FURTHER INVESTIGATION INTO MULTILINGUAL TRAINING AND ADAPTATION OF STACKED BOTTLE-NECK NEURAL NETWORK STRUCTURE

František Grézl, Ekaterina Egorova and Martin Karafiát

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
{grezl,karafiat}@fit.vutbr.cz, xegoro00@stud.fit.vutbr.cz

ABSTRACT

Multilingual training of neural networks for ASR is widely studied these days. It has been shown that languages with little training data can benefit largely from multilingual resources. We have evaluated possible ways of adaptation of multilingual stacked bottle-neck hierarchy to target domain. This paper extends our latest work and focuses on the impact certain aspects have on the performance of an adapted neural network feature extractor. First, the performance of adapted multilingual networks preliminarily trained on different languages is studied. Next, the effect of different target units – phonemes vs. triphone states – used for multilingual NN training is evaluated. Then the impact of an increasing number of languages used for multilingual NN training is investigated. Here the condition of constant amount of data is added to separately control the influence of larger language variability and larger amount of data. The effect of adding languages from a different domain is also evaluated. Finally a study is performed where a language with the phonetic structure similar to the target's one is added to multilingual training data.

Index Terms— multilingual training, neural networks, stacked bottle-neck, neural network adaptation

1. INTRODUCTION

One of the challenges in speech recognition community is to build an ASR system with limited in-domain data. Thus the data hungry speech recognition training algorithms have to be modified to handle such limits. This also applies to neural networks (NNs) which are part of essentially any state-of-the-art ASR system today. They serve either as features extractors or as acoustic models estimating probabilities of sub-phoneme classes. NNs have to be trained on a large amount of in-domain data in order to perform well. Today's deep neural networks are largely under-trained from the perspective of the research done on tasks with rich resources [1]. The need for more training data can be alleviated by layer-wise training [2] or unsupervised pre-training [3]. Another techniques such as dropout [4] and maxout [5] effectively reduce the number of parameters in the neural network.

This work supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. The work was also supported by the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. M. Karafiat was supported by Grant Agency of the Czech Republic post-doctoral project No. P202/12/P604.

But even these techniques cannot cope with limited data from target domain. Leveraging the out-of-domain data is unavoidable in order to improve the performance behind the limits posed by the amount of data. Porting NN from language with rich resources to target language is possible when NN is used as feature extractor [6, 7]. But even then the different phonetic structure of target language limits the performance. To use such network in hybrid scheme one needs to solve problem with missing phonemes. Vu et al. [8] suggest to approximation of such phonemes by several phonemes from the source languages that, in combination, have the characteristic of the target phoneme.

The adaptation technique that eliminates the necessity of phoneme mapping was proposed in [9]. It removes the final layer of multilingual NN and replaces it by a random one which is then trained on target data. This technique can improve the performance of multilingual system. In [10] it was evaluated on two languages when other four were available for multilingual training. Our recent work [11] presents several strategies of adaptation of Stacked Bottle-Neck (SBN) (originally called "Universal Context") NN hierarchy [12] and thoroughly evaluated in [13] using five target languages and two language sets for multilingual training. Similar NN topology is used in [14], but the adaptation is not properly described and the evaluation is done on one language only.

This work is the extension of the [11, 13]. It addresses several points when doing the adaptation of multilingual training:

- Stability of the results with respect to the randomness in NN training.
- Differences in adapted system performance as a result of different languages used for multilingual NN training.
- Dependence of adapted system performance on units used as targets for multilingual NN training.
- Adapted system performance with respect to the number of languages used for multilingual NN training. Here we also evaluated the role of amount of data for training.
- Effect of adding several out-of-domain languages.
- Effect of adding "close" language to multilingual NN training set.

We have done also phonetic analysis which brings further insight into the multilingual training and adaptation procedure problem.

2. EXPERIMENTAL SETUP

2.1. Data

The IARPA Babel Program data simulates a case of what one could collect in limited time from a completely new language. Two training scenarios are defined for each language – Full Language Pack (FLP), where all collected data are available for training; and Limited Language Pack (LLP) consisting only of one tenth of FLP. Vocabulary and language model (LM) training data are defined with

Table 1. *Statistics of the data. The LM and dictionary statistics are taken from LLP which is used to train the HMM system. The OOV rate is reported with respect to LLP.*

Language	AS	BE	HA	LA	ZU
LLP hours	7.8	8.9	7.9	8.1	8.4
LM sentences	11814	11763	9861	11577	10644
LM words	75610	84334	93131	93328	60832
dictionary	8729	9497	5333	3856	14962
# tied states	1179	1310	1257	1453	1379
dev hours	6.4	6.9	7.4	6.6	7.4
# words	51931	56221	81087	81661	50053
OOV rate [%]	8.3	8.5	4.1	1.8	22.4
baseline WER	68.5	69.7	65.9	63.6	74.2

respect to the Language Pack. They basically consists of transcripts of the given data pack.

The following language collection releases were used in this work: Cantonese IARPA-babel101-v0.4c (CA), Pashto IARPA-babel104b-v0.4aY (PA), Turkish IARPA-babel105-v0.6 (TU), Tagalog IARPA-babel106-v0.2g (TA), Vietnamese IARPA-babel107b-v0.7 (VI), Assamese IARPA-babel102b-v0.5a (AS), Bengali IARPA-babel103b-v0.4b (BE), Haitian Creole IARPA-babel201b-v0.2b (HA), Lao IARPA-babel203b-v3.1a (LA) and Zulu IARPA-babel206b-v0.1e (ZU).

The characteristics of the languages can be found in [15]. The FLP data of IARPA-babel10* (CA, PA, TU, TA, VI, AS, BE) languages are used for multilingual NN training. The LLP data are used for NN adaptation and for training of GMM-HMM system. The Haitian, Lao and Zulu are considered as target languages in majority of presented experiments.

Haitian is a French Creole language spoken on Haitian. It is based mainly on French, but is also influenced by other European languages, such as Spanish and Portuguese, and West African languages. The phoneme set is relatively simple, with just 32 phonemes, all of them typical to the aforementioned European languages.

Lao is a tonal language from the Tai-Kadai family, which is spoken in Laos and also in parts of Thailand. With the total of 132 phonemes, Lao has a very complicated vowel system. Apart from tones, vowels are also distinguished according to their length. Moreover, there are three diphthongs. As for consonants, some of them can be aspirated.

Zulu is spoken in South Africa and belongs to the Niger-Congo language family. The phonetic set used in our data consists of 66 phonemes and differentiates between stressed and unstressed vowels and voiced consonants. Apart from this, vowel system is quite simple, whereas consonants pose some problems for multilingual training, as Zulu has clicks, and they are unique for our set of languages. Moreover, Zulu shows a wide variety of non-pulmonic consonants and also have aspiration.

Assamese and Bengali are used as target languages in Sec. 3.6 and are presented in more detail there. Statistics for target languages are given in Tab. 1. The amounts of data refer to the speech segments after dropping the long portions of silence.

2.2. NNs for feature extraction

The NN feature extraction is exactly the same as in [11]. Please refer to it for more details.

The NN input features are composed of logarithmized outputs of 24 Mel-scaled filters applied on squared FFT magnitudes (CRBE) and 10 F0-related coefficients . The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames of *CRBE+F0s* are stacked together. Hamming window followed by DCT consisting of 0th to 5th base are applied on the time trajectory of each parameter. The whole data set is mean and variance normalized.

A structure of two 6-layer NNs is employed according to [12]. The first stage NN in Stacked Bottle-Neck (SBN) hierarchy has four hidden layers. The 1^{st}, 2^{nd} and 4^{th} layers have 1500 units with sigmoid activation function. The 3^{rd} is the BN layer having 80 units with linear activation function. The BN layer outputs are stacked (hence Stacked Bottle-Neck) over 21 frames and downsampled by factor of five before entering the second stage NN. The second stage NN is the same as the first one with exception of BN layer size. In this NN, it has 30 units. Outputs of the second stage NN BN layer are the final outputs forming the BN features for GMM-HMM recognition system.

The forced alignments were generated with provided segmentations. Resegmentation stripping off long silence parts was done afterwards. Tied triphone states are used as NN targets.

2.3. Recognition system

The evaluation system is based on BN features only and thus directly reflects the changes in neural networks we made. The BN features are BN outputs transformed by Maximum Likelihood Linear Transform (MLLT), which considers HMM states as classes. The models are trained by single-pass retraining from an HLDA-PLP initial system. 12 Gaussian components per state were found to be sufficient for MLLT-BN features. 12 maximum likelihood iterations are done to settle HMMs in the BN feature space.

The final word transcriptions are decoded using 3gram LM trained only on the transcriptions of LLP training data – this is coherent to BABEL rules, where *the provided data only* can be used for system training.

2.4. Multilingual SBN training and adaptation

The multilingual NN in SBN system are trained with the last layer – softmax – split into several blocks. Each block accommodates training targets from one language. This was found superior to having NNs with one softmax accommodating either full or compacted target set [16]. The context-independent phoneme states were used as training targets in multilingual NN training.

Trained NN is adapted to target language in two steps:
1. *Training of the last layer.* The last layer of multilingual NN is dropped and a new one is initialized randomly with number of outputs given by the number of tied states in the target language. Only this layer is trained keeping the rest of the NN fixed.
2. *Retraining of the whole NN.* The remaining layers are released and the whole NN is retrained. The starting learning rate for this phase is set to one tenth of the usual value.

Two adaptation schemes which provided best results in our previous work [11] are considered here:
adapt-LLP scheme adapts first NN in the hierarchy while the second one is completely trained on the LLP data of the target language from the random initialization.
adapt-adapt scheme adapts both NNs in SBN hierarchy. Adapting the first NN basically changes the inputs to the second one so it could have problems with adaptation. But it appears that NN can adapt also to slight changes in input features.

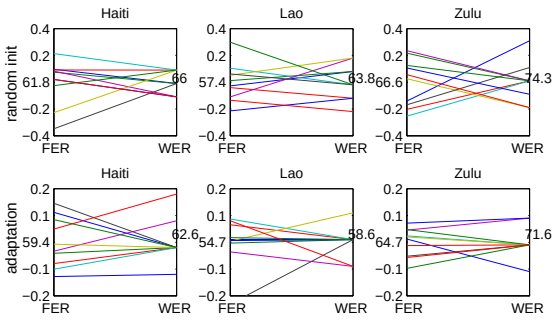

Fig. 1. FER and WER [%] for ten repeated runs of training and adaptation of NNs for individual languages.

3. EXPERIMENTS

3.1. Stability analysis

One of the first questions which arises when comparing results from different experiments is, whether the results obtained from two systems are really different. In case of neural networks, one usually gets different results by repeated training of the same neural network – in terms of input features, structure, training data and targets – due the random initialization of the network. Even if this randomness is eliminated, different results can be obtained due to the different sizes of training batch or differences in the hardware.

We decided to run multiple training of SBN hierarchy with random initialization to see the spread in recognition results. The evaluation was done for two scenarios – complete training of SBN on LLP data, where the randomness affects the whole NN, and the adaptation from multilingual network, where there is random initialization of the last layer. Although the effect of random initialization should be much smaller in the second case, we would like to evaluate it as the adaptation from multilingual NNs is the working mode for the remainder of this paper. Here we evaluated the most effective *adapt-adapt* adaptation scheme where both NNs are initialized by multilingual NN and adapted to the target language domain.

Fig. 1 shows the frame error rate (FER) and WER on ten runs of the same training setup with random initialization of NN weights and with adaptation of multilingual NN. The FER refers to the second NN in SBN hierarchy and each line connects FER and WER belonging to each other. The numbers on y-axis are average values to which we normalized both, FER and WER. The differences are in decimals of percents and are same for both y-axis. The following observations can be made:

- Lower FER does not necessarily lead to lower WER.
- The spread of WER depends on language.
- As assumed, the spread of WER after adaptation is smaller than the spread of WER of SBN systems trained on LLP data only – random initialization of whole NNs.

The results provide us a guidance in judging further results – the difference of 0.2% can be obtained just by different random initialization of the last layer in SBN adaptation.

3.2. Languages from multilingual SBN training

Here, the effect of language selection for training of multilingual NN training is evaluated on the performance of adapted system. Six combinations of five languages were generated from the IARPA-babel10* languages – CA, AS, BE, PA, TU, TA, VI. The ensembles are given in the upper part of Tab. 2. The middle part of the table shows how well various phonetic phenomena are represented in different ensembles. In the rows there are different phonetic entities that are represented in the phonetic structure of a language. For example, a language may or may not have tones or may or may not differentiate between long and short vowels. The individual cells show how many of the languages in a given ensemble have this particular phonetic characteristic. For example, set **a** contains one language with tones. Intuition suggests that the bigger is the number of languages that possess a certain characteristic, the bigger the chance that the NN will learn it and work better when adapted to a target language which possesses the same phonetic characteristic.

Table 2. *Language combinations for multilingual SBN training and number of languages in given combination having specific phonetic characteristic.*

combination	**a**	**b**	**c**	**d**	**e**	**f**
languages	CA AS BE PA TA	CA BE PA TU VI	AS BE PA TU TA	AS BE TU TA VI	CA PA TU TA VI	CA AS BE TU VI
tones	1	2	0	1	2	2
stresses	2	1	2	1	1	0
long vowels	2	4	2	2	4	3
diphtongs	3	1	2	2	2	2
nasalized vowels	2	1	2	2	0	2
non-pulmonic consonants	0	1	0	1	1	1
aspirated conso.	2	1	2	2	0	2
total phonemes	404	417	287	317	452	392
Haitian WER [%]	62.6	62.0	62.0	62.2	62.0	62.4
Lao WER [%]	59.4	58.4	59.9	59.1	58.3	58.7
Zulu WER [%]	71.9	71.3	72.0	72.0	71.2	71.8

Comparing the experimental results in the lower part of Tab. 2 with the numbers in the middle helps to understand which phonetic characteristics are important for training and which do not affect the result. The strongest correlation can be seen between the results and the presence of languages with tones in a multilingual NN which is then adapted for a language with tones. The same applies for the long/short opposition of the vowels. The settings with 2 tonal languages and 4 languages with long/short vowel opposition work best on Lao, which possesses both characteristics and has a well-developed vowel system (108 vowels). Experiments on both Lao and Zulu show that pre-training on languages with aspirated consonants is not important for test languages that also have aspiration. For Zulu, for example, having information about stresses on vowels and voiced consonants overrides the information about aspiration. There is also a strong correlation between the final accuracy and the total number of output phonemes in a multilingual NN: the bigger it is, the better the result. Presence or absence of other phonetic characteristics seems to have no apparent effect on the final results.

Over all, the differences in results are between 0.6% and 1.6% WER absolute. The smallest difference is seen on Haitian, the largest on Lao. The same can be said about the phonetic complexity of the two languages – Haitian has the smallest phonetic complexity and Lao the largest.

Thus it is obvious that languages with smaller phonetic complexity are less sensitive to language selection for multilingual training of NN to be adapted to this language. This also corresponds with the fact that phonetic characteristic of a small complexity language can be more widely covered by given set of languages than it is possible for high complexity language.

Table 3. *Output layer size, number of parameters and training time needed for one epoch of 2^{nd} stage NN training of multilingual SBN.*

targets	monophone states	triphone states
3 languages		
output layer size	813	14064
parameters in SBN	8.4 M	48.1 M
time for 1 epoch	3.5 h	10.0 h
5 languages		
output layer size	1368	25270
parameters in SBN	10.1 M	81.8 M
time for 1 epoch	7.0 h	30.0 h

3.3. Target units for multilingual neural network

This section addresses the discrepancy between targets used for multilingual neural networks and those used for adaptation to target language. The multilingual neural networks are trained towards phoneme state targets and then they are adapted using triphone states as targets. This is done mainly for the practical reasons because the number of triphone state targets gets huge for multilingual training. As the last NN layer gets large, it governs the overall number of parameters. With increasing number of parameters, the NN training time also increases.

The effect of different target units in multilingual SBN on adapted system performance is evaluated using two multilingual SBN. One trained on three languages (CA, PA, TU) and the other on five languages (+ TA, VI; set **e** in Sec. 3.2). On each language set the NNs were trained with both - monophone states and triphone states as targets. Table 3 presents the output layer size, number of training parameters in SBN and training time needed for one epoch of 2^{nd} stage NN training of multilingual SBN. The times are averages as time for particular training epochs vary due to the cluster load (here, mainly the data transfer from storage to computing nodes plays role), but all time are obtained using the same GPU. As can be seen, the size of the NNs and the time needed for training grows enormously.

Figure 2 presents the results. They are drawn in FER-WER axis so it is possible to compare the effects of switching the target units in multilingual NN on both kinds of results. The individual experiments are marked either by stars ($*$) when phoneme states are used as targets for multilingual NN training or by diamonds ($\diamond$) when triphone states are used. Red color indicates use of 3 languages for multilingual NN training, 5 languages are indicated by black color. The lines then connect adaptation experiments where only the kind of targets in multilingual NN training have changed. The solid line connect experiments where *adapt-adapt* adaptation scheme was applied, dotted line connect *adapt-LLP* scheme.

From the plots we can see that using triphone states as targets for multilingual NNs training mostly leads to decrease in FER. But this improvement does not transfer to WER. Conversely, it mostly leads to WER increase. Only exception is the Zulu three-language *adapt-adapt* case, where FER and WER decrease. But there, the phoneme state case has rather poor performance compared with where the other results are located.

Thus we can say that using phoneme states as targets for multilingual NN training is not only satisfactory for obtaining good performance after its adaptation to target domain, but rather preferable. Regardless the increased demand the tied states targets lays on the training, the best results are always obtained by adapting NNs with phoneme states as targets.

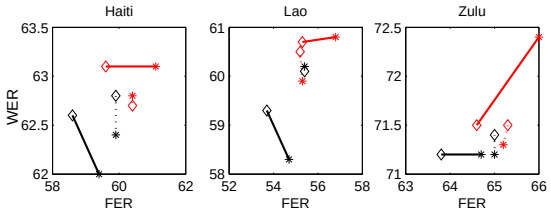

Fig. 2. FER and WER of adapted systems. Black marks refers to multilingual NNs wit 5 languages, red marks with 3 languages. Star ($*$) marks monophone states targets, diamond ($\diamond$) marks tied states targets. Solid line connect *adapt-adapt* and dotted *adapt-LLP* adaptation schemes.

Table 4. *Overview of languages, number of phoneme state targets and amount of data in hours for used data sets.*

data set	languages	# targets	duration
3 lang	CA, PA, TU	813	186.4 h
5 lang	+ TA, VI	1368	304.5 h
7 lang	+ AS, BE	1656	405.0 h

3.4. Number of languages in multilingual training

In this section we investigate the role the number of languages used for multilingual training plays in performance of the adapted system. The intuition says that the adaptation results should improve with increasing number of languages as the number of unique phonemes in the target language decreases. Moreover, increasing number of languages increases also the amount of data used for training. It has been shown in [16] that the training of NNs on large amount of data from different domains can bring better performance than training on small amount of in-domain data. The contrastive experiment thus would limit the amount of data when adding more languages. This would show whether it is more important to have better coverage of phoneme space with rather small amount of data or whether it is better to cover less phonemes with more data.

Again, the 10* languages were used. Data sets consisting of three, five and seven languages were created. The basic characteristic of created data sets are summarized in Tab. 4. For data sets with five and seven languages, reduced versions with the size equal to three-languages set were created. They were created by random segment selection from the full training set. Then, the multilingual networks have been trained on all the sets and adapted to the target language.

The results obtained using multilingual and adapted neural networks are shown in Fig. 3. Solid line connects system results where multilingual NNs were trained using full data sets (Tab. 4) and dashed line connect results obtained with NNs trained on reduced data sets.

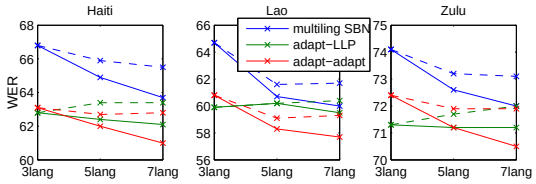

Fig. 3. WER of systems with multilingual and adapted networks. Solid line connects systems where multilingual NNs are trained using all data, dashed line connect systems with data reduced to the size of three-languages set.

First, let us discuss results obtained by individual adaptation techniques on top of multilingual systems trained on different number of languages. As can be seen, the *adapt-LLP* scenario achieves better performance on three-languages multilingual NN, whereas the *adapt-adapt* preforms better when multilingual NNs are trained on more languages.

The difference between the two adaptations is in the approach used for the second NN of the SBN hierarchy. Once it is trained from scratch and once it is adapted from multilingual NN. It seems, that for the adaptation of the second stage NN, it is important to have multilingual NN with good phonetic coverage. Since the ability of the adaptation procedure to shift the weights is limited, it might have problems with introducing new phonemes from the target language. In case a multilingual NN does not have sufficient coverage of phoneme space (for given target language) it is preferable to train the second stage NN on the target data only. The first stage NN seems to be far less vulnerable to this problem. That is because its role is mainly to extract phoneme-related clues from raw features. Clues are then recombined into final features by the second stage NN. The smaller importance of the first NN's ability to well model the acoustic space of target language can be seen from our previous experiments [13], where good performance was achieved when this NN was kept multilingual and only the second one was adapted.

Now, let us compare how the performance changes with increasing number of languages and also in reduced data condition:

• Using full training sets, WER decreases with increasing number of languages. In this case, the phoneme coverage is increased and all phonemes have good coverage in training data. The only exception is the case of 5 language NNs adapted to Lao by *adapt-LLP* scenario. This results are slightly worse than the one obtained with 3 language NNs.

• Using multilingual neural networks trained on reduced set (reduced multilingual NNs) directly for feature extraction, the WER still decreases with increased number of languages. But the decrease is much smaller compare to full training set.

• Adapting the reduced multilingual NNs by *adapt-LLP* scenario increases the WER with increasing number of languages. This suggests that the first stage NN is vulnerable to reduction of data per phoneme class.

• The *adapt-adapt* scenario on top of the reduced multilingual NNs still brings improvement when going from three to five languages, but it does not improve further when going for seven languages. It seems that the adaptation of second stage NN is able to benefit from larger coverage of phoneme space and shift the weights towards actual phones in target language, but the degrading effect of the first stage NN limits further improvements.

3.5. Adding data from different domain

In this section, the effect of adding data from different domain to multilingual training is evaluated. The GlobalPhone database [17] was used. The database consists of about 20 hours of read newspaper speech for each language. The same languages as in our previous work [16] have been used: Czech, German, Portuguese, Russian, Spanish, Turkish, Vietnamese and English taken from Wall Street Journal database. Together, the set has 146.1 hours of training data and 933 phoneme state targets. Please refer to [16] for more details.

The GlobalPhone data were added to the five-languages set – set **e** in Sec. 3.2. Fig. 4 shows results obtained on this combined BABEL+GlobalPhone set and compares them with results obtained on five and seven language set form Sec. 3.4. We can see that additional languages from different domain perform differently on each test set.

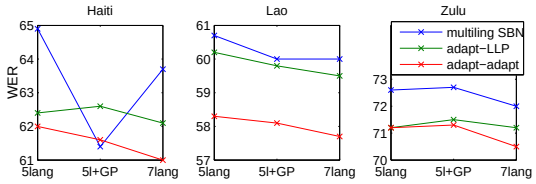

Fig. 4. WER of systems with multilingual and adapted networks. Comparing three and seven languages training set to combined five language BABEL plus GlobalPhone training set

On Haitian, we see huge improvement when the multilingual NN is used directly for feature extraction. But this does not carry on through the adaptation step. From the performance of purely multilingual NN features, it can be said that the GlobalPhone languages are close to Haitian and cover well the Haitian phoneme space. But as the adaptation fails to improve further, it seems that the GlobalPhone acoustic space is too far from the target one. This is apparent from the decrease in performance after the *adapt-LLP* adaptation which fails to find the relation between the GlobalPhone and BABEL acoustics. The *adapt-adapt* scenario can undo some of the damage caused by adaptation of the first stage NN but is still slightly behind the performance of the multilingual SBN.

The GlobalPhone languages seems not to be of particular importance for Lao, as the improvements are only slight and the results are between the one obtained on five- and seven-languages training set.

On Zulu, we see slight degradation when GlobalPhone is added to five-languages training set. It shows that the GlobalPhone data do not bring either acoustic or phonetic information useful for target domain represented by Zulu.

3.6. Adding a language with similar phonetic structure

Finally, we have performed a series of experiments on a pair of languages with similar phonetic structure. This experiment should show how much benefit it is possible to get if there is a possibility to include a language close to the target one to multilingual training set. From the languages provided by BABEL project, such closely related ones are Assamese and Bengali. Both of them are are Eastern Indo-Aryan languages and therefore it is not surprising that their phoneme sets are similar. Their vowel systems both have diphthongs and nasalization, although Assamese has more vowels due to the presence of /U/ and /E/ and their nasalized versions. Consonant sets of both languages are also similar, but not completely alike. Bengali has more retroflex consonants and a more systematic aspiration. Moreover, Bengali has more affricates, which also participate in the aspiration pattern.

The close language (AS/BE) was added to three and five language set used in Sec. 3.4 and the resulting SBN hierarchy was adapted to target language (BE/AS). The performance of the original three and five language multilingual systems adapted to target language was evaluated too. The SBN was trained also on the close language only and then adapted. The performance of the multilingual SBN NNs without adaptation was evaluated as well. The results, together with the with the one obtained on LLP of target language, are shown in Fig. 5.

We see that adapting the SBN trained on close language brings substantial improvement over the LLP baseline. Adaptation of the three language system improves further thanks to the larger coverage of acoustic space, but the phoneme space is not represented well - the *adapt-adapt* scenario is behind the *adapt-LLP*. Adding the close language increases the phoneme coverage and the *adapt-adapt* scenario

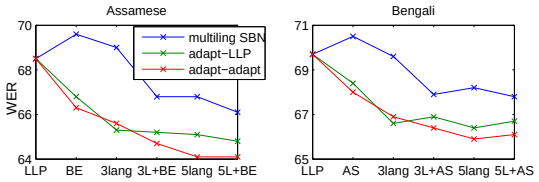

Fig. 5. WER of systems trained on multilingual sets with addition of close language.

improves further. The performance after the *adapt-LLP* adaptation is about the same, which suggests already saturated coverage of acoustic space. Quite large improvement in performance of non-adapted SBN should be also noted. The smaller difference in results from adapted and non-adapted system also suggests quite good match in acoustic and phonology of training and test languages. Further increase of multilingual training set to five and five + close languages is useful for Assamese, which has smaller phoneme inventory and the adaptation is thus able to leverage the increased phoneme inventory. Bengali, with its richer phonology benefits from five language training set, but adding the close language to it does not have further positive effect. This also points to the fact that although both languages have some phonemes missing in the other, Bengali retroflex consonants are quite unique. On the other hand, the vowels present in Assamese and missing in Bengali can be easily pre-trained from other languages.

This results further confirmed our conclusion that the first NN of the SBN hierarchy is extracting acoustic clues which are then combined in phoneme oriented manner by the second stage NN.

4. CONCLUSIONS

This work addresses analysis of different aspect of training multilingual SBN NN structure when adapted to the target language. The results are obtained with two adaptation schemes which performed the best in our previous work on three target languages which differ considerably in phonological complexity.

First, we analyzed the effect of language selection for multilingual NN on the final performance. We have shown, that language set which covers more of the phonetic phenomena in target language gives better performance. Generally, more phoneme classes given language has, more useful it will be in the multilingual training set. If there is possibility, then choosing languages for multilingual training cleverly, e.g. according to their language family, phoneme coverage, etc. can lead to the same or better performance with smaller number of languages then training blindly on all languages.

Next, the difference between monophone and triphone states as targets for multilingual training was evaluated. Although the triphone states are categories used during adaptation to the target language, using them for multilingual training is detrimental.

The effect of number of languages in multilingual training was studied next including the case of constant amount of data in the training set. Results shows, that when amount of data is limited the effect of adding more languages does not have to be always positive.

When adding data from different domain, the acoustic space of original features plays important role as the adaptation may not by able to transfer it to target domain. So even though the phoneme coverage would be very good, the results after adaptation may be behind the expectations, as it happened with Haitian in our case.

Finally, a close language, in terms of phonetic structure, to the target one, was added to the set of languages for multilingual training. The positive effect can be seen mainly for sets with few (three in our case) languages. The improvement diminishes when the close language was added to the 5-languages set.

The experiments have also confirmed that the two stages in SBN system process different information – the first NN works in acoustic space and produces phonetic clues which are then recombined by the second stage NN. Independence of these processing steps can be beneficial as the optimal adaptation strategy can be chosen for given multilingual setup.

5. REFERENCES

[1] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP 1999*, Phoenix, Arizona, USA, Mar. 1999, pp. 1013–1016.

[2] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, 2007, pp. 153–160.

[3] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Mar. 2010.

[4] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[5] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio, "Maxout networks," in *ICML*, 2013.

[6] Stefano Scanzio, Pietro Laface, Luciano Fissore, Roberto Gemello, and Franco Mana, "On the use of a multilingual neural network front-end," in *Proceedings of INTERSPEECH-2008*, 2008, pp. 2711–2714.

[7] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Proc. of ICASSP*, Vancouver, Canada, May 2013, pp. 7349–7353.

[8] N.T. Vu, W. Breiter, F. Metze, and T. Schultz, "An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance," in *Proc. Interspeech*, 2012.

[9] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc of ICASSP*, 2012, pp. 4269–4272.

[10] Z. Tüske, D. Nolden, R. Schluter, and H. Ney, "Multilingual MRASTA features for low-resource keyword search and speech recognition systems," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 5607–5611.

[11] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. of ICASSP*, Florence, Italy, May 2014.

[12] F. Grézl, M. Karafiát, and L. Burget, "Investigation into bottle-neck features for meeting speech recognition," in *Proc. Interspeech 2009*, Sep 2009, pp. 294–2950.

[13] František Grézl and Martin Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proc. of SLTU*, St. Petersburg, Russia, May 2014.

[14] Q.B. Nguyen, J. Gehring, M. Muller, S. Stuker, and A. Waibel, "Multilingual shifting deep bottleneck features for low-resource ASR," in *Proc. of ICASSP*, Florence, Italy, 2014, pp. 5607–5611.

[15] M. Harper, "The BABEL program and low resource speech technology," in *Proc. of ASRU 2013*, Dec 2013.

[16] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottleneck features in multilingual environment," in *Proceedings of ASRU 2011*, 2011, pp. 359–364.

[17] Tanja Schultz, Martin Westphal, and Alex Waibel, "The globalPhone project: Multilingual LVCSR with janus-3," in *in Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, Czech Republic*, 1997, pp. 20–27.