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 and IT4I Center of Exellence
Brno, Czech Republic

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.

Index Terms— Automatic speech recognition, Multilingual neural networks, Bidirectional Long Short Term Memory

1. INTRODUCTION AND PRIOR WORK

Quick delivery of an automatic speech recognition (ASR) system for a new language is one of the challenges in the community. Such scenarios call not only for automated construction of systems, that have been carefully designed and crafted "by hand", but also for effective use of available resources. Without any question, the data collection and annotation are the most time- and money-consuming processes.

The recently finished IARPA Babel program focused on fast development of ASR systems, while the amount of per-language data was decreasing from year to year. The data from 24 low-resource languages were collected, which led to numerous multilingual experiments.

For humans, borrowing the information from other sources when learning a new language is very natural. We all share the same vocal tract architecture and phonetic systems of languages overlap, therefore automatic systems should be able to have the universal and language-independent low-level components (feature extraction and partially also acoustic models), that would be built with various sources of data. In the past, we have verified that the multilingual pre-training is an important technique for feature extraction, especially if limited amount of training data is available [1, 2], similarly to [3, 4]. We have also performed an analysis of combining semi-supervised and multi-lingual training of NN-based bottleneck feature extractors [5]. The hybrid Deep Neural Networks-Hidden Markov Models (DNN-HMM) systems benefit from the multi-lingual training too [6]. Recently in [7], we extended this idea to Bi-directional Long-Short Term Memory Recurrent Neural Networks (BLSTM-RNN) acoustics models and show significant effect of the multilingual pre-training for low resource languages.

The amount of training data over the languages in the Babel project is more or less consistent (50-80h) and limited. Therefore, more detailed analysis of multilingual techniques is not possible. In this paper, we added also English Switchboard and Fisher corpora to investigate BLSTM acoustic models for larger amounts of training data.

2. DATA

The IARPA Babel program data simulate a situation, in which the data for a new language are collected in a limited time. The data consists mainly of conversational telephone speech (CTS) but some scripted recordings and far field recordings are present too. During the 4-year project, datasets of 25 languages were created: **Year 1**: Cantonese (CA), Pashto (PA), Turkish (TU), Tagalog (TA), Vietnamese(VI). **Year 2**: Assamese (AS), Bengali (BE), Haitian Creole (HA), Lao (LA), Zulu (ZU), Tamil (Tam). **Year 3**: Kurdish (KU), Cebuano (CE), Kazakh (KA), Telugu (TE), Lithuanian (LI), TokPisin (TP), Swahili (SW). **Year 4**: Pashto progress set (about 40h subset of Year 1) (PA2), Javanese (JA), Igbo (IG), Mongolian (MO), Dholuo (DH), Guarani (GU), Amharic (AM), Georgian (not used in this work) (GE). **Non-Babel - English** used in this work includes: Switchboard-1 Release 2 (SWB), Fisher English Training Speech Part 1+2 (FSH)

The work was supported by Technology Agency of the Czech Republic project No. TJ01000208 "NOSICI" and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

Y1 Langs.	CA	PA	TU	TA	VI		
Hours	65	65	57	44	53		
Y2 Langs.	AS	BE	HA	LA	ZU	Tam	
Hours	47	54	55	57	58	56	
Y3 Langs.	KU	CE	KA	TE	LI	TP	SW
Hours	37	38	40	38	41	26	34
Y4 Langs.	PA2	JA	IG	MO	DH	GU	AM
Hours	32	40	39	39	38	39	39
Non-Babel	SWB				FSH		
Hours	270				1700		

Table 1. Amounts of data used for the training.

The amounts of data can be found in Table 1. Note, that the data sizes are summarized after trimming the silence to 150 ms on the edges of speech segments, according to a forced alignment. More details about Year 1–3 languages can be found in [2].

For Babel languages, we limited the language model training corpus to the transcriptions of the training audio we received from the Babel program. Pronunciation dictionaries were not provided, so we relied on graphemic lexicons. For English, a 3-gram language model based on FSH+SWB transcription was used. The pronunciation dictionary was based on CMU dictionary [8].

Several data-sets based on packs from Table 1 were generated for the multi-lingual acoustic model training. We simulated a real situation with the data "growing" over time (see section 4.4). The experiments were evaluated with Javanese (JA), Pashto (PA) and Amharic (AM) from the 4th year of the program. In addition, to simulate varying sizes of target language training data, we experimented also with English (SWB), where the training set was reduced to various subsets (10h, 50h, 100h, 150h, 200h, Full). The hub5 eval 2000 set was used for testing. Note that, unlike many other sites who report results only on SWB subset of hub5 eval 2000 set, we report results on the whole set in order to prevent overoptimistic results caused by the significant overlap of speakers in training set and the SWB subset.

3. SYSTEM DESCRIPTION

Our systems were built with several toolkits: We used STK/HTK [9] toolkit[1] Kaldi [10] for maximum likelihood (ML) Gaussian mixture model (GMM) training.Finally, we trained BLSTM networks using CNTK [11].

For sake of simplicity, all the results in this work are coming from cross-entropy trained systems with no further use of sequence discriminative criteria, (for example sMBR).

[1]STK is BUT's variant of HTK: `http://speech.fit.vutbr.cz/software/hmm-toolkit-stk`

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

Table 2. %WER of GMM Babel systems used as the alignment system.

	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

Table 3. %WER of GMM SWB systems used as the alignment system.

3.1. GMM system

First, GMM based acoustic models are trained to produce phoneme alignments as the labels for the following NN training. These models are based on cross-word tied-states trained from scratch using standard ML algorithm. The baseline GMM systems had approximately 4000 cross-word triphone tied states for Babel and 9100 for full SWB. They were trained on multilingual Region Dependent Transform features trained on 17 Babel languages (Y1-3) as the alignments were found to lead to better NN performance over alignments coming from PLP/MFCC based systems [12] [2]. The initial GMM results can be found in Table 2 for Babel languages and in Table 3 for various subsets of SWB.

3.2. BLSTM systems

The BLSTM acoustic-models were trained with last layer producing posterior probabilities of tied-states for HMM models. The latency-controlled BLSTM architecture [13] contains 3 bi-directional layers, for each direction there are 512 memory units and 300 dimensional projection layer.The training is done with truncated back-propagation through time (BPTT) algorithm [14]. Each update is based on $T_{bptt} = 20$ time-steps of recurrent forward-propagations and back-propagations. For detailed description of the procedure used in our training see [13]

3.3. Feature extraction

The BLSTM NN input is fed with filter bank based features. It contains of 24 log-Mel-filter-bank features concatenated with different pitch features: "BUT F0" has 2 coefficients (F0 and probability of voicing), "snack F0" is a single F0 estimate and "Kaldi F0" has 3 coefficients (F0 normalized with a sliding window, probability of voicing and F0 delta). Fundamental frequency variation (FFV) produces a 7 dimensional vector. The whole feature vector has 24+2+1+3+7=37 coefficients

[2]The scripts for MultRDT features generation can be found in `http://speech.fit.vutbr.cz/software`

Features	Mult-NN	Javanese	Amharic	Pashto
11FBANK_F0	None	54.4	44.0	50.7
11FBANK_F0	24L	**49.2**	**39.8**	**46.1**
FBANK_F0	None	**54.0**	44.0	**49.0**
FBANK_F0	24L	52.1	42.2	47.7

Table 4. Comparison of %WER with various initialization and feature extraction.

(see [15] for details on pitch features). These features will be called "FBANK_F0".

After a conversation-side mean subtraction, we apply a Hamming window and Discrete cosine transform to the feature trajectories spanning 11 frames. We retain 0^{th} to 5^{th} DCT coefficients for each of the original 37 features resulting in 37×6=222 coefficients. These features will be called "11FBANK_F0".

4. MULTI-LINGUAL EXPERIMENTS

4.1. Multilingual architecture

All multilingual models in this work were trained with a 'block-softmax' output layer, which consists of per-language softmaxes [16]. The training targets were the context-independent phoneme states, otherwise the size of the final layer would be excessively large.

The NNs were trained with standard cross-entropy objective function and Stochastic Gradient Descent (SGD) approach. Whenever objective degraded on cross-validation data, the learning rate was halved and the previous (so far best performing model) was loaded.

The procedure of porting multilingual models into target language can be described in the following steps:

1. The final multilingual layer (context-independent phoneme states for all languages) is stripped and replaced with a layer specific to target-language (tied-state triphones) with random initialization.

2. This new layer is trained for 8 epochs with a standard learning rate, while the rest of the NN is fixed.

3. Finally, the whole NN is fine-tuned with 10 epochs, the initial value of learning-rate schedule is set to 0.5 of the original value.

4.2. Analysis of feature extraction

Here, we were interested in optimal feature extraction for Multilingual and Monolingual BLSTM architectures. In our recent work [7], we have shown significant gain from using 11FBANK_F0 over bottle-neck features in BLSTM systems. As BLSTM can naturally incorporate context information, any feature stacking might not be necessary, so we experimented with FBANK_F0 features as well.

	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	-	-

Table 5. %WER obtained with fine-tuned NNs, which were pre-trained using different number of training epoch.

According to Table 4, the plain FBANK_F0 features are the most suitable in monolingual systems, so the models can naturally learn context information. But advantage of multilingual pre-training is partly lost with NN trained on this features (comparing to system pre-trained on 11FBANK_F0 features). This is very interesting outcome, it shows that context information is advantageous for multilingual systems. Therefore, 11FBANK_F0 features were used for all multilingual systems and FBANK_F0 were used for plain monolingual baseline systems.

4.3. Analysis of number of training epochs

The final multilingual NN is used for the pre-training of a final language specific system, therefore training into "ultimate" minima of the objective function could not be optimal. The multilingual NN has to be able to change its parameters to different languages, therefore early stopping should be taken into account.

For this experiments, the Y1-3 (17L) and Y1-4 (24L) languages were chosen. The first language set simulates adaptation of a multilingual NN to an unknown new language and the second one is showing the case where the target language is contained in multilingual training data.

The first and second columns of Table 5 present significant drop of accuracy when NN is well trained on multilingual data. The first halving of learning rate during the training process was observed on 20th epoch for Y1-3 NN and on 19th for Y1-4 NN. Therefore the final multilingual NN should be taken before this point.Well trained multilingual NN is suitable only in cases where target language is part of multilingual training set, see third and fourth column.

4.4. Multilingual training data

Next, the amount of multilingual training data was investigated. Table 6 presents positive effect of adding more training data into multilingual training, which is consistent with our previous work on feature extraction [2, 7].

In addition, we are presenting the effect of adding rich resource English data sets (last raw and column of the table). It shows that having 11 small resource languages is giving better performance than a lot of training data from single language

Languages	Javanese	Amharic	Pashto	SWB
0 (monoling)	54.0	44.0	48.7	18.1
5	52.2	42.1	46.8	17.5
11	50.1	40.6	46.2	17.4
17	50.9	40.6	46.2	17.5
24	49.2	39.6	46.0	17.1
Fsh	51.5	41.4	47.1	16.5

Table 6. Comparison of %WER for BLSTM systems with various multilingual initialization.

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

Table 7. Comparison of SWB %WER on various data size and type of training.

(compare rows "11" and "Fsh"). The last column (SWB), shows that improvement from multilingual pre-training is noticeable also on significantly bigger corpus than Babel data. Obviously, the NN pre-trained on huge amounts of target data (last column and raw) is giving the best results, but it is not a sensible scenario and the NN could be simply trained on the combined pre-training and fine-tuning data (i.e. Fisher and SWB).

4.5. Multilingual pre-training on SWB

Next, we were interested in the improvement brought by multilingual pre-training as function of the size of target language training data. Table 7 shows huge improvement for low-amount of training data (10h). After 100h, the gain is starting to saturate. Interestingly, the improvement never completely disappears and even for full SWB, it is still 1% WER absolute. Therefore, multilingual pre-training can play significant role even for rich-resource task.

4.6. Large NNs

Another possible improvement brought by multilingual training is a possibility to pre-train larger NNs that it would be possible on small target language data. We were experimenting with increasing the amount of layers for various SWB data sizes. Surprisingly, monolingual experiments in Table 8 show that 5 layers are better than 3 layers even for 10h of training data. When more than 100 hours of training data is used, 6 layers are giving the best performance.

Similarly to SWB, 5 to 6 layers are more suitable even for Babel languages, see Table 9. Therefore, we decided to retrain multilingual NN on 24languages with 6 layers as well. The results are in Table 10.

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

Table 8. Comparison of SWB %WER on various data size and number of layers.

n.Layers	Javanese	Amharic	Pashto
3L	54.0	44.0	49.0?
4L	53.5	42.7	48.3
5L	53.2	42.4	**48.2**
6L	**52.6**	**42.2**	49.2
7L	53.2	43.6	49.5

Table 9. %WER for various number of layers on Babel data.

5. CONCLUSION

This paper analyzes multi-lingual training of BLSTM systems. We have shown clear advantage of multi-lingual training of acoustic models in low-resource scenarios. Small but consistent gains are also present on rich resources scenario.

With multilingual pre-training, we have found essential to include context information into multilingual systems even for BLSTM which can naturally learn it.

The optimum size of acoustic model NN was also investigated and we found that even low resource systems (10h) can be trained with 5 layers. The rich resource language can advantageously exploit a more complex system, therefore we are presenting additional gain from building multilingual system on 6 layers.

System	Javanese	Amharic	Pashto
Mono 6L	52.6	42.2	49.2
Multi 3L	49.2	39.8	46.1
Multi 6L 24Lang	**49.0**	**39.3**	**45.8**

Table 10. %WER for multilingual 6 Layer Babel system.

6. REFERENCES

[1] F. Grézl, M. Karafiát, and M Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proceedings of ASRU 2011*, 2011, pp. 359–364.

[2] František Grézl and Martin Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU-2014. St. Petersburg, Russia, 2014*. 2014, pp. 39–45, International Speech Communication Association.

[3] Markus Mller, Sebastian Stker, Zaid Sheikh, Florian Metze, and Alex Waibel, "Multilingual deep bottle neck features: A study on language selection and training techniques," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 4-5 2014.

[4] Zoltan Tuske, David Nolden, Ralf Schluter, and Hermann Ney, "Multilingual MRASTA features for low-resource keyword search and speech recognition systems," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, May 2014, IEEE, pp. 5607–5611.

[5] F. Grezl and M. Karafiat, "Combination of multilingual and semi-supervised training for under-resourced languages," in *Proceedings of Interspeech 2014*, Singapore, 2014, pp. 820–824.

[6] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks.," in *ICASSP*. 2013, pp. 7319–7323, IEEE.

[7] Martin Karafiát, K. Murali Baskar, Pavel Matějka, Karel Veselý, František Grézl, and Jan Černocký, "Multilingual blstm and speaker-specific vector adaptation in 2016 but babel system," in *Proceedings of SLT 2016*. 2016, pp. 637–643, IEEE Signal Processing Society.

[8] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Tech. Rep., Mountain View, CA, USA, 2004.

[9] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK book*, Entropics Cambridge Research Lab., Cambridge, UK, 2002.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[11] Amit Agarwal, Eldar Akchurin, Chris Basoglu, Guoguo Chen, Scott Cyphers, Jasha Droppo, Adam Eversole, Brian Guenter, Mark Hillebrand, Ryan Hoens, Xuedong Huang, Zhiheng Huang, Vladimir Ivanov, Alexey Kamenev, Philipp Kranen, Oleksii Kuchaiev, Wolfgang Manousek, Avner May, Bhaskar Mitra, Olivier Nano, Gaizka Navarro, Alexey Orlov, Marko Padmilac, Hari Parthasarathi, Baolin Peng, Alexey Reznichenko, Frank Seide, Michael L. Seltzer, Malcolm Slaney, Andreas Stolcke, Yongqiang Wang, Huaming Wang, Kaisheng Yao, Dong Yu, Yu Zhang, and Geoffrey Zweig, "An introduction to computational networks and the computational network toolkit," Tech. Rep. MSR-TR-2014-112, August 2014.

[12] Martin Karafiát, Miloš Janda, Jan Černocký, and Lukáš Burget, "Region dependent linear transforms in multilingual speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing 2012*. 2012, pp. 4885–4888, IEEE Signal Processing Society.

[13] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yao, Sanjeev Khudanpur, and James R. Glass, "Highway long short-term memory RNNS for distant speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5755–5759.

[14] Ronald J. Williams and Jing Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural Computation*, vol. 2, pp. 490–501, 1990.

[15] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, Igor Szoke, and Jan "Honza" Černocký, "BUT 2014 Babel system: Analysis of adaptation in NN based systems," in *Proceedings of Interspeech 2014*, Singapure, September 2014, IEEE.

[16] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*. 2012, pp. 336–341, IEEE Signal Processing Society.