ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse249942020.pdf
https://doi.org/10.30534/ijatcse/2020/249942020

Development of Automatic Speech Recognition for Kazakh Language using Transfer Learning

Amirgaliyev Beibut[1], Kuanyshbay Darkhan[2,3], Baimuratov Olimzhan[3], Kutubayeva Madina[4]
[1]Professor, Astana IT University, Kazakhstan, amirgaliyev@gmail.com
[2]Institute of Information and Computational Technologies, Kazakhstan, darkhan.kuanyshbay@sdu.edu.kz
[3]Suleyman Demirel University, Kazakhstan, olimzhon.baimuratov@sdu.edu.kz
[4]L.N. Gumilyov Eurasian National University, Kazakhstan, mada96@mail.ru

ABSTRACT

Development of an automatic speech recognition system for the Kazakh language is a challenging task due to the lack of audio data and specificity and complexity of the language itself. In this paper, we propose a new method which gets a pre-trained model of the russian language and uses the weight values of the pre-trained model in the proposed neural network. The main reason for choosing the Russian language model is that the pronunciation of the Kazakh and Russian languages is very similar in many respects, because they account for 78% of the total letters and there is a rather large corpus of the Russian speech dataset. The dataset of Kazakh speech with transcriptions was formed by the university's faculty. In general, 50 native speakers were involved who generated about 400 sentences. A special technology has been created for the automatic expansion of the database. The data was extracted from well-known Kazakh books such as "Abai zholy", "Kara sozder", etc.

Key words : Automatic speech recognition, transfer learning, neural networks, connectionist temporal classifier, recurrent neural networks.

1. INTRODUCTION

Automatic Speech Recognition tasks (ASR) are very challenging, although the results are improving and growing due to the raise of required data, advancement of graphical processors (GPU) and “fine-tuning” of neural network. The most accurate and state-of-the-art ASR systems [1-4] for English language and mandarin that has been developed recently uses Switchboard, Fisher, TIMIT dataset with over 2000 hours of continuous speech data. This amount of data is clearly enough to set a particular experiment and gain promising results. Despite the usage of recent neural network techniques, the main influence of desired output is the dataset capacity, which for popular languages like English is growing fast. Another important influence on accurate ASR system is the selection of right neural network structure and tuning parameters, which leads to fast training and resolving the over fitting problem.

In digital signal processing, speech processing is one of the areas that is used in many type of applications. It is one of an intensive field of research. The major criterion for good speech processing system is the selection of feature extraction technique, which plays a major role in achieving higher accuracy. In this paper, most commonly used techniques for feature extraction such as Linear Predictive Coefficient (LPC), Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), Relative Spectral Perceptual Linear Prediction (RASTA-PLP) and Wavelet Transform (WT) are presented [5-7].

End-to-end ASR systems have already overcome the traditional HMM and DNN systems due to its simplicity and convenience, where there is no need to have the usage of language model, pronunciation model etc. This [8] model have been built with the help of the technique, which is called CTC (Connectionist Temporal Classifier). CTC makes the automatic segmentation of audio signal and maps the audio wave directly to transcriptions. Neural network structure based on RNN (recurrent neural network) where, each neuron returns the probability distribution of all characters including the blank space for each segment of an audio wave (Figure 1). To find the CTC loss they sum all corresponding sequence distributions. Decoding part is done by applying an algorithm called beam search or greedy search.

The model that has been developed here [9] uses the same technique with a small advancement, which is called attention based CTC model. Basically, they combine the CTC loss function with attention function that is used in sequence-to-sequence model to build an effective and robust ASR system. This method improved the accuracy of the simple CTC model up to 17%, which was a huge leap forward.

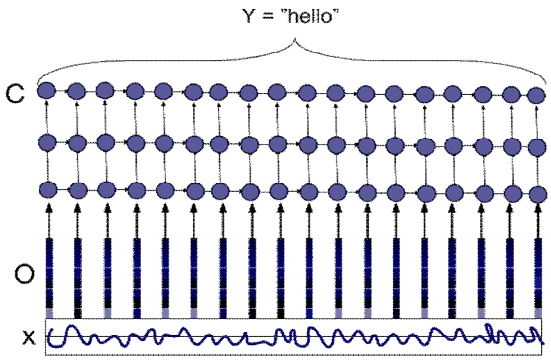

Figure 1: Connectionist Temporal Classifier

Recent usages of deeper LSTM networks show that they noticeably outperform the fully connected neural networks [10]. The approach that was presented here [11] using LSTM on speech recognition task reduces the character error rate down (CER) to 14% on Shenma voice search data. They were able to train a deep 7 layer LSTM network with using a layer-wise training and exponential moving average methods. In this paper [12] authors present the projection layer between output layer and LSTM layer. The presented method improves the performance of this LSTM over the traditional LSTM.

The main disadvantage of these models [8-9], they require a large amount of data, which is a big problem for Kazakh language, because it hasn't been investigated and researched well. The datasets that exist today in Kazakh language are mostly private and not available for free. Even so, these datasets are not big enough to get a good result.

Considering all the disadvantages and obstacles, the problem with a lack of data has been partially solved. Firstly, a convenient website tool was built which contains a lot of sentences in Kazakh language and allows the user to pronounce the record and save these sentences in a comfortable way. Such a way, almost 20 hours of data has been collected and it continues to grow. Secondly, the existing [13] speech recognition model was taken for Russian language based on VoxForge dataset with over 100 hour of speech data with corresponding text transcriptions and the weights of first 2 LSTM layers with 128 neurons were copied and pasted to our neural network with the same amount of layers and neurons. The weights of the last layer were deprecated because it's size doesn't match with layer's size, since the number of characters in Russian alphabet is less than in Kazakh alphabet. As a result, the LER was lowered down to 19% with the help of external Russian language model, which is a great leap forward.

The paper is organized as follows: Section 2 explains the transfer learning in speech recognition systems. Section 3 discusses about datasets and preprocessing that has been done. Section 4 contains an experiment with 4 scenarios with its construction illustrations and the results. After makes the conclusion of the experiment and the future plans related to ASR system of Kazakh language and data collection methods.

2. CONNECTIONIST TEMPORAL CLASSIFICATION

For an automatic labeling problem where each label taken from an alphabet A, CTC has softmax layer with one symbol more than there are in alphabet A. For a given input sequence at the particular time the activations of the |A| symbols are probabilities of corresponding labels. The activation of an extra unit is a probability of "blank symbol". Each complete network outputs used to define a distribution over all possible label sequences that match to the original input with the same length. The whole sequence of outputs are used to find out a distribution of all possible label sequences with the same length as input sequences.

Labeling the alphabet $A' = A \cup \{blank symbol\}$, the output k at the particular time t with an activation function y_t^k is a probability of outputting the element k of a given A' at the time t with a given length T and input x. Let us define the A'^T as a set of sequences with a length T over A'. Assuming that the probability output at each time is independent of other times, we will get the following conditional distribution:

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t$$

In order to differentiate π sequences over A' from the label sequences l over A, we refer them as paths. After that, we define the function $F: A'^T \rightarrow A^{\leq T}$, which converts the set of paths into set of possible label sequences of x. This is can be achieved by removing the blanks and repeated elements in the paths. For instance, $F(a\_b\_b\_cc) = abbc$. We will get the probability of some label sequence $l \in A^{\leq T}$ by summing up the probabilities of all paths matched to the label sequence:

$$p(l|x) = \sum_{\pi \in F^{-1}(l)} p(\pi|x)$$

This way of matching the set of paths onto the same label sequence makes the unsegmented data work well or CTC, because it makes it possible for network to predict the labels not knowing their occurrence place.

Initial CTC model did not require blank symbols, where all the repeated symbols removed in the process. This process had 2 main problems. The first problem happened when there are 2 consecutive elements, which cannot be handled, because the transition appears only between different elements. The second problem happened when there are unlabeled data (usually pauses or noises) in between the labels, because can be extremely expensive task, since it requires to predict the first label until the second one began. Therefore, CTC model changed adding the blank space in each label sequence in order to prevent these problems.

2.1 Feature extraction

Feature extraction step provides the numerical representation of the speech signal. This representation should provide the minimal information loss compared to the original speech. It can be interpreted as dimensionality reduction technique, which transforms the original full-size data into reduced-size data without influencing ending outcome.

Feature extraction widely used and applied in a lot of tasks like speaker verification, gender recognition, language recognition etc. For an automatic speech recognition tasks there are 2 popular feature vector representations which are linear predictive coding and Mel-frequency cepstrum coefficient.

2.2 Mel-frequency cepstrum coefficient (MFCC)

MFCC are one of the most popular and often used feature vectors in Automatic speech recognition (ASR) system.

The MFCC try to imitate the behavior of the human ear's bandwidths having frequencies below 1 kHz. It has calculated by dividing the audio signal into overlapping frames. If the frames divided into N samples and for each frame, hamming window is multiplied and the equation is:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

In the next step, signal will be converted from time domain into frequency domain by using Fourier transform. The equation of Discrete Fourier transform (DFT) for the signal is the following:

$$X_k = \sum_{i=0}^{N-1} x_i e^{-\frac{j2\pi ki}{N-1}}$$

After DFT calculation step, the Mel for a defined frequency is calculated by converting the frequency domain into Mel frequency scale, which is suitable for human perception. This is handled by triangular filters in order to make the approximation of a Mel scale. The equation for calculating the Mel scale with a given frequency is the following:

$$M = 2595\ log_{10}\left(1 + \frac{f}{700}\right)$$

The next step converts the Mel scale into time domain using discrete cosine transform (DCT). The calculation of DCT is done by the equation below:

$$X_k = \alpha \sum_{i=0}^{N-1} x_i \cos\left\{\frac{(2i+1)\pi k}{2N}\right\}$$

After the step above, we end up with acoustic vectors, which are called Mel frequency Cepstrum Coefficient. The MFCC feature vectors are good for representing the individual characters of each speech for further recognition tasks like speaker identification.

3. SOME COMMON MISTAKES

Transfer learning is a novel approach that makes the training a lot better and accurate by transferring the knowledge from different task to a current task (Figure 2).

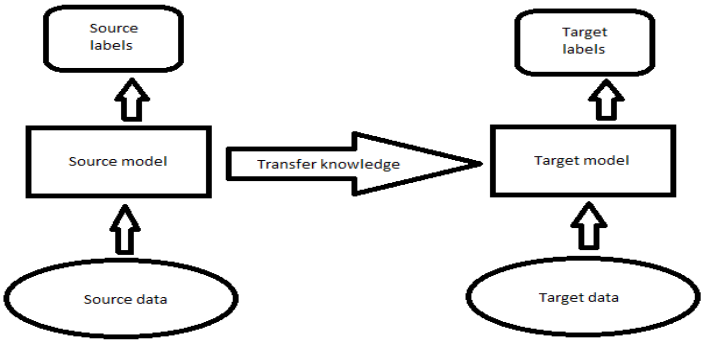

Figure 2: Transfer learning illustration

Just as human beings, any model can be trained, learned significantly faster and efficient, if it had a previous experience on related tasks. Usually, source model is the model that has been trained on a massive amount of data, whereas the target model can be trained on small amount. Therefore, transfer learning can solve a problem with a lack of data. Nowadays, there are a lot of pre-trained models that are open-sourced and can be accessed very easily. The main idea behind transfer learning is to transfer the features and parameters (weights) from source task to target task. Basically, source model is considered as a starting point for a target model.

There are a lot of researches being done on multilingual speech recognition tasks [14-18]. The most common problem that they are facing till these days is that languages are specific. Language adaptive acoustic models can be built if languages share acoustic, pronunciation, phonetic properties. This [19] paper presents the interesting approach on different region specific Indian 9 Languages, where each of these languages share the same phonetic, acoustic features and properties. They have combined all grapheme sets and trained on a sequence-to-sequence model. The result they have gained was 21% improvement on performance compared to the same model trained individually for each language.

For English, German and Spanish data from Cortana there has been an experiment presented here [20]. They have built a multilingual ASR system by using the universal character set shared around all languages. Each language had 150 hours of training set and 10 hours of validation and test set. By using [21], they have resulted 81 English labels, 93 German label and 97 Spanish labels. Therefore, they have built a universal

label set (108) with 75% shared overlapping labels. By creating the mechanism which is called the language specific gating mechanism they have trained their model, which outperforms the monolingual approach.

In this [22] paper they applied the transfer learning to a Text-to-Speech task, which is able to generate an audio with data, that has never been seen before. They have used 3 different pre-trained components: 1) *speaker encoder networ*k that has been trained on noisy speech dataset with no transcriptions; 2) *synthesis network* which is trained to generate mel spectrogram to text; 3) *vocoder network* which converts mel spectrogram to the waveform with time domain.

This [23] paper applies the transfer learning approach on developing ASR system for German language. With a limited training data, they have adapted a Wav2Letter model, which is originally trained on English language. The paper [24] presented by Vu and Schultz have developed multilingual Multilayer Perceptrons (MLP). This MLP later was applied as a starting point on target languages like Vietnamese, Czech and Hausa. As a result, they have obtained the improvement on error rate up to 22%.

The proposed approach that was taken is almost the same as [23, 24], in which the development of an ASR system for Kazakh language with the help of Russian language was applied. The pre-trained model for Russian language has been trained on VoxForge dataset having a neural network structure with 2 LSTM layers with 128 neurons each and dense layer. Neural network has been trained with 700 epochs and has the same model with Bidirectional LSTM. It uses the same CTC loss function (Figure 3).

Basically, the pre-trained model was used to extract the weight matrix and copy to our exactly the same neural network structure with 2 LSTM layers. After application of "fine-tuning" to our neural network, we start to train 20 hour of Kazakh speech dataset.

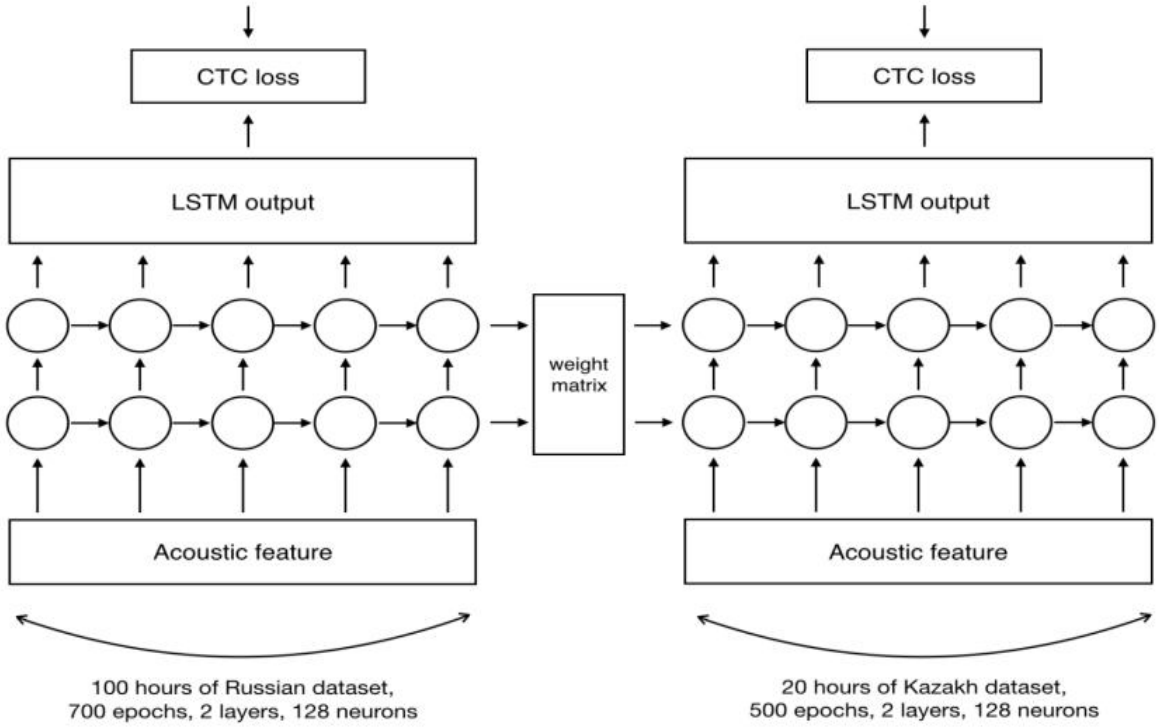

Figure 3: Weight Transferring

4. DATASETS AND PREPROCESSING

The dataset for Kazakh language has been collected in the base of Suleyman Demirel University. By using special tool (website), 50 native speakers have been involved in pronouncing and saving their utterances. The sentences have been collected from famous Kazakh books "Kara sozder", "Abay zholy" etc. Each speaker has pronounced approximately 400 sentences. Audio files with duration longer than 15 seconds have been omitted, in order to have a strong dependency between transcriptions and audio files. Generally, there has been gained around 20 hours of data. All audio files have been sampled to 16 kHz. Using a librosa library, feature extraction has been done by the Mel Frequency Cepstral Coefficients (MFCC). Text files have been normalized by removing all the unnecessary characters and representing in lower case.

The overall duration of our speech corpus, slightly inferior compared to corpuses that has been described here [25, 26], where they have biggest speech corpus for Kazakh language. They have collected around 30 hours of data with 200 different speakers with different genders and ages. Since, our dataset has

been collected in limited amount of time with special developed website; we will soon pass their corpus in term of duration.

5. EXPERIMENTS AND RESULTS

For the network the RNN based architecture was built with 2 Long-Short Term Memory layers and 1 dense layer. The training has been done on several Graphical Processors (GPU) Tesla K80. The environment that has been selected for this task is Jupiter Notebook on Python language (Tensorflow library). Dataset has been cloned from github repository and split 80%, 10% and 10% for training, validation and testing respectively. The parameters of neural network are the following:

- 2 layers of LSTM and BiLSTM separately
- 128 neuron for each layer
- 500 epochs
- Dropout layer after each LSTM layer with 50% probability
- Batch size is 4
- Momentum value for MomentumOptimizer is 0.9
- Learning rate is 0.0005
- CTC loss function

Metric is Label Error rate (LER)

Table 1: Results of Training

RNN type	Training cost	Training LER	Validation cost	Validation LER	Epochs
LSTM	15.603327	0.054680485	3.98756	0.01569438	500
LSTM with Russian model	14.426534	0.056311063	4.78974	0.01453637	500
BiLSTM	18.366533	0.062855324	4.25945	0.01602836	500
BiLSTM with Russian model	13.924501	0.042720266	3.87567	0.014177615	500

4 different scenarios have been considered (Table 1): 1) LSTM neural network without Russian model; 2) LSTM neural network with Russian model; 3) Bidirectional LSTM without Russian model; 4) Bidirectional LSTM with Russian model. The result of each recurrent neural network actually very close, but it can be observed that the architectures with transfer learning clearly make an improvements on everything.

LSTM layered neural network with external model has improved the training cost up to 8%, whereas Label error rate has increased up 4%. Bidirectional LSTM has showed very promising results, improving the training cost up to 24% and decreasing the label error rate down to 32%.

The experiments above showed that using an external Russian ASR system model to transfer its knowledge to Kazakh language system improves the performance decently.

5. CONCLUSION

As a conclusion, it is clear that using an external Russian language model can partially solve the lack of data problem. Our model has been trained using 2 different neural networks (LSTM, BiLSTM) and each of them trained by transferring the weights from external model. This external model was trained on VoxForge dataset with 100 hours of Russian speech. For Kazakh language dataset, around 20 hours of Kazakh speech was gathered using famous Kazakh books. Results showed that BiLSTM model with external Russian model improved the performance very well, lowering the training cost and LER down to 24% and 32% respectively.

The dataset of kazakh speech with transcriptions in the base of Suleyman Demirel University with 50 native speakers each having around 400 sentences has been collected. Data has been chosen from famous kazakh books like “Abay zholy”, “Kara sozder” etc.

In the experiment 4 different scenarios have been proposed. First, the neural network was trained without using a pre-trained russian model with 2 LSTM (Long-Short-Term Memory) layers and 2 BiLSTM (Bidirectional Long-Short-Term Memory). Second, the same 2 LSTM layered and 2 BiLSTM layered networks have been trained using a pre-trained model. As a result, model’s training cost and Label Error Rate (LER) were improved by using external Russian speech recognition model up to 24% and 32% respectively. Pre-trained Russian language model has trained on 100 hours of data with the same neural network architecture.

In the future, for a more expanded presentation of the Kazakh speech recognition capabilities in order to increase the efficiency of the automatic Kazakh speech recognition system, we plan to solve the following tasks: collect a large amount of data in the Kazakh language using our web tool; to improve the performance and quality of the recognition system developed by us, consider integration with the most trained models from other languages; propose different neural networks for use, including improving the mathematical and

computer models we use in comparison with the best speech recognition models.

ACKNOWLEDGEMENT

This work was supported by "The best teacher" grant funded by the Ministry of Education and Science (Kazakhstan); and the grant of Institute of Information and Computational Technologies (grant holder) under Grant No.AP05132648 "Creation of verbal-interactive robots based on modern speech and mobile technologies" funded by the Ministry of Education and Science (Kazakhstan).

REFERENCES

1. Awni Hannun , Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng, **Deep Speech: Scaling up end-to-end speech recognition**, arXiv:1412.5567v2 [cs.CL], 2014
2. Long Wu , Ta Li, Li Wang, Yonghong Yan, Improving Hybrid CTC/Attention Architecture with Time-Restricted Self-Attention CTC for End-to-End Speech Recognition, Appl. Sci. 2019, 9, 4639; doi:10.3390/app9214639
3. Senmao Wang ,Pan Zhou , Wei Chen , Jia Jia , Lei Xie, **Exploring Rnn-Transducer For Chinese Speech Recognition**, arXiv:1811.05097v2 [cs.CL], 2019
4. Xu Tian, Jun Zhang, Zejun Ma, Yi He, Juan Wei, Peihao Wu, Wenchang Situ, Shuai Li, Yang Zhang, **Deep Lstm For Large Vocabulary Continuous Speech Recognition**, arXiv:1703.07090v1, 2017
5. Aditya Dhinavahi, Prashant R Nair , Sundeep V V S Akella, Aneeswar K S **Speech Recognition based Billing System: A multi-model design and implementation**, International Journal of Advanced Trends in Computer Science and Engineering, Volume 9 No.2, March -April 2020, https://doi.org/10.30534/ijatcse/2020/101922020
6. Mohammed Arif Mazumder, Rosalina Abdul Salam , **Feature Extraction Techniques for Speech Processing: A Review**, International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019, https://doi.org/10.30534/ijatcse/2019/5481.32019
7. Felizardo Reyes Jr., Arnel Fajardo, Alexander Hernandez **Convolutional Neural Network for Automatic Speech Recognition of Filipino Language**, International Journal of Advanced Trends in Computer Science and Engineering , Volume 9, No.1.1, 2020, https://doi.org/10.30534/ijatcse/2020/0791.12020
8. Paribesh Regmi, Arjun Dahal, Basanta Joshi, **Nepali Speech Recognition using RNN-CTC Model,** International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 31, July 2019
9. Takaaki Hori, Shinji Watanabe, John R. Hershey, **Joint CTC/attention decoding for end-to-end speech recognition,** Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 518–529
10. Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton, **Speech Recognition With Deep Recurrent Neural Networks**, arXiv:1303.5778v1 [cs.NE], 2013
11. Xu Tian, Jun Zhang, Zejun Ma, Yi He, Juan Wei, Peihao Wu, Wenchang Situ, Shuai Li, Yang Zhang, **Deep Lstm For Large Vocabulary Continuous Speech Recognition**, arXiv:1703.07090v1 [cs.CL], 2017
12. Has¸im Sak, Andrew Senior, Franc¸oise Beaufays, Long Short-Term Memory Based Recurrent Neural Network Architectures For Large Vocabulary Speech Recognition, arXiv:1402.1128v1 [cs.NE], 2014
13. O.O. Iakushkin , G.A. Fedoseev, A.S. Shaleva, A.B. Degtyarev, O.S. Sedova, **Russian-Language Speech Recognition System Based On Deepspeech,** Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018), Dubna, Moscow region, Russia, September 10 - 14, 2018
14. Arnab Ghoshal, Pawel Swietojanski, Steve Renals, **Multilingual Training Of Deep Neural Networks**, Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference, 2013 https://doi.org/10.1109/ICASSP.2013.6639084
15. Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, Chin-Hui Lee, A **Study On Multilingual Acoustic Modeling For Large Vocabulary** ASR, Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference, 2009
16. Samuel Thomas, Sriram Ganapathy, Hynek Hermansky, **Multilingual Mlp Features For Low-Resource Lvcsr Systems**, Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference, 2012 https://doi.org/10.1109/ICASSP.2012.6288862
17. Jui-Ting Huang, Jinyu Li, Dong Yu, **Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, Acoustics**, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference, 2013
18. Yajie Miao, Florian Metze, Improving Language-Universal Feature Extraction with Deep Maxout and Convolutional Neural Networks, Fifteenth Annual Conference of the International Speech Communication Association, 2014
19. Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, Kanishka Rao, **Multilingual Speech Recognition With A Single End-To-End Model**, arXiv:1711.01694v2 [eess.AS], 2018
20. Suyoun Kim, Michael L. Seltzer, **Towards Language-Universal End-To-End Speech Recognition**, arXiv:1711.02207v1 [cs.CL], 2017

21. Geoffrey Zweig, Chengzhu Yu, Jasha Droppo and Andreas Stolcke, **Advances In All-Neural Speech Recognition**, arXiv:1609.05935v2 [cs.CL], 2017
22. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, **Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis**, arXiv:1806.04558v4 [cs.CL] 2 Jan 2019
23. Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, Sebastian Stober, **Transfer Learning for Speech Recognition on a Budge**t, Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 168–177, 2017
24. Ngoc Thang Vu, Tanja Schultz, Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families, Interspeech. ISCA, pages 515–519.
25. Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev, Kuralay Mukhsina, Alimukhan Keylan, Bagher BabaAli, Gulnaz Nabieva, Aigerim Duisenbayeva, Bekturgan Akhmetov, **Continuous speech recognition of kazakh language**, itm Web of Conferences 24, 01012 (2019) https://doi.org/10.1051/itmconf/20192401012
26. Yedilkhan Amirgaliyev, Darkhan Kuanyshbay, Didar Yedilkhan, Shoiynbek A. **Automatic speech recognition system for kazakh language using connectionist temporal classifier**. // Journal of Theoretical and Applied Information Technology. -2020 -- Vol. 98. No. 04 -- 2020