neural networks were used to extract features from the original signal [6]. And the application of these modifications has led to an enhancement in the efficiency of speech recognition systems. Different ANN architectures can be used at all stages of recognition, and this makes it effective in terms of performance compared to other popular systems. This approach is end-to-end. The end-to-end structure represents the system as one whole element, in contrast to the traditional one, which has several independent elements. The end-to-end system provides a direct mapping of acoustic signals in a sequence of labels without intermediate states, without the need for post-processing at the output, making it easy to implement.

For the first time the end-to-end model was mentioned in the work of Alex Graves [7], he presented the model based on the Connectionist temporal classification. But the application of the E2E model was implemented after 2013 [8]. This was due to the absence of computing power. After the emergence of powerful techniques with support for parallel computing, end-to-end models began to be used in speech recognition in popular languages that have a fairly large amount of training data. Also, the appearance of other models of end-to-end architecture influenced the result of indicators of speech recognition systems. Thus, to build an end-to-end model, it is enough to observe three basic rules: 1) the presence of high-performance technology, a supercomputer; 2) the presence of a large speech corpus for several thousand hours of speech, which will consist of audio data and their transcriptions, 3) the selected architecture of the E2E model with the appropriate settings that is suitable for speech recognition of the desired language or a group of languages of the same structure (inflectional, agglutinative, etc.).

The E2E system realizes direct reflection of acoustic indicators in a sequence of marks without intermediate states, which does not require further processing at the output. These processes make the system easy to implement. There are several basic types of E2E models, such as connectionist temporal classification (CTC) [7, 8], encoder-decoder with attention models [9, 10], and Conditional Random Fields (CRF) [11]. In the CTC-based model, there is unnecessary frame-level alignment between acoustics and transcription, since a special token is allocated, like an empty “label”, which defines the start and end of one phoneme. In encoder models based on the attention mechanism, the

encoder is an acoustic model (AM), the decoder is similar to language model (LM) – it works autoregressive, predicting each output token depending on previous predictions [9, 10]. The CRF-based model allows you to combine local information to predict the global probabilistic model by sequences [11]. These models greatly simplify the speech recognition process. The increase in training data refined the quality of the ASR systems compared to the HMM [12]. But data reduction increases recognition errors. However, E2E models have the flexibility to combine them to mitigate their specific disadvantages.

E2E models require a large amount of speech data for training, which is problematic for languages with limited training data. And one of these languages is the Kazakh language. The Kazakh language has an agglutinative character, in which the dominant type of inflection is agglutination [13], opposite to the inflectional one. Some research works [14–16] have shown that the combined use of E2E models like CTC and attention can be trained from start to finish, while this combination gave a very good result, which almost came close to the accuracy of the human level [16]. Based on these studies, we study the end-to-end system of the joint CTC and CRF models. Until now, systems have been developed on E2E CTC and Transformer models [17, 18] for the recognition of Kazakh speech with different sets of training data. At the moment, E2E CRF based systems for the recognition of Kazakh speech have not been investigated, and we decided to build this model. The effectiveness of this model has been realized in natural language processing.

In this research work, we have built a hybrid model based on two end-to-end methods, CRF and CTC for Kazakh speech recognition.

The structure of the research work is given in the following order: Sect. 2 provides a brief analytical review on scientific topics. Section 3 shows the principle of operation of models based on CTC and CRF. Further in Sect. 4 our experimental data, speech corpus and model settings are described, the results obtained are analyzed. The final section summarizes the findings.

2 Literature Review

Conditional Random Fields (CRF) is a model that allows you to combine local information to predict the global probabilistic model from sequences. This model is considered to be a kind of Markov random field. This model was first proposed in [11] for speech recognition. J. Lafferty et al. (2001) proposed an algorithm for estimating parameters for conditional random fields and showed that CRF has a greater advantage over HMM and MEMM (maximum entropy Markov models) for natural language data. In [19, 20], the CRF model was applied to assess the measurement of accuracy in the problem of phonetic recognition, as well as the accuracy of detecting boundaries between them. The results show that when using transition functions in the CRF-based recognition structure, recognition performance is significantly improved by reducing the number of phoneme deletions. Show that when using transition functions in the recognition structure based on CRF, recognition performance is significantly improved by decreasing the number of phone deletions. Bounding efficiency is also improved, mainly for transitions between the phonetic classes of silence, stop, and clicks. In addition, the CRF model gives a lower error rate than the HMM and Maxent models in the task of detecting the boundaries of sentences in speech.

Currently, the most common in speech recognition are linear and segmental CRF (linear chain & segmental CRF) models. This model is most often used to solve the problems of marking and segmenting sequences.

Keyu An et al. (2019) [21] demonstrated a new CAT toolkit that represents an implementation of CTC-CRF E2E models. For the experiment, Chinese and English tests, such as Switchboard and Aishell, were applied, thus obtaining the most modern results among the existing end-to-end models with fewer parameters and competitive in comparison with the hybrid models DNN-HMM. In addition, the same authors in [22] (2020) proposed a new technique called contextualized soft forgetting that allows the CAT tool to perform streaming ASR without compromising accuracy and has performed well with limited datasets compared to existing ones.

In [23] (2017), an end-to-end model was built based on segmented conditional random fields (SCRf) and connection time classification (CTC). SCRf uses a globally normalized joint label and segment length model, and CTC classifies each frame as either an output symbol. Through experimentation with the TIMIT dataset, a multitasking approach to training improved the recognition accuracy of CTC and SCRf models. In addition, it has been illustrated that CTC can be used to pre-train an RNN encoder, accelerating the training of a collaborative model.

Hongyu Xiang et al. [24] (2019) developed a single-stage acoustic simulation based on a CRF with a state topology based on CTC. Evaluation experiments were conducted with WSJ, Switchboard and LibriSpeech datasets. In direct comparison, the CTC-CRF model using simple bidirectional LSTMs consistently outperformed SS-LF-MMI (lattice-free maximum-mutual-information) on all three benchmark datasets and in both monophones and mono symbols. And it was revealed that the CTC-CRF model avoids some special operations in SS-LF-MMI.

In [25] (2021), methods were investigated to apply the newly developed text word modeling modules and Conformer neural networks in the CTC-CRF. Research observations are conducted on Switchboard and LibriSpeech, and the German CommonVoice dataset. Experimental results show that Conformer can significantly improve recognition quality; verbal systems perform slightly worse than telephone systems for a target language with a low graphemphoneme match, while both systems can perform equally well when that match is high for the target language.

Yang, Li et al. (2019) [26] propose a text processing model after Chinese speech recognition that combines a bidirectional long-term short-term memory (LSTM) network with a conditional random field (CRF) model. The need to process the text after recognition is associated with the appearance of a problem with the dialect and accent, since it is necessary to correct the text after speech recognition before displaying it. The objective is divided into two steps: detecting text errors and editing text errors. In this article, a bi-directional long-term short-term memory (Bi-LSTM) network and a conditional random field are used in two stages of text error detection and text error correction, respectively. Through validation and system testing of the SIGHAN 2013 Chinese Spelling Check (CSC) dataset, experimental results show that the model can effectively improve the accuracy of text after speech recognition.

The reviewed works shows us that the joint use of E2E models developed the productivity of the ASR system than using them separately.

Unfortunately, there is very little new research on this topic specifically in the study of speech recognition. But we tried to consider the publications that were in the public domain to the maximum.

3 Methodology of E2E Models

This section describes the CTC and CRF models and their joint model.

3.1 Connectionist Temporal Classification (CTC)

The CTC function is used to train a neural network in sequence recognition. Let's say we have an output sequence $y = S_w(x)$. Let each element of this sequence have a probability distribution vector for each symbol V' at time t . Therefore, we must define y_k^t , which is the probability of pronouncing the character k from the alphabet V' at time t . If μ is a sequence of symbols and a "space" for the given input x , then the probability $P(\mu|x)$ can be defined as follows (1):

$$P(\mu|x) = \prod_t y_{\mu t}^t \quad (1)$$

From the above equation, you can see that the components of the output sequence are independent of each other. To align data, you need to add an auxiliary character that will remove duplicate letters and spaces. Let's denote it as B . Thus, the total probability of the output sequence can be expressed by the following (2):

$$P(y|x) = \sum_{\mu \in B^{-1}(y)} P(\mu|x) \quad (2)$$

The given above equation determines the sum over all alignments using dynamic programming, and helps to train the ANN on unlabeled data (3):

$$CTC(x) = -\log P(y|x) \quad (3)$$

From the above it follows that ANN can be trained on any gradient optimized algorithm. In the CTC architecture, any kind of ANN can be used as an encoder, such as LSTM and BLSTM.

To decode the CTC-model, the assumption was presented in [7] (4):

$$\operatorname{argmax} P(y|x) \approx B(\mu^\circ) \quad (4)$$

where $\mu^\circ = \operatorname{argmax} P(y|x)$.

CTC eliminates the need for data alignment and allows for quite a few layers, a simple network structure to implement a model that maps audio to sequence of utterances.

3.2 Conditional Random Fields (CRF)

Conditional random fields (CRF) are a class of statistical modeling techniques that are usually used in machine learning and are used for structured prediction. While the classifier predicts a label for one sample without considering “adjacent” samples, the CRF can take context into account. For this, the forecast is modeled as a graphical model that implements the relations between the forecasts.

Other examples of the use of CRF are: labeling or analysis of sequential data for natural language or biological sequence processing, POS marking, and object recognition and image segmentation in computer vision [27].

Conditional Random Field (CRF) is a discriminative undirected probabilistic graphical model [28]. This method, in contrast to the Markov model of maximum entropy, does not have a label bias problem [29, 30]. CRF and its various modifications have found applications in areas such as natural language processing, computer vision, speech recognition, etc. (see Fig. 1).

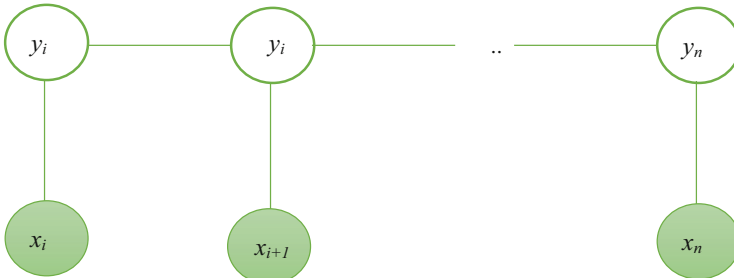


Fig. 1. General structure of the model CRF

Here $X = (x_1, \dots, x_n)$ is the input sequence to be recognized, and $Y = (y_1, \dots, y_n)$ is the recognized set of character set labels, where $i = 1$.

In the following paragraphs, the widely used linear chain and segmental CRF will be discussed. These models are successfully used to solve machine learning problems, where the precedent is a sequence of random variables with labels assigned to them. These are the so-called sequence marking and segmentation tasks. Therefore, this model is popular in areas that are characterized by a sequence of variables, for example, in natural language processing tasks.

Linear-Chain CRF. The linear CRF model is a discriminative model, and this is where it resembles the popular Maximum entropy Markov model (MEMM). In contrast to CRF, the maximum entropy model has a disadvantage called label bias [29, 30]. The essence of the problem is that, due to the peculiarities of learning, the maximum entropy model tends to give preference to those latent states that have a lower entropy of distribution of subsequent states.

The linear CRF model put in to solve many natural language processing problems, mostly for the English language, for example, to the problem of POS-tagging, to the problem of surface parsing, or to the problem of resolving anaphora [31].

The disadvantage of the linear CRF model is the used complex algorithm of the analysis of the training sample, that makes it hard to constantly update the model when new training data arrives [32].

Segmental Conditional Random Fields. Segmental CRFs are built on the basis of segmented recurrent neural networks, namely, it is semi-Markov CRF [33].

SCRf generalizes conditional random fields to operate at the segment level rather than the traditional frame level. Basically, each segment is labeled with a word. Features are then extracted, each measuring some form of consistency between the underlying sound and the word hypothesis for the segment. They are combined in a log-linear model to obtain the posterior probability of a sequence of words, including sound.

In the context of speech recognition, taking into account the sequence of input vectors X and the corresponding sequence of output labels Y CRF of the linear chain of zero order determines the conditional probability at the level of sequence (5) using auxiliary segments in the following way

$$P(y, s|x) = \frac{1}{Z(x)} \prod_{i=1}^n \exp f(y_i, s_i, x_i) \quad (5)$$

where $Z(X)$ is the normalization term and the complete calculation of this value is given in [34], where the segmental level functions were studied using RNN and the segmental RNN (SRNN) model was applied as an implementation of the SCRf acoustic model for multitasking learning.

In [34] SRNNs were applied for speech recognition. On the TIMIT speech corpus, a PER score of 17.3% was obtained. Contemporaneously, the implemented model did not use LM.

3.3 Joint Training Models for Kazakh Speech Recognition

Joint training of E2E models can be expressed as follows (6):

$$L_{CTC/Att} = \tau L_{CTC}(x) + (1 - \tau) L_{SCRf}(x) \quad (6)$$

where τ is an adjustable parameter and satisfies the condition $-0 \leq \tau \leq 1$. Neural networks with short and long-term memory LSTM and bidirectional LSTM and ResNet architecture were used as neural networks.

In the CTC model, it uses monotonic alignment between speech and tag sequences and trains the network quickly. And besides, the proposed model will be effective in recognizing speech in long sequences if training took place in short training data. In addition, CTC helps speed up the process of assessing the desired alignment without the help of rough estimates of this process, which is labor-intensive and time-consuming.

4 Experiments and Results

This section contains descriptions of the case, preliminary model settings, experimental data, as well as the results obtained and a comparative analysis of the data obtained.

4.1 Data Preparation

The corpus of speech for the Kazakh language was implemented by the researchers of the laboratory “Computer Engineering of Intelligent Systems” of the Institute for Computer Engineering of the Ministry of Education and Science of the Republic of Kazakhstan [35]. This corpus consists of pure speech and speech of telephone conversations.

For the recording, 250 speakers took part (55% of them were men and 45% were women). The corpus mainly includes young and middle-aged people. Thus, the group of speakers has relatively small changes in age, profession and education. The recording was made 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 efficiency, we have chosen phonetically rich words in which consonants dominate vowels. The base includes the read text, consisting of 94 267 words in 1200 sentences. All speech files were named with a unique identification code.

Given the recent advances in speech recognition, it is necessary to add data recorded under various conditions. Thus, we added to our corpus the recordings of telephone conversations, which were transcribed by young volunteers who were involved from the Higher School - these are undergraduates, undergraduates and doctoral students of Almaty national universities.

As a result, the total volume of Kazakh-language speech data was 300 h of speech, with 90% of the audio data used for training, and 10% for model validation.

All audio materials were in .wav format. All audio data has been converted to single channel. The PCM method was used to convert the data into digital form. Discrete 44.1 kHz, 16-bit.

4.2 Presetting Models

End-to-end models with different variations of neural networks were implemented, as well as the implementation of models separately and jointly.

Feature extraction is an obligatory part of the system, and in this part, convolutional neural networks were used, since they have a stronger anti-interference ability than the use of cepstral coefficients with a mel-frequency frequency [36]. The activation function ReLU was applied for convolutional layers [37]. Next, a maxpooling layer was added to filter the low frequencies of speech. In addition, layers of compression and normalization of the extracted features were additionally introduced.

During training, the gain of the weight coefficient of the CTC model was set: $\lambda = 0.2$. Also, the external language model was leveraged.

To the basic CTC model, the RNN varieties were applied, such as LSTM and BLSTM with five layers containing 1024 cells for each layer, the ResNet network - a method for eliminating the gradient fading effect for CNN. ResNet consists of 8 convolutional layers and a max-pooling layer with batch normalization [38].

The initial learning rate coefficient was set to 0.001. Dropout was used for each output of the recurrent layer as a regularization and is equal to 0.5. For our model, we used a gradient descent optimization algorithm based on Adam [39].

To measure the quality of the Kazakh speech recognition system based on E2E models, the CER metric was used - the number of incorrectly recognized characters, as well as on the basis of the word error rate (WER), which is calculated using the Levenshtein distance [40].

4.3 Results and Analysis of Experimental Studies

Audio recordings with transcription from news sites in the Kazakh language (<https://qazaqstan.tv/live>, <https://24.kz/kz/>), audiobook sites (<https://kitap.kz/>, <https://e-history.kz/ru/audio/>) and which is 1 600 separate phrases of different speakers and none of the speakers was used simultaneously in both parts.

The end-to-end model, when using the CTC function without a language model, reached a CER of 17.45% and a WER of 29.01% (see Fig. 2). Integration of the external language model into the CTC end-to-end system improved the CER and WER indicators by 13% and 18%, respectively. Our joint model of CTC and CRF with ResNet architecture showed good results without the use of LM, and CER reached 11.57% and WER - 18.32%.

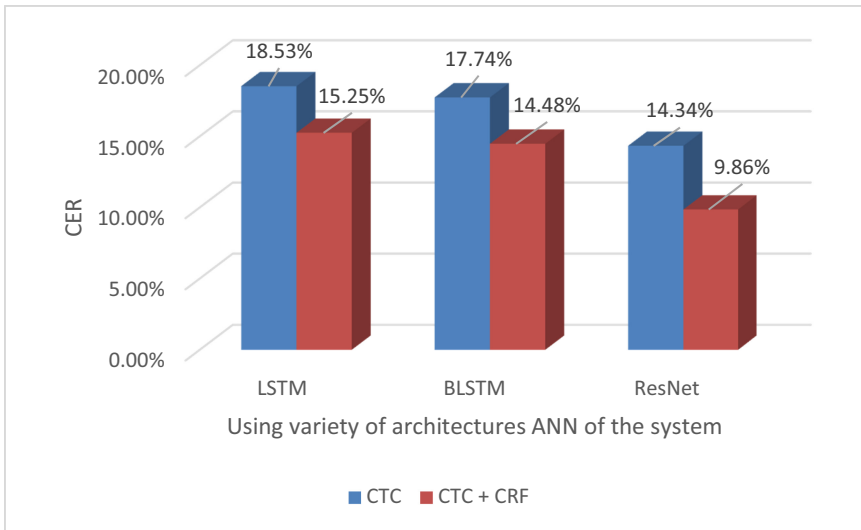


Fig. 2. Comparative graphs of the basic and joint model based on different architectures of the NN with LM in terms of CER

And after adding LM, the model slightly improved its quality, by almost 1.5%. The gain values are shown in Table 1.

Table 1. Results of experiments on the use of E2E models.

| Model | WER, % | CER, % |
|-----------------------|--------|--------|
| CTC without LM | 34,48 | 21,56 |
| LSTM | 32,59 | 20,56 |
| BLSTM | 29,01 | 17,45 |
| ResNet | | |
| CTC with LM | 26,63 | 18,53 |
| LSTM | 24,31 | 17,74 |
| BLSTM | 23,65 | 14,34 |
| ResNet | | |
| Without LM | 31,92 | 18,20 |
| CTC (LSTM) + CRF | 27,19 | 16,82 |
| CTC (BLSTM) + CRF | 18,32 | 11,57 |
| CTC (ResNet) + CRF | | |
| With LM | 26,35 | 15,25 |
| CTC (LSTM) + CRF | 23,61 | 14,48 |
| CTC (BLSTM) + CRF | 16,75 | 9,86 |
| CTC (ResNet) + CRF | | |

Table 1 shows that the models using the ResNet network showed the best result in terms of the coefficients of correctly recognized words and characters.

5 Conclusion

The work considered the joint end-to-end models CTC and CRF for the recognition of Kazakh speech. To implement this model, RNN variations were applied, such as LSTM and BiLSTM, as well as the ResNet. Convolutional neural networks were used for feature extraction. The practice works were conducted using the Kazakh language corpus with a volume of 300 speech hours, and the result demonstrated that the system can achieve high results using the ResNet and the use of RNN-based language model. Decoding based on these models does not increase the computational cost, and due to this, the decoding speed does not slow down. Thus, the best CER indicator reached 9.86%, which is a competitive result today. The proposed method is flexible enough and does not require conditional independence of variables. In addition, we can realize that proposed model can be used to recognize other languages with limited training data, which are part of the Turkic languages.

Now, we target to study insertion-based models for recognizing agglutinative languages.

Acknowledgement. This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP08855743).

References

1. Gales, M., Young, S.: 2007. The application of hidden Markov models in speech recognition. *Found. Trends Signal Process.* **1**(3), 195–304 (2008). <https://doi.org/10.1561/20000000004>
2. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**(6), 82–97, (2012). <https://doi.org/10.1109/MSP.2012.2205597>
3. Maas, A., Qi, P., Xie, Z., Hannun, A., Lengerich, C., Jurafsky, D., Ng, A.: Building DNN acoustic models for large vocabulary speech recognition. *Comput Speech Lang.* **41** (2016). <https://doi.org/10.1016/j.csl.2016.06.007>
4. Fohr, D., Mella, O., Illina, I.: New Paradigm in speech recognition: deep neural networks. In: *IEEE International Conference on Information Systems and Economic Intelligence, Marrakech, Morocco*. fihal-01484447f (2017)
5. Shi, Y., Zhang, WQ., Liu, J., et al.: RNN language model with word clustering and class-based output layer. *J. Audio Speech Music Proc.* **22** (2013). <https://doi.org/10.1186/1687-4722-2013-22>
6. Huang, S., Tang, J., Dai, J., Wang, Y.: Signal status recognition based on 1DCNN and its feature extraction mechanism analysis. *Sensors (Basel)* **19**(9) (2018). <https://doi.org/10.3390/s19092018>
7. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376 (2006). <https://doi.org/10.1145/1143844.1143891>
8. Mamyrbayev, O., Oralbekova, D.: Modern trends in the development of speech recognition systems. *News Nat. Acad. Sci. Republic of Kazakhstan*, **4**(32), 42 – 51 (2020). <https://doi.org/10.32014/2020.2518-1726.64>
9. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.L.: Attend and Spell. *ArXiv*, abs/1508.01211. (data of request: 14.09.2021) (2015)
10. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 4945–4949 (2016)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the International Conference on Machine Learning (ICML 2001)*, Williamstown, MA, USA, pp. 282–289 (2001)
12. Garcia-Moral, A., Solera-Ureña, R., Peláez-Moreno, C., Díaz-de-María, F.: Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems. *IEEE Trans. Audio Speech Lang. Process.* **19**. 468 - 481 (2011). <https://doi.org/10.1109/TASL.2010.2050513>
13. Agglutinating language - http://www.glottopedia.org/index.php/Agglutinating_language, (data of request: 27 Sep 2021)
14. Hori, T., Watanabe, S., Zhang, Y., Chan, W.: *Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM* (2017)
15. Kim, S., Hori, T., Watanabe, S.: *Joint CTC-attention based end-to-end speech recognition using multi-task learning* (2016)
16. Mamyrbayev, O., Alimhan, K., Oralbekova, D., Bekarystankyzy, A., Zhumazhanov, B.: Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. *Eastern-Euro. J. Enter. Technol.* **1**(9)(115), 84–92 (2022). <https://doi.org/10.15587/1729-4061.2022.252801>

17. Mamyrbayev, O., Kydyrbekova, A., Alimhan, K., Oralbekova, D., Zhumazhanov, B., Nuranbayeva, B.: Development of security systems using DNN and i & x-vector classifiers. *Eastern-Euro. J. Enter. Technol.* **4** (9 (112)), 32–45 (2021). <https://doi.org/10.15587/1729-4061.2021.239186>
18. Orken, M., Dina, O., Keylan, A., Tolganay, T., Mohamed, O.: A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci Rep* **12**, 8337 (2022). <https://doi.org/10.1038/s41598-022-12260-y>
19. Dimopoulos, S., Fosler-Lussier, E., Lee, C., Potamianos, A.: Transition features for CRF-based speech recognition and boundary detection. In: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 99–102 (2009). <https://doi.org/10.1109/ASRU.2009.5373287>
20. Liu, Y., Stolcke, A., Shriberg, E., Harper, M.: Using Conditional Random Fields for Sentence Boundary Detection in Speech (2005). <https://doi.org/10.3115/1219840.1219896>
21. An, K., Xiang, H., Ou, Z.: CAT: CRF-based ASR Toolkit. arXiv: abs/1911.08747, <https://arxiv.org/abs/1911.08747> (2019)
22. An, K., et al.: CAT: A CTC-CRF based ASR Toolkit Bridging the Hybrid and the End-to-end Approaches towards Data Efficiency and Low Latency. In: NTERSPPEECH (2020)
23. Lu, L., Kong, L., Dyer, C., Smith, N.A.: Multitask Learning with CTC and Segmental CRF for Speech Recognition In: Interspeech (2017)
24. Xiang, H., Ou, Z.: CRF-based single-stage acoustic modeling with CTC topology. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5676–5680 (2019)
25. An, K., Xiang, H., Ou, Z.: CAT: A CTC-CRF based ASR Toolkit Bridging the Hybrid and the End-to-end Approaches towards Data Efficiency and Low Latency. In: INTERSPEECH (2020)
26. Yang, L., Li, Y., Wang, J., Tang, Z.: Post Text Processing of Chinese Speech Recognition Based on Bidirectional LSTM Networks and CRF. *Electronics* **8**(11) 1248 (2019). <https://doi.org/10.3390/electronics8111248>
27. Abney S.: Parsing by chunks. In: Berwick, R., Abney, S., Tenny, C., (eds.) *Principle-based Parsing*. Kluwer Academic Publishers, pp. 257–279 (1991)
28. Sutton, C., McCallum, A.: *An Introduction to Conditional Random Fields for Relational Learning*. MIT Press (2006)
29. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, Massachusetts, pp. 282–289 (2001)
30. Bottou, L.: Une approche theorique de l'apprentissage connexionniste: Applications a la reconnaissance de la parole. Doctoral dissertation, Universite de Paris XI (1991)
31. Culotta, A., Wick, M., Hall R., McCallum, A.: First-order probabilistic models for coreference resolution. In: *Proc. of HLT-NAACL* (2007)
32. Markovnikov, N.M., Kipyatkova, I.S.: An analytic survey of end-to-end speech recognition systems. *Tr. SPIIRAN* **58**, 77–110 (2018)
33. Kong, L., Dyer C., Smith, N.A.: Segmental recurrent neural networks. arXiv: 1511.06018, <https://arxiv.org/abs/1511.06018>. (Accessed 02 Oct 2021) (2015)
34. Lu, L., Kong, L., Dyer, C., Smith, N., Renals, S.: Segmental recurrent neural networks for end-to-end speech recognition. In: *Proc. INTERSPEECH* (2016)
35. Laboratory of computer engineering of intelligent systems – <https://iict.kz/laboratory-of-computer-engineering-of-intelligent-systems/> (data of request: 02 Aug 2021)
36. Li, F., et al.: Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J. Adv. Signal Process.* **2019**(1), 1–11 (2019). <https://doi.org/10.1186/s13634-019-0651-3>

37. Zhao, G., Zhang, Z., Guan, H., Tang, P., Wang, J.: Rethinking ReLU to Train Better CNNs. 603–608 (2018). <https://doi.org/10.1109/ICPR.2018.8545612>
38. Ioffe, S., Szegedy, C.: Proceedings of the 32nd International Conference on Machine Learning, PMLR, vol. 37, pp. 448–456 (2015)
39. Kingma D. P., Ba J. Adam: A method for stochastic optimization. <http://arxiv.org/abs/1412.6980> (data of request: 01.11.2021) (2014)
40. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys. Doklady **10**, 707–710 (1996)