INVESTIGATION ON CROSS- AND MULTILINGUAL MLP FEATURES UNDER MATCHED AND MISMATCHED ACOUSTICAL CONDITIONS

Zoltán Tüske[1], *Joel Pinto*[2], *Daniel Willett*[2], *Ralf Schlüter*[1]

[1]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany

[2]Nuance Communications Deutschland GmbH, Kackertstraße 10, 52072 Aachen

{tuske, schlueter}@cs.rwth-aachen.de, {joel.pinto,daniel.willett}@nuance.com

ABSTRACT

In this paper, Multi Layer Perceptron (MLP) based multilingual bottleneck features are investigated for acoustic modeling in three languages — German, French, and US English. We use a modified training algorithm to handle the multilingual training scenario without having to explicitly map the phonemes to a common phoneme set. Furthermore, the cross-lingual portability of bottleneck features between the three languages are also investigated. Single pass recognition experiments on large vocabulary SMS dictation task indicate that (1) multilingual bottleneck features yield significantly lower word error rates compared to standard MFCC features (2) multilingual bottleneck features are superior to monolingual bottleneck features trained for the target language with limited training data, and (3) multilingual bottleneck features are beneficial in training acoustic models in a low resource language where only mismatched training data is available — by exploiting the more matched training data from other languages.

Index Terms— MLP, bottleneck, multilingual, mismatched acoustical condition

1. INTRODUCTION

As manually transcribed speech data is one of the substantial cost factors to develop an automatic speech recognition (ASR) system for a new language, there is an increasing interest in reusing multilingual resources to ease the model training. Neural networks have become a major component of the frontend techniques of recent ASR systems [1]. Beside the hybrid acoustic modeling [2], the MLP based posterior estimations can be integrated into the Gaussian Mixture based Hidden Markov Model (GMM-HMM) framework through either probabilistic or bottleneck TANDEM approach [3, 4]. Since it was first observed in [5] that the concatenated cepstral and MLP based posterior features trained on English data significantly improved the MFCC based systems in entirely different languages, Arabic or Mandarin, many studies investigated the cross-lingual portability of NNs.

In [6], cross-lingual portability of long-term bottleneck features having complex hierarchical structure was investigated in concatenation with MFCC, and it was shown that the topology of the NN is more important than the language on which the MLP is trained on. The results indicate the universal, language independent feature extraction properties of MLP. It also demonstrated the possibility of fast development of TANDEM GMM-HMM systems without the time consuming training of MLPs. However, with modern GPUs the constraint of training complex MLPs on hundreds of hours of speech has become a less limiting factor.

To mitigate the requirement of a significant amount of data to train large MLPs from scratch e.g. for languages with low resources, NNs of other languages can be used for initialization. In the case of MLPs, only the weights before the language specific output layer are randomly initialized, and then adapted with limited target data [7]. Before adapting the net, the mapping of the target phoneme set to the "borrowed" language is another possible approach [8]. In order to train a NN on multiple languages, similar sounds across different languages can be unified knowledge based, such as IPA [9, 10], or by data driven approaches [11]. To avoid the mapping to a common phone set, more fundamental units such as articulatory features can be used [12].

Without limiting our investigation to low resource languages, this study focuses on multilingual training of MLPs, especially on BN features. Since in real-time applications the long term features having about 500ms delay is not acceptable, our investigation is limited to short-term MLP features only. Due to the fact that the available lexicons for ASR are usually simplified (e.g. by phone folding), mapping phones of multiple languages on a common set is often ambiguous or inaccurate. Following a similar approach as in [13], we investigate novel multilingual bottleneck MLP features, which force the BN layer to learn a low dimensional language independent representation of the speech.

Besides mutually testing the cross-lingual portability of short-time MLP features between three different languages, the multilingual BN features are also extensively evaluated in single pass recognition experiments, for the most part, using GMM-HMMs trained with maximum likelihood criteria. Furthermore, in a constrained scenario where we assume that only acoustically unmatched recordings are available on the target language, the multilingual bottleneck features are tested whether they are able to benefit from matched data at hand from other languages.

The paper is organized as follows. After the overview of the related work in Section 2, the Section 3 introduces the proposed multilingual bottleneck MLP to extract more language-independent BN features. The details of our experimental setups are given in Section 4. Section 5 reports the results. The study finishes with conclusions in Section 6.

2. RELATION TO PRIOR WORK

The incorporation of MLP based posterior estimation as additional features for GMM-HMM was introduced in [3]. This TANDEM approach was further improved by the bottleneck concept of [4]. To initialize the bottleneck MLP for low resource languages, [14] applied data from multiple languages after each other. In order to avoid the mapping onto a common phone set, the last layer was changed

and randomly initialized after training on each language. The multilingual MLP training applied in this paper was first proposed in [13] for hybrid acoustic modeling, and this work generalizes this idea for BN features with TANDEM approach. Compared to [14], instead of using the data of different languages sequentially, our BN are trained on the merged multilingual resources jointly. Similarly, unifying similar sounds of different languages is still not necessary. In contrast to [13], the language dependent class posteriors are generated by non-linear transformation from a low-dimensional feature space using a much wider hidden layer before the output layer.

3. MULTILINGUAL TRAINING OF BOTTLENECK MLP

In order to extract roboust MLP features from multilingual resources, we apply a recently proposed training method [13]. The three languages are denoted as GER, FRA, ENU, and they do have different numbers of phonemes. The output layer of the MLP is the joint phoneme set, where each phoneme is appended with language identity, similar as in [11]. The raw feature vectors from the three languages are merged, randomized and presented to the MLP for training along with the phonetic and language labels. In contrast to standard MLPs, the network applies language specific softmax function as output non-linearity, thus the output sums up to three in case of three languages. Exploiting the language-ID of the feature vector, back propagation is initiated only from the language specific subset of the output. In this way the often inaccurate mapping of all the phonemes of different languages to a common set is avoided. The subsets of the output still can be considered as phoneme posterior estimation of the given languages. Except the output layer, the network is shared between the languages.

In Fig. 1 the multilingual training of NNs is summarized using a special 5-layer bottleneck MLP structure. The key idea is that the bottleneck layer is shared between the languages, and the multilingual training forces the net to extract a more language-independent representation from the input. The posteriors of the different languages are generated by a non-linear transformation from a common, low dimensional feature space (bottleneck). Although in this study only the last layer is considered language dependent, the last hidden-layer can also be split up to language specific parts. However, for the proof of this type of multilingual BN concept we restrict the language dependent part of the MLP to the last layer only.

4. EXPERIMENTAL SETUP

4.1. Corpus description

Our investigation of how acoustic data from other languages can be reused for MLP based feature extraction is limited on short message dictation recognition task of three languages: German (GER), American English (ENU), and French (FRA). As we focus our study on multilingual aspects, we choose data sets with comparable acoustic conditions in each language. Thus, for testing automotive data collected in driving cars was used. Although the recording environments are similar, there are slight differences regarding the type of cars and driving noise conditions related to the specific country. In order to mitigate the potential effect of different amounts of language specific data on the cross- and multilingual investigation, corpora with similar size are selected for all the three languages to train acoustic models and MLPs. Therefore, for our research the corpus size per language is chosen as approximately 150 hours, resulting in a total of about 450 hours of speech. Every training corpus consists of two types of recording. The first part is acoustically *matched* with the test data, whereas recordings in the second group are *unmatched*. Table 1 summarizes the distribution of speech data between languages and types of recordings selected for this study.

Table 1. *Amount of training and testing data for different languages*

Corpus		Type	Language		
			GER	ENU	FRA
Amount of speech [h]	Training	matched	93	96	86
		unmatched	77	61	66
	Test		9.3	11	26
#phonemes			49	51	39

4.2. Acoustic modeling

In order to evaluate the cross- and multilingual BN features, we set up a contemporary single pass ASR system with online adaptation for our research purpose. The acoustic models are trained on cepstral features (MFCC). The pre-emphasized power spectrum is computed every 10ms. With the derivatives augmented mean normalized MFCCs are projected by LDA to a lower dimensional subspace. The acoustic models for all systems are based on cross-word triphone HMM, and the emission probabilities are estimated by Gaussian mixture distributions using maximum likelihood criteria. Discriminatively trained acoustic models whenever used are trained using minimum classification error (MCE) criterion.

For training the 5-layer bottleneck MLPs, the number of nodes in the first and last hidden layer is always fixed to 5000, whereas the bottleneck layer consists of 50 nodes. As input, 9 consecutive MFCC frames and their derivatives are fed to the MLP. The randomly initialized, fully connected MLPs are trained using the cross-entropy criterion to approximate phoneme posterior probabilities. In order to prevent overfitting and adjust the learning rate, ten percent of the training set is used for cross validation. All activations of the nodes within the output layer are transformed by the softmax function, whereas the sigmoid transfer function is applied in all other layers. For multilingual training we refer to Section 3. To extract the BN features, the linear output of the bottleneck layer is taken. Moreover, the BN features are further transformed by PCA and concatenated to the MFCC.

5. EXPERIMENTAL RESULTS

5.1. Baseline

In the first experiment, only MFCC based acoustic models were trained. The recognition results in Word Error Rate (WER) can be seen in the first column of Table 2. Moreover, we also trained the BN features using only target language data. The recognition performance on the given language is shown in the diagonal of the next three columns of the table. It can be seen, the short-time BN features improved the MFCC baseline by about 10% relative, and it is consistent with the results reported in the literature [15]. Besides that a significant real-time factor improvement was also observed.

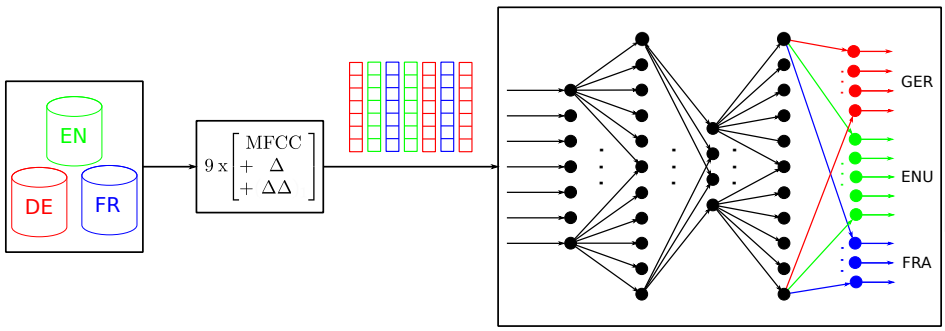

Fig. 1. *The joint training of bottleneck MLP on multiple languages (GER, ENU, FRA). The different colors indicate different languages, and language dependent back-propagation from the output layer. The other parts of the network including the bottleneck layer are shared between the languages.*

Table 2. *Baseline MFCC results in Word Error Rate (WER) are compared with the performance of the target and cross-lingual bottleneck (MFCC+BN) features. The relative improvements over the MFCC system of the target language are indicated in round brackets.*

WER [%]		MFCC	MFCC+BN Bottleneck trained on GER	ENU	FRA
Test language	GER	29.97	*27.50* (8.2)	29.63 (1.1)	30.38 (-1.4)
	ENU	21.69	21.31 (1.8)	*18.85* (13.1)	22.63 (-4.3)
	FRA	37.78	37.76 (0.1)	38.72 (-2.5)	*33.95* (10.1)

5.2. Cross-lingual portability of mono-lingual BN features

In the second experiment, the cross lingual portability of the MLPs trained in the previous experiment was investigated. Comparing the off-diagonal entries of Table 2 to the first column, we observed only a slight maximal 2% relative improvements compared to MFCC alone. There exists cross-lingual portability between German and English to a certain extent (1-2% relative), but using French BN features for the remaining two languages or BN features from other languages on the French task shows WER increase. As a summary, the cross-lingual portability of BN could help, but the performance remained far behind that was achieved by using target language data to train the BN. Our short-time BN features are much simpler as the long-term features applied in [6], thus our observation is similar to [7], where short-time MLP features did not lead to performance improvement without additional weight adaptation.

5.3. Results with multilingual BN features

In the third experiment, we investigated the multilingual BN features trained according to Section 3. In the first tests the multilingual BNs were trained on two languages other than the target one. E.g. BN features trained on US English and French were tested in German ASR experiments. The results are presented in the first column of Table 3. Although the cross-lingual French BN deteriorated the recognition performance, the multilingual training on the merged French and English data improved the German system more than 5% relative. The improvement does not reach the target language BN performance, but clearly – 4% relative – outperformed the best results of cross-lingual BN. Similar observations can be made on English and on French using German+French or German+English multilingual BN features respectively. The results indicate that through the multilingual training the BN features capture more language-independent representation of the speech, and are better suited for cross-lingual porting to new languages.

In the next experiment, multilingual bottleneck features were trained using target language data with other languages. The results can be seen in the 2nd-4th columns of Table 3. It is encouraging to see that adding additional data from a non-target language further improved the performance. To obtain common BN features for the three languages, we also trained a network on all the 450 hours of data. Remarkably, this single net outperformed all the above results in Table 2.

Experimental results indicate that multilingual BN feature estimation is superior compared to the monolingual case despite possible differences in the type of cars and noise conditions specific to the country (and therefore language). Since in our experiments, we used only about 150 hours of data per language, we attribute this to availability of larger training data. Therefore, if more data were available in individual languages, the trend could be different.

To investigate the effect of language dependent softmax and back-propagation, BN features using a unified phoneme set as in [11] were also tested. On German task this BN showed 27.57% WER which is 2.5% relative worse than the proposed multilingual training. In order to have a better understanding of the multilingual BN features and the effect of the amount of data, the previous experiment was repeated with a multilingual BN trained on one third (chosen randomly) of the merged corpora resulting in about the same amount of speech data from each language. This multilingual BN achieved 27.90% on the German task, which is slightly worse than using same amount of source language speech. The previous results prove the effectiveness of the multilingual training, and underline the importance of target language data.

The results so far are obtained using an GMM-HMM system trained using the ML criterion. Table 4 shows WERs of the discriminatively trained German GMM-HMM systems. It can be seen that the gain we observed previously with ML models are not diminished by MCE. Again, the multilingual BN achieved the best performance outperforming the BN trained on target language data only. The BN trained on French and English (without seeing any German

Table 3. *Recognition results achieved with multilingual BN features. The relative improvements over the MFCC (in Table 2) are indicated in round brackets.*

WER [%]		MFCC+BN			
		BN trained on			
Test language	GER	ENU+FRA	GER+FRA	GER+ENU	GER+ENU+FRA
		28.37	27.06	26.89	26.90
		(5.3)	(9.7)	(10.3)	(10.2)
	ENU	GER+FRA	ENU+FRA	ENU+GER	GER+ENU+FRA
		20.29	18.21	17.99	17.89
		(6.5)	(16.0)	(17.1)	(17.5)
	FRA	GER+ENU	FRA+GER	FRA+ENU	GER+ENU+FRA
		35.88	33.52	33.45	33.61
		(5.0)	(11.3)	(11.5)	(11.0)

data) improved the MFCC system more than 7% relative, whereas the monolingual English BN hardly resulted in better performance than baseline MFCC.

Table 4. *Recognition results after discriminative training of GMM-HMM on the German task*

Features			WER [%]	rel.imp [%]
MFCC			29.10	-
MFCC+BN	BN trained on	GER	26.40	9.3
		ENU	28.78	1.1
		ENU+FRA	27.06	7.0
		GER+ENU	25.68	11.8
		GER+ENU+FRA	25.61	12.0

5.4. Multilingual BN in mismatched acoustical conditions

The BN features were also investigated in an experiment where it is assumed that only acoustically *mismatched* training data is available on the target language. However, *matched* data from other languages is available, and the multilingual MLP is applied to take advantage of them. Column 1 in Table 5 shows the WERs obtained by training GMM-HMM acoustic models using baseline MFCC features on *mismatched* data. Comparing to the results in Table 2, the baselines become 15% worse because of the acoustical difference between training and test recordings. Concatenating MFCC with BN trained on the target language showed less improvement than in the matched case (2nd column). In this special ASR experiment, using non-target monolingual BN (3rd column) led to more improvement than previously, since it had seen *matched* data, but in another language. Moreover, porting acoustically matched knowledge from two other languages through multilingual BN improved the results further. However, as the last column of Table 5 shows, the best results were achieved when the mismatched data available in the target language and all matched data from other languages was used to train the BN. In this case, the amount of target language data is less than 1/5 during BN training. The final systems achieved comparable results as the MFCC system trained on matched data (Table 2).

Table 5. *Baseline (MFCC), cross-, and multilingual results using only* mismatched *data in the test language. Bold font indicates the availability of both* matched *and* mismatched *data in the language*

WER [%]	MFCC	MFCC+BN BN trained on			
Test language	GER	34.58	GER	**ENU**	**ENU** + **FRA**	GER + **ENU**+**FRA**
			33.39	34.07	32.74	31.72
			(3.4)	(1.5)	(5.3)	(8.3)
	ENU	26.14	ENU	**GER**	**GER** + **FRA**	**GER** + ENU+**FRA**
			23.54	24.81	23.68	21.79
			(9.9)	(5.1)	(9.4)	(16.6)
	FRA	43.52	FRA	**GER**	**GER** + **ENU**	**GER** + **ENU**+FRA
			40.51	43.65	41.96	39.98
			(6.9)	(-0.3)	(3.6)	(8.1)

5.5. Discussion

Although the experiments were designed to have similar acoustic conditions for all languages, there is a slight driving condition and car noise characteristic mismatch between them. Consequently, the neural networks in the multilingual experiments were trained not only on more languages but also on more types of noises, which contribute to better generalization. However, as the results on cross- and multilingual porting of BN for another languages showed, the improvements increased only slightly even in completely mismatched training and testing conditions. This could also indicate that the improvement is mainly related to the better cross-language generalization property of multilingual MLP.

Since our research was limited to three languages and phoneme sets, as a future direction, we intend to carry out experiments with more languages and corresponding MLP output targets.

6. CONCLUSIONS

A recently introduced multilingual MLP training was extensively evaluated within the bottleneck TANDEM framework. Applying the multilingual technique for bottleneck MLP to extract more language independent features, it was experimentally shown that the multilingual BN features offered better cross-lingual portability. Moreover, we also showed, that through the multilingual approach a single BN net can be trained for three languages, and in all cases it outperformed the BN features trained only on target language data. Finally, the multilingual BN was successfully applied to reduce the mismatch between training and testing acoustical conditions reusing matched data from other languages.

Acknowledgement
This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. [213850]. 11, Speech Communication with Adaptive Learning - SCALE.

7. REFERENCES

[1] M. Sundermeyer *et al.*, "The RWTH 2010 QUAERO ASR Evaluation System for English, French, and German," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 2212–2215.

[2] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[3] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

[4] F. Grézl *et al.*, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 757–760.

[5] A. Stolcke *et al.*, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2006, pp. 321–324.

[6] C. Plahl *et al.*, "Cross-lingual Portability of Chinese and English Neural Network Features for French and German LVCSR," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 371–376.

[7] L. Tóth *et al.*, "Cross-lingual Portability of MLP-Based Tandem Features–A Case Study for English and Hungarian," in *Proc. of Interspeech*, 2008, pp. 2695–2698.

[8] S. Thomas *et al.*, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in *Proc. of Interspeech*, 2010, pp. 877–880.

[9] D. Imseng *et al.*, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

[10] N. T. Vu *et al.*, "An Investigation on Initialization Schemes for Multilayer Perceptron Training Using Multilingual Data and Their Effect on ASR Performance," in *Proc. of Interspeech*, 2012.

[11] F. Grézl *et al.*, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 359–364.

[12] Y. Qian *et al.*, "Strategies for using MLP based features with limited target-language training data," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 354–358.

[13] S. Scanzio *et al.*, "On the Use of a Multilingual Neural Network Front-End," in *Proc. of Interspeech*, 2008, pp. 2711–2714.

[14] S. Thomas *et al.*, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 4269–4272.

[15] F. Valente *et al.*, "Analysis and Comparison of Recent MLP Features for LVCSR Systems," in *Proc. of Interspeech*, 2011, pp. 1245–1248.