MULTILINGUAL DEEP NEURAL NETWORK BASED ACOUSTIC MODELING FOR RAPID LANGUAGE ADAPTATION

Ngoc Thang Vu,[1], *David Imseng*[2], *Daniel Povey*[3], *Petr Motlicek*[2], *Tanja Schultz*[1], *Hervé Bourlard*[2]

[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Idiap Research Institute, Martigny, Switzerland
[3]Johns Hopkins University, Baltimore, USA

{thang.vu, tanja.schultz}@kit.edu, {dimseng, motlicek, bourlard}@idiap.ch, dpovey@gmail.com

ABSTRACT

This paper presents a study on multilingual deep neural network (DNN) based acoustic modeling and its application to new languages. We investigate the effect of phone merging on multilingual DNN in context of rapid language adaptation. Moreover, the combination of multilingual DNNs with Kullback–Leibler divergence based acoustic modeling (KL-HMM) is explored.

Using ten different languages from the Globalphone database, our studies reveal that crosslingual acoustic model transfer through multilingual DNNs is superior to unsupervised RBM pre-training and greedy layer-wise supervised training. We also found that KL-HMM based decoding consistently outperforms conventional hybrid decoding, especially in low-resource scenarios. Furthermore, the experiments indicate that multilingual DNN training equally benefits from simple phoneset concatenation and manually derived universal phonesets.

Index Terms— Multilingual DNN, phone merging, rapid language adaptation, KL-HMM

1. INTRODUCTION

HMM/DNN hybrid systems that use deep neural networks (DNNs) to estimate the emission probabilities of the hidden Markov model (HMM) states [1–4] were successfully applied to large vocabulary continous speech recognition (LVCSR) and led to a significant improvement in various tasks with different data sets.

Many recent studies [5–8] exploited multilingual data during DNN training in different unsupervised and supervised ways to improve the monolingual ASR performance. In these studies, it was shown that the shared hidden layer is language independent and can be used to bootstrap the DNN for a new language. This result was also confirmed in multilingual LVCSR using Tandem approach with bottle-neck features [9–12].

To train a multilingual acoustic model, there are several possible ways [13]: on a merged universal phoneset based on the international phonetic alphabet (IPA) chart, i.e. the same IPA symbols are merged across languages or on a universal phoneset without merging strategy. In this paper, we compare these two methods in the context of multilingual DNN.

Recently, it was also shown that such multilingual DNNs work particularly well in combination with Kullback–Leibler divergence based hidden Markov modeling (KL-HMM) if only small amounts of data are available for the new language [14]. However, in [14], only DNNs with three hidden layers were used, pre-training was not applied and the setup was bilingual (Afrikaans and Dutch).

In this paper, we investigate the effect of IPA based phoneme merging on the multilingual DNN and its application to new languages. We also study multilingual DNNs in combination with KL-HMM on a large scale, involving up to five hidden layers, up to 6,000 MLP outputs and DNNs trained on up to six languages. We also investigate how different pre-training methods influences cross-lingual DNN based acoustic modeling in the context of rapid language adaptation.

Compared to previous studies, the two main contributions of this paper are: (1) investigating the effect of phone merging on multilingual DNNs, and (2) exploration of DNN based acoustic modeling in the context of rapid language adaptation on a variety of languages.

This paper is organized as follows: Section 2 describes the DNN training procedure which was used in our study. In Section 3, we present the multilingual DNN along with the creation of the universal phoneset, the cross-language adaptation and KL-HMM implementation. Sections 4 and 5 describe the experimental setup and results. The study is concluded in Section 6 with a summary.

2. DNN TRAINING

This section describes some key features of the Kaldi DNN training recipe [15] - part of the Kaldi ASR toolkit [16] - which we used in our study. Currently Kaldi contains two parallel implementations for DNN training. Both recipes

support DNN training which is done on top of the standard HMM/GMM training recipe. That means, the context dependent decision tree, the audio alignment and the feature transform (in the case, if it is used) are adopted from the HMM/GMM system. The neural net is trained to predict the posterior probability of each context-dependent state. During decoding, the output probabilities are divided by the priors of each state to form a "pseudo-likelihood" that is used in place of the state emission probabilities in the HMM [17].

The first implementation is as described in [18]. This implementation supports Restricted Boltzmann Machines (RBM) pre-training [19] - generative pre-training, stochastic gradient descent (SGD) training using NVidia graphics processing units (GPUs) and discriminative training.

The second Kaldi DNN training recipe supports parallel training on multiple CPUs. Instead of using RBM pre-training, the greedy layer-wise supervised training [20] or the "layer-wise backpropagation" of [1] is used. A parameter was defined to set the number of iterations in which the network should be trained before the new hidden layer is inserted between the last hidden layer and the softmax layer. This is repeated until a desired number of layers is reached. The parallelization of the neural network training is done in two levels: on a single machine, and also across machines. The parallelization method on a single machine is to have multiple threads simultaneously updating the parameters while simply ignoring any synchronization issues. This is similar to the Hogwild! approach [21]. Furthermore, on different machines, multiple training processes are done independently using SGD, on different random subsets of the data. After processing a specified amount of data, each machine writes its model to disk. Afterwards, the averaged model parameters become the starting point for the next iteration of training. The training recipe does also support different methods to stabilize the training such as preconditioned SGD and enforcing the maximum change in the parameters per minibatch. The initial and final learning rates in the training recipe must be specified by hand. During training we decrease the initial learning rate exponentially to reach the final learning rate for a few epochs at the end. The learning rate is unchanged during these last epochs. After the final iteration of training, the models from the last n iterations are combined via a weighted-average operation into a single model. The weights are determined via nonlinear optimization, optimizing the cross-entropy on a randomly selected subset of the training data.

3. MULTILINGUAL DEEP NEURAL NETWORKS

For our studies, we use multilingual DNNs. We train the multilingual DNNs in two steps: (1) training on multilingual data using a universal phoneset, and (2) performing cross-language model transfer by re-training the output layer on target language data. To further exploit the (limited amount of) target language data, we also perform Kullback–Leibler divergence based HMM (KL-HMM) decoding.

3.1. Universal phoneset

To train the multilingual DNN, we investigate two different kinds of universal phone sets. The first kind of multilingual phone set, *MUL*, is created by simply concatenating all involved monolingual phone sets with a language identification prefix to ensure that all the phones are distinct between languages. To create the second kind of universal phone set, *MUL-IPA*, we merge all the monolingual phones which share the same symbol in the IPA table.

To obtain the tied-state targets for the training of the multilingual DNN, we used the Kaldi toolkit. More specifically, for both universal phonesets, we trained multilingual HMM/GMM systems and build multilingual decision trees to generate tied-state alignments.

3.2. Cross-language model transfer

To bootstrap the acoustic model for a new language using multilingual DNN, the hidden layers of the multilingual DNN are shared and transferred to the new language. The multilingual softmax layer is simply replaced with a new output layer corresponding to the target language. All the weights which connect the neurons of the last hidden layer and the bias are randomly initialized.

3.3. KL-HMM

In a recent study [14], it was shown that KL-HMM decoding is particularly useful if ASR systems for low-resourced languages are improved by using out-of-language data. Therefore, in this paper, we also apply KL-HMM decoding as an alternative to conventional hybrid decoding. Conventional hybrid systems directly use the MLP output to estimate the emission probability of the HMM states, hence each HMM state only considers one MLP output dimension. In (deep) Tandem systems [5] on the other hand, each HMM state considers the whole MLP output vector. However, since Tandem systems model the HMM states with Gaussian mixtures, the MLP output vector needs to be post processed and usually the dimensionality is reduced as well. The KL-HMM acoustic modeling technique can directly model high dimensional MLP output vectors. The HMM states are parametrized with reference posterior distributions (categorical distributions) that can be trained by minimizing the Kullback–Leibler divergence between the categorical distributions and the MLP output. More details about training and decoding in the KL-HMM framework can be found in, for instance, [22].

4. EXPERIMENTAL SETUP

4.1. GlobalPhone database

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [23]. It contains more than 400 hours of speech spoken by more than 1,900 adult native speakers. In this study, we used the Bulgarian (BG), Czech (CZ), French(FR), Japanese (JP), German (GE), Hausa (HA), Mandarin (MAN), Portuguese (PO), Spanish (SP), and Vietnamese (VN) datasets from the GlobalPhone corpus. In addition, we also used English speech data from the Wall Street Journal corpus (WSJ0). The trigram language models that we used are publicly available [24].

4.2. Setup

We conducted two different sets of experiments by varying the relation between the source and the target languages. Furthermore, to verify the generalization of the study, the experiments were performed with different implementations which support two state-of-the-art techniques for deep neural network training namely RBM pretraining and greedy layer-wise supervised training (see section 2).

In the first set of experiments, we experimented with four Indo-European languages. Three source languages, namely FR, GE and SP are used to train the multilingual DNN which is then adapted to PO. Note that in this case the target language is related to the source languages and the training is faster since only three languages are involved in the multilingual DNN training.

The second set of experiments was conducted with speech data from different language families. We use BG, GE and SP as representatives of Indo-European languages, Mandarin as a Sino-Tibetan language and Japanese from the Altaic language family for the multilingual DNN training. The multilingual DNN is then adapted to three different target languages, namely CZ, HA and VN which are from three different language families. CZ and VN belong to the Indo-European and Sino-Tibetan languages, respectively. Both language families are represented in the source languages. HA on the other hand is a language from the Afro-Asiatic language family which is not related to any of the source languages.

5. RESULTS

This section presents all the experimental results of our study. Different DNNs were trained using different initialization schemes, namely generative pre-training (*Gen-PT*) and greedy layer-wise supervised training (*GL-sup*), and served as baseline system.

Furthermore, we used different universal phonesets (described in Section 3 - *MUL* and *MUL-IPA*) to train the multilingual DNNs that were then used to bootstrap the monolingual DNNs, which we refer to as *DNN-MUL* and *DNN-MUL-IPA* respectively. We also performed KL-HMM decoding as an alternative to conventional hybrid decoding, referred to as *DNN-MUL + KL* and *DNN-MUL-IPA + KL*.

5.1. Related languages

The first set of experiments was carried out on similar languages and we always evaluated on the PO test set. All the DNNs were trained using the first DNN implementation of Kaldi. We assumed to have different amounts of PO data available: the full training set (17 h), and randomly selected 5 h and 1 h subsets. All the results are summarized in Table 1.

Table 1. *Word error rates (WER) on the PO test data of different DNNs trained with RBM pre-training*

Amount of PO data	17 h	5 h	1 h
DNN (Gen-PT)	13.2	15.2	20.5
DNN-MUL	12.9	13.9	17.8
DNN-MUL + KL	**12.6**	13.8	17.7
DNN-MUL-IPA	12.9	**13.7**	17.4
DNN-MUL-IPA + KL	12.7	**13.7**	**17.1**

System *DNN* was pre-trained on the PO data. For all the other systems, we used multilingual data for the pre-training. Afterwards, to obtain the PO DNN, the cross-language model transfer is applied.

All the DNNs used in this set of experiments had three hidden layers, each consisting of 2,000 units and were trained from 9 consecutive frames (4 preceding and 4 following frames) of 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC) including deltas and double deltas. System *DNN* corresponds to the baseline system that only used the PO data. The Portuguese DNN was trained to estimate posterior probabilities of 2,252 tied-state triphone targets. We then also evaluated cross-language model transfer by bootstrapping the DNNs with hidden layers trained on FR, GE and SP data, using *MUL* and *MUL-IPA* phone sets. The *MUL*-DNN and the *MUL-IPA*-DNN were trained to estimate posterior probabilities of 3,338 and 3,139 tied-state targets, respectively, obtained from the multilingual decision trees. We also evaluated KL-HMM based decoding for each scenario. For the experiments on the whole PO training set, we fixed the number of KL-HMM states to 20,000. For the subsets of 5 h and 1 h, we used 8,000 and 6,000 KL-HMM states, respectively.

Table 1 reveals the following trends: The cross-language model transfer based on multilingual DNN consistently outperforms the PO baseline system. Moreover, using KL-HMM, the performance is same or better. The ASR performance tends to improve more in case of small amounts of training data while only marginal performance differences are observed if the whole PO training set is used. The lowest WER on the PO test set was obtained by using KL-HMM

Table 2. *Word error rates (WER) on BG, EN, GE, JP, MAN, and SP test data using greedy layer-wised supervised training DNN and DNNs pre-trained with multilingual DNNs.*

Systems	BG	EN	GE	JP	MAN	SP
DNN (GL-sup)	17.4	9.9	6.2	16.8	12.3	14.9
DNN-MUL	16.8	9.5	**5.8**	16.2	**11.8**	**14.3**
DNN-MUL-IPA	**16.7**	**9.2**	**5.8**	**16.1**	**11.8**	**14.3**

in combination with a multilingual DNN. The difference between MUL and MUL-IPA seems to be rather small, but in the case of less training data, using IPA seems to be beneficial.

5.2. Non-related languages

5.2.1. Multilingual DNN

In the second set of experiments, we used the second DNN implementation of Kaldi to train two different multilingual DNN AMs with *MUL* and *MUL-IPA* phone set using the training data of six different languages (BG, EN, GE, JP, MAN, and SP). We applied the greedy layer-wise supervised training to train the multilingual DNN. MFCC features with the first 13 coefficients concatenated with 5 left and 5 right neighbors were used directly as input of the DNN after fMLLR transformation. For each multilingual DNN, 6,000 tied-state triphones were trained. The DNN had 5 hidden layers, each consisting of 1,500 units. We then applied crosslingual model transfer[1] and retrained the DNN for each target language. Table 2 shows the results. Crosslingual model transfer consistently improved WER compared to the greedy layer-wise supervised training and fine-tuned DNN that used the monolingual data only. The *DNN-MUL-IPA* systems yielded slightly better performance than the *DNN-MUL* system in the case of BG, EN and JP. For GE, MAN and SP, the WER was almost the same.

5.2.2. Rapid language adaptation to new languages

For language adaptation experiments, we conducted two different experiments on the CZ, HA and VN GlobalPhone data set: with the full amount of training data and with only small amount of training data. Based on the result of the first set of experiments, we applied KL-HMM based decoding only with small amounts of training data. In the first experiment, all the training data was used to train the DNN for CZ, HA and VN. Table 3 summarizes the WER on CZ, HA and VN test data. Again, the crosslingual model transfer yielded consistent improvements compared to the baseline system which was greedy layer-wise supervised trained and fine-tuned only with monolingual data. In this set of experiments, using IPA to merge the phoneset of the multilingual DNN seems to slightly improve the ASR system in the case of CZ and HA. However, the syllable ER increases a bit in the case of VN. Note that, in the case of HA, even though the target language and the source languages are completely unrelated, we observed up to 6% relative improvement.

[1]Note that in this context, the target language was already part of the multilingual DNN training, hence the term crosslingual model transfer may be misleading. However, the re-training procedure is as described in Section 2.

Table 3. *ASR performance on CZ, HA, and VN test data trained with all the training data.*

Systems	CZ	HA	VN
DNN (GL-sup)	9.9	10.1	10.0
DNN-MUL	9.3	9.8	**8.6**
DNN-MUL-IPA	**9.2**	**9.5**	8.8

In the second experiment, we assume that only a small amount of training data - one hour - for each target language is available. The results in Table 4 show that by using multilingual DNN, we observed larger improvements over the baseline system than the improvements in the previous experiment. It indicates that multlingual DNN is very useful if the amount of training data is rather small. The *DNN-MUL-IPA* is slightly better than the *DNN-MUL* system in the case of HA. In the case of CZ and VN, the ASR performance is marginally different. However, if we use KL-HMM based decoding, we consistently obtained better ASR performance by using *DNN-MUL*.

Table 4. *ASR performance on Czech, Hausa and Vietnamese test data trained with one hour of training data.*

Languages	Czech	Hausa	Vietnamese
DNN (GL-sup)	16.9	16.1	32.1
DNN-MUL	14.0	13.6	27.1
DNN-MUL + KL	**13.1**	**12.0**	**26.6**
DNN-MUL-IPA	13.9	13.3	27.0
DNN-MUL-IPA + KL	13.4	12.3	26.8

6. CONCLUSION

This paper presented an investigation of multilingual DNN based acoustic modeling in the context of rapid language adaptation. On different languages, we found that crosslingual model transfer through multilingual DNN in combination with KL-HMM decoding yields the best performance. The performance improvement is more pronounced in low-resource scenarios. Our experiments also suggest that it is not needed to manually derive IPA based universal phonesets for multilingual DNN training.

7. REFERENCES

[1] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks.," in *Proc. of Interspeech*, 2011, pp. 437–440.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[3] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[4] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 246–251.

[6] J.T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of ICASSP*, 2013.

[7] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. of ICASSP*, 2013.

[8] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. of ICASSP*, 2013.

[9] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource lvcsr systems.," in *Prof. of Interspeech*, 2010, pp. 877–880.

[10] N.T Vu, F. Metze, and T. Schultz, "Multilingual bottleneck features and its application for under-resourced languages," *Proc. of SLTU*, vol. 12, 2012.

[11] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. of SLT*, 2012, pp. 336–341.

[12] Z. Tuske, J. Pinto, D. Willett, and R. Schluter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in *Proc. of ICASSP*, 2013, pp. 7349–7353.

[13] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[14] D. Imseng, P. Motlicek, P. N. Garner, and H. Bourlard, "Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition," in *Proc. of ASRU*, 2013.

[15] Z. Xiaohui, T. Jan, D. Povey, and K. Sanjeev, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. of ICASSP*, 2014.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[17] N. Morgan and H. Bourlard, "Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[18] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. of Interspeech*, 2013.

[19] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, pp. 625–660, 2010.

[20] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, pp. 153, 2007.

[21] F. Niu, B. Recht, C. Ré, and S.J. Wright, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," *arXiv preprint arXiv:1106.5730*, 2011.

[22] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Communication*, 2013.

[23] T. Schultz, N.T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *Proc. of ICASSP*, 2013.

[24] LM-BM, "Benchmark globalphone language models," Retrieved November 3rd, 2013, `http://csl.ira.uka.de/GlobalPhone/`.