ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

An Investigation on Initialization Schemes for Multilayer Perceptron Training Using Multilingual Data and Their Effect on ASR Performance

Ngoc Thang Vu[1], *Wojtek Breiter*[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

10.21437/Interspeech.2012-12

Abstract

In this paper we present our latest investigation on initialization schemes for Multilayer Perceptron (MLP) training using multilingual data. We show that the overall performance of a MLP network improves significantly by initializing it with a multilingual MLP. 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. Furthermore, by applying Bottle-Neck feature (BN) initialized with multilingual MLP the ASR performance increases for both, the languages which were used for multilingual MLP training, and the new language. Our experiments show a word error rate improvements of up to 16.9% relative on a range of tasks for different target languages (Creole and Vietnamese) with manual and automatic transcribed training data.
Index Terms: multilingual multilayer perceptron, 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. [3] explored the behaviour of portable phone- and articulatory feature based tandem features in a new 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 SpeechDat(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 gives 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 the training and testing languages 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 explore the potential of using an existing MLP trained with multilingual data for initializing MLP training. We compare the performance of different MLPs which were initialized with random values, an existing monolingual MLP and a multilingual MLP and their impact on ASR performance. 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 ones with small amount of data. Finally, we check the robustness of our approach by applying it with automatic transcription which contains some transcription errors.

2. Bottle-Neck feature (BN)

In this short section, we present our scheme to extract Bottle-Neck feature for an ASR system using MLP. Figure 1 shows the layout of our 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 BN systems, the LDA transform is replaced by the first 3 layers of the Multi Layer Perceptron using a 143-1500-42 feed-forward architecture (Bottle-Neck), followed by stacking of 5 consecutive 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].

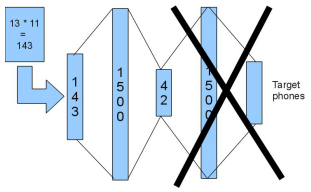

Figure 1: Bottle-Neck feature

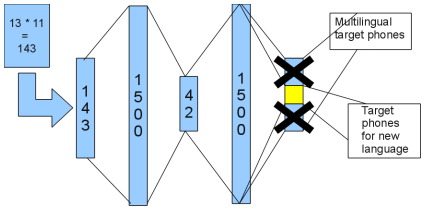

Figure 2: Initialization for MLP training or adaptation using a multilingual MLP

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 thereafter the alignment for the complete data set. In this work 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 the 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 uncovered target phoneme. Since all phonemes in IPA are described by their articulatory features, we used the data from the phoneme that have at least one articulatory feature of the target phoneme randomly. 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 the training data for all uncovered target phonemes is selected, 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 uncovered target phonemes with the all selected data. Due to the fact that the uncovered target phonemes do not have real training data, 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.

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. In addition, we used the Creole speech data in [9] as target language. To retrieve large text corpora for the language model building, we used our Rapid Language Adaptation Toolkit [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 standard front-end was used 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]. Table 1 gives a breakdown of the trigram perplexities (PPL), Out-Of-Vocabulary (OOV) rate, vocabulary size, and error rate (ER) for the selected languages.

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

Languages	PPL	OOV	Vocabulary	ER
Creole (CR)	46	1.0%	22k	12.3%
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 MLP

Using English, French, German and Spanish training data we trained a multilingual MLP (ML-MLP) which has 5 layers and has a topology 143-1500-42-1500-81. We used a learning rate of 0.008 and a scale factor of successive learning rates of 0.5. The initial values of this network were chosen randomly. On the cross validation data (10% of the training data) we observed a frame accuracy of 67.61%. To make a comparison between different initialization scheme, we trained different monolingual MLPs with the same topology (only the number of target phones differs) but initialized them with random values (*Random- Init*) or with the values of the ML-MLP (*Multilingual-Init*). For all the MLP training, we used the same parameter setup as the ML-MLP training. Table 2 shows the MLP performance on the cross validation data. We observed an overall improvement by using multilingual MLP as initialization compared to random initialization. In addition, we observed an overall speed improvement for the training (up to 40% of training time). Furthermore, different ASR systems were trained with BN features extracted

Table 2: *Multi Layer Perceptron performance*

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

from different MLPs (one with random initialization and one with multilingual MLP initialization) for all languages. The results in Table 3 show that the ASR systems using BN features universally outperform the baseline system which was trained with traditional MFCC features. Moreover, the *Multilingual-Init* system is relatively up to 9% better than the *Random-Init* for all languages which indicates that an MLP network trained with multilingual MLP initialization is more robust.

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
Multilingual-Init	10.2	20.0	9.7	11.2

4.3. Language adaptation for Creole

In this section we describe our experiments on Creole and compare the different initialization scheme for MLP training: random initialization, using monolingual MLP, and multilingual MLP and their impact on the ASR system. We chose French (FR) in this case due to the fact that Creole is related to French. We applied our approach to train the "open target language" MLP with only 80 hrs French data from BREF database [14] (*Monolingual- Init*) and used it for the MLP training for Creole. Furthermore, we also applied the ML-MLP (in 4.2) to initialize the MLP training. Table 4 shows the frame-wise classification accuracy for all MLPs trained with different initializations on cross-validation data and WER on the Creole data set. Using the MLP trained with French data for initialization, we observed a small improvement in terms of WER (0.2% absolute), but the final performance is still worse than the system trained with multilingual MLP initialization which gave 1.9% absolute improvement. We suggest that the reason lies in the fact that using multilingual data we could increase the phoneme coverage of the target language and moreover, the knowledge could be shared between languages and transformed to the new language in this case.

Table 4: *Frame-wise classification accuracy for all MLPs on cross-validation data and WER on Creole database*

Systems	CVAcc	WER	Delta
Baseline	-	12.3	
Random-Init	73.36	11.6	-0.7
Monolingual-Init (FR)	75.15	11.4	-0.9
Multilingual-Init	75.38	10.4	-1.9

4.4. Language adaptation for Vietnamese

4.4.1. Data selection for MLP training

Since not all Vietnamese phonemes could be covered by the multilingual universal phone set, we had to select some multilingual data to train the weights and bias for those phones. Table 5 shows all uncovered Vietnamese phonemes and their phonetic features. For uncovered Vietnamese vowel and consonants we used the training data from the phoneme that have at least one articulatory feature of the target phoneme 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 almost the same procedure for diphthongs, but using the first and the second vowel.

Table 5: *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

4.4.2. 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 BN 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 "open target language" MLP which trained with the multilingual data and the selected data for uncovered phones. Using this initialization scheme we could train an MLP for Vietnamese with 67.09% accuracy on the cross validation set. In terms of SyllER we observed 10% relative improvement compared to the BN system which used the MLP trained with random initialization.

Table 6: *Frame-wise classification accuracy for all MLPs on cross-validation and SyllER from ASR trained with 22.5h VN*

Systems	CVAcc	SyllER
Baseline	-	12.0
Random-Init	65.13	11.4
"Open target language" MLP	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 (*ML-MLP.Direct*). The SyllER was improved by 0.7% absolute which indicates that some useful language independent information was learnt as a result of the 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 (*ML-MLP.Adapt*). 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 7: *Frame-wise classification accuracy for all MLPs on cross-validation and SyllER from ASR trained with 2h VN*

Systems	CVAcc	SyllER
Baseline	-	26.0
ML-MLP.Direct	37.23	25.3
ML-MLP.Adapt	57.54	22.8
"Open target language" MLP	58.32	21.6

4.5. Integration of multilingual bottle-neck feature in Multilingual Unsupervised Training Framework

4.5.1. Multilingual Unsupervised Training Framework

In [13] we presented our Multilingual Unsupervised Training Framework (MUT) which enable training an acoustic model without any transcribed audio data. We use several acoustic models from different source languages to generate iteratively automatic transcriptions and apply "multilingual A-stabil" confidence score to select accurate transcriptions for acoustic model adaptation. By using this process we could enlarge the amount of automatic transcriptions with a high precision on one hand and select data from many different contexts due to the multilingual effect on the other hand. Finally we use the multilingual inventory which was trained earlier from seven GlobalPhone languages [11] to write the alignment for the selected data and train the acoustic model. The final acoustic model is the one with the best performance on the development set.

4.5.2. Experiments with Automatic Transcriptions

On top of the Vietnamese ASR which trained using 4 different source languages (English, French, German and Spanish) [13], we applied our approach "open target language" MLP to improve accuracy. Using MUT we were able to select 10 hours of training data with automatic transcriptions which have 16% SyllER and trained the Baseline ASR system for this experiment with 18.6% SyllER. Using this data we trained 2 different MLPs: one using random initialization and the one using the multilingual initialization proposed in Section 4.4. Table 8 shows the frame-wise classification accuracy for all MLPs on cross-validation data and SyllER from all systems trained with MUT. The results indicate that initialization of an MLP training with random value is quite critical with automatically transcribed data (SyllER decreases 0.4% absolute) while the multilingual initialization is much more robust (2.0% absolute improvement).

Table 8: *Frame-wise classification accuracy for all MLPs on cross-validation data and SyllER from all systems trained with Multilingual Unsupervised Training Framework*

Systems	CVAcc	SyllER	Delta
Baseline	-	18.6	
Random-Init	61.5	19.0	+0.4
"Open target language" MLP	65.0	16.6	-2.0

5. Conclusion and Future Work

This paper presents our latest investigation on initialization schemes for MLP training using multilingual data and their effect on ASR performance. Based on a range of experiments we are able to draw four principal conclusions:

- Multilingual MLP is a good initialization for MLP training especially for a new language and therefore we could save up to 40% training time in our experiments.
- "Open target language" MLP is a new method to train more flexible model for rapid language adaptation.
- Using "Open target language" MLP as initialization, the resulting model is robust against transcriptions errors.
- Multilingual MLP is a better initialization technique than monolingual MLP for MLP training even if the source language and target language are related.

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. For the task with automatic transcription, we observed about 11% relative improvement . On the Creole speech data corpus, the WER was improved from 12.3% to 10.4%.

6. Acknowledgments

The authors would like to thank Prof. Alan Black for providing us the Creole speech database. 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 Proc. 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. ASRU, 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 Proc. ASRU, 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] http://www.speech.cs.cmu.edu/haitian.

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

[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 Proc. ASRU, Italy, 2009.

[14] L.F. Lamel, J.L. Gauvain, M. Eskenazi. BREF, a Large Vocabulary Spoken Corpus for French. In Proc. EuroSpeech 1991, Italy.

[13] N.T. Vu, F. Kraus, T. Schultz. Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training. In Interspeech 2011, Italy, 2011.