BOOSTING PERFORMANCE ON LOW-RESOURCE LANGUAGES BY STANDARD CORPORA: AN ANALYSIS

František Grézl and Martin Karafiát

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic

{grezl,karafiat}@fit.vutbr.cz

ABSTRACT

In this paper, we analyze the feasibility of using single well-resourced language – English – as a source language for multilingual techniques in context of Stacked Bottle-Neck tandem system. The effect of amount of data and number of tied-states in the source language on performance of ported system is evaluated together with different porting strategies. Generally, increasing data amount and level-of-detail both is positive. A greater effect is observed for increasing number of tied states. The modified neural network structure, shown useful for multilingual porting, was also evaluated with its specific porting procedure. Using original NN structure in combination with modified porting *adapt-adapt* strategy was fount as best. It achieves relative improvement 3.5–8.8% on variety of target languages. These results are comparable with using multilingual NNs pretrained on 7 languages.

Index Terms— DNN topology, Stacked Bottle-neck, feature extraction, multilingual training, system porting, low resource

1. INTRODUCTION

ASR systems are working very well for the main world languages. Many mobile telephone applications are well documenting the maturity of the systems. However, outside the group of about 100 most spoken languages, the speech technology is inaccessible. The main reason for missing reliable ASR system is unavailability of a good transcribed speech databases. Hand in hand with this problem might come complicated phonology or syntax and uneducated population missing phoneticians and linguists specialists able to describe their own language.

All these aspects are becoming or have already become interesting areas for research. Missing phonetic transcription can be overcome by automatically generating phoneme-like units [1, 2]. Missing notion about words can be bridged by translation to well-described language [3]. This usually applies when a regional language is gradually replaced by the dominating one (official language of the country or former colonial language). These two areas are just attracting interest and so far, they are bringing more questions than answers.

On the other hand, researchers have dealt with limited resources for building ASR system for some time already. Subspace Gaussian mixture model [4, 5] can efficiently leverage data from other sources to build model for target language with very little speech data. As the neural networks (NNs) took over the Gaussian models, techniques dealing with insufficient training data had to be developed too. The layer-wise training [6] of neural network limits the amount of trainable parameters at a time. The trainable parameters are also reduced by the drop-out technique [7] which also prevents co-training of neurons.

To prevent over-training of neural networks, the above mentioned drop-out technique can be used. It was also successfully combined with the maxout neuron structure [8]. The over-training can be also prevented by introducing a regularization term into the objective function [9, 10].

The problem of insufficient data can be also eased by so called data augmentation [11, 12], where modified versions of the original recordings are generated. This can be achieved by artificial (de)noising and/or (de)reverberating of the original signal [13] or by modifying the features by, for example, vocal tract length perturbation [14].

The data from other sources can be used for unsupervised pre-training [15] or for multilingual NN training and subsequent porting to target language [16, 17, 18, 19]. The multilingual training and porting attracted attention namely with IARPA BABEL program with many more publication on these topics. It has been the most efficient technique developed in the context of low resource ASR system training.

The drawback of multilingual training and porting is the need of language collection usable for multilingual training. Although there are databases for many languages, they significantly differ in amount of data, quality of audio and transcription. The IARPA BABEL project is unique in a way researchers were provided with more-or-less homogeneous databases of 23 languages. However, this collection is not publicly available and to put together a similar collection from public resources would be an enormous effort.

1.1. Goals of the paper

In this paper we would like to provide guidelines for researchers working with low resource languages who does not have access to multilingual corpora. Instead, a large collection of single language – English Fisher – will be used as source language for multilingual training. We want to show that training the NN on a single (and distinct) language and porting it to target language can bring improvement in final system performance. This analysis will also show how

This work was 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 Technology Agency of the Czech Republic project No. TA04011311 "MINT" and Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

to handle a single language for multilingual training to get maximum benefit from it.

The hierarchy of two NNs known as Stacked Battle Neck (SBN) [20] is used in this analysis. This structure exhibits high performance in place of feature extractor as well as acoustic model and was adopted by other researchers [21, 16, 22, 17].

Revisiting the multilingual porting procedures [23] for single source language scenario will reveal optimal porting strategy. The effect of the amount of source language training data on ported system performance will be shown. The impact of different "detailness" (number of triphones) of source language alignment for NN training will also be evaluated. The latest findings on optimal multilingual NN structures [24] will be re-evaluated for single language case.

The evaluation will be done on five languages from IARPA BABEL program. The selected languages belong to different language families thus providing a representative sample.

2. DATA

The source language is English taken from the Fisher database[1]. English belongs to Germanic branch of the Indo-European language family. In our setup, 39 phonemes (26 consonants, 13 vowels including diphthongs) are used.

The forced alignment of the data was done using simple PLP-based GMM-HMM system. The features are formed by HLDA transform of 13 PLP coefficients with their first, second and third order derivatives. The resulting feature vector has 39 coefficients. The conversation side mean and variance normalization followed.

The model was trained from scratch using mix-up maximum likelihood training. Three-state cross-word tied-states triphones models were used, each state has 18 Gaussian mixture components. The model was trained on randomly selected 1000 hours and it has 9824 tied triphone states.

After forced alignment, the segmentation was changed so that each segment has a maximum 150 ms of silence on the each end and pause between two speech parts cannot be longer than 300 ms. If longer silence region occurred, the segment was split in two. The resulting segmentation contains 1710 hours of audio data with approximately 13% of silence. Doing the re-segmentation to reduce the amount of silence turned out to be beneficial for NN training and has a non-negligible effect on final system performance.

The languages selected as target ones are the following BABEL languages:

Telugu – TE – IARPA-babel303b-v1.0a – is a Dravidian language spoken in the south-eastern part of India. Telugu phoneme set used for the experiments contains 39 phonemes, vowels showing long/short dichotomy and containing two diphthongs. Consonant set contains quite a few retroflex phonemes.

Lithuanian – LI – IARPA-babel304b-v1.0b – language belongs to the family of Baltic languages, and the phoneme set used for the experiments consists of 110 phonemes. On vowels and voiced consonants, it contains markings of stress and of falling or rising tone where applicable. Apart from that, vowels have long and short versions. Nearly every consonant in the Lithuanian consonant set has two versions: palatalized and non-palatalized.

Haitian Creole – HA – IARPA-babel201b-v0.2b – a French Creole language spoken in Haiti. It is based mainly on French, which belongs to Romance branch of the Indo-European family, 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.

[1]Fisher 1,2; LDC2004S13, LDC2005S13 for speech data; LDC2004T19, LDC2005T19 for transcripts

Language	TE	LI	HA	LA	ZU
LLP hours	8.6	9.6	7.9	8.1	8.4
LM sentences	11935	10743	9861	11577	10644
LM words	68175	83157	93131	93328	60832
dictionary	14505	12722	5333	3856	14962
# phonemes	39	110	32	132	66
# tied states	1370	1763	1257	1453	1379
dev hours	7.8	8.1	7.4	6.6	7.4
# words	59340	77790	81087	81661	50053
OOV rate [%]	16.1	11.4	4.1	1.8	22.4
baseline WER	78.7	60.3	65.9	63.6	74.2

Table 1. *Statistics of the data for target languages.*

Lao – LA – IARPA-babel203b-v3.1a – 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 – ZU – IARPA-babel206b-v0.1e – a South Africa language belonging 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, the 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 has aspiration.

For training, the defined Limited Language Pack (LLP) was used. This means that the dictionary contains only the words appearing in the training part and the data for language model consists only of training data transcription. Dictionaries coming with the language pack were used. The forced alignments were created in the same way as for English data including the silence handling. Statistics for target languages are given in Tab. 2. The baseline system (described below) results, where all parts are trained on target language data only, are on the last line of the table.

3. SYSTEM DESCRIPTION

The evaluation system is a tandem system where the features for the final GMM-HMM classifier are the Bottle-Neck (BN) features obtained by Stacked Bottle-Neck (SBN) Neural Network hierarchy.

3.1. SBN neural network hierarchy

The SBN is a two-stage structure of 6-layer NNs as described in [20]. Both NNs have Bottle-Neck layer with linear activation function as the 3^{rd} hidden layer. The first stage NN has 80 units in its BN layer the second stage NN uses 30 units. The 1^{st}, 2^{nd} and 4^{th} hidden layers have 1500 units with sigmoid activation function.

The BN layer outputs of the first stage NN are stacked (hence Stacked Bottle-Neck) over 21 frames and downsampled by factor of five before entering the second stage NN.

The NN input features are composed of critical band energy (CRBE) and fundamental frequency features. As critical band energy features, we use logarithmized outputs of 24 Mel-scaled filters

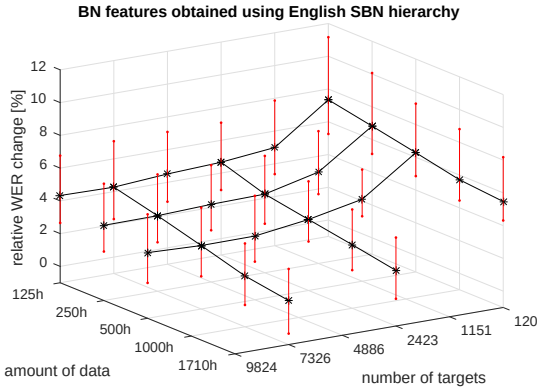

Fig. 1. Average relative WER change over five target languages as a function of training data amount and number of NN targets. The reference are language specific BN features. Red bars connect minimum and maximum for given setting.

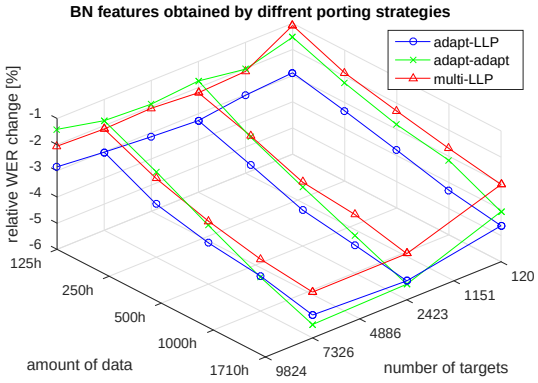

Fig. 2. Comparison of different porting strategies. Average relative WER change over five target languages as a function of training data amount and number of NN targets. The reference are language specific BN features.

applied on squared FFT magnitudes. The fundamental frequency features consist of F0 and probability of voicing estimated according to [25] and smoothed by dynamic programming, F0 estimates obtained by Snack tool[2] function *getf0* and seven coefficients of Fundamental Frequency Variations spectrum [26, 27]. Together, there are 10 F0 related coefficients.

Conversation-side based mean subtraction is applied on the whole feature vector and 11 frames are stacked together. Hamming window followed by DCT consisting of 0th to 5th base are applied on the time trajectory of each parameter resulting in $(24+10)\times 6 = 204$ coefficients on the first stage NN input. Global mean and variance normalization is applied on this 204 coefficients vector.

3.2. GMM-HMM decoder

GMM-HMM acoustic model is a simple maximum-likelihood trained model without any speaker adaptation obtained by single-pass retraining from PLP-based model used for forced alignment. The acoustic features are formed by BN outputs of the SBN hierarchy transformed by Maximum Likelihood Linear Transform (MLLT). For the MLLT computation, each HMM state is considered as class.

The number of Gaussian components per state found sufficient for MLLT-BN features is 12. There are 12 iterations of maximum likelihood training to settle the GMMs in the MLLT-BN feature space.

The final word transcriptions are decoded using 3gram LM trained only on the transcriptions of LLP training data.

4. EXPERIMENTS AND ANALYSIS

One of the goals is to analyze the significance of source language data amount and alignment detailness for ported system performance. For this purpose, subsets with 125, 250, 500, 1000 hours of English training data were created. The selection of data was done randomly on a segment level. To alter the detailness of forced alignment, the decision tree was climbed up to create clusterings with different numbers of states. The clusterings containing 75%, 50%, 25% and 10% of original triphone tied states were created and phoneme state clustering was added. Thus the number of targets for NN training is 9824 for the full triphone tied states, 7326, 4886, 2423, and 1151 for the reduced triphone states and (39+1)x3=120 for phoneme states.

For the training sets with 125, 250 and 500 hours, NNs for each clustering were trained. For larger training sets, the clusterings using phoneme states and 25% and 75% of original triphone tied states were used. This grid should provide solid insight in the implications of the data amount and triphone-state detailness.

4.1. Monolingual SBN features

In case of the multilingual SBN hierarchy, provided multilingual BN features (i.e. NNs without any porting) lead to better performance than the language specific features obtained from NNs trained on the LLP data only [23, 28]. In case of English as the only source language, the chances of "monolingual" BN features being better than the language-specific ones are much smaller. Figure 1 shows relative WER change averaged over all five test languages. The baseline results are those obtained by language-specific SBN shown in Tab.2. The positive change means worse results, the negative change means improvements. The red bars connect the minimal and maximal relative change.

It can be seen that the wort performance is obtained by the minimal setting - 125 hours of training data and phoneme states as targets. The results improve with increasing both the amount of data and the number of NN targets, the second having larger effect. On average, results obtained by features generated using English NNs are worse than from language specific NNs. The worst performing language is Lao with the smallest relative change 3,6%, on the other side are Zulu with only 0.6% relative degradation and Lithuanian with 0.3% relative improvement in the best scenario.

4.2. SBN porting revisited

First, let us shortly review the two-step NN porting procedure:
1. *Training of the last layer.* The last layer of trained 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

[2] www.speech.kth.se/snack/

trained keeping the rest of the NN fixed.
2. *Retraining of the whole NN.* The whole NN is retrained, starting learning rate value is one tenth of the usual one.

The porting of multilingual SBN hierarchy was thoroughly evaluated in [23]. The following approaches are reevaluated for single language case:

adapt-adapt – port the first and also the second NN. This scenario worked the best for multilingual NNs.

adapt-LLP – port the first NN, train the second one on the LLP data. This was the second best scenario. The detailed analysis of the results revealed advantages of this scenario in case the phoneme set of target language is far from phoneme sets of source languages.

multi-LLP – keep the first NN multilingual, train the second one on LLP data only. This scenario gives information about the ability of the first stage NN to extract relevant acoustic information from the inputs for the second stage NN. Since the second NN is trained only on LLP data, the results will directly reflect this ability.

The results in form of averaged relative WER change for all three porting strategies are shown in Fig. 2(only subsed of all combinations reported in Fig. 1). We can see that the average relative WER reduction is between 1% and 5%. The best results are mostly achieved by *adapt-LLP* porting strategy. *Adapt-LLP* strategy ports the first NN from English SBN to target language. This ported NN thus knows about target language acoustics but the exact phoneme classification is not important because the second NN follows. The second NN in this strategy is trained on LLP data of target language only, thus not being tied to any pretrained weights.

The limitation of the second source NN trained on single language is evident from the *adapt-adapt* scenario. Unless there are enough triphone targets trained on enough data, the porting procedure is not able to shift source weights to classify well phonemes of the target language. The "enough" is apriori not known and it is possible that for some target languages, it is not reachable.

The third porting strategy – *multi-LLP* – shows that the first NN trained on a different language can provide better inputs than the one trained on the same target language. The differences between this porting strategy and *Adapt-LLP* one show us how much improvement is reachable by porting the first NN and thus exposing it to target language acoustic. Comparison with *adapt-adapt* for settings with 125 and 250 hours of training data and triphone state targets exposes further the limitation of the second NN porting – here the second NN trained on small data performs better on outputs of the first NN which did not sees any of the target language data.

For the best setting (1710 hours, 9824 targets), the relative WER change per language are shown in Fig. 3. The first thing that can be spotted is that the poor performance with BN features from English SBN hierarchy does not mean poor performance after porting and vice versa. Lao, which has biggest performance degradation by using features from English SBN, achieved the second best relative improvements after SBN porting. On the other hand Zulu, with minor degradation caused by English SBN feature extraction, did not improve much after porting.

The second thing is that the *adapt-adapt* porting scheme is not always the best choice. For Zulu and Telugu, the *adapt-LLP* porting strategy is preferable.

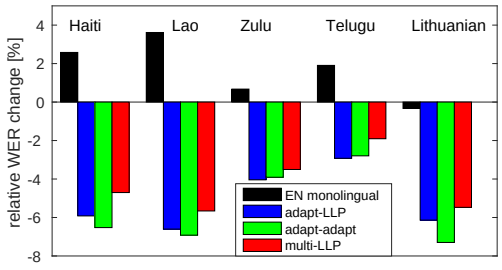

Fig. 3. Comparison of different porting strategies for best performing setting (1710 hours, 9824 targets). The reference are language specific BN features.

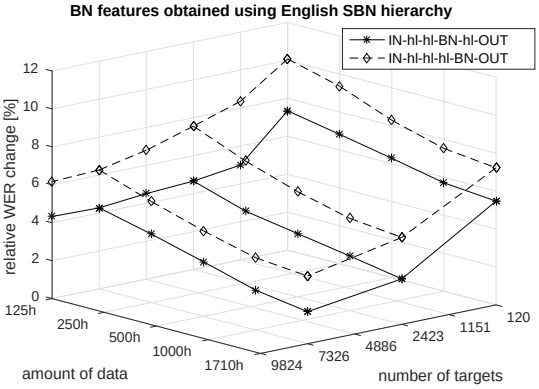

Fig. 4. Average relative WER change over five target languages as a function of training data amount and number of NN targets. The reference are language specific BN features. Results for original (IN-hl-hl-BN-hl-OUT (*2+1*) and modified (IN-hl-hl-hl-BN-OUT) NN structure.

4.3. Changing NN topology

We have recently shown that simple modifying NN topology leads to system improvement [24]. The modification lies in removing the large hidden layer between bottle-neck layer and final softmax layer in the ported NN. This can be achieved by two ways:

1. Removing the layer during NN porting. Having the source NN with hidden layer between bottle-neck and output ones the porting procedure is altered so that all layers after bottle-neck one are removed and direct bottle-neck–to–output layer is initialized. The rest of the porting is unchanged. The NNs after porting will have just two hidden layers between input and bottle-neck one, and then the output layer (*IN-hl-hl-BN-OUT* or *2+0* NN structure).
2. Training source NN with modified topology – direct bottle-neck–to–output layer. Here, the left-out hidden layer can be moved in front of the bottle-neck layer. The porting procedure is unchanged. The source NN and the ported NN will both have three hidden layers between input and bottle-neck one, and then the output layer (*IN-hl-hl-hl-BN-OUT* or *3+0* NN structure).

Both ways have theirs advantages: the first one provides better

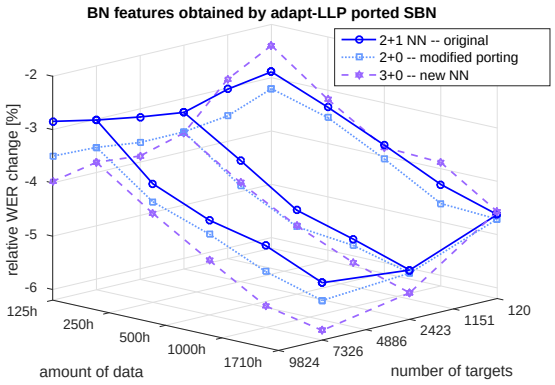

Fig. 5. Average relative WER change over five target languages as a function of training data amount and number of NN targets for *adapt-LLP* porting and different final NN structures. The reference are language specific BN features.

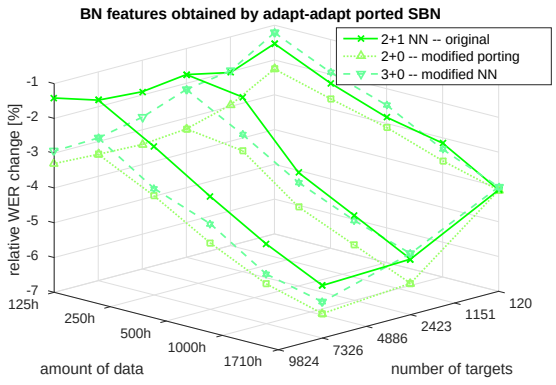

Fig. 6. Average relative WER change over five target languages as a function of training data amount and number of NN targets for *adapt-adapt* porting and different final NN structures. The reference are language specific BN features.

performance when phoneme state targets are used for multilingual training, the second one is necessary when tied-state triphone targets are used as multilingual NN targets.

It is unclear how these two ways of final NN structure modification will work in combination with different porting strategies on a single language source NN. We decided to run two most efficient porting strategies over both modification ways. For this purpose, another set of English SBN hierarchies was trained. In this set, both NNs have three hidden layers between input and bottle-neck one, each with 1500 units. The bottle-neck layer is directly connected to output softmax layer.

First, the performance of BN features obtained from English SBN hierarchies with different versions of NN structures is compared in Fig 4. The averaged relative WER change shows that the NNs with modified structure (denoted as *IN-hl-hl-hl-BN-OUT*) performs much worse than the original structure (*IN-hl-hl-BN-hl-OUT*). The average degradation over all experiments is 2% relative.

The per language behavior of modified NN structure results is similar to the original structure. For the best setting, the worst performing language is Lao with 6.1% relative degradation, the least degrading languages are Zulu and Lithuanian both with 2.4% relative degradation.

Next, the porting strategies are compared for both ways of obtaining the final NN structure with direct bottle-neck–to–output layer connection.

Figure 5 compares the *adapt-LLP* porting strategy over original source NN structure and porting procedure (*2+1* target NN structure), original source NN structure and modified porting (*2+0* target NN structure) and modified source NN structure and original porting (*3+0* target NN structure). It can be seen that the modified porting improves over the original one in all settings. The smallest improvements are seen for low number of targets and large data used for source NN training. The results improve more with increasing number of targets and for decreasing training data. The modified NN structure seems to exhibit different pattern over the source NN amount-of-data – number-of-targets settings. The improvements for smaller amount of data and number of outputs used for training source NNs are smaller than for the original NN structure. However, with increasing the number of source language targets, the ported NN improves well over the original *2+1* target NN.

Since the second NN in the SBN hierarchy is trained on LLP data only and the behavior of BN features from original and modified NN structure from English SBNs are similar (similar shape of the plane over the various settings), the difference in behavior can be ascribed to the porting of the first NN. In case of modified NN porting, it seems that the retraining process finds it easier to move the weights towards target language acoustics. More targets in source language allow easier mapping to new targets. Skipping hidden layer after the bottle-neck one puts the new language targets closer to the bottle-neck, thus influencing directly its outputs.

When the source NN is trained with direct bottle-neck–to–output layer connection, the bottle-neck outputs are directly influenced by the source language targets. New language targets thus need to be closer to some existing targets to be able to efficiently influence the weights during retraining. This can be seen from the smaller improvements achieved for low number of source language targets.

The results obtained from *adapt-adapt* porting scheme are summarized in Fig. 6. The modified porting again improves over the original one. The lowest improvement (only 0.01%) is achieved for source NN being trained with 1710 hours of data and phoneme state targets. The biggest difference, 1.88% is achieved for the smallest training set with full triphone clustering (125 hours, 9824 targets). The observation from *adapt-LLP* porting results thus can be confirmed – removing the hidden layer between bottle-neck and output ones makes influencing the weights by new language acoustic easier. When the modified NN structure is used for source NN training, the porting procedure is not as efficient. Still there is an advantage over the original NN structure and porting, specially for the source NN being trained with more targets. But it does not reach the performance of original source NN with modified porting.

An interesting observation is that the performance of system ported from source NNs trained on full data with phoneme state targets is about the same over different NNs topologies. It shows that the well trained weights from source NN are equally hard to move towards new language acoustic no matter if the output layer is closer or farther from the bottle-neck layer.

The per language results for both porting strategies and all three NN structure combinations are shown in Fig. 7. Comparing *adapt-LLP* and *adapt-adapt* porting strategies on the original (2+1) and modified (3+0) NN topologies, we see that the trends are the same

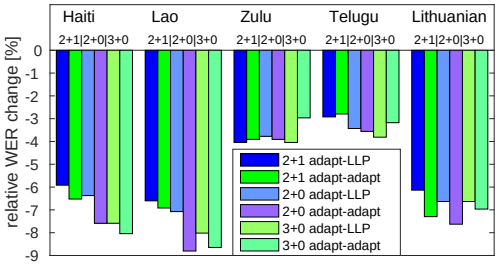

Fig. 7. Comparison of different porting strategies on different NN structure for best performing setting (1710 hours, 9824 targets). The reference are language specific BN features.

– if *adapt-adapt* porting performed better for 2+1 NNs, it performs better also for 3+0 NNs. However, if the trend is opposite, the gap between *adapt-LLP* and *adapt-adapt* porting is bigger for the modified 3+0 NN topology. The difficulty of diverse NN targets in source and target language is amplified when NN with direct bottle-neck–to–output connection is used. Since the second NN does the final classification (even though in hybrid system it "only" provides features for final classification done by GMMs) the errors made here matter much more. In fact, the positive effect from the first NN ported, which can be seen from *adapt-LLP* porting results, is completely lost in attempt to port also the second NN. Thus for very distinct phoneme sets such is in Zulu, the *adapt-adapt* porting on NN with modified structure leads to the worst of ported results.

Comparing the modified porting strategies on the original NN structure, the *adapt-adapt* porting is in all cases better than the *adapt-LLP* one. It is not always the best, but it is never the worst, thus presents optimal porting strategy for unknown language.

5. CONCLUSIONS

In this paper, we have evaluated the multilingual techniques for single source-language scenario. Since it is hard to obtain coherent multilingual corpora usable for multilingual training, using single, well resourced, language instead is quite attractive.

The English Fisher database, containing about 2000 hours of audio, was used for our analysis. It was aligned on the level of context dependent triphone states. Climbing up the decision tree, several sets of tied triphones were created to study the porting efficiency as function of the "detailness" of source language data alignment. To see the effect of the amount of source language training data, several training sets were created as well.

For the majority of number-of-triphones – data-amount pairs, source Stacked Bottle-Neck NN hierarchy was trained and subsequently ported to five languages. We have used the limited language packs of Haiti, Lao, Zulu, Telugu and Lithuanian distributed through the IARPA BABEL program. The ported SBN hierarchy was used to generate BN features for simple GMM-HMM decoder. The performance of ported systems was compared to the SBN trained on the target language data only.

First, the non-ported English NNs were used to generate features for target languages. Opposed to the multilingual SBN, this leads to performance degradation. Next, three porting strategies (*adapt-LLP, adapt-adapt, multi-LLP*), proposed earlier for multilingual SBN, were evaluated. In average, the *adapt-adapt* porting strategy performed the best, but for some really distinct languages, such as Zulu and Telugu, *adapt-LLP* strategy worked better.

Language	TE	LI	HA	LA	ZU
baseline – language specific SBN	78.7	60.3	65.9	63.6	74.2
EN modified *adapt-adapt* porting	75.9	**55.7**	60.9	58.0	71.3
Mult 5L *adapt-adapt* porting	76.2	57.9	62.4	58.7	71.8
Multil 7L *adapt-adapt* porting	75.5	57.3	61.0	57.7	**70.5**
modif Mult 5L *adapt-adapt* porting	75.1	56.2	60.8	57.1	70.8
modif Mult 7L *adapt-adapt* porting	**74.5**	56.0	**59.9**	**56.6**	**70.5**

Table 2. *WER resutls on test languages*

The modified NN structure, which was found to be beneficial for multilingual NNs, was evaluated next. Here, we have two options to achieve the desired NN topology for target language: to modify the porting procedure or to train the source NN with desired structure. Both options were evaluated exhibiting different patterns when ported.

The observed effect of porting individual NNs with different structures is the following:
Porting the first NN is always beneficial. Since it provides features for the second NN in hierarchy, it does not have to do precise classification. Thus the source NN with modified structure can be safely used. Adaptation of the second stage NN is more delicate since it provides features for the HMM. For successful porting, the target language triphones should be close to some source language triphones. The modified porting process, when the hidden layer between bottle-neck and the output one is omitted is easing the retraining to target language. In this case, the error from target layer is propagated directly to bottle-neck layer, affecting directly its outputs. Training the second stage NN on target language only might also be an option, especially for languages with distinct acoustic characteristic such as Zulu.

The effect of increasing amount of source language data and number of triphones on ported language is mostly positive. The general recommendation should be to use number of triphones appropriate for the available training data.

Over all, the modified *adapt-adapt* porting strategy is recommended a as safe option, giving in average the best results for the best performing combination of source data amount and number of triphone states. It was also better than *adapt-LLP* version on all of our test languages. The average relative improvement achieved was 6.2% spanning from 3.5% for Telugu to 8.8% for Lao.

The achieved WERs are given in Tab. 5. The first line shows the baseline results where features are generated using SBN trained on target language data only. Second line results were obtained with features from English NN trained on 1710 hours towards 9824 targets ported using the modified *adapt-adapt* porting (2+0) to target language. The results on the following lines are from [24]: the *Mult 5L* and *Mult 7L* are multilingual networks with original NN structure (2+1) and phoneme state targets. The last two lines, *modif Mult 5L* and *modif Mult 7L* are multilingual networks with modified NN structure (3+0) and tied triphone state targets. All multilingual networks were ported using *adapt-adapt* scheme to the target language.

The results obtained from just one well resourced source language are better than having multilingual system trained on 5 languages, and are close to 7 language system. The modified NN structure trained on 7 languages gives an additional 1% absolute WER reduction, but training such multilingual NN with triphone targets is computationally demanding. We have thus shown that single language can be successfully used for multilingual techniques.

6. REFERENCES

[1] Cheng-Tao Chung, Cheng-Yu Tsai, Hsiang-Hung Lu, Chia-Hsiang Liu, Hung-yi Lee, and Lin-Shan Lee, "An iterative deep learning framework for unsupervised discovery of speech features and linguistic units with applications on spoken term detection," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 245–251.

[2] Lucas Ondel, Lukáš Burget, and Jan Černocký, "Variational inference for acoustic unit discovery," in *Procedia Computer Science*. 2016, vol. 2016, pp. 80–86, Elsevier Science.

[3] Gilles Adda, Sebastian Stker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hlne Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Nol Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, Franois Yvon, and Sabine Zerbian, "Breaking the unwritten language barrier: The BULB project," *Procedia Computer Science*, vol. 2016, no. 81, pp. 8 – 14, 2016.

[4] Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, K. Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard Rose, Petr Schwarz, and Samuel Thomas, "Subspace gaussian mixture models for speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*. 2010, vol. 2010, pp. 4330–4333, IEEE Signal Processing Society.

[5] Lukáš Burget, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, K. Nagendra Goel, Martin Karafiát, Daniel Povey, Ariya Rastrow, Richard Rose, and Samuel Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Proc. International Conference on Acoustictics, Speech, and Signal Processing*. 2010, vol. 2010, pp. 4334–4337, IEEE Signal Processing Society.

[6] 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.

[7] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

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

[9] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7893–7897.

[10] Vikrant Singh Tomar and Richard C Rose, "Manifold regularized deep neural networks," *Proceedings on Interspeech - on line*, vol. 2014, no. 9, pp. 348–352, 2014.

[11] Anton Ragni, Kate M Knill, Shakti P Rath, and Mark JF Gales, "Data augmentation for low resource languages.," in *Proceedings of Interspeech 2014*. 2014, vol. 2014, International Speech Communication Association.

[12] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.

[13] Martin Karafiát, František Grézl, Lukáš Burget, Igor Szőke, and Jan Černocký, "Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the aspire challenge," in *Proceedings of Interspeech 2015*. 2015, vol. 2015, pp. 2454–2458, International Speech Communication Association.

[14] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

[15] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, "Why does unsupervised pre-training help deep learning?," *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Mar. 2010.

[16] Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Multilingual hierarchical MRASTA features for ASR," in *Interspeech*, Lyon, France, aug 2013, pp. 2222–2226.

[17] K.M. Knill, M.J.F.Gales, S.P. Rath, P.C. Woodlandand C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. of ASRU 2013*, Dec 2013.

[18] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked Bottle-Neck neural network structure for new language," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, May 2014, IEEE.

[19] Ngoc Thang Vu, Jochen Weiner, and Tanja Schultz, "Investigating the learning effect of multilingual bottle-neck features for ASR," in *Interspeech*, Singapore, Sept. 2014.

[20] 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.

[21] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 4165–4168.

[22] Jonas Gehring, Wonkyum Lee, Kevin Kilgour, Ian R Lane, Yajie Miao, and Alex Waibel, "Modular combination of deep neural networks for acoustic modeling.," in *Proceedings of Interspeech 2013*, 2013, number 8, pp. 94–98.

[23] František Grézl and Martin Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proc. of The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, St. Petersburg, Russia, May 2014.

[24] František Grézl and Martin Karafiát, "Bottle-neck feature extraction structures for multilingual training and porting," in *Procedia Computer Science*. 2016, vol. 2016, pp. 144–151, Elsevier Science.

[25] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. Paliwal, Eds., New York, 1995, Elseviever.

[26] Kornel Laskowski, Mattias Heldner, and Jens Edlund, "The fundamental frequency variation spectrum," *in Proceedings of FONETIK 2008*, pp. 29–32, 2008.

[27] Kornel Laskowski and Jens Edlund, "A Snack implementation and Tcl/Tk interface to the fundamental frequency variation spectrum algorithm," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[28] František Grézl, Ekaterina Egorova, and Martin Karafiát, "Further investigation into multilingual training and adaptation of stacked Bottle-Neck neural network structure," in *Proceedings of 2014 Spoken Language Technology Workshop*. 2014, pp. 48–53, IEEE Signal Processing Society.