IRSTI 16.31.21

Y. Kalmurzayev[1], D. Kuanyshbay[2], M. Othman[3]
[1,2]Suleyman Demirel University, Kaskelen, Kazakhstan
[3]Universiti Putra Malaysia

THE METHODS AND ALGORITHMS FOR RECOGNIZING KAZAKH LANGUAGE FEATURES

Abstract. Despite the importance of automatic speech recognition (ASR), it is difficult to find freely available models, especially for languages with few speakers. This paper describes a method for training Kazakh models based on end-to-end ASR architecture using open-source data. We put the models to the test, and the results are promising. However, much more training data is required to perform well in noisy environments. We make available to the public our trained Kazakh models and training configurations.

Keywords: Automatic speech recognition, neural networks, recurrent neural networks.

***

Аннотация. Несмотря на важность автоматического распознавания речи (ASR), трудно найти свободно распространяемые модели, особенно для языков с небольшим количеством носителей. В данной статье описывается метод обучения казахских моделей на основе сквозной архитектуры ASR с использованием данных из открытых источников. Мы протестировали модели, и результаты оказались многообещающими. Однако для хорошей работы в шумной среде требуется гораздо больше данных для обучения. Мы выкладываем в открытый доступ наши обученные казахские модели и конфигурации для обучения.

Ключевые слова: Автоматическое распознавание речи, нейронные сети, рекуррентные нейронные сети.

***

Аңдатпа. Сөйлеуді автоматты түрде танудың (ASR) маңыздылығына қарамастан, ашық кодты модельдерді табу қиын, әсіресе ана тілінде сөйлейтіндер аз тілдер үшін. Бұл мақалада ашық көздерден алынған деректерді қолдана отырып, ASR архитектурасына негізделген қазақстандық модельдерді оқыту әдісі сипатталған. Біз модельдерді сынап көрдік және нәтижелер үміт күттірді. Алайда шулы ортада жақсы жұмыс істеу үшін әлдеқайда көп дайындық деректері қажет. Біз ашық қол жетімдікте оқыту үшін өзіміздің оқытылған қазақстандық модельдер мен конфигурацияларды жариялаймыз.

Түйін сөздер: Сөйлеуді автоматты түрде тану, жүйке желілері, қайталанатын жүйке желілері.

1. Introduction

Automatic Speech Recognition (ASR) is the translation of an audio recording or spoken language into a text transcript. It is a key component of voice assistants such as Siri, Alisa, Google, speech translation devices, or for automatic transcription of audio and video files. For any language except German, French, English, available pre-trained models are still rare. For Kazakh language [1] we know only models trained on HMM (Hidden Markov Model) [2], [3] and Gaussian Mixture Model (GMM) [4]. For the recently introduced Mozilla DeepSpeech framework [5], [6], the Kazakh model is still missing. This is a serious obstacle for applied research of Kazakh speech data, since there are no services available. Therefore, we use publicly available speech data to train the Kazakh DeepSpeech model.

2. Methods and materials

In this paper, we focused on Mozilla's DeepSpeech framework because it is an end-to-end neural system that is fairly easy to train, unlike other frameworks that require more knowledge in their domain. Mozilla DeepSpeech (v0.9.3) was based on TensorFlow implementation of Baidu's end-to-end ASR architecture. Since it is under active development, the current architecture differs significantly from the original version. Figure 1 provides an overview of the v0.9.3 architecture. DeepSpeech is a deep recurrent neural network (RNN) [5], [6] at the symbol level, which can be trained end-to-end using supervised learning. It extracts Mel-Frequency Cepstral Coefficients as features and outputs the transcription directly, without the need for forced input alignment or any external knowledge source such as the Graphemeto Phoneme (G2P) converter. In general, the network consists of six layers: the speech features arrive at the three tightly coupled (dense) layers, followed by the unidirectional RNN layer, then the fully coupled (dense) layer, and finally the output layer, as shown in Figure 1. The RNN layer uses LSTM cells, and the hidden fully connected layers use the ReLU activation function. The network outputs a character probability matrix, i.e., for each time step, the system outputs a probability for each character in the alphabet, which is the probability that that character matches a sound. Further, the CTC loss function is used to maximize the probability of a correct transcription.

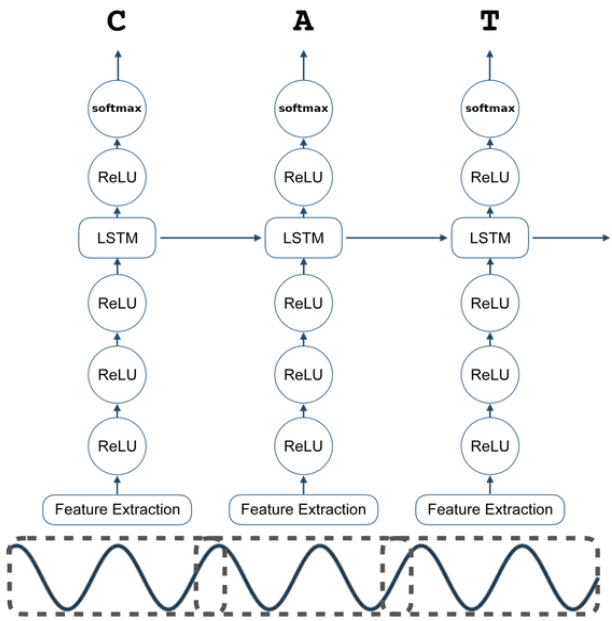

Figure 1: DeepSpeech architecture

3. Model training

In this section, we describe in detail our setup for training the Kazakh model in order to facilitate later attempts to train DeepSpeech models [7].

Datasets. We use publicly available datasets to train the Kazakh Deep Speech model. About 39 hours of audio recordings and transcriptions were used.

	wav_filename	wav_filesize	transcript
0	5fa504fc60eb0.wav	305644	ғайып құран мен тәпсірге сәйкес діни тағзым ет...
1	5fa504bd775db.wav	236204	жеке қаламгерлер арқылы болса да өзге ел әдеби...
2	5fa503abb5ba9.wav	127980	ханның қабылдауларына қатысып отырған деп те а...
3	5fa5018a756f5.wav	187884	театрландырылған қойылымдарымыз жәрмеңкелерімі...
4	5fa5010db4950.wav	235884	мәселен сіңбелік су жер қыртысының жоғарғы қа...
...	...	...	...
29994	5f4e4f5de8df0.wav	210306	қыран құсы көкті шарлап қалықтаған құлдилайды ...
29995	5f4e4f549d43c.wav	210306	ел ішінде көп тараған түрлері альфа пвп спайс ...
29996	5f4e4f4b43289.wav	139308	апа жыламаңызшы өздеріңіз аман есенсіздер ме
29997	5f4e4f422c8bf.wav	188460	өсімдік тіршілігіне қажетті құбылыстың бірі су...
29998	5f4e4f39b97ef.wav	253996	бүлдіршіндерді мектепке дейінгі мекемемен қамт...

Figure 2: The input file format

Preprocessing. For DeepSpeech [7], it is necessary to prepare audio and transcriptions in a certain format so that they can be read (Figure 2). We cleaned the transcriptions of unnecessary characters and converted everything to lowercase. We also checked that all audio clips were in.wav format. The data were divided into training data (60%), validation data (20%), and test data (20%).

Hyperparameter Setup. We chose a training rate of 0.0001 and a training/test batch size of 1.0. The number of hidden layers in the network was set to 100. The number of epochs has also been set at 100.

```
%cd /content/DeepSpeech/
! python3 DeepSpeech.py \
  --train_files /content/kz/kz_30000/clips/train.csv \
  --dev_files /content/kz/kz_30000/clips/dev.csv \
  --test_files /content/kz/kz_30000/clips/test.csv \
  --train_batch_size 1 \
  --test_batch_size 1 \
  --n_hidden 100 \
  --epochs 100 \
  --checkpoint_dir ../checkpoint \
  --export_dir ../model \
  --alphabet_config_path /content/kz/kz_30000/alphabet.txt \
  --scorer data/lm/kenlm.scorer
```

Figure 3: Hyperparameters used in the experiments

Language Model. We use the KenLM-trained probabilistic language model on the pre-processed corpus provided by Dauren Chapaev [8]. The corpus was created for the Kazakh language based on the Wikipedia database. It consists of 21 million words. Almost 600 thousand words have different derivations.

Server and Runtime. We trained and tested our model on a computing server with Tesla K80 Graphical Processors (GPUs). The environment that was chosen for this task is Jupyter Notebook(Python).

4. Experiments and results

Table 1 shows word error rate (WER) obtained by DeepSpeech training and testing on available Kazakh datasets. According to the results of language model testing, the accuracy of our model is 0.718301. Figure 4 shows that with each iteration of the model training, Loss decreases.

Table 1: Results of Training

	WER	CER	Loss	Source	Result
	0.718301	0.404967	91.590378		
Best	0.000000	0.000000	63.202705	германия федеративтік республикасының бавария жерінде орналасқан муниципалитет	германия федеративтік республикасының бавария жерінде орналасқан муниципалитет
Median	0.769231	0.385714	51.263573	тек ең жақын құрбым ғана қолымнан тартып есіңді жи деп бәйек боп жатыр	кеткен жақын құрылған қолымнан тартып есімді жетекші
Worst	1.428571	0.523810	130.729034	командирлерге батальон командирімізге мұғалімдерімізге барлықтарына алғысымыз шексіз	қара жерлер де бата бітірген қаражемістер аға емдемесе барлықтарына ағысымен

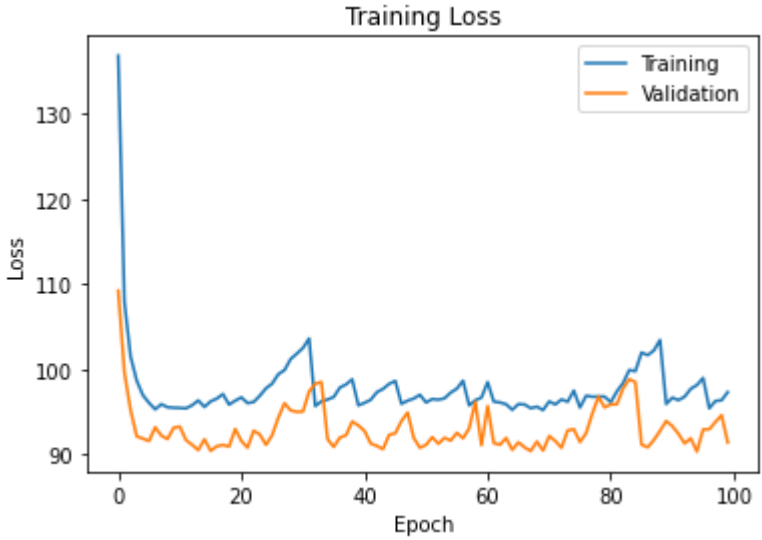

Figure 4. Loss vs. number of Epochs graph

5. Conclusion

This paper presents the results of building a Kazakh speech recognition model using DeepSpeech. Our model achieved a WER of 0.71. Our results support the view that Mozilla Deep Speech can be easily adapted to new languages. On new datasets, the model can be easily retrained and optimized. The trained model does not require special hardware and can be run on an ordinary desktop or laptop computer. The model is also easily adaptable to the Android operating system.

References

1 Ying Shi, Askar Hamdulla, Zhiyuan Tang, Dong Wang, Thomas Fang Zheng, A Free Kazakh Speech Database and a Speech Recognition Baseline, kazak_speech_recognition_add_jj.pdf (tsinghua.edu.cn).

2 Hongbing Hu, Stephen A. Zahorian, Dimensionality reduction methods for HMM phonetic recognition, (15) (PDF) Dimensionality reduction methods for HMM phonetic recognition (researchgate.net).

3 Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Gerald Penn, Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition https://www.researchgate.net/publication/261119155_Applying_Convolutional_Neural_Networks_concepts_to_hybrid_NNHMM_model_for_speech_recognition.

4 Jia Pan, Cong Liu, Zhiguo Wang, Yu Hu, Hui Jiang, Investigation of

Deep Neural Networks(DNN) for Large Vocabulary Continuous Speech Recognition: WhyDNN Surpasses GMMsin Acoustic Modeling. URL: https://wiki.eecs.yorku.ca/user/hj/_media/publication:dnn_asr_v4.pdf.

5 Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu, Deep Speech 2: End-toEnd Speech Recognition inEnglish and Mandarin URL: https://arxiv.org/pdf/1512.02595.pdf.

6 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. URL: https://arxiv.org/pdf/1412.5567.pdf

7 Deep speech documentation. URL: Deep Speech Model — DeepSpeech 0.9.3 documentation

8 Language Resources and Tools for Turkic Languages. URL: Turkic Languages Resources and Tools (itu.edu.tr).