Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 13, MARCH 2023

DOI: 10.37943/13IMII7575

Bekarystankyzy Akbayan
Master of IS, PhD student
akbayan.b@gmail.com, orcid.org/0000-0003-3984-2718
Satbayev University, Narxoz University, Kazakhstan

Mamyrbayev Orken Zhumazhanovich
PhD, Professor, Deputy Director
morkenj@mail.ru, orcid.org/0000-0001-8318-3794
Institute of Information and Computational Technologies,
CS MSHE RK, Kazakhstan

END-TO-END SPEECH RECOGNITION SYSTEMS FOR AGGLUTINATIVE LANGUAGES

Abstract: With the improvement of intelligent systems, speech recognition technologies are being widely integrated into various aspects of human life. Speech recognition is applied to smart assistants, smart home infrastructure, the call center applications of banks, information system components for impaired people, etc. But these facilities of information systems are available only for common languages, like English, Chinese, or Russian. For low-resource language, these opportunities for information technologies are still not implemented. Most modern speech recognition approaches are still not tested on agglutinative languages, especially for the languages of Turkic group like Kazakh, Tatar, and Turkish Languages.

The HMM-GMM (Hidden Markov Models - Gaussian Mixture Models) model has been the most popular in the field of Automatic Speech Recognition (ASR) for a long time. Currently, neural networks are widely used in different fields of NLP, especially in automatic speech recognition. In an enormous number of works application of neural networks within different stages of automatic speech recognition makes the quality level of this systems much better.

Integral speech recognition systems based on neural networks are investigated in the article. The paper proves that the Connectionist Temporal Classification (CTC) model works precisely for agglutinative languages. The author conducted an experiment with the LSHTM neural network using an encoder-decoder model, which is based on the attention-based models. The result of the experiment showed a Character Error Rate (CER) equal to 8.01% and a Word Error Rate (WER) equal to 17.91%. This result proves the possibility of getting a good ASR model without the use of the Language Model (LM).

Keywords: agglutinative languages, integral approach, CTC, LSTM neural network, speech recognition.

Introduction

Speech refers to a whole system of sound signals, symbols, and written signs that are used by a person to store, represent, process, and transmit data. In addition, speech is a tool that is necessary for a machine and a person to interact with each other [1]. In order for the voice interface to be implemented, it is necessary to resort to the help of many specialists: a Deep Neural Network (DNN) programmer, a computer linguist, and so on. The standard speech recognition system is divided into a number of modules: language and acoustic models, as well as decoding [2]. The main thing for the modularity design is a large number of irrespective

DOI: 10.37943/13IMII7575

suggestions, and the standard acoustic model is trained on the set of frames that depend on the Markov Model. In various ASR systems, the most popular models of speech signals, as well as a list of Gaussian distributions of probability densities in order to distribute signals over a stationary short time interval, usually correspond to a pronunciation unit.

The most popular learning approach is the hybrid HMM–DNN architecture, in which the structure of the HMM remains unchanged, and the GMM goes away, and its place is taken by the deepest neural network. This neural network can simulate the characteristics of speech signals. For example, Recurrent Neural Network (RNN) was used to train LMs in a significant number of studies [3]; in [4], authors use the Long Short-Term Memory (LSTM) networks to obtain a dictionary. In turn, in the work [5], deep neural networks demonstrated sufficiently high results for the formation of acoustic models, and in the work [6], a method for determining signs due to limited Boltzmann machines was demonstrated. Therefore, ideas arose for the use of artificial neural networks at various periods of speech recognition.

Deep learning methods, thanks to high-performance GPUs, are used in practice, and this approach is called the integral method. With this approach, in the process of training a neural network, only one of the models can demonstrate the necessary result without using other elements — an integral model. These or other integrated networks can be formed due to the fact that a number of recurrent and integral layers will be added, which are language and acoustic models, and thanks to them, speech data at the output are compared with each other using transcription. Currently, there are a number of ways to implement integral models – encoder-decoder models, as well as connective time classification. They are based on the mechanism of attention. In a variety of speech recognition tasks, special emphasis is placed on the integral approach [7]. In a large number of papers, it was proved that the success in performance of the integral approach depends on the number of training data: a large amount of data gives better resulting model. There are many applications that implement on the integral approach: BaiduDeepSpeech, GoogleListen, Attend, Spell, SpeechtoTranslatorTTS, and VoicetoTextMessenger. The main reason for this is the availability of a sufficient amount of big data to train end-to-end systems. Based on the analysis above, it is possible to notice the basic issue associated with the recognition of low-resource languages that are part of the agglutinative group. There are no common training data corpora for these languages.

In order to improve the integral approach in various models – encoder-decoder, as well as CTC, which are based on the mechanism of attention, various variants of networks were carried out. To apply local correlations to speech signals, complex encoders, which include ultra-precise neural networks, were used. These models apply certain advantages to all sub-models and affect the occurrence of restrictions for the model. The analysis above has a positive effect on the degree of performance of various integrated systems. Past studies have revealed that deep learning models in a variety of languages are the most reliable, and the multitasking approach is the most optimal for integral learning [8].

Within the framework of this work, the recognition of agglutinative languages has the goal of solving the problems of a limited speech resource in an integrated architecture.

Related works

Models that are based on a co-sectional time classification and are necessary for speech recognition carry out their activities, excluding the initial alignment of output and input sequences [9]. The CTC is formed in order to decode the language. Hannun and his team used Baidu for speech recognition, which uses a parallel learning algorithm with the use of CTC.

In [10], it is proposed to use deeper recurrent convolutional, as well as deep residual networks together with CTC. The most optimal result was obtained using residual networks

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 13, MARCH 2023

with butch normalization. Thanks to this, it was possible to obtain a PER, which on the TIMIT speech corpus was equal to 17.3%.

An alternative version of CTC is Sequence to Sequence with attention [11]. These models include both a decoder and an encoder. Thanks to the encoder, audio frame data is compressed into the most compact representation due to the reduction in the number of neurons in the layers. In turn, the decoder, based on a compressed representation, as well as a recurrent neural network, is working to restore the stages of words, symbols and phonemes.

In [12], the CTC model is proposed using the deepest convolutional networks instead of recurrent networks. The most optimal model, which is based on convolutional networks, had ten convolutional layers, and three fully connected layers. The best PER is 18.2%, given that the best PER for bidirectional LSTMs is 18.3%. Testing was carried out on the basis of TIMIT. It was possible to determine that convolutional networks make it possible to make the learning rate faster, and they are the most optimal for learning on phoneme sequences.

Within the framework of the CTC network, the output values of the neural network are the transition probabilities. Neural network architecture – bidirectional LSTM networks. Three models were compared among themselves: RNN-CTC model, RNN-CTC model, as well as a minimized retrained WER and a basic hybrid model, which was written thanks to Kaldi tools [13].

Soltau and a number of others [14] performed context-dependent phoneme recognition, which trained the model based on the CTC in the YouTube video caption task. In Sequence-to-sequence models, there is an insufficient amount of recognition by about 13-35%, when compared with basic systems. There is a "generalization" of CTC models, namely, a transformative RNN, on the basis of which two RNNs are combined with each other into a specific transformative system [15]. One network is similar to the CTC network, and is engaged in processing the same time period, which is processed by the input sequence. In turn, thanks to the second RNN, the probability of future meteors is modeled, taking into account past ones [16]. In CTC networks, dynamic programming is used in order to calculate algorithms, as well as forward and reverse transition algorithms, while taking into account the limitations present in the two RNNs. If compared with CTC networks, thanks to the use of RNN converter, it is possible to form output sequences that are the longest than the input ones [17]. RNN converters have demonstrated fairly good results in the process of recognizing phonemes with PER, which is 17.7%, and is based on the TIMIT corpus.

Proposed end-to-end ASR system

In the work, the methodology was carried out in the following way:

CTC function. The CTC in the training of neural networks participates as a loss function. The sequence determined as an output of a neural network is described by the following formula: $y = f_w(x)$ [18]. The Neural network's output layer includes one block for each of the characters of the output sequence, as well as an additional <blank> character. All components of the output sequence are probability distribution vectors for all symbols G' in a specific time period t**.** So, component y^t is the probability that in a specific time interval t the character will be pronounced in the input sequence k from a set of characters G' [19].

So, let α be a sequence of characters of length T and indexes blanks, according to x. Probability $P(\alpha|x)$ is represented as follows: the product of the probabilities of the appearance of symbols in each time period:

$$P(\alpha|x) = \prod_{t=1}^{T} y_{a_t}^t, \forall \alpha \epsilon G^t \quad (1)$$

Let B – is an operator that removes repetitions of characters and blanks.

DOI: 10.37943/13IMII7575

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

The above formula is calculated using dynamic programming, and the neural network will be trained to minimize the CTC function:

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

Decoding is based on the next assumption:

$$\arg max P(y|x) \approx B(\alpha*) \quad (4)$$

wherever $\alpha^* = \overset{w}{\arg max_\alpha} P(\alpha|x)$.

Attention model:

Attention is a mechanism of the Encoder-Decoder, which was formed in order to improve the level of performance of RNN in the process of speech recognition. The encoder is a neural network, which includes DNN, BLSTM, and CNN. It changes the input sequence $x = (x_1, \dots, x_L F)$, to define features in a specific intermediate representation: $h = (h_1, \dots, h_L)$ [20].

$$h = Encoder(x_1, \dots, x_L F) \quad (5)$$

Decoder – this is the usual RNN, which uses an intermediate representation to generate output sequences:

$$P(y|x) = AttentionDecoder(h, y) \quad (6)$$

In the form of a decoder, a recurrent sequence generator was used, which is based on the mechanism of attention [21].

Dataset

The information for the analysis was presented thanks to the laboratory of "Computer Engineering of Intelligent Systems". In this regard, the following were used: professional recording studio Vocalbooth.com, which is noise-insulating [22].

There were people among the announcers with no problems with pronunciation. A total of 380 speakers of different genders and ages participated in the recording. The process of voicing and recording one speaker lasted about 50 minutes. Each speaker read his own text, which included one hundred sentences recorded in separate files. Each sentence included an average of seven words with a fairly rich phoneme. Textual information was collected from news sites that write in Kazakh, but other data that is stored electronically was also used. In total, data recording reached 123 hours [23]. During the recording period, transcriptions were formed, which means a description of each of the audio files in text form. The formed case allows, first of all, to work with fairly large amounts of data, to offer system characteristics. In addition, it is possible to investigate how database extensions affect the speed of text recognition [24].

Audio materials have identical characteristics:

1. PCM is a method of forming a file into a digital form;
2. 16 bit file size;
3. One audio channel;
4. Wav is a file extension;
5. The frequency is 441.1 kHz.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 13, MARCH 2023

As part of the training of the integrated system, two corpuses were used, which include:

- Turkish Language Corpus: http://www.tnc.org.tr/
- The Tatar language Corpus: https://commonvoice.mozilla.org

The implementation of the integrated speech recognition system using the CTC function was implemented using TensorFlow. The Essen tool in TensorFlow was used in this system. This system makes it possible to use language models that were formed in the Kaldi format, but no additional conversion was applied. Tensor2Tensor5 was used to carry out experiments.

The experiments were provided using the server with the next characteristics: 100 CPU/1TBRAM/1TB SSD. This server has a high-performance graphics card, the NVIDIA TESLA P100.

Experiments

In the experiments [25], to extract the features, we used low frequency kepstral coefficients (MFCC) with the first 13 calculated coefficients. All training data was divided into training (90%) and cross-validation (10%). At the second stage of the experiment, we will describe the results of the model based on the CTC loss function. The results of the corresponding CTC models are presented in Table 1. In our research, we used several types of neural networks: ResNet, LSTM, MLP, and Bidirectional LSTM. Pre-tuning neural networks without a language model gave us the best results:MLP: there were 6 hidden layers with 1024 nodes, when using the Relay activation function with an initial learning rate of 0.007 and an attenuation coefficient of 1.5.

LSTM: there were 6 layers with 1024 units in each with a dropout equal to 0.5 s, an initial learning rate equal to 0.001, and a attenuation coefficient equal to 1.5.

- ConvLSTM: used one two-dimensional convolutional layer with 8 filters, ReLU activation function. Then it drops out with a retention probability equal to 0.5.
- LSTM: used 6 layers with 1024 units and dropped out with a retention probability equal to 0.5.
- ResNet had 9 residual blocks with normalization (batch-normalization).

In the first experiment for encoder-decoder models based on the attention mechanism (attention-based models), to extract the features, we used an algorithm MFCC.

In the first experiment for encoder-decoder models based on the attention mechanism (attention-based models), to extract the features, we used an algorithm MFCC, The CTC function was used in the neural network. We did not use language models in this model. In the second experiment, we used the MFCC and language models.

Table 1. Results of CTC models.

Model	CER%	WER%	Decode	Train
Models that do not use language models.				
MLP	48.12	59.25	0.2031	131.3
LSTM	36.42	46.5	0.2150	421.2
Conv+LSTM	34.91	39.30	0.2686	465.1
BLSTM	33.6	37.65	0.2721	491.5
ResNet	32.52	36.56	0.2656	192.6
Models which use language models and MFCC.				
MLP	39.14	63.29	0.0199	156.2
LSTM	24.33	46.41	0.0162	511.3
Conv+LSTM	22.91	39.30	0.0080	465.8
BLSTM	13.60	20.64	0.0026	591.3
ResNet	11.53	19.59	0.0054	242.1

In the following experiment, we tested LSTM and BLSTM neural networks. In our model, 6 layers of 256 units were used with an initial dropout reduction with a probability of saving 0.7 in the encoder. As a decoder, we used LSTM and an encoder-decoder models based on the attention mechanism.

Discussion

The results are given in Table 2. Provided experiments show that the CTC model works well for agglutinative languages, without integration of LMs, but ResNet is still the best with a result of CER equal to 11.52% and WER equal to 19.57% using a language model. It shows the importance of LMs as a part of speech recognition.

Table 2. Results of the models based on the attention mechanism.

Model	CER%	WER%
LSTM	8,62	17,57
BLSTM	8,02	17,90

The CTC model can make mistakes while constructing words and other tokens, like sentences, from recognized symbols. But after the experiment, we explored whether using the encoder-decoder model based on the attention mechanism for agglutinative languages without LMs allowed us to achieve good results. The encoder-decoder based Bi-LSTM neural network based on the attention mechanism showed a result of CER equal to 8.02% and WER equal to 17.90%.

Conclusion

In this paper, we considered the task of recognizing agglutinative languages applying an end-to-end approach, such as the CTC and the attention-based models. In experiments, different types of modern neural network were used: MLP, LSTM, Bi-LSTM and ResNet. By our experiments, it was proved that it is possible to apply existing neural network types for LSTM and BLSTM; moreover, the integration of language models is not necessary to obtain reasonable performance from ASRs. The results obtained with ResNet were the best; they were better than even the results of the basic hybrid models. The plan for future works included the use of other models, modification of attention mechanisms, and application of various feature extraction mechanisms. Moreover, in the future, the model of conditionally random fields will be applied (ConditionalRandomFile).

References

1. Perera, F.P., Tang, D., Rauh, V., Tu, Y.H., Tsai, W.Y., Becker, M., ... & Lederman, S.A. (2007). Relationship between polycyclic aromatic hydrocarbon–DNA adducts, environmental tobacco smoke, and child development in the World Trade Center cohort. *Environmental Health Perspectives*, *115*(10), 1497-1502. https://doi.org/10.1289/ehp.10144
2. Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A., & Turdalykyzy, T. (2019, March). Automatic recognition of Kazakh speech using deep neural networks. In *Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8–11, 2019, Proceedings, Part II* (pp. 465-474). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14802-7_40
3. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). Recurrent neural network based language model. In *Interspeech* (Vol. 2, No. 3, pp. 1045-1048).
4. Rao, K., Peng, F., Sak, H., & Beaufays, F. (2015, April). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on*

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 13, MARCH 2023

Acoustics, Speech and Signal Processing (ICASSP) (pp. 4225-4229). IEEE. https://doi.org/10.1109/ICASSP.2015.7178767

5. Jaitly, N., & Hinton, G. (2011, May). Learning a better representation of speech soundwaves using restricted boltzmann machines. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5884-5887). IEEE. https://doi.org/10.1109/ICASSP.2011.5947700
6. Joshua, B. (2019). *Effective Java* (3rd ed.). Addison-Wesley.
7. Vaněk, J., Zelinka, J., Soutner, D., & Psutka, J. (2017). A regularization post layer: An additional way how to make deep neural networks robust. In *Statistical Language and Speech Processing: 5th International Conference, SLSP 2017, Le Mans, France, October 23–25, 2017, Proceedings 5* (pp. 204-214). Springer International Publishing. https://doi.org/10.1007/978-3-319-68456-7_17
8. Kim, S., Hori, T., & Watanabe, S. (2017, March). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4835-4839). IEEE. https://doi.org/10.1109/ICASSP.2017.7953075
9. Speech & Voice. https://labs.mozilla.org/learn/speech/
10. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., & Courville, A. (2017). Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*. https://doi.org/10.48550/arXiv.1701.02720
11. Dragon Naturally Speaking Solutions. (2018). http://www.dragonsys.com
12. CMUSphinx Wiki Tutorial. (2018). http://cmusphinx.sourceforge.net/wiki/
13. Bot API Tutorial. (2018). https://tlgrm.ru/docs/bots/api
14. FFmpeg Filters Tutorial. (2018). https://www.ffmpeg.org/ffmpeg-filters.html#afftdn
15. Journal of Engineering Trends and Technology. (2018). *4*(2).
16. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597
17. Englund, C. (2004). Speech recognition in the JAS 39 Gripen aircraft-adaptation to speech at different G-loads. *Centre forSpeech Technology, Stockholm*, 2.
18. Dongsuk, Y. (2019). *Robust speech recognition using neural networks and hidden markov models*. [Doctoral thesis, The State University of New Jersey].
19. Giampiero, S. (2017). *Mining speech sounds, machine learning methods for automatic speech recognition and analysis*. [Doctoral thesis, Stockholm: KTH school of computer science and communication].
20. (2019). An Open-Source Machine Learning Framework for Everyone. Tensorflow. https://www.tensorflow.org/
21. Benesty, J., Sondh, M., & Huang, Y. (2008). Springer handbook of speech recognition. *NY: Springer*, *1176*.
22. Vyas, G., & Kumari, B. (2013). Speaker recognition system based on MFCC and DCT. *International Journal of Engineering and Advanced Technology (IJEAT)*, *2*(5), 145-148.
23. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.
24. Nilsson, M., Ejnarsson, M. (2018) *Speech recognition using hidden Markov model*. Karlskrona: Kaserntryck-eriet AB.
25. Bekarystankyzy, A., Mamyrbayev, O. (2023). Integral'naja sistema avtomaticheskogo razpoznavanija slitnoj rechi dlja aggljutinativnyh jazykov [Integral automatic fusion speech recognition system for agglutinative languages]. Proceedings of the National Academy of Sciences of Kazakhstan. Physics and Mathematics Series, (1), 37–49. https://doi.org/10.32014/2022.2518-1726.167