Investigating Cross-lingual Multi-level Adaptive Networks: The Importance of the Correlation of Source and Target Languages

Alexandros Lazaridis[a,*], *Ivan Himawan*[b], *Petr Motlicek*[a], *Iosif Mporas*[c] *and Philip N. Garner*[a]

[a]Idiap Research Institute, Martigny, Switzerland
[b]Science and Engineering Faculty, Queensland University of Technology, Australia
[c]School of Engineering and Technology, University of Hertfordshire, Hertfordshire, UK
alaza@idiap.ch

Abstract

The multi-level adaptive networks (MLAN) technique is a cross-lingual adaptation framework where a bottleneck (BN) layer in a deep neural network (DNN) trained in a source language is used for producing BN features to be exploited in a second DNN in a target language. We investigate how the correlation (in the sense of phonetic similarity) of the source and target languages and the amount of data of the source language affect the efficiency of the MLAN schemes. We experiment with three different scenarios using, i) French, as a source language uncorrelated to the target language, ii) Ukrainian, as a source language correlated to the target one and finally iii) English as a source language uncorrelated to the target language using a relatively large amount of data in respect to the other two scenarios. In all cases Russian is used as target language. GLOBALPHONE data is used, except for English, where a mixture of LIBRISPEECH, TEDLIUM and AMIDA is available. The results have shown that both of these two factors are important for the MLAN schemes. Specifically, on the one hand, when a modest amount of data from the source language is used, the correlation of the source and target languages is very important. On the other hand, the correlation of the two languages seems to be less important when a relatively large amount of data, from the source language, is used. The best performance in word error rate (WER), was achieved when the English language was used as the source one in the multi-task MLAN scheme, achieving a relative improvement of 9.4% in respect to the baseline DNN model.

1. Introduction

We are interested in general in multilingual automatic speech recognition (ASR), and in particular in its use when combined with machine translation for broadcast data monitoring. As news becomes pertinent in new parts of the world, different languages and dialects become relevant. It is necessary for already multilingual ASR systems to adapt to these new environments as quickly as possible.

When news is happening in an under-resourced language (or dialect), data may be available, or may be collected, in a closely related (correlated) language. It seems reasonable to assume that such data would be useful in ASR for the under-resourced language.

One successful class of approaches for cross-lingual adaptation has made use of posterior features derived from neural networks in tandem and hybrid ASR systems [1, 2, 3]. In these systems, features derived from a neural network trained as a phone classifier (i.e., bottleneck features) are concatenated with the traditional spectral features (e.g., MFCCs, PLPs) in order to train ASR systems. It has been shown in many studies that the bottleneck features derived from a multilingual network (i.e., an MLP trained using multiple languages) are transferable across different languages [3, 4, 5]. This is useful in a cross-lingual adaptation scenario to alleviate the problem of requiring a significant amount of data to train neural networks from scratch, where a model trained from a resource-rich language can be adapted with limited target data [3]. In the context of DNN/HMM, model adaptation can also be achieved by replacing and retraining the existing layer of the network (i.e., the last hidden layer) using alignments derived from the in-domain data. In this approach, the hidden layers are shared but the output layer is made language specific [6, 7].

In this work, we investigate the importance of the correlation of the source and target languages in the framework of cross-lingual adaptation. The correlation of the source and target languages is measured in the sense of phonetic similarity between the two languages. For cross-lingual adaptation, the multi-level adaptive network (MLAN) [2, 8] and multi-task MLAN [5] architectures are used. We hypothesize that a source language more correlated to the target language, is going to be more beneficial in the MLAN schemes. An other issue that is investigated in this paper, is the importance of the amount of data of the source language. In all our experiments, the Russian language is used as target one. As source languages, three different scenarios are investigated. In the first scenario, French, a language uncorrelated to the target one, is chosen. In the second scenario, Ukrainian, a language correlated to the target one, is chosen. The first two scenarios are aiming at validating our main hypothesis. In the third scenario, the English language is chosen. The reasoning be-

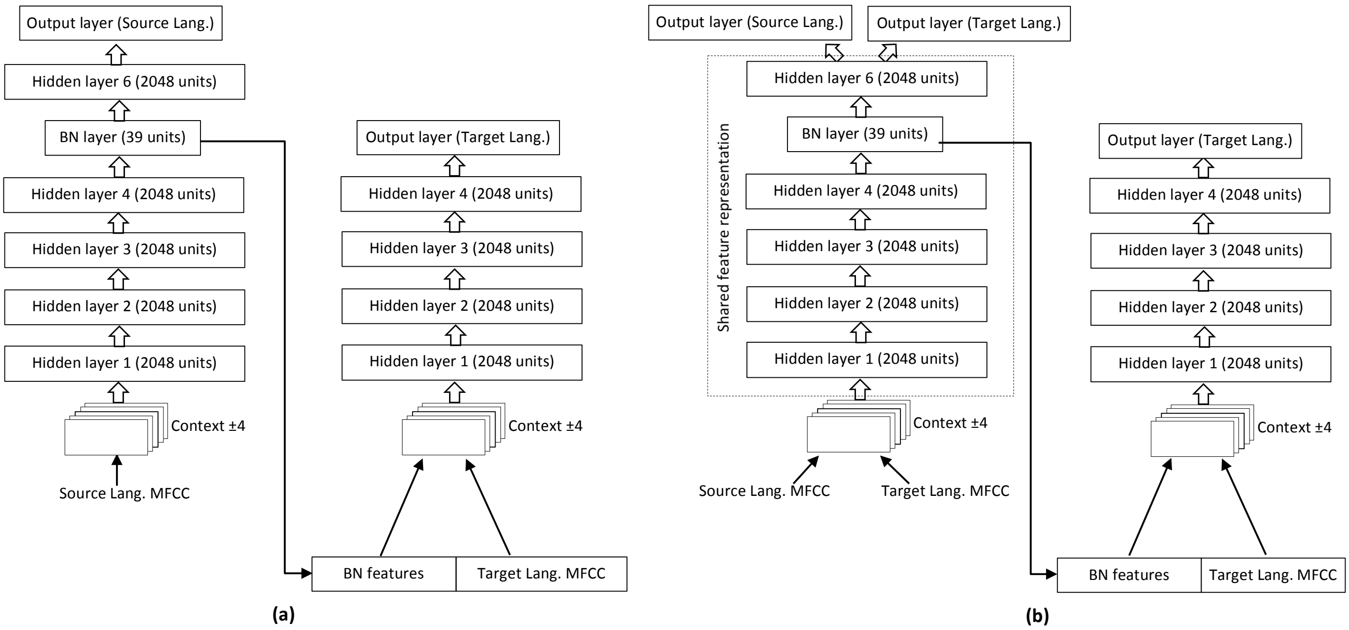
Output layer (Source Lang.)
Hidden layer 6 (2048 units)
BN layer (39 units)
Hidden layer 4 (2048 units)
Hidden layer 3 (2048 units)
Hidden layer 2 (2048 units)
Hidden layer 1 (2048 units)
Context ±4
Source Lang. MFCC
Output layer (Target Lang.)
Hidden layer 4 (2048 units)
Hidden layer 3 (2048 units)
Hidden layer 2 (2048 units)
Hidden layer 1 (2048 units)
Context ±4
BN features
Target Lang. MFCC
(a)
Output layer (Source Lang.)
Output layer (Target Lang.)
Shared feature representation
Hidden layer 6 (2048 units)
BN layer (39 units)
Hidden layer 4 (2048 units)
Hidden layer 3 (2048 units)
Hidden layer 2 (2048 units)
Hidden layer 1 (2048 units)
Context ±4
Source Lang. MFCC
Target Lang. MFCC
Output layer (Target Lang.)
Hidden layer 4 (2048 units)
Hidden layer 3 (2048 units)
Hidden layer 2 (2048 units)
Hidden layer 1 (2048 units)
Context ±4
BN features
Target Lang. MFCC
(b)

Figure 1: *The MLAN (a) and multi-task MLAN (b) schemes.*

hind the third scenario, is to be able to use relatively large amount of data, in respect to the other two scenarios, from a language uncorrelated to the target one, for investigating the importance of the amount of data over the aspect of correlation of the two languages.

The remainder of the paper is organized as follows. In Section 2, the relation to prior work is presented. The MLAN and multi-task MLAN systems are described in Section 3. The experimental setup is given in Section 4. In Section 5, the evaluation results are presented and discussed. Finally the conclusions are given in Section 6.

2. Relation to prior work

The standard approach for exploiting out-of-domain (OOD) knowledge is by performing adaptation of the existing model trained with OOD data to the target domain using maximum a posteriori (MAP) [9] or maximum likelihood linear regression (MLLR) [10] in GMM/HMM frameworks. One popular approach for DNN adaptation is a transformation based method. This method is originally employed for speaker adaptation by augmenting the existing neural net with an extra input layer with a linear activation function [11]. The adaptation layer could also be inserted before the final output activation functions (i.e., softmax) [12]. This new layer can be trained to be condition specific. For example, by adapting the multilingual DNN model to the new language [13], or for adapting multi-condition network to the new acoustic condition [14, 15].

Another approach is improving the training method of the target language by exploiting resource-rich language to better initialize the nets or by sharing parameters. This regularization technique has been successfully applied in multilingual DNN training (i.e., by sequentially training target languages while swapping the output layer with each language) [13, 5]. The regularization effect may be achieved by using multitask learning, where the final layer (i.e., softmax layer used to estimate the posterior probabilities of the senones) varies between languages during training [6]. This allows hidden layers to be shared across multiple languages and used to improve the performance of other languages [1, 6].

The bottleneck (BN) features extracted from multilingual network have been shown to posses cross-lingual properties and transferable accross languages [3, 16]. The tandem features obtained by concatenating BN features derived from the OOD network and the in-domain acoustic features can be used to improve the performance of the target language in tandem GMM and hybrid systems. This domain adaptation procedure, called multi-level adaptive networks (MLAN), aims to take advantage of both regularization and feature-space approaches by exploiting the language independent bottleneck features as relevant features for discrimination in the target language [2, 8]. The extension of this method used multi-task learning in order to generate BN features has been proposed in [5]. The overall adaptation procedure consists of training two DNNs. The multi-task learning is used to train the first DNN with BN layer by exploiting OOD and in-domain data simultaneously. The BN features extracted from the first DNN are then combined with the in-domain spectral features (i.e., PLP or MFCC) for training the second DNN.

3. MLAN and multi-task MLAN schemes

The MLAN architecture combines regularization and feature-space approaches for exploiting resource-rich languages [5]. The overall MLAN framework consists of training two DNNs. In the conventional MLAN approach, the first DNN with BN layer is trained using OOD. The extracted BN features are then combined with in-domain spectral features for training the second DNN where the in-domain alignment is used. The second-level DNN training is used to discriminatively select OOD features that are important for classification. The extended version of MLAN has used multi-task training for developing the first DNN (with BN layer). In multi-task learning, the primary task is solved jointly with additional closely related tasks using shared feature representation in order to improve the generalization of the model. This is implemented by sharing the input layer of the network and effectively increases the amount of training data for each task. Figure 1 shows the MLAN and multi-task MLAN architectures used for language adaptation in this study.

4. Experimental setup

4.1. Experimental scenarios and datasets

In all our experiments, the Russian language is used as the target one. Specifically, the part of the GLOBALPHONE database [17, 18] consisting of Russian speech is used. The training set of the database, consisting of approximately 21 hours of speech, was split to 90% and 10% sets used for training and cross-validation respectively. For evaluation, the Russian test set of the GLOBALPHONE database is used, containing 1.6 hours of speech from 10 speakers.

As source languages, three different scenarios are investigated. In the first scenario, French, a language uncorrelated to the target one, was chosen. The French part of the GLOBALPHONE database was used for this scenario. It consists of approximately 25 hours of speech. As above, the training set of the database was split to a 90% set used for training and a 10% set used for cross-validation. As mentioned earlier, the correlation/uncorrelation of the source and target languages is measured in the sense of phonetic similarity between the two languages. In the case of French language, there is approximately a 47% overlap of phones in the phonesets of Russian and French languages.

In the second scenario, Ukrainian, a language correlated to the target one, was chosen. The Ukrainian part of the GLOBALPHONE database was used for this scenario. It consists of approximately 11.5 hours of speech. The sets were split in the same way as described above. Since the Ukrainian data of GLOBALPHONE have approximately half the size of the French data, an additional case where half the size of the French data (12.5 hours) were used as source data. This was done in order to compare the two systems with the two different source languages with the same amount of source training data. In the case of Ukrainian language, there is a approximately a 79% overlap of phones in the phonesets of Russian and Ukrainian languages.

In the third scenario, the English language was chosen. A combination of three partial English databases was used. The LIBRISPEECH dataset contains 1000 hours of read speech recordings based on texts from Project Gutenberg [19]. The ICSIAMI corpus is obtained by combining both ICSI and AMI meeting corpus with a total of 140 hours of meeting recordings [20, 21]. The TED-LIUM dataset is derived from TED talks which contains 118 hours of TED recordings recorded from a close-talking microphones of a high-quality [22]. In this case, we randomly selected 50 hours of speech from each dataset and combined them. This gives a total of 150 hours of English data. In the case of English language, there is a approximately a 45% overlap of phones in the phonesets of Russian and English languages.

4.2. Acoustic modeling

The Kaldi toolkit is used to build DNN/HMM system [23]. The acoustic model is trained on 39-dimensional MFCC features, including their delta and acceleration versions without speaker adaptive training (SAT). The DNN used a 9-frame temporal context, enriched with cepstral mean normalization, employing 4 hidden layers of 2048 neurons each. In the case of the systems trained on English data, the pronunciation dictionary was built based on publicly available CMU dictionary and include vocabularies in the training text from the LIBRISPEECH, ICSIAMI and TED-LIUM datasets. In the cases of French, Ukrainian and Russian systems, the respective dictionaries and phonesets of GLOBALPHONE were used. In the cases of MLAN and multi-task MLAN, the BN layer was composed of 39 units. The state alignments for training DNN were obtained from GMM/HMM system. Since the target language is always Russian, for evaluation, the decoding is performed using a 3-gram language model developed based on GLOBALPHONE.

5. Results

In Table 1, the word error rate (WER) in percentage, of the baseline DNN trained on the Russian dataset (“Baseline”), can be seen. Additionally the MLAN and multi-task MLAN results for the three cases i.e., using as source languages, French (both cases, using half and full data), Ukrainian and English are presented. Finally for reasons of comparison, the “Adaptation” cases for each of the source languages are shown in the table. In this case, the baseline DNN trained on each of the source languages, is retrained using the training set of Russian language. No “freezing” of any layer is performed in this case.

As general remarks, seen from the results in the table, all three cross-lingual adaptation schemes (in all scenarios) managed to outperform the Baseline DNN system trained on Russian data. Additionally, in each case of the different source languages, the MLAN case improves the accuracy in

Table 1: Word error rates (WER) in percentage of the baseline DNN system in Russian and the three adaptation schemes: i) adaptation of entire network initially trained on the source language, ii) cross-lingual MLAN adaptation and iii) cross-lingual multi-task MLAN adaptation. In parenthesis, the hours of training/cross-validation data of the source languages are presented. The target language is Russian.

System	Source Language	WER(%)
Baseline	Russian (21h)	30.50
Adaptation	French (12.5h)	30.51
MLAN		28.86
multi-task MLAN		28.20
Adaptation	French (25h)	30.38
MLAN		28.66
multi-task MLAN		27.96
Adaptation	Ukrainian (11.5h)	30.33
MLAN		28.56
multi-task MLAN		28.00
Adaptation	English (150h)	29.83
MLAN		27.71
multi-task MLAN		27.62

respect to the Adaptation case and the multi-task MLAN outperforms both the other adaptation schemes.

At this point it should be denoted that in all three adaptation schemes, in the two scenarios of using the full French data and of using the Ukrainian data, the performance is very similar. The Adaptation cases outperform the Baseline one by 0.4% and 0.6% relative improvement respectively for French and Ukrainian cases. The MLAN cases outperform the Baseline one by 6% and 6.4% relative improvement respectively for French and Ukrainian cases. The multi-task MLAN cases outperform the Baseline one by 8.3% and 8.2% relative improvement respectively for French and Ukrainian cases.

On the other hand, when half of the French data are used, matching approximately the amount of Ukrainian data used in the second scenario, the performance of all three adaptation schemes was decreased in respect to the full French data case. These results validate our hypothesis, showing clearly the importance of the correlation of the source and target languages, in respect also to the amount of data of the source language, used. Nonetheless, in this scenario, the Adaptation method didn't manage to outperform the Baseline model, achieving the same WER with it.

Finally, the third scenario, using English as source language and using 150 hours of training data, was aiming at investigating the importance of the amount of training data used in the source language. The highest improvement, over the Baseline system, was achieved by the MLAN and the multi-task MLAN schemes of this scenario, showing 9.2% and 9.4% relative improvement. This shows clearly the importance of using adequate amount of data from the source language. Finally, it can be seen that in this case, the difference between the two MLAN schemes is very small.

6. Conclusions

The MLAN and multi-task MLAN schemes were investigated in this preliminary work. French, Ukrainian and English data were used in three different scenarios, using Russian as target language. In the scenario where French was used as the source language, the MLAN schemes needed to be trained with double the size of data of the ones used in the Ukrainian case, in order to achieve the same performance as in the case where Ukrainian was used as the source language. When less French data were used, the performance was decreased in respect to the scenario where the Ukrainian data were used. These results showed the importance of the correlation between the source and target languages. Furthermore, in the case when English was used as source language, using 6 times and 13 times more training data, for producing the BN features, than in the French and the Ukrainian scenarios, the MLAN and multi-task MLAN schemes achieved the highest relative improvement of 9.2% and 9.4% respectively.

These results have shown the importance of both the investigated factors. Specifically, as follows from the results, perhaps counter-intuitively, it is worth training using large amounts of uncorrelated data. Further, the results show that correlated data is also helpful. It follows that, to take on an under-resourced language or dialect, it is worth collecting and using as much correlated data as possible. In the future, the use of multilingual data for producing the bottleneck features will be investigated by the authors.

7. Acknowledgements

This work has been conducted with the support of the European Community H2020 “Research and Innovation Action” funding, under “Scalable Understanding of Multilingual Media” (SUMMA) project and SESAR funding under “Machine Learning of Speech Recognition Models for Controller Assistance” (MALORCA) project.

8. References

[1] S. Thomas, S. Ganapathy, and H. Hermansky, “Multilingual MLP features for low-resource LVCSR systems.” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012, pp. 4269–4272.

[2] P. J. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland., “Transcription of multi-genre media archieves using out-of-domain data,” in *IEEE Spoken Language Technology Workshop*, 2012.

[3] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, “Investigation on cross- and multilingual MLP features un-

der matched and mismatched acoustical conditions," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2013.

[4] J. Li, R. Zheng, and B. Xu, "Investigation of cross-lingual bottleneck features in hybrid ASR systems," in *Proceedings of Interspeech*, 2014.

[5] P. Bell, J. Driesen, and S. Renals, "Cross-lingual adaptation with multi-task adaptive networks," in *Proceedings of Interspeech*, 2014.

[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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[7] P. Motlicek, D. Imseng, B. Potard, P. N. Garner, and I. Himawan, "Exploiting foreign resources for DNN-based ASR," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, pp. 1–10, 2015.

[8] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2013.

[9] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, 1994.

[10] C. Legetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[11] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Proceedings Eurospeech*, 1995, pp. 2183–2186.

[12] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proceedings of Interspeech*, 2010.

[13] A. Ghoshal and P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[14] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[15] Y. Huang, M. Slaney, M. L. Seltzer, and Y. Gong, "Towards Better Performance with Heterogeneous Training Data in Acoustic Modeling using Deep Neural Networks," in *Proceedings of Interspeech*, 2014.

[16] M. Müller, S. Stüker, Z. Sheikh, F. Metze, and A. Waibel, "Multilingual deep bottle neck features - A study on language selection and training techniques," in *International Workshop on Spoken Language Translation (IWSLT)*, 2014.

[17] T. Schultz, "Globalphone: A multilingual speech and text database developed at karlsruhe university," in *Proceedings of the ICSLP*, 2002, pp. 345–348.

[18] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *The 38th International Conference on Acoustics, Speech, and Signal Processing*, 2013, iCASSP 2013.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2015, pp. 5206–5210.

[20] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI spring 2007 meeting and lecture recognition system," *Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science*, vol. 4625, pp. 450–463, 2008.

[21] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 486–498, 2012.

[22] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in *in Proceedings of the Eight International Conference on Language Resources and Evaluation*, 2012.

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