Impact of Using a Bilingual Model on Kazakh–Russian Code-Switching Speech Recognition

Dmitrii Ubskii[1,2,3][0000−0003−1760−6837], Yuri Matveev[2][0000−0001−7010−1585], and Wolfgang Minker[3][0000−0003−4531−0662]

[1] STC-innovations Ltd, St. Petersburg, Russia
ubskiy@speechpro.com
[2] ITMO University, St. Petersburg, Russia
matveev@mail.ifmo.ru
[3] Ulm University, Ulm, Germany
wolfgang.minker@uni-ulm.de

Abstract. Due to the prevalence of bilingualism among Kazakh speakers, code-switching to Russian is common in Kazakh speech. That presents a challenge for monolingual Kazakh-language ASR systems that struggle to transcribe the embedded Russian words.
This paper attempts to determine the benefit of bilingual training on matrix language (Kazakh) and embedded language (Russian) monolingual data, as opposed to training on code-switched data only. Specifically, we evaluate the model's performance on matrix language words and embedded words separately.
We make use of two datasets: Kazakh speech with code-switching and Russian speech with no code-switching. We train a monolingual model on each dataset, and a bilingual model on a mixture of the two. The main objective of the experiments is to compare the performance of a model trained on code-switched speech with that of a model trained on full utterances in both languages.
Experimental results suggest that bilingual training improves the model's performance on matrix words, and greatly improves its performance on embedded words. We observe an absolute WER improvement of 14.69% in the code-switched words.

Keywords: speech recognition · code-switching · Kazakh language

1 Introduction

Previous attempt at building a bilingual Kazakh–Russian speech recognition system by Khomitsevich et al. [1] has uncovered two main challenges: lack of Kazakh language resources and large amounts of code-switching to and borrowing from Russian, a very phonotactically different language.

Code-switching (also referred to as code-mixing [2]) is a practice of alternating languages within an utterance that is common in bilingual and multilingual

communities. The dominant language in code-switched speech is often referred to as the *matrix language*, while the language whose elements are inserted into the dominant one is referred to as the *embedded language* [3]. Since it mostly occurs in informal conversations [4], difficulty of recognition of code-switched speech is compounded by the difficulty of conversational speech recognition.

Although most state-of-the-art ASR systems are monolingual, the impact of code-switching on ASR performance has recently sparked research interest [5–9]. The success so far has, however, been limited, largely due to the challenges outlined above.

Due to the majority of Kazakh speakers being bilingual [10], code-switching occurs commonly in Kazakh conversations. Because of this, it is important that any automatic speech recognition system deployed for Kazakh language is able to handle code-switching.

In this paper we attempt to determine the impact of training on both Kazakh and Russian language data on the quality of speech recognition of the embedded Russian segments in Kazakh speech.

The rest of the paper is organized as follows: Section 2 describes the dataset used in this work. Section 3 describes the model architecture and reports the experimental results. Finally, Section 4 concludes the paper and discusses the results.

2 Data

We make use of a proprietary Russian–Kazakh dataset consisting of Kazakh call centre operator recordings. No data augmentation techniques were used in the course of this work. Data statistics by language and subset are presented in Table 1.

Table 1: Data breakdown.

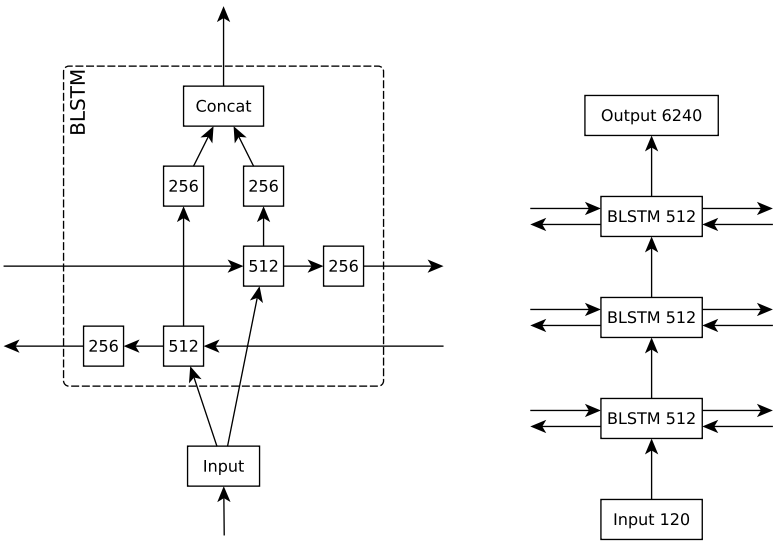
Language	Subset duration, hrs	
	training	evaluation
Kazakh	97.4	1.2
Russian	58.8	1.0

The domain of the data is very narrow, containing a significant amount of stock phrases and domain-specific words. Approximately 10% of words in the Kazakh language data are code-switched speech. The observed cases of code-switching include intra-sentential code-switching (insertion of Russian phrases into otherwise Kazakh sentences), as well as intra-word switching (Russian words conjugated as if they were Kazakh) [11]. Conversely, the amount of code-switching in the Russian language data is negligible.

3 Experiment

For training we use 40-dimensional log Mel-scale filter bank energy features with CMN with first- and second-order derivatives.

All the ASR systems are built using the Kaldi speech recognition toolkit [12]. For each set of data (code-switched Kazakh, Russian, and combined training set) we train a Deep Neural Network Hidden Markov Models (DNN-HMM) acoustic model [13]. The experiments are carried out using the *nnet3* setup of the Kaldi toolkit.



(a) Single BLSTM layer (b) Acoustic model

Fig. 1: Full BLSTM architecture.

For language modeling, all transcripts available for each set of data are merged and used to train a 3-gram language model. We use graphemic pronunciation dictionaries when compiling the language model into a WFST-decoder.

Acoustic models based on deep Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks have been demonstrated to be highly effective in various ASR tasks [14–16]. We use identical BLSTM architecture for the acoustic model for each set of data. Each has three hidden BLSTM layers with projections [17]. The dimension of each cell is 512, and the dimensions of the recurrent and non-recurrent projections are set to 256. The output layer consists of 6240 units. (See Fig. 1).

Each acoustic model is then trained using Natural Gradient for Stochastic Gradient Descent [18] and evaluated on appropriate evaluation data sets. Evaluation results are presented in Table 2.

Table 2: WER results for monolingual and bilingual models on Kazakh and Russian evaluation sets.

	WER, %	
	kaz	rus
Monolingual (kaz)	52.38	—
Monolingual (rus)	—	31.91
Bilingual	49.42	37.42
Improvement	2.96	–5.51

As seen in Table 2, the bilingual model displays higher performance on the Kazakh evaluation set at the expense of significant performance loss on the Russian evaluation set.

As the Russian language evaluation set contains no code-switched sentences, it and the Russian monolingual model are not considered further. Instead, we focus on the Kazakh evaluation set for closer examination.

To determine the impact of bilingual training on code-switching, we've collected per-word statistics used in WER calculation (Table 3). Each error—substitution (S), insertion (I), or deletion (D)—is classified based on the language the word belongs to. Note that substitutions are thus split into two classes: substitution with a Kazakh word or a Russian word.

Table 3: Per-word statistics by model and language.

	Target word language	Correct, #	Errors, #			
			S (kaz)	S (rus)	I	D
Monolingual (kaz)	Kazakh	3573	2173	103	490	529
	Russian	303	235	47	44	55
Bilingual	Kazakh	3508	2106	135	311	629
	Russian	377	189	30	24	44

We then calculate WER for matrix and embedded language words separately (Table 4). For the purposes of this calculation, all substitutions are considered to belong to the language of the token of the reference transcription.

The results shown in Table 4 show clear improvement in recognition of the embedded language words.

Table 4: WER by model and language of the word in reference transcription.

	WER, %	
	kaz	rus
Monolingual (kaz)	51.66	59.53
Bilingual	49.87	44.84
Improvement	1.79	**14.69**

4 Conclusions

In this paper we presented a bilingual Kazakh–Russian speech recognition system. We observe significant WER improvement in the matrix (Kazakh language) data, and 14.69% absolute WER improvement on the embedded (Russian language) data. It is worth noting that this is not the case of more data trivially yielding better results. The bilingual model performs significantly worse on the Russian language data alone.

The results indicate that multilingual speech recognition systems are inherently better capable of recognizing code-switched speech than monolingual systems trained on code-switched speech itself. Future directions include investigating end-to-end multilingual systems from the point of view of code-switched segments, developing more sophisticated language modeling for code-switching, and introducing more than two languages.

5 Acknowledgments

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08), and by the grant of Ministry of Education and Science of the Russian Federation Goszadanie No. 2.13462.2019/13.2.

References

1. Khomitsevich O., Mendelev V., Tomashenko N. et al.: A Bilingual Kazakh-Russian System for Automatic Speech Recognition and Synthesis. In: Ronzhin A., Potapova R., Fakotakis N. (eds) Speech and Computer. SPECOM 2015. Lecture Notes in Computer Science, vol 9319. Springer, Cham (2015)
2. Muysken, P., Díaz, C. P., Muysken, P. C.: Bilingual speech: A typology of code-mixing (Vol. 11). Cambridge University Press. (2000)
3. Myers-Scotton, C.: Duelling Languages: Grammatical Structure in Codeswitching Oxford: Clarendon Press, 20 (1993)
4. Sitaram, S., Chandu, K. R., Rallabandi, S. K., Black, A. W.: A Survey of Code-switched Speech and Language Processing. arXiv preprint arXiv:1904.00784 (2019)
5. Vu, N.T., Lyu, D., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Siong, C.E., Schultz, T., Li, H.: A first speech recognition system for Mandarin-English code-switch conversational speech. IN: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4889–4892 (2012)

6. Modipa, T., Davel, M.H., Wet, F.D.: Implications of Sepedi/English code switching for ASR systems. In: Conference Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa, Johannesburg, South Africa (2013)
7. Lyudovyk, T., Pylypenko, V.: Code-Switching speech recognition for closely related languages. SLTU (2014)
8. Yilmaz, E., Heuvel, H.V., Leeuwen, D.A.: Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech. SLTU (2016)
9. Biswas, A., Wet, F.D., Westhuizen, E.V., Yilmaz, E., Niesler, T.R.: Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech. INTERSPEECH (2018)
10. Pavlenko, A.: Russian in post-Soviet countries Russ. linguist. **32**(1), 59–80 (2008)
11. Myers-Scotton, C.: Codeswitching with English: types of switching, types of communities. World Englishes 8, 333–346 (1989)
12. Povey, D. et al.: The Kaldi Speech Recognition Toolkit. In: IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 1–4. Big Island (2011)
13. Hinton, G., Deng, L., Yu, D. et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, **29**(6), 82–97 (2012)
14. Hochreiter, S. and Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
15. Graves, A., Jaitly, N., Mohamed, A.: Hybrid speech recognition with deep bidirectional LSTM. In: IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 55–59. Scottsdale (2015)
16. Mohamed, A., Seide, F., Yu, D., Droppo, J., Stolcke, A., Zweig, G., Penn, G.: Deep Bi-directional Recurrent Networks Over Spectral Windows In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278. Olomouc (2013)
17. Sak, H., Senior, A., Beaufays, F.: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128 (2014)
18. Povey, D., Zhang, X., Khudanput, S.: Parallel Training of DNNs with Natural Gradient and Parameter Averaging. arXiv preprint arXiv:1410.7455 (2014)