BRNO UNIVERSITY OF TECHNOLOGY

ANALYSIS OF MULTILINGUAL BLSTM ACOUSTIC MODEL ON LOW AND HIGH RESOURCE LANGUAGES

Martin Karafiát, Murali Karthick Baskar, Karel Veselý, František Grézl, Lukáš Burget, Jan Černocký

Brno University of Technology, Speech@FIT group, Czech Republic

e-mail: {karafiat,baskar,matejka,iveselyk,grezl,burget,cernocky}@fit.vutbr.cz

Abstract

The paper provides an analysis of automatic speech recognition systems (ASR) based on multilingual BLSTM, where we used multi-task training with separate classification layer for each language. The focus is on low resource languages, where only a limited amount of transcribed speech is available. In such scenario, we found it essential to train the ASR systems in a multilingual fashion and we report superior results obtained with pre-trained multilingual BLSTM on this task. The high resource languages are also taken into account and we show the importance of language richness for multilingual training. Next, we present the performance of this technique as a function of amount of target language data. The importance of including context information into BLSTM multilingual systems is also stressed, and we report increased resilience of large NNs to overtraining in case of multi-task training.

1 Introduction

- *Multilingual pre-training has huge importance on low resource data:* how does it perform on various amounts of data?
- *Incorporating context into BLSTM features.*
- *System complexity:* multilingual pretraining should allow to train more complex architectures.

2 Data

Mainly conversational telephone speech (CTS).
Y1 Babel languages (50-60h/lang.): Cantonese, *Pashto*, Turkish, Tagalog, Vietnamese.
Y2 Babel languages (50-60h/lang.): Assamese, Bengali, Haitian Creole, Lao, Zulu, Tamil.
Y3 Babel languages (30-40h/lang.): Kurdish, Cebuano, Kazakh, Telugu, Lithuanian, TokPisin, Swahili.
Y4 Babel languages (30-40h/lang.): *Pashto*, *Javanese*, Igbo, Mongolian, Dholuo, Guarani, *Amharic*.
Non-Babel languages: Switchboard (270h), Fisher English (1700h), hub5 test set.
Babel target languages: *Javanese*, *Amharic* and *Pashto*. Note, all are coming from Y4.

3 GMM system

- Used to produce phoneme alignments for NN training.
- GMM features are based on multilingual Region Dependent Transform trained on 24 Babel languages (Year 1-4) (Released on `http://speech.fit.vutbr.cz/software/`).

Features	Javanese	Amharic	Pashto
PLP	66.4	56.2	61.1
MultRDT	55.9	46.2	51.2

%WER of Babel GMM systems used for alignment.

	SWB training data size h					
Features	10	50	100	150	200	Full
MFCC	39.6	32.7	31.4	30.4	29.8	29.4
MultRDT	32.7	27.5	26.0	25.4	24.9	24.4

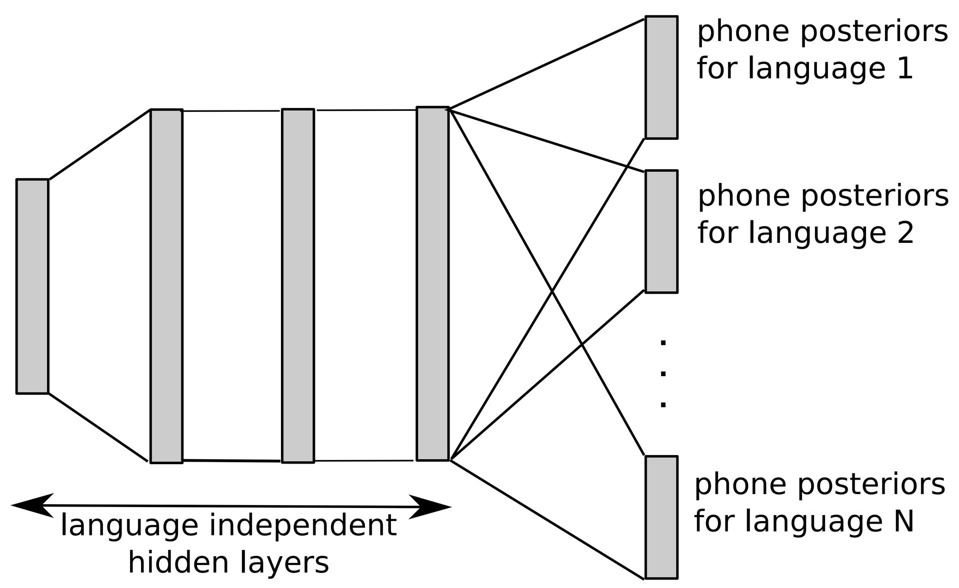
%WER of SWB GMM systems used for alignment.

4 BLSTM systems

- Standard hybrid DNN-HMM acoustic models
- **Feature extraction**:
  Single: 24 log-mel-filter-bank + different pitch features **FBANK_F0**.
  Contextual: **FBANK_F0** feature trajectories spanning 11 frames with Hamming window and Discrete cosine transform - **11FBANK_F0**.

4.1 Multilingual architecture

- Trained on Y1-Y3 = 17 languages or Y1-Y4 = 24 languages.
- 'block-softmax output layer with context-independent phoneme states targets.



- Porting of multilingual models into target language:
  1. The final multilingual layer is replaced with randomly initialized target-language layer.
  2. Only this new layer is trained with a standard learning rate.
  3. The whole NN is fine-tuned with low learning-rate.

4.2 Analysis of feature extraction

Features	Mult-NN	Javanese	Amharic	Pashto
11FBANK_F0	None	54.4	44.0	50.7
11FBANK_F0	24L	**49.2**	**39.6**	**46.0**
FBANK_F0	None	**54.0**	44.0	**48.7**
FBANK_F0	24L	52.1	42.2	47.7

WERs [%] obtained with monolingual vs. multilingual training and various feature extractions

- Context information is advantageous for multilingual systems.

4.3 Training epochs

	17L Mult.NN		24L Mult.NN	
n. epoch	Javanese	Amharic	Javanese	Amharic
5	50.8	41.2	50.6	41.0
10	50.4	40.6	49.9	40.2
15	**50.1**	**40.3**	49.2	39.8
20	50.5	40.4	49.2	40.3
25	50.5	40.5	**48.9**	**39.5**
30	50.9	40.6	-	-

WERs [%] obtained with fine-tuned NNs, which were pre-trained using different number of training epoch.

- Final multilingual NN should be taken around the first halving of learning rate (20th epoch for 17L, 19th for 24L NN).
- Well trained multilingual NN is suitable only if target language is part of multilingual training set.

4.4 Training data analysis

Pre-trained on		Javanese	Amharic	Pashto	SWB
Monoling.	0 h	54.0	44.0	48.7	18.1
5L	294 h	52.2	42.1	46.8	17.5
11L	621 h	50.1	40.6	46.2	17.4
17L	841 h	50.9	40.6	46.2	17.5
24L	1076 h	**49.2**	**39.6**	**46.0**	**17.1**
Fsh	1700 h	51.5	41.4	47.1	**16.5**

WERs [%] for BLSTM systems multilingually pre-trained on different data sets.

- More diverse data leads to better results.

Data size	Monoling.	Multiling. (24L)
10 h	35.5	26.0 (-9.5)
50 h	24.8	21.2 (-3.6)
100 h	22.4	19.6 (-2.8)
150 h	20.3	18.9 (-1.4)
200 h	18.9	17.9 (-1.0)
270 h	18.1	**17.1** (-1.0)

WERs [%] on SWB for various target language data sizes.

- Multilingual pre-training can play a significant role even for the high-resource tasks.

4.5 Experiments with larger BLSTMs

Data Size	3 layers	4 layers	5 layers	6 layers	7 layers
10 h	35.5	33.8	**33.0**	33.4	35.3
50 h	24.8	23.9	**23.1**	24.8	23.6
100 h	22.4	20.8	21.5	**20.4**	21.4
150 h	20.3	20.1	20.4	19.5	**19.4**
200 h	18.9	18.6	**18.1**	18.3	18.6
270 h	18.1	17.1	**16.8**	**16.8**	17.0

WERs [%] on SWB for various training data sizes and number of BLSTM layers.

- 5 BLSTM layers perform better than 3 layers even with 10 hours of SWB training data.

System	Javanese	Amharic	Pashto
Mono 3L	54.0	44.0	49.0
Mono 6L	52.6	42.2	49.2
Multi 3L 24 lang.	49.2	39.6	46.0
Multi 6L 24 lang.	**48.5**	**39.3**	**45.8**

WERs [%] for monolingual and multilingual Babel system with 3 and 6 BLSTM layers.

- Additional gain with adding more parameters into systems.

5 Conclusion

- Analysis of improvement from multilingual approaches for large scale of training data - significant gain even for 270h of training data.
- Important contextual information in multilingual system features.
- Multilingual NN should be taken from training process around the first halving of learning rate.

Acknowledgment

This work was supported by Technology Agency of the Czech Republic project No. TA04011311 "MINT", European Union's Horizon 2020 project No. 645523 BISON and Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science" - LQ1602.