Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families

Ngoc Thang Vu and Tanja Schultz

Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)

thang.vu@kit.edu, tanja.schultz@kit.edu

Abstract

In this paper, we present our latest investigations of multilingual Multilayer Perceptrons (MLPs) for rapid language adaptation between and across language families. We explore the impact of the amount of languages and data used for the multilingual MLP training process. We show that the overall system performance on the target language is significantly improved by initializing it with a multilingual MLP. Our experiments indicate that the more languages we use to train a multilingual MLP, the better is the initialization for MLP training. As a result, the ASR performance is improved, even if the target language and the source languages are not in the same language family. Our best results show an error rate improvement of up to 22.9% relative for different target languages (Czech, Hausa and Vietnamese) by using a multilingual MLP which has been trained with many different languages from the GlobalPhone corpus. In the case of very few training or adaptation data, an improvement of up to 24% relative in terms of error rate is observed.

Index Terms: multilingual speech processing, multilingual Bottle-Neck feature, rapid language adaptation

1. Introduction

The performance of speech and language processing technologies has been improved dramatically over the past decade with an increasing number of systems being deployed in a large variety of languages and applications, such as spoken dialog systems, speech summarization, spoken information retrieval and speech translation. However, most efforts to date are still focused on a small number of languages. With more than 6,900 languages in the world and the need of supporting multiple input and output languages, the most important challenge today is the rapid portation of speech processing systems to new languages with little manual effort, with few data, and at reasonable costs. In the last years, the use of neural networks to improve ASR performance earned a lot of attention in the speech community. One application of them are Multilayer Perceptrons (MLP) for feature extraction [1, 2]. Instead of the traditional Mel-Frequency Cepstral Coefficients (MFCC), the values of the output layer (Tandem features) or the values of the hidden layer (Bottle-Neck features) are used in the preprocessing step. In many setups and experimental results, MLP features proved to be a high discriminative power and very robust against speaker and environmental variations. Furthermore, cross-lingual and multilingual studies indicated that MLP features are language independent. In many papers, it was shown that features extracted from an MLP which was trained with one language can be used for another language. For example, the authors of [3] showed that features extracted from an English-trained MLP improve Mandarin and Arabic ASR performance over the spectral feature (MFCC) baseline system. In [4], cross-lingual portability of MLP features from English to Hungarian was investigated by using English-trained phone and articulatory feature MLPs for a Hungarian ASR system. Furthermore, a cross-lingual MLP adaptation approach was performed, in which the input-to-hidden weights and the hidden biases of the MLP corresponding to the Hungarian language were initialized by English-trained MLP weights, while the hidden-to-output weights and output biases were initialized randomly. The results indicated that cross-lingual adaptation often outperforms cases in which the MLP features are extracted from a monolingual MLP. In [5], the authors explored the portability of phone- and articulatory feature based tandem features to a different language without retraining. Their results showed that articulatory feature based tandem features are comparable to the phone-based ones if the MLPs are trained and tested on the same language. However, the phone based approach is significantly superior in application to a new language without retraining. Imseng et al. [6] investigated multilingual MLP features on five European languages, namely English, Italian, Spanish, Swiss French, and Swiss German from the SpeechDat(II) corpus. They trained a multilingual MLP to classify context-independent phones and integrated it directly into the preprocessing step for monolingual ASR. Their studies indicate that shared multilingual MLP feature extraction gives the best results. Plahl et al. [7] trained several Neural Networks (NNs) with a hierarchical structure with and without bottle neck topology. They showed that the topology of the NN is more important than the training language, since almost all NN features achieve similar results, irrespective of whether training and testing languages match. They obtained the best results on French and German by using the (cross-lingual) NN which has been trained on Chinese or English data without adaptation. In [8, 9], Thomas et al. demonstrated how to use data from multiple languages to extract features for an under-resourced language and hence improve ASR performance. They referred to using a data-driven approach in which no knowledge about the phone set of the target languages was needed. In [10], the language independent character of bottle neck features was demonstrated on the GlobalPhone database. Improvements were observed by using multilingual bottle-neck features. Our latest research in [11] presented first experiments on using a multilingual MLP for initializing MLP for new languages. The approach showed a substantial improvement in terms of ASR performance and proved its robustness against transcription errors [12].

In this paper, we focus on the investigation of the impact of the amount of languages and data used for the MLP training process. Moreover, its potential to improve the MLP training process and its influence on ASR performance using multilingual MLP between and across language families will be explored.

Futhermore, the use of multilingual MLP to initialize the MLP training for the new language offers an opportunity to retrain the network with the data of the new language. Hence, it raises the important question about the difference of the ASR performance with and without retraining. Therefore, we explore if the multilingual MLPs are accurate enough and so retraining may no longer be required.
The remainder of the paper is organized as follows: In section 2 we describe our data resource and the baseline system. Section 3 describes our multilingual MLP and its application to new languages. In section 4, we present our experiments and results. The study is concluded in Section 5 with a summary and future work.

2. Data Resource and Baseline System

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [13]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work, we selected French, German, Spanish, Bulgarian, Polish, Croatian, Russian, Czech, Portuguese, Mandarin, Korean, Thai, Japanesse, Hausa and Vietnamese from the GlobalPhone corpus. In addition, we used the English speech data from WSJ0. In our experiments, we used Czech, Hausa, and Vietnamese as target languages and the remaining ones as source languages. We splitted the source languages into three different catogeries in order to perform our experiments: The first one called *Big4* contains European, resource-rich languages like English, French, German, and Spanish. The second one consists of four different *Slavic* languages, namely Bulgarian, Croatian, Polish and Russian. The last one is composed of the four *Asian* languages Chinese, Japanese, Korean and Thai.
The baseline recognizer for the target languages can be described as follows: the language model was built with a large amount of text data which were crawled using the Rapid Language Adaptation Toolkit [14]. For acoustic modeling, we applied the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory trained from seven GlobalPhone languages [15]. To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds. The standard front-end was used by applying a Hamming window of 16ms length every 10ms. Each feature vector has 143 dimensions resulting from stacking 11 adjacent frames of 13 MFCC coefficients each. A Linear Discriminant Analysis transformation reduces the feature vector size to 42 dimensions. For Vietnamese ASR, we merged monosyllable words to bi-syllable words to enlarge the context in acoustic modeling and the history of the language model [16]. Table 1 presents the trigram perplexities (PPL), Out-Of-Vocabulary (OOV) rates, vocabulary sizes, and error rates (ER) for the selected languages.

Table 1: *PPL, OOV, vocabulary size, and ER for Czech, Hausa, and Vietnamese*

Languages	PPL	OOV(%)	Vocabulary	ER (%)
Czech (CZ)	1361	4.0	267k	19.5
Hausa (HAU)	77	0.5	41k	14.6
Vietnamese (VN)	176	0	30k	12.1

3. Multilingual Multilayer Perceptron and its Application to New Languages

To train a multilingual Multilayer Perceptron (ML-MLP) for context-independent phones, we used the knowledge-driven approach to create a universal phone set, i.e. the phone sets of all languages were pooled together and then merged based on their IPA symbols. Afterwards, some training iterations were applied to create the multilingual model and, thereafter, the alignment for the complete data set. In this work, we used audio data from different languages, such as English, French, German, Spanish, Bulgarian, Polish, Croatian, Russian, Mandarin, Korean, Thai, and Japanese to train the multilingual Multilayer Perceptron. Figure 1 shows the layout of our MLP architecture which is similar to [17]. As input for the MLP network, we stacked 11 adjacent MFCC feature vectors and used phones as target classes. A five layer MLP was trained with a 143-1500-42-1500-81 feed-forward architecture. All neural networks were trained using ICSI QuickNet3 software [18]. We used a learning rate of 0.008 and a scale factor of successive learning rates of 0.5. The initial values of the network were chosen randomly. In this work, we trained five different multilingual MLPs based on different languages in the same language family and also the combination between them. Table 2 presents the frame-wise classification accuracy of the multilingual MLPs on cross-validation data (10% of our training data).

Table 2: *Frame-wise classification accuracy of the multilingual MLPs on their cross-validation data*

Languages	CVAcc
Big4	67.61
Slavic	67.55
Asian	66.34
Big4 + Slavic	60.86
Big4 + Slavic + Asian	60.15

In the preprocessing of the bottleneck (BN) systems, the LDA transformation is replaced by the first 3 layers of the Multilayer Perceptron using a 143-1500-42 feed-forward architecture (Bottle-Neck), followed by stacking 5 consecutive output frames. Finally, a 42-dimensional feature vector is generated by an LDA, followed by a covariance transformation.

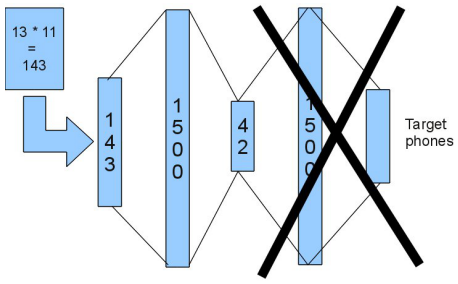

Figure 1: Bottle-Neck features

Figure 2 illustrates the initialization scheme. For the new language, we select the output from the ML-MLP based on the IPA table and use it as initialization of the MLP training. All

the weights from the ML-MLP but only the output biases from the selected targets are employed.

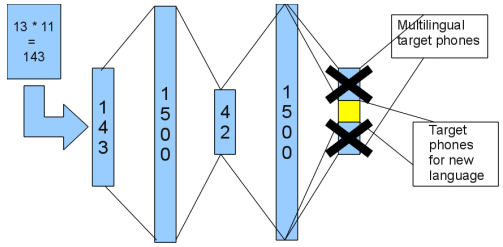

Figure 2: Initialization scheme for MLP training or adaptation using a multilingual MLP

4. Experiments and Results

For language adaptation, we conducted two different experiments: using all training data and employing only a small amount of training data of the Czech, Hausa, and Vietnamese GlobalPhone data set. In both cases, we applied different multilingual MLPs for the MLP training initialization and also experimented with and without retraining.

4.1. Using full database

In the first experiment, we applied different multilingual MLPs for the MLP training initialization and used all the training data to train the monolingual MLP for each target language. Table 3 shows the frame-wise classification accuracy on the cross-validation data for all MLPs trained with different initializations. We observed large improvements in comparison to the MLP trained with random initialization. The more languages we used to train the multilingual MLP, the better was the final performance of the target language MLP. However, the difference between the MLP performance of the target language was minor by varying the group of the source languages.

Table 3: *Frame-wise classification accuracy of the target language MLPs with different initialization on cross-validation data*

Initialization	Czech	Hausa	Vietnamese
Random	72.34	73.47	65.13
Big4	76.62	76.49	67.09
Slavic	76.28	76.38	66.94
Asian	76.05	76.61	67.05
Big4 + Slavic	77.13	76.70	67.56
Big4 + Slavic + Asian	77.62	76.92	68.08

After finishing the MLP traing, all the MLPs were used to extract the BN features for the ASR experiments. Table 4 shows the ASR performance for Czech, Hausa, and Vietnamese with MFCC features and BN features which were initialized with different multilingual MLP trained on speech data between and across language families after retraining. The results show that we got overall significant ASR performance improvements over the MFCC and the MLP with random initialization even if the source languages and the target language are not in the same language family. However, for the case of Czech and Vietnamese, we obtained the best results by using the source languages which belong to the same language family as the target language. Using the multilingual MLP trained on *Asian* data, we obtained the best performance for the Hausa ASR system.

Table 4: *ER for Czech, Hausa, and Vietnamese ASR using MFCC features and BN features with different multilingual MLPs between and across language families for initializations.*

Systems	Czech	Hausa	Vietnamese
MFCC	19.5	14.6	12.1
Random Init	19.2	15.1	11.4
Big4 Init	16.8	14.2	10.1
Slavic Init	**16.3**	14.2	10.7
Asian Init	17.1	**14.1**	**10.0**

Next, we successively increased the number of languages and thereby the amount of data to train different multilingual MLPs and used it to initialize the MLP for our target languages. Figure 3 illustrates the ASR performance on Czech, Hausa, and Vietnamese using those different BN features. The results show

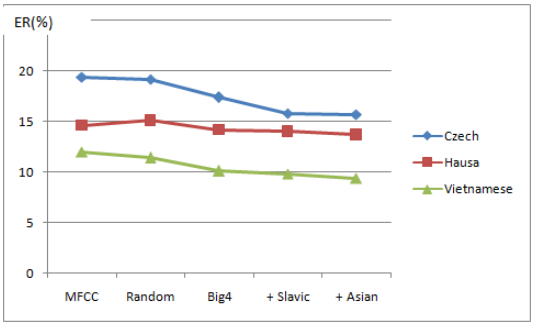

Figure 3: ER for Czech, Hausa, and Vietnamese ASR trained on all the training data using MFCC features, and BN features with different initializations.

that the more languages we used to train the multilingual MLP, the better was the final ASR performance. We observed some improvements by adding more languages to train the multilingual MLP, especially, if the source languages and the target language are in the same language family. The results in [12] showed that using multilingual MLP for the initialization of an MLP training led to better results than using monolingual MLP trained with the same amount of data. Hence, the improvements by using more languages might be a result of the diversity of languages used to train the multilingual MLP.

For the case of Hausa, we also observed improvements, although all the source languages are quite different from the Hausa language. Hence, we performed a further analysis of the similarity between the Hausa language and the source languages. We applied the polyphone average, in this case monophone, triphone, and quintphone, as a criteria for language similarity which was successfully used in [15]. Figure 4 shows the monophone, triphone and quintphone coverage of Hausa by twelve languages. We observed a quite high correlation between the polyphone coverage and the ASR improvement by adding more source languages to train the multilingual MLP. Furthermore, the coverage was increased by adding the *Asian* languages. This may be the reason why we obtained the best performance for the Hausa system by using *Asian* languages as sources.

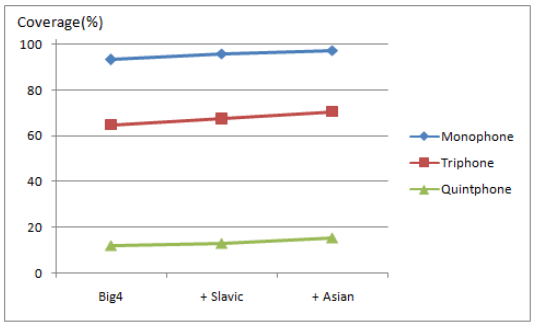

Figure 4: Hausa polyphone coverage by twelve languages

4.2. Using very small amount of data

In the second experiment, we assumed that we have very little training data (about 10% of the full training data) for Czech, Hausa, and Vietnamese. We trained the baseline system using MFCC features and obtained an ER of 27,5%, 24,9% and 26% on the Czech, Hausa, and Vietnamese test set respectively. Due to the fact that two hours are not enough data for an MLP training, we directly used the multilingual MLPs which were trained in the previous experiment to extract the Bottle-Neck features. We also trained an oracle system for each target language by using the best MLP which was trained with the full training data from the previous experiments. Figure 5 illustrates the ASR performance for Czech, Hausa, Vietnamese using different multilingual MLPs. Again, the more languages we used to train the multilingual MLP, the better was the final ASR performance. As in our experiments with the full database, we observed a substantial improvement everytime we added more data of other languages to train the multilingual MLP. In contrast to the case of Hausa and Vietnamese, the best performance for Czech is close to the oracle result. Since the ASR performace increases almost proportional with the number of languages used to train the multilingual MLP, it seems to be very promising to achieve similar results to the oracle system with more languages.

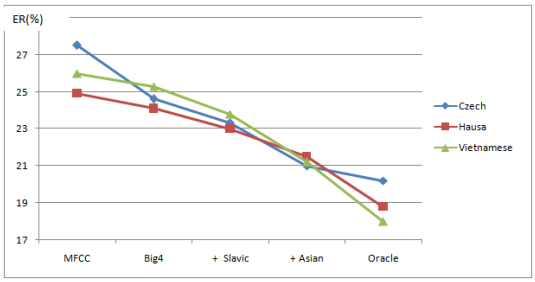

Figure 5: ER for Czech, Hausa, and Vietnamese ASR trained on a very small amount of training data using MFCC features, and BN features with different initializations wihout any retraining.

Furthermore, we retrained the multilingual MLP using the available data to improve the MLP accuracy. Table 5 presents frame-wise classification accuracy of the target language MLPs with different initializations on cross-validation data after retraining. We observed a significant improvement of the MLP performance of the target languages by adding more training data from other languages to train the multilingual MLP. It can be observed that even if the source and target languages are not related, we obtained some improvements of the MLP performance in most cases.

Table 5: *Frame-wise classification accuracy of the target language MLPs with different initializations on cross-validation data*

Initialization	Czech	Hausa	Vietnamese
Big4	70.58	71.12	58.32
Big4 + Slavic	72.18	72.56	60.12
Big4 + Slavic + Asian	72.38	73.42	62.38

Using BN features extracted from the retrained MLP, we retrained the AM and observed an overall improvement compared to the system without MLP retraining. In average, an improvement of around 4% relative was obtained. Table 6 summarizes the ER for Czech, Hausa, and Vietnamese ASR using MFCC and BN features with different multilingual MLPs for initialization after retraining.

Table 6: *ER for Czech, Hausa, and Vietnamese ASR using MFCC features, and BN features with different initializations after retraining.*

Systems	Czech	Hausa	Vietnamese
MFCC	27.5	24.9	26.0
Big4	23.8	23.7	22.8
+ Slavic	22.0	22.4	21.7
+ Asian	**20.9**	**21.3**	**20.3**
Oracle	20.2	18.8	18.0

5. Conclusions

The paper presented our latest investigations of multilingual bottle-neck features and their application to rapid adaptation to a new language at feature level. Based on the experiments on the GlobalPhone data set, we are able to draw four principal conclusions:

- Multilingual MLP is a good initialization for MLP training, especially for a new language.
- More languages in the training of the multilingual MLP lead to a superior MLP performance for a new language. As a result, the ASR performance improves, even if the target language and the source languages are not in the same language family.
- Multilingual bottle-neck features are language independent and can be used for a new language without retraining to improve ASR performance.
- Even with a very small amount of training data, retraining of the multilingual MLP improves the accuracy.

Our best results showed an error rate improvement of up to 22.9% relative for different target languages (Czech, Hausa and Vietnamese) by using a multilingual MLP trained with many different languages of the GlobalPhone database. In the case of a small amount of data, an improvement of up to 24% was observed.

6. References

[1] H. Hermansky, D. Wellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMMsystems. In Proc. ICASSP, Turkey, 2000.

[2] F. Grezl et al.. Probabilistic and bottle-neck features for LVCSR of meetings. In Proc. ICASSP, USA, 2007.

[3] A. Stolcke, F. Grezl, M-Y Hwang, X. Lei, N. Morgan, D. Vergyri. Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons. In Proc. ICASSP 2006.

[4] L. Toth, J. Frankel, G. Gosztolya, S. King. Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian. In Proc. Interspeech, 2008.

[5] O. Cetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel. Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs. In Proc. ASRU, 2007.

[6] D. Imseng, H. Bourlard, M. Magimai.-Doss. Towards mixed language speech recognition systems. In Proc. Interspeech, Japan, 2010.

[7] C. Plahl, R. Schlueter and H. Ney. Cross-lingual Portability of Chinese and English Neural Network Features for French and German LVCSR. In Proc. ASRU, USA 2011.

[8] S. Thomas, S. Ganapathy and H. Hermansky. Multilingual MLP Features For Low-resource LVCSR Systems. In Proc. ICASSP, Japan, 2012.

[9] S. Thomas, S. Ganapathy, A. Jansen and H. Hermansky. Data-driven Posterior Features for Low Resource Speech Recognition Applications. In Proc. Interspeech, USA, 2012.

[10] K. Vesely, M. Karafiat, F. Grezl, M. Janda, E. Egorova. The language-independent bottleneck features. In Proc. SLT, USA, 2012.

[11] N.T. Vu, F. Metze, T. Schultz. Multilingual bottle-neck feature for under resourced languages. In Proc. SLTU, South Africa, 2012.

[12] N.T. Vu, W. Breiter, F. Metze, T. Schultz. An Investigation on Initialization Schemes for Multilayer Perceptron Training Using Multilingual Data and their Effect on ASR Performance. In Proc. Interspeech, USA, 2012.

[13] T. Schultz, N.T. Vu, T. Schlippe. GlobalPhone: A Multilingual Text & Speech Database in 20 Languages. In Proc. ICASSP, Canada, 2013.

[14] N.T. Vu, Tim Schlippe, Franziska Kraus, Tanja Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Proc. Interspeech, Japan, 2010.

[15] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication, Volume 35, Issue 1-2, pp 31-51, 2001.

[16] N.T. Vu, T. Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In Proc. ASRU, Italy, 2009.

[17] F. Metze, R. Hsiao, Q. Jin, U. Nallasamy, and T. Schultz. The 2010 CMU GALE Speech-to-Text System. In Proc. Interspeech, Japan, 2010.

[18] http://www.icsi.berkeley.edu/Speech/qn.html