MULTILINGUAL BOTTLE-NECK FEATURES AND ITS APPLICATION FOR UNDER-RESOURCED LANGUAGES

Ngoc Thang Vu[1], Florian Metze[2], Tanja Schultz[1]

[1]Cognitive Systems Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)
[2]Language Technologies Institute, Carnegie Mellon University (CMU), Pittsburgh, PA, USA
thang.vu@kit.edu

ABSTRACT

In this paper we present our latest investigation on multilingual bottle-neck (BN) features and its application to rapid language adaptation to new languages. We show that the overall performance of a Multilayer Perceptron (MLP) network improves significantly by initializing it with a multilingual MLP. Furthermore, ASR performance increases on both, on those languages which were used for multilingual MLP training, and on a new language. We propose a new strategy called "open target language" MLP to train more flexible models for language adaptation, which is particularly suited for small amounts of training data. The final results on the Vietnamese GlobalPhone database gave 15.8% relative improvement in terms of Syllable Error Rate (SyllER) for the ASR system trained with 22.5h data and 16.9% relative gains for the system trained with only 2h data.

***Index Terms*—** multilingual bottle-neck feature, language adaptation

1. INTRODUCTION

The performance of speech and language processing technologies has improved dramatically over the past decade with an increasing number of systems being deployed in a large variety of languages and applications. However, most efforts are still focused on a small number of languages. With more than 6,900 languages in the world, the biggest challenge today is to rapidly port speech processing systems to new languages with little manual effort and at reasonable costs. In the last few years the use of multi layer perceptron (MLP) for feature extraction showed impressive ASR performance improvements. In many setups and experimental results, MLP features proved to be of high discriminative power and very robustness against speaker and environmental variation. Furthermore, some interesting cross-lingual and multilingual studies exist. In [1], it was shown that features extracted from an English-trained MLP improves Mandarin and Arabic ASR performance over the spectral feature (MFCC) baseline system. 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 in [2]. Furthermore, a cross-lingual MLP adaptation approach was investigated, where the input-to-hidden weights and hidden biases of the MLP corresponding to Hungarian language were initialized by English-trained MLP weights, while the hidden-to-output weights and output biases were initialized randomly. These results indicated that cross-lingual adaptation often outperforms cases, in which the MLP feature is extracted from a monolingual MLP. In [3] was explored how portable phone- and articulatory feature based tandem features are in a different language without any 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. But the phone-based approach is significantly better on a new language without retraining. Imseng et al. [4] investigated multilingual MLP features on five European languages, namely English, Italian, Spanish, Swiss French, and Swiss German from the Speech-Dat(II) corpus. They trained a multilingual MLP to classify context-independent phones and integrated it directly into preprocessing step for monolingual ASR. Their studies indicate that shared multilingual MLP feature extraction give the best results. Plahl et al. [5] trained several Neuronal 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 language match. They obtained the best results on French and German by using the (cross-lingual) NN which trained on Chinese or English data without any adaptation.

In this paper we investigate multilingual bottle-neck feature and the potential to use it for initializing MLP training. Furthermore, we investigate its application to rapid language adaptation of new languages at the feature level. We propose a new strategy called "open target language" MLP to train more flexible models for language adaptation, particularly with small amount of data. The paper is organized as follows: Section 2 describes the bottle-neck feature approach. In section 3 we introduce the multilingual multi layer perceptron network training and its application for rapid language adaptation at feature level. Section 4 presents the results on the GlobalPhone data set. The study is concluded in Section 5 with a summary and future work.

2. BOTTLE-NECK FEATURES

Figure 1 shows the layout of our bottle-neck MLP architecture which is similar to [6]. As input for the MLP network we stacked 11 adjacent MFCC feature vectors and used phones as target classes. A 5 layer MLP was trained with a 143-1500-42-1500-81 feed-forward architecture. In the pre-processing of the bottle-neck systems, the LDA transform is replaced by the first 3 layers of the Multi Layer Perceptron using a 143-1500-42 feed-forward architecture, followed by stacking of 5 consecutive bottle-neck output frames. Finally, a 42-dimensional feature vector is generated by an LDA, followed by a covariance transform. All neural networks were trained using ICSIs QuickNet3 software [7].

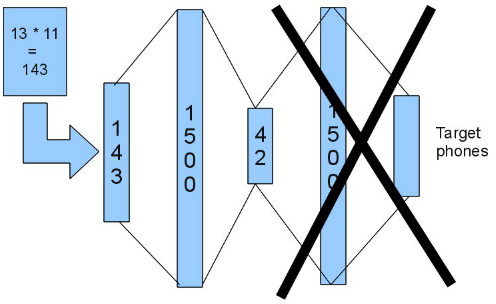

Fig. 1. Bottle-Neck feature

3. "OPEN TARGET LANGUAGE" MULTILAYER PERCEPTRON

To train a multilingual multilayer perceptron (ML-MLP) for context-independent phones, we used the knowledge-driven approach to create an universal phone set, i.e., the phone sets of all languages were pooled together and then merged based on their IPA symbols. After that some training iterations were applied to create the multilingual model and therefore the alignment for the complete data set. In this case we used English, French, German, and Spanish to train the multilingual acoustic models. The universal phone set has 81 phonemes which cover only about 30% of the IPA symbols. This leads to the fact, that we have some difficulties applying this multilingual MLP to a new language especially when the amount of training data is limited. So, we propose a new strategy to train an "open target language" MLP network and its application for language adaptation at feature level. Our idea is to extend the target classes so that we can cover all the phoneme in the IPA table. So the first thing that should be done is to select the training data for the new open target class. Since all phonemes in IPA are described by their articulatory features, we used the data from the phoneme with the same articulatory features as new target phoneme. For some special phonemes like aspirated phoneme or diphthong, the following steps could be applied:

1) If the phoneme is an aspirated phoneme then we use the frames of the main phone (e.g. A A-b, A-m) and /h/-e

2) else if the phoneme is a diphthong, vowel-1 vowel-2 (V1V2) then we use the frames of V1-b, V1-m and V2-e.

After finishing the training data selection of all new target phonemes, we first trained a normal MLP with a subset of the training data to save time and learn a rough structure of the phone set which can be covered in our training set. After that, we used this MLP as initialization to train weights for the new target phonemes with all selected data. Due to the fact, that the new target classes are not real, it is possible that the MLP network after this step does not match our real target phones anymore. So we retrained the whole network using all of the training data. For the new language, we select the output from the ML-MLP based on the IPA table and use it for initialization of the MLP adaptation or training. Figure 2 illustrates the idea of our approach. All the weights from the ML-MLP were taken and only the output biases from the selected targets were used.

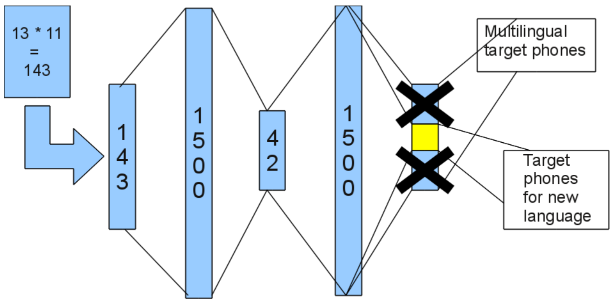

Fig. 2. Initialization for MLP training or adaptation using a multilingual MLP

4. EXPERIMENTS AND RESULTS

4.1. Data corpora and baseline system

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [8]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work we selected Vietnamese, English, French, German, and Spanish from the GlobalPhone corpus. To retrieve large text corpora for language model building, we used our Rapid Language Adaptation Toolkit [9] for an up to twenty days crawling process [10]. For acoustic modeling, we applied the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory trained from seven GlobalPhone languages [11]. 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 closest match is derived from an IPA-based phone mapping. In this work, we did a phone mapping for each language and trained five different acoustic models, using the standard front-end by applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 143 dimensions resulting from stacking 11 adjacent frames of 13 MFCC coefficient 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 [12]. The model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. Table 1 gives a breakdown of the trigram perplexities (PPL), Out-Of-Vocabulary (OOV) rate, vocabulary size, and word error rate (WER) for the selected languages.

Table 1. *PPL, OOV, vocabulary size, and ER for Vietnamese, English, French, German, and Spanish*

Languages	PPL	OOV	Vocabulary	ER
Vietnamese (VN)	323	0%	35k	12.1%
English (EN)	284	0.5%	60k	11.5%
French (FR)	352	2.4%	65k	20.4%
German (GE)	148	0.4%	41k	10.6%
Spanish (SP)	224	0.1%	19k	11.9%

4.2. Multilingual bottle-neck feature

We trained a multilingual MLP with English, French, German and Spanish training data. The MLP has 5 layers and has a topology 143-1500-42-1500-81. To make a comparison we trained also different monolingual MLPs with the same topology (only the number of target phones is changed). For all the MLP training, we used a learning rate of 0.008 and a scale factor of successive learning rates of 0.5. Table 2 shows the frame-wise classification accuracy for all MLPs using random and multilingual MLP initialization on cross-validation data. We observed overall improvement by using multilingual MLP as initialization compared to random initialization. Moreover, we observed an overall speed improvement for the MLP training (about 40%).
Furthermore, different ASR systems were trained using different MLP features (one with random initialization (*Random-Init*) and one with multilingual MLP initialization (*MultiLing-Init*)) for all languages. The results in Table 3 show that the MLP systems overall outperform the baseline system which trained with traditional MFCC feature. Furthermore, the *MultiLing-Init* systems outperform the *Random-Init* systems by up to 10% relative in terms of WER for all languages which indicates that a MLP network trained with multilingual MLP initialization is more robust.

Table 2. *Frame-wise classification accuracy for all MLPs using random and multilingual MLP initialization on cross-validation data*

Languages	Random Init	Multilingual Init
Multilingual (ML)	67.61	-
English (EN)	70.98	73.46
French (FR)	76.73	78.57
German (GE)	63.93	68.87
Spanish (SP)	71.75	74.02

Table 3. *WER on the GlobalPhone development set*

Systems	English	French	German	Spanish
Baseline	11.5	20.4	10.6	11.9
Random-Init	11.1	20.3	10.5	11.6
MultiLing-Init	10.2	20.0	9.7	11.2

4.3. Language adaptation for a new language

4.3.1. Data selection for MLP training

Since not all Vietnamese phonemes could be covered by the multilingual universal phone set, we had to train some open phonemes using the multilingual training data. Table 4 shows all uncovered Vietnamese phonemes and their phonetic features. For uncovered Vietnamese vowel and consonants we used the training data from the phoneme with the same articulatory features e.g. Plosive, Palatal for consonant /ch/ or Close, Back for vowel /o3/. For the case of aspirated phones like /th/, we used the frames of the first two states (-b and -m) of the main phoneme (in this case, /t/) and the frames of the last state h-e. We did also almost the same for diphthongs, but using the first and the second vowel.

Table 4. *Vietnamese phones not covered by the universal phone set and their articulatory features*

VN	Articulatory features
/d2/	Plosive, DAP
/tr/	Plosive, Retroflex
/s/	Fricative, Retroflex
/r/	Fricative, Retroflex
/ch/	Plosive, Platal
/th/	t-b, t-m, h-e
/o3/	Close, Back
/ie2/	i-b, i-m, e2-e
/ua/	u-b, u-m, a-e
/ua2/	ir-b, ir-m, a-e

4.3.2. Results

For language adaptation experiments we conducted two different experiments on the Vietnamese GlobalPhone data set. In the first experiment we used all the training data and trained an ASR system using the MLP feature. By using random initialization, we achieved 65.13% accuracy on the cross validation training set by MLP training and a SyllER of 11.4% on the Vietnamese development set. To get a better initialization we applied the multilingual MLP from the previous experiment, which lead to 67.09% accuracy on the cross validation training set and 10% relative improvement in terms of SyllER compared to the MLP system with random initialization.

Table 5. *Frame-wise classification accuracy for all MLPs on cross-validation and SyllER from a system trained with 22.5h Vietnamese data*

MLP	CVAcc	SyllER
Baseline	-	12.0
Random-Init	65.13	11.4
MultiLing-Init	67.09	10.1

In the second experiment, we assumed that we have very little training data (about 2 hours) for Vietnamese. We trained the baseline system using MFCC feature and observed a SyllER of 26% on the Vietnamese development set. Due to the fact that two hours are too small for a MLP training, we directly used the multilingual MLP which was trained in the previous experiment to extract the bottle-neck feature. The SyllER was improved by 0.7% absolute which indicates that something language independent useful was learned by MLP training. To make a comparison with our new approach we adapted the MLP with 2h of Vietnamese data using the approach in [2] when the hidden-to-output weights and output biases were initialized randomly. The results were improved significantly (about 20% of cross validation accuracy and 2.5% absolute in term of SyllER). After that, we applied our method "open target language" MLP, in which we can use all the weights and output biases of the multilingual MLP. We observed 0.8% improvement after adaptation in MLP training and 1.2% absolute improvement in terms of SyllER.

Table 6. *Frame-wise classification accuracy for all MLPs on cross-validation and SyllER from a system trained with 2h Vietnamese data*

MLP	CVAcc	SyllER
Baseline	-	26.0
ML-MLP	37.23	25.3
+ Adaptation	57.54	22.8
"Open target language" MLP	58.32	21.6

5. CONCLUSION AND FUTURE WORK

The paper presents our latest investigation on the multilingual bottle-neck feature and its application for rapid language adaptation for a new language at feature level. Based on the experiments on the GlobalPhone data set, we are able to draw three principal conclusions:

- Multilingual MLP is a good initialization for MLP training especially for a new language.
- Using multilingual MLP to initialize MLP training we could reduce training time by about 40% in our experiments.
- "Open target language" MLP is a new method to train a more flexible model for rapid language adaptation at feature level especially with very little data.

In the final performance on the Vietnamese GlobalPhone database we achieved 15.8% and 16.9% relative improvement in term of SyllER for the ASR system trained with 22.5h and 2h audio data respectively.

6. ACKNOWLEDGMENTS

The authors would like to thank Franziska Kraus for her support. This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

7. REFERENCES

[1] A. Stolcke, F. Grzl, M-Y Hwang, X. Lei, N. Morgan, D. Vergyri. Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons. In International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006.

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

[3] 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. IEEE Workshop on Automatic Speech Recognition and Understanding, December 2007.

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

[5] C. Plahl, R. Schlueter and H. Ney. Cross-lingual Portability of Chinese and English Neural Network Features for French and German LVCSR. In IEEE Workshop on Automatic Speech Recognition and Understanding, USA 2011.

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

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

[8] T. Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[9] T. Schultz and A. Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc. ICASSP Las Vegas, USA 2008.

[10] N.T. Vu, T. Schlippe, F. Kraus, and T. Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Interspeech, Makuhari, Japan, 2010.

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

[12] N.T. Vu, T. Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In Automatic Speech Recognition and Understanding, Italy, 2009.