Automatic recognition of Slovak-English bilingual speech

Matus Pleva[1], Yuan-Fu Liao[2], Daniel Hladek[1], Jan Stas[1],
Martin Lojka[1], Jozef Juhar[1], and Stanislav Ondas[1]

[1] Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovakia
Matus.Pleva@tuke.sk, Daniel.Hladek@tuke.sk,
Jan.Stas@tuke.sk, Martin.Lojka@tuke.sk,
Jozef.Juhar@tuke.sk, Stanislav.Ondas@tuke.sk,
WWW home page: http://kemt.fei.tuke.sk

[2] Department of Electronic Engineering,
College of Electrical Engineering & Computer Science
National Taipei University of Technology, Taiwan
yfliao@ntut.edu.tw,
WWW home page: https://sites.google.com/site/speechlabx/home

Abstract: This article describes the progress of a joint project on Multilingual Automatic Speech Recognition using Deep Neural Networks, in which the Technical University works together with National Taipei University of Technology in Taiwan. During the last year, we managed to train multilingual models of combinations of Slovak - English and Slovak - English - Chinese/Mandarin languages. In this paper, we are presenting the results of the Slovak - English model based on deep learning with and without language detection. Furthermore we present new bilingual Slovak-English code-switching database for bilingual systems training and testing. The results indicate that the use of the language detection module can lower the error rate of the multilingual model to the result similar to monolingual models that are generally better for the monolingual tasks.

1 Introduction

Thanks to globalization, open culture and easy access to information on the Internet, users are more exposed to the multi-language environment than they were in the past. As a result, foreign words began to appear in spontaneous spoken language with foreign pronunciation (rather than being adapted to Slovak pronunciation). This forced developers to test multiple language models that would be able to recognize multiple languages at the same time they often mix.

Multilingual LVCSR (Large Vocabulary Continuous Speech Recognition) has made great progress in recent years, notably by introducing Deep learning in Neural Networks (DNNs), [1, 2, 3]. In these works, DNNs were taught separately to recognize many different languages or to perform one primary role of speech recognition by using several helper functions.

Shared-hidden layer (SHL) [1] architecture where hidden layers are shared between languages, but the output layer is language-dependent. Another approach is used for Multi-component Recurrent Neural Networks (MRNNs) [2] implementation, where bilingual automatic speech recognition systems with a large LVCSR dictionary and LIDs (Language IDentification) have been combined and run in parallel to assist each other. In this work, an alternative was used when using a Linguistically Universal / independent End-to-End model (LUE) [3]. Our proposed method uses a language-specific gate mechanism that allows the internal representation of a network to be modulated in a language-specific manner. Similar approach was presented for Cantonese/Turkish/Vietnamese language specific gate units described in [4]

2 Training bilingual speech models

2.1 Monolingual databases and acoustic models

In our previous work we used Julius for English [5] and Slovak [6] local automatic speech recognition tasks. We have decided to use the Kaldi tool [1] [7] (open source Apache License v2.0) to create an English and Slovak language recognizer based on Deep Neural Networks. TaipeiTech Lab already has experience in implementing and testing Mandarin / English DNN bilingual recognizer [8].

The acoustic corpuses used for the training process of the acoustic model of the English and Slovak languages is shown in Table 1. In addition, a multilingual hidden layer sharing recognizer (SHL) has been used as shown in Fig. 1.

The acoustic corpuses were divided to 3 parts: Training, Development and Evaluation. More details are depicted in Table 2

[1] http://kaldi-asr.org/doc/about.html

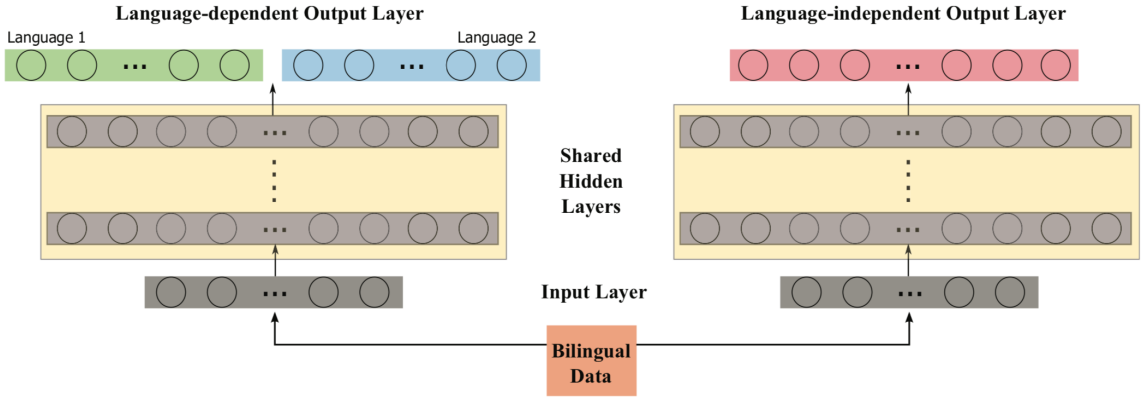

Figure 1: Block diagram of multilingual recognizer with hidden layers sharing.

Table 1: Acoustical corpuses used for monolingual acoustic models training

DB	Language	Hours
LibriSpeech [9]	English	475
TUKE-BNews-SK [10]	Slovak	280

Table 2: Acoustical corpuses division for development and evaluation purposes

Language	Speakers	Utterances
En-Train	2,682	132,553
En-Dev	97	2,703
En-Test	87	2,620
Sk-Train	8,230	112,039
Sk-Dev	1,000	31,824
Sk-Test	1,000	39,274

2.2 Code Switching

To train a bilingual speech recognizer, databases with sentences where words in a foreign language are spoken with a foreign accent are required. It means that commonly the most words of the sentence is in the speaker's native language but he use also different language words as non-native speaker. But it is not necessary, because for example for children raised in bilingual family the native language is difficult to find out.

The so called Code Switching occurs when a speaker alternates two or more languages or language dialects within a single conversation. This is a well-known meeting behavior in global technology-oriented companies where English or German is the official language. During meetings, the native language is mixed with English words or phrases used in the company for specific tasks, processes, equipment, etc.

Other examples of bilingual speech can be found in air traffic control when part of the communication between Slovak flight dispatchers with Slovak pilots is according to international English standards and then a few polite phrases are heard in Slovak. Another example is technology-oriented lectures or product and car tests, and reviews and tutorials on English software. However, we do not have to stick to a technology sub-group during lectures, but English phrases often also appear in finance, social sciences or climate conferences. [11]

Table 3: Acoustical corpus of bilingual Slovak-English data

Source	% of English	# of words	Minutes
Tech. speech	13.71%	1174	8.2
Game review	8.7%	1528	13.1
Financial	8.8%	3843	28.9
Bodybuild.	4.2%	429	3.4
Pilot conv.	62.6%	484	11.8
Sk/En course	60.3%	2043	55.2
Total	23%	>9500	121.9

From such sources we create the first Slovak English bilingual database, [11], which is currently annotated to increase the accuracy of the language recognition system (LID) but also to test the resulting bilingual system. Current state of the database is depicted in Table 3.

2.3 Textual corpuses and language models

The language models were built to test the monolingual and bilingual LVCSR engines on the Development (Dev Set) and Evaluation sets (Test Set). The language model for this research was provided by the project entitled "Automatic Subtitling of Audiovisual Content for Hearing Impaired"[2] [12]. But this model was finally not used in the tests presented in this paper, but it will be used in further tests until the end of 2019.

In this work we presented results of the Slovak language model trained using neural networks and from the train part of the KEMT-BN corpus (112 thousand utterances) not used in testing. That is also a reason the results are

[2] http://access.kemt.fei.tuke.sk/

Table 4: Results of the monolingual LVCSR engines

TDNN Model	Dev Set WER [%]	Test Set WER [%]
Slovak	17.46	17.76
English	8.05	8.63

Table 5: Results of the bilingual DNN LVCSR engines without the language identification module (LID)

Sk/En Model	Dev Set WER [%]	Test Set WER [%]
Slovak	17.35	16.15
English	8.71	9.19

Table 6: Results of the bilingual DNN LVCSR engines with the language identification module (LID) used for the gating mechanism

Sk/En Model	Dev Set WER [%]	Test Set WER [%]
Slovak	17.32	16.13
English	8.64	8.97

worse then our broadcast news engine baseline [13], but the important contribution is the comparison of monolingual and bilingual systems results.

For English language the same approach was chosen so 132 thousand utterances from LibriSpeech database were used for English language model training and also for phonetic vocabulary generating.

2.4 Bilingual LVCSR

For Slovak/English bilingual speech recognition we built a bilingual LVCSR (as shown in Fig. 2). For this purpose we shared the input and hidden layers across languages, while the output layer is grouped, i.e., $[O_{en}^{LVCSR}, O_{sk}^{LVCSR}]$. For better discrimination of similar phonemes all neurons in the output layer are tuned at the same time (i.e., single task goal) in this work [14].

To help discriminate difference languages, an LID-based gating mechanism was used to control the bilingual LVCSR's outputs [14]. Finally, the multilingual LVCSR output scores are multiplied by their corresponding Language IDentification - LID module scores.

For LVCSR training we used 43 MFCC feature vector together with 100 dimensional i-vector. The TDNN with 850 neurons for each of the 6 hidden layers following the AiShell/nnet3 recipe [15]. The time splicing was chosen (-2,-1,0,1,2), (-1, 0, 2), (- 3,0,3), (-7,0,2) and (-3,0,3)).

For Language Identification training the same features and time splicing was used to have even longer-term language cues. The only difference is in number of neurons, where we used 425 for 6 hidden layers.

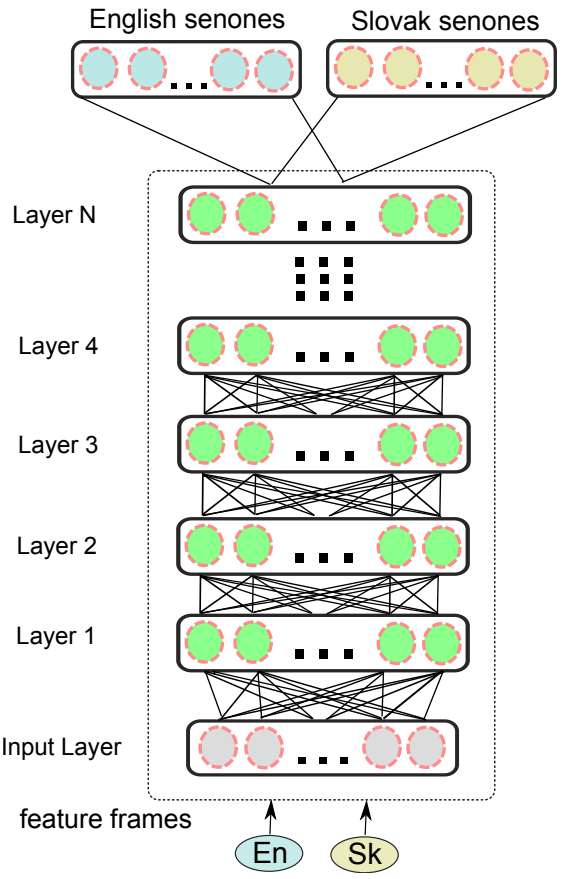

Figure 2: The block diagram of the bilingual share-hidden-layer (SHL) neural networks.

3 Achieved results and future work

In Tables 4. 5. and 6 we can see the results achieved with the proposed bilingual speech recognition system. The results are presented in WER (Word Error Rate), which is the ratio of successfully recognized words to all words in the test database and is expressed as a percentage [16].

It can be stated that thanks to the proposed multilingual LVCSR and LID running in parallel and integrated with the gate mechanism, it was possible to achieve results comparable (see Table 6) with a single-language recognizer (see Table 4.). These types of engines are generally better at recognizing single-language data than bilingual recognizer without the language identifier and gate mechanism (see Table 5). It should be noted that the bilingual test database is still under development, so it was not possible to assess the results in a bilingual test when the words of both languages and pronunciations are in one sentence.

This article describes the ongoing work of our two-year project (2018 to 2019). We have now conducted the first tests of bilingual and monolingual large speech dictionary recognition (LVSCR) based on Kaldi [7] (TDNN, ResNet or even DenseNet) and we are also planning to test new

technologies such as DeepSpeech[3] [17] or TensorFlow[4].

Acknowledgment

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 108-2911-I-027-501 and 108-2221-E-027-067.

References

[1] Huang, J. T., Li, J., Yu, D., Deng, L., Gong, Y.: *Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,* in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. , Vancouver, pp. 7304—7308, 2013.

[2] Tang, Z., Li, L., Wang, D: *Multi-task recurrent model for true multilingual speech recognition,* in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, Dec 2016, pp. 1—4.

[3] Kim, S. and Seltzer, M. L. : *Towards language-universal end-to-end speech recognition,* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary. IEEE, pp. 4914–4918, 2018.

[4] Liu, D., Wan, X., Xu, J., Zhang, P. Multilingual speech recognition training and adaptation with language-specific gate units (2019) 2018 11th International Symposium on Chinese Spoken Language Processing, ISCSLP 2018 - Proceedings, art. no. 8706584, pp. 86–90.

[5] Pleva, M., Juhar, J., Cizmar, A., Hudson, C., Carruth, D. W., Bethel, C. L.: *Implementing English speech interface to Jaguar robot for SWAT training*, In proceedings: Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 15th International Symposium on, Herlany, Slovakia, IEEE, pp. 105–110, 2017.

[6] Ondas, S., Juhar, J., Pleva, M., Lojka, M., Kiktova, E., Sulir, M., Cizmar, A., Holcer, R.: "Speech technologies for advanced applications in service robotics," Acta Polytechnica Hungarica, 10 (5), pp.45–61, 2013.

[7] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: *The Kaldi speech recognition toolkit.*, In Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding - ASRU 2011, Hilton Waikoloa Village, Big Island, Hawaii. IEEE Signal Processing Society, 2011.

[8] Lin, C. T., Wang, Y. R., Chen, S. H., Liao, Y. F.: *A preliminary study on cross-language knowledge transfer for low-resource Taiwanese Mandarin ASR,* 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Bali, IEEE, pp. 33–38, 2016.

[9] Panayotov, V., Chen, G., Povey, D., Khudanpur, S. : *Librispeech: an ASR corpus based on public domain audio books.* In Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, South Brisbane, pp. 5206–5210. IEEE, 2015.

[10] Pleva, M., Juhar, J.: *TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation*, In: LREC 2014 : Ninth International Conference on Language Resources and Evaluation : May 26-31, 2014, Reykjavik, Iceland. - Paris : ELRA, 2014, pp. 1709–1713, 2014.

[11] Hudak, L.: *Methods for bilingual automatic speech recognition*, 2019 Bachelor thesis in Slovak, KEMT, FEI,TUKE, Kosice, 2019.

[12] Stas, J., Viszlay, P., Lojka, M., Koctur, T., Hladek, D., Juhar, J.: Automatic Transcription and Subtitling of Slovak Multi-genre Audiovisual Recordings. In In Human Language Technology. Challenges for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015, Revised Selected Papers (Vol. 10930, p. 42). Springer. pp. 42–56). 2018.

[13] Lojka M., Viszlay, P., Stas, J., Hladek, D., Juhar, J.: Slovak Broadcast News Speech Recognition and Transcription System. International Conference on Network-Based Information Systems. In: Barolli L., Kryvinska N., Enokido T., Takizawa M. (eds) Advances in Network-Based Information Systems. NBiS 2018. Lecture Notes on Data Engineering and Communications Technologies - LNDECT, vol 22. Springer, Cham, pp. 385–394, 2019.

[14] Liao, Y. F., Pleva, M., Hladek, D., Stas, J., Viszlay, P., Lojka, M., Juhar, J.: "Gated Module Neural Network for Multilingual Speech Recognition," 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, pp. 131–135, 2018.

[15] Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline. In 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, pp. 1–5. IEEE. 2017.

[16] Mackova, L., Cizmar, A., Juhar, J.: A study of acoustic features for emotional speaker recognition in i-vector representation. Acta Electrotechnica et Informatica, 15 (2), pp. 15–20, 2015.

[17] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z. et al.: *Deep speech 2: End-to-end speech recognition in English and Mandarin.* In: International Conference on Machine Learning, pp. 173–182, 2016, June.

[3] `https://github.com/mozilla/DeepSpeech`

[4] `https://www.tensorflow.org/`