Towards Slovak-English-Mandarin Speech Recognition Using Deep Learning

Matus Pleva[1], Yuan-Fu Liao[2], Wuhua Hsu[2], Daniel Hladek[1], Jan Stas[1],
Peter Viszlay[1], Martin Lojka[1] and Jozef Juhar[1]
[1] Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics,
Technical university of Kosice, Letna 9, 04120 Košice, Slovakia.
[2] Department of Electronic Engineering,
National Taipei University of Technology, Zhongxiao E. Rd. 406, 10608 Taipei, Taiwan.
matus.pleva@tuke.sk

***Abstract*—This paper describes the progress of the development of multilingual speech enabled interface by exploring the state-of-the-art deep learning techniques in the frame of the bilateral project named "Deep Learning for Advanced Speech Enabled Applications". The advancement is especially expected in automatic subtitling of broadcast television and radio programs, databases creation, indexing and information retrieval. This implies investigation of deep learning techniques in the following sub-tasks: a) multilingual large vocabulary continuous speech recognition, b) audio events detection, c) speaker clustering and diarization, d) spoken discourse, speech, paragraph and sentence segmentation, e) emotion recognition and f) microphone array/multi-channel speech enhancement, g) data mining, h) multilingual speech synthesis, and i) spoken dialogue user interfaces. This paper describes the current work, description of the available data in the project and achieved results in the first task of Slovak speech recognition Kaldi module using deep learning algorithms.**

***Keywords*—Large Vocabulary Continuous Speech Recognition (LVCSR); Human Computer Interface (HCI); Deep Neural Networks (DNNs); Code-Switching; Bilingual Language Switching**

I. Introduction

The project "Deep Learning for Advanced Speech Enabled Applications" focuses on the development, testing and implementation of novel techniques utilizing Deep Neural Networks (DNNs) for multilingual speech enabled applications. The targeted applications include:

1) Transcribing TV/Radio programs into accessible multimedia/digital talking books/databases,
2) Genre clustering of TV/Radio programs,
3) Spoken dialog user interface,
4) Speaker clustering and diarization,
5) Emotion recognition.

Interfaces that enable speech communication are usually language-dependent, what can be seen as a disadvantage that leads to increasing costs of adaptation to a new language. Within this project, we focus on multilinguality of such modules aiming to minimize the financial cost and time required to adapt the system to other languages. This project will maximize the use of multi-language modules.

The scientific advancement is expected in automatic subtitling of broadcast television and radio programs, databases creation, indexing and information retrieval from audio-visual documents. This implies investigation of deep learning techniques in the following sub-tasks:

(a) multilingual large vocabulary continuous speech recognition,
(b) audio events detection,
(c) speaker clustering and diarization using speaker verification algorithms [1],
(d) spoken discourse, speech, paragraph, and sentence segmentation,
(e) emotion recognition [2]
(f) microphone array/multi-channel speech enhancement,
(g) data mining,
(h) multilingual speech synthesis and
(i) spoken dialogue user interfaces.

Specifically, this project will focus on the

- semi-supervised, unsupervised, reinforcement and generative adversarial learning of DNNs,
- end-to-end structures of DNNs and
- deep latent space modeling using various types of DNNs.

During the research and development, the following objectives are examined:

- The new and progressive English/Chinese/Slovak speech recognition module for audio recordings automatic subtitling will be designed in close cooperation between both research teams. This module will be used for testing the performance of DNN technologies on Slovak language and tri-lingual systems. Both research teams will share their language resources to handle this task.
- A new web based application for automatic subtitle generation from recordings will be designed to evaluate the tri-lingual and compare with uni-lingual speech modules. This application will be used for dissemination of the project activities and results.
- A new interface that uses the speech recognition with wireless audio devices (like BT headsets, etc.) will be designed, deployed and evaluated by end-users.

Expected results of the project are:

- Development of an interface for DNN-based speech recognition – depending on the needs of a specific language, direct result of the bilateral Taiwan-Slovak cooperation will be the advanced tri-lingual module implemented in server based application.
- Development of a web-based tri-lingual demo of automatic subtitling application and an evaluation of the scalability of a server based application.
- Development of a testbed for evaluating various types of equipment for wireless audio communication with the server based speech recognition module.
- Design of a new evaluation database for cross-lingual voice commands recognition in Slovak, English and Chinese. The English recognizer will be realized and evaluated throughout a collaboration with different U.S. universities (MSU [3], [4], etc.) and planned Erasmus student exchanges in the years 2018/2019.

II. Deep Neural Networks Training

A. Monolingual databases

For establishing the English/Chinese/Slovak speech recognizer we decided to use the Kaldi speech recognition toolkit[1] [5] (open source with Apache License v2.0). The TaipeiTech SpeechLab has already experience with the implementation and testing of a Mandarin/English bilingual recognizer [6]. Several monolingual acoustical corpora needed for Mandarin, English and Slovak language model training are depicted on Table I. Moreover, a shared hidden layer-based approach will be adopted for multilingual speech processing as depicted in Figure 1.

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

[2]Catalog Reference: ELRA-U-S 0034 Mandarin Speech Corpus TCC300 corpus: http://universal.elra.info/product_info.php?products_id=1672

[3]http://bn.kemt.fei.tuke.sk/

[4]https://nlp.web.tuke.sk/pages/tedx

B. Code Switching databases

For training a multi-lingual speech recognizer, some code-switching corpora are needed. Code-switching occurs when a speaker alternates between two or more languages, or language varieties in the context of a single conversation. This is a well-known behaviour in technical meetings in a global company where the official l anguage i s E nglish o r German, etc. During the meetings the native language is mixed with English words or phrases used in company for specific tasks, process, equipment, etc.

For Slovak code-switching database we use Slovak Cisco technology lectures [15] given by TUKE-RCNA[5] (Regional Cisco Network Academy) lectures on Technical University of Kosice, where a huge number of English terms are embedded in Slovak sentences. For automatic speech recognition of such lectures, a Slovak recognizer will not work. So a bilingual or tri-lingual automatic speech recognition solution could be an ideal solution. Code-switching databases available for Deep Neural Network-based training are presented in Table II.

[5]Computer networks laboratory (CNL) which contains TUKE-RCNA http://cnl.sk/en/about-us/

[6]SEAME (South East Asia Mandarin-English) on LDC https://catalog.ldc.upenn.edu/LDC2015S04

TABLE I. Single Language Acoustical corpora Available for Training

Database	Language	Hours
MATBN [7]	Mandarin	127
TCC300[2]	Mandarin	26
AiShell [8]	Mandarin	100
NER [9]	Mandarin	300
LibriSpeech [10]	English	960
Wall Street Journal [11]	English	73
TUKE-BNews-SK[3] [12], [13]	Slovak	165
TEDxSK & JumpSK[4] [14]	Slovak	58

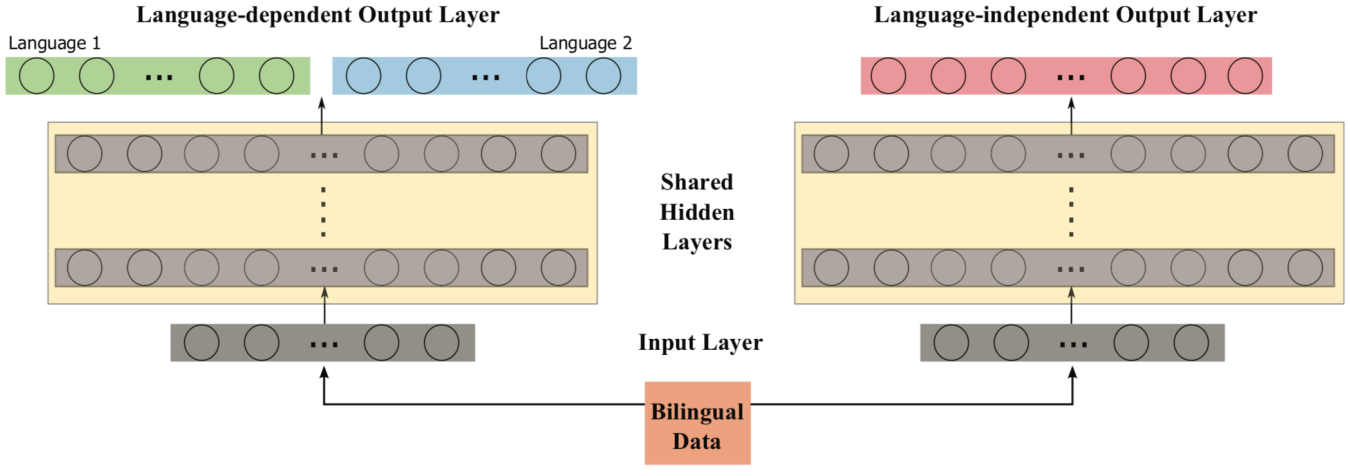

Figure 1. The block diagram of a shared hidden layers-based approach for multilingual speech processing

C. Slovak Phonetics and X-SAMPA

The TaipeiTech Speech Communication Laboratory[7] had developed Mandarin/English bilingual recognizer based on the X-SAMPA phoneme set. So we decided to map the Slovak phonetic dictionaries from SAMPA-SK to X-SAMPA phonemes/symbols. The initial phoneme mapping is presented on Table V and we are open for discussion about this topic.

D. Available phonetic dictionaries - lexicons

Also, several Slovak, English and Chinese dictionaries are shared in this project (see Table IV). As can be seen from the Table IV, the Slovak dictionary is much larger than the others, due to high inflection and rich morphology in the language that cause a large number of unique word forms.

E. Text databases and language models

For testing and evaluation purposes the system needs language models. The Slovak language model was generated only from manually annotated transcriptions of speech recordings obtained from available acoustic database. The Slovak language modeling is based on trigrams restricted to the vocabulary with smoothing using Witten-Bell algorithm [21].

Another option is to train a new language model from available databases listed in Table III, where the OC16-CE80, SEAME, Chinese GigaWord, Wall Street Journal, or APD corpora will be used.

[7]https://sites.google.com/site/speechlabx/home

[8]SEAME (South East Asia Mandarin-English) on LDC https://catalog.ldc.upenn.edu/LDC2015S04

[9]Chinese Gigaword database: https://catalog.ldc.upenn.edu/LDC2003T09

[10]BLLIP 1987-89 WSJ Corpus Release 1: https://catalog.ldc.upenn.edu/ldc2000t43

TABLE II. CODE-SWITCHING MULTI-LANGUAGE ACOUSTICAL CORPORA AVAILABLE FOR TRAINING

Database	Languages	Hours
OC16-CE80 [17]	Mandarin/English	63.8
SEAME [18] [6]	Mandarin/English	95.1
CNL-SK-EN	Slovak/English	19.8

TABLE III. TEXTUAL CORPORA AVAILABLE FOR LANGUAGE MODEL TRAINING

Database	Language	K-words
OC16-CE80 [17]	Mandarin/English	~ 400
SEAME [18] [8]	Mandarin/English	~ 500
Chinese Gigaword[9] [19]	Mandarin	1,118,380
Wall Street Journal[10] [11]	English	30,000
APD [13]	Slovak	2,105,970

TABLE IV. PHONETIC DICTIONARIES/LEXICONS AVAILABLE

Database	Language	Words
OC16-CE80 [17]	English	6,879
OC16-CE80 [17]	Mandarin	4,653
SEAME [20]	English	8,415
SEAME [20]	Mandarin	6,923
APD [13]	Slovak	573,576

III. SLOVAK CONTINUOUS SPEECH RECOGNITION RESULTS FROM MODELS TRAINED ON SHARED DATA

The TaipeiTech SpeechLab team from NTUT has made great effort to train and test the Slovak Large Vocabulary Continuous Speech Recognition (LVCSR) system based on Kaldi engine using shared data. From 10,230 speakers in the database, 1,000 speakers were extracted for validation (Dev Set) and also 1,000 speakers for evaluation (Test Set). From the remaining 8,230 speakers and 112,039 utterances, different Slovak acoustic models were trained. The language model for testing was built only from training transcriptions with approximately 119,795 words in the phonetic lexicon. The results are summarized in the Table VI.

The results are presented as Word Error Rate - WER and Character Error Rate - CER values, because the CER is important for comparison with languages, such as Mandarin. The results are promising because the Broadcast News database provided is rather small for Slovak language modeling and in the future we will test the results on common evaluation set, so the results will be comparable with the TUKE team.

TABLE V. THE PROPOSED SLOVAK X-SAMPA PHONETIC SET

SAMPA-SK	X-SAMPA	Phoneme usage example
i	i	**p**ivo, trenk**y**
e	e	m**e**no, m**ä**so
a	a	k**a**pitola, p**a**pier
o	o	n**o**ha, p**o**pol
u u_^	u	b**u**bon, ko**v**, pra**v**da
i:	i:	vít**a**l, b**ý**k
e:	e:	g**é**n
a:	a:	poh**á**r, p**á**s
o:	o:	katal**ó**g, p**ól**
u:	u:	m**ú**r, p**ú**pava
i_^a	i a	p**ia**tok
i_^e	i e	m**ie**r
i_^u	i u	pan**iu**
u_^o	u o	k**ôň**
r r= r=:	r	pa**r**a, p**r**st, v**ŕ**ba
l l= l=: L	l	ska**l**a, v**l**k, v**ĺ**ča, **ľ**ad
m F	m	**m**a**m**a, a**m**fiteáter
n N\N	n	ra**n**a, Slove**n**sko, ba**n**ka, de**nn**ý
J	J	va**ň**a, de**nn**e
v w	v	slo**v**o, **v**dova
j i_^	j	**j**ama, kra**j**
p	p	**p**o**p**ol
b	b	ža**b**a
t	t	va**t**a
c	c	Ma**ť**o, pla**ti**ť
d	d	vo**d**a
J_	J\	há**ď**a
k	k	pá**k**a
g	g	**g**uma, a**g**át
f	f	**f**iga, **f**ajka
s	s	o**s**a, o**s**em
S	S	**š**ek, vyš**š**í
z	z	**z**ima
Z	Z	ve**ž**a
x	x	**ch**ata
h G	h	Pra**h**a, vr**ch**_hory
ts	ts	**c**ena
tS	tS	o**č**i, ma**č**ka
dz	dz	me**dz**a, há**dz**ať
dZ	dZ	**dž**ungľa, **dž**em

TABLE VI. RESULTS OF THE SLOVAK LVCSR BUILT FROM SHARED RESOURCES BY SPEECHLAB (FROM TAIPEITECH)

Models	Dev Set		Test Set	
	WER [%]	*CER* [%]	*WER* [%]	*CER* [%]
Mono	43.68	19.53	41.36	17.71
Tri1	28.34	10.37	28.20	10.27
Tri2	28.37	10.41	28.29	10.29
Tri3a	28.04	10.02	27.76	9.65
Tri4a	25.39	8.52	25.63	8.52
Tri5a	23.27	7.43	23.73	7.53
Tdnn	17.46	4.50	17.76	4.50
Chain	16.52	4.12	16.89	4.15

IV. CONCLUSION AND FUTURE WORK

This paper describes the work in progress of our two-year project (2018-2019) named "Deep Learning for Advanced Speech Enabled Applications". Currently, the TUKE Laboratory of Speech and Mobile Technologies in Telecommunications and TaipeiTech SpeechLab shared the resources, where the license not prohibit the sharing. Both will use the available data for training monolingual, bilingual and finally trilingual (Slovak-English-Mandarin) speech recognition systems based on Kaldi [5] (TDNN, ResNet or even DenseNet) and later also DeepSpeech[11] [22] (TensorFlow[12]) recognition toolkits.

The TUKE team shared the new Slovak evaluation BN database for comparing the results of the Slovak LVCSR module between the teams. Meanwhile, the TUKE team is working on code-switching Slovak-English testing database for evaluation of the trained combined Slovak-English models and systems. The Slovak-Mandarin and Slovak-English-Mandarin systems will be also built and evaluated during the project. The results of the project will be used also for building multilingual spoken dialogues [23].

ACKNOWLEDGMENT

The research in this paper was supported by the Taiwan's Ministry of Science and Technology MOST-SRDA PPP under the contract number 107-2911-I-027-501 and partially by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17 and the Slovak Research and Development Agency under the contracts SK-TW-2017-0005 and APVV-15-0517.

REFERENCES

[1] R. Jarina, J. Polacký, P. Pocta, M. Chmulik, "Automatic speaker verification on narrowband and wideband lossy coded clean speech", *IET Biometrics*, vol. 6, no. 4, pp. 276–281, 2017.

[2] M. Hric, M. Chmulik, I. Guoth, R. Jarina, "SVM based speaker emotion recognition in continuous scale" *Proceedings of 25th International Conference Radioelektronika, RADIOELEKTRONIKA 2015*, art. no. 7129063, pp. 339–342, 2015.

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

[4] Ch. Hudson, D. Carruth, C. Bethel, M. Pleva, J. Juhar, A. Cizmar, "A training tool for speech driven human-robot interaction applications", In proceedings: *Emerging eLearning Technologies and Applications (ICETA)*, Stary Smokovec, Slovakia, IEEE, pp. 167–172, 2017.

[11] https://github.com/mozilla/DeepSpeech

[12] https://www.tensorflow.org/

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, " 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.

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

[7] Wang, H.M., Chen, B., Kuo, J.W. and Cheng, S.S., "MATBN: A Mandarin Chinese broadcast news corpus." *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 10, no. 2, Special Issue on Annotated Speech Corpora, pp. 219–236, 2005.

[8] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, " AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline." arXiv preprint arXiv:1709.05522. 2017.

[9] Y. h. S. Chang et al., "Development of a large-scale Mandarin Radio Speech Corpus," *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, Taipei, IEEE, pp. 359–360, 2017.

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

[11] D.B. Paul, and J.M. Baker, "The design for the Wall Street Journal-based CSR corpus." In Proceedings of the *Workshop on Speech and Natural Language*, Association for Computational Linguistics - ACM, pp. 357–362, 1992, February.

[12] M. Pleva and J. Juhar, "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.

[13] M. Rusko et al. " Advances in the Slovak Judicial Domain Dictation System." In: Vetulani Z., Uszkoreit H., Kubis M. (eds) *Human Language Technology. Challenges for Computer Science and Linguistics.* LTC 2013. LNCS, vol 9561. Springer, Cham, pp 55–67, 2016.

[14] J. Stas, D. Hladek, P. Viszlay, T. Koctur, "TEDxSK and JumpSK: A new Slovak speech recognition dedicated corpus," *Journal of Linguistics*, vol. 68, no. 2, pp. 346–354, 2017.

[15] L. Derjaninova, R. Hajduk, M. Michalko, F. Jakab, J. Sekerak and D. Sveda, "IT academy project: An opportunity for the IT industry in Slovakia," *15th Int. Conference on Emerging eLearning Technologies and Applications (ICETA)*, Stary Smokovec, IEEE, pp. 1–7, 2017.

[16] D. Zlacky, J. Stas, J. Juhar, A. Cizmar, "Slovak Text Document Clustering," *Acta Electrotechnica et Informatica*, vol. 13, no. 2, pp. 3–7, 2013.

[17] D. Wang, Z. Tang, D. Tang and Q. Chen, "OC16-CE80: A Chinese-English mixlingual database and a speech recognition baseline," *Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Bali, IEEE, pp. 84–88, 2016.

[18] G. Lee, T. N. Ho, E. S. Chng and H. Li, "A review of the Mandarin-English code-switching corpus: SEAME," *International Conference on Asian Language Processing (IALP)*, Singapore, IEEE, pp. 210–213, 2017.

[19] J.F. Hong, and C.R. Huang, "Using Chinese Gigaword corpus and Chinese word sketch in linguistic research." In Proceedings of the *20th Pacific Asia Conference on Language, Information and Computation*, pp. 183–190, 2006.

[20] D.C. Lyu, T.P. Tan, E.S. Chng, and H. Li, "Mandarin–English code-switching speech corpus in south-east Asia: SEAME." *Language Resources and Evaluation*, vol. 49, no. 3, Springer, pp. 581–600, 2015.

[21] J. Staš and J. Juhár, "Modeling of Slovak language for broadcast news transcription" *Journal of Electrical and Electronics Engineering*, vol. 8, no. 2, pp. 43–46, 2015.

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

[23] S. Ondas and M. Gurcik, "Domain-specific language models training methodology for the in-car infotainment" *Intelligent Decision Technologies*, vol. 11, no. 4, pp. 417–422, 2017.