Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 15 – 22

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Study of Large Data Resources for Multilingual Training and System Porting

František Grézl*, Ekaterina Egorova, Martin Karafiát

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

Abstract

This study investigates the behavior of a feature extraction neural network model trained on a large amount of single language data ("source language") on a set of under-resourced target languages. The coverage of the source language acoustic space was changed in two ways: (1) by changing the amount of training data and (2) by altering the level of detail of acoustic units (by changing the triphone clustering). We observe the effect of these changes on the performance on target language in two scenarios: (1) the source-language NNs were used directly, (2) NNs were first ported to target language.

The results show that increasing coverage as well as level of detail on the source language improves the target language system performance in both scenarios. For the first one, both source language characteristic have about the same effect. For the second scenario, the amount of data in source language is more important than the level of detail.

The possibility to include large data into multilingual training set was also investigated. Our experiments point out possible risk of over-weighting the NNs towards the source language with large data. This degrades the performance on part of the target languages, compared to the setting where the amounts of data per language are balanced.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Stacked Bottle-Neck; feature extraction; multilingual training; large data; Fisher database

1. Introduction

Multilingual resources are of great help in case the data from the target language are not sufficient to train good acoustic model. In such case, the multilingual model, which is usually trained beforehand, is ported to the target language. Such multilingual models outperform the model trained only on limited target data [1,2,3]. The same holds for neural network model used for feature extraction [4,5,6].

It has been shown that increasing the number of languages used to train the multilingual model decrease the WER after porting to target language [5]. However, in multilingual training, it is desired to use the languages which are potentially close to the target one. A study has been carried out to show that careful selection of languages used to

* Corresponding author. Tel.: +420-541-141-280 ; fax: +420-541-141-270.
E-mail address: grezl@fit.vutbr.cz

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.024

train the multilingual NN leads to an improvement on target language[7]. Improvements can carry past the porting stage, when the NN is retrained on the target language data. The selection does not have to stop on the level of the language as an atomic unit. It is possible to select only appropriate sentences or even frames.

The disadvantage of such language selection is the necessity to know the target language a-priori and subsequent training of the multilingual NN on potentially large amount of data. Moreover, the optimal thresholding for data selection may differ depending on the target language. Also, very distinct languages may not benefit from this technique at all.

Although there is a lot of studies comparing different strategies of multilingual NN training and theirs effect on the target language e.g.[8,9,10,11,12] a detailed analysis of the important issues of training language handling is missing. By a handling we mean properties, which can be altered. For example, the acoustic characteristic of the language cannot be changed, but we can change the modeling or labeling granularity of the given acoustic space.

Such study can be better made on a single training language, thus eliminating interaction between training languages during multilingual NN training. The advantage of multilingual training is a rich phoneme set seen over several languages, but variety in used recording device, which can be also language dependent as certain locations may tend to use specific handsets is a clear drawback. There is a danger of conditioning certain phonemes rather on the audio channel than on the underlying acoustic information. A big collection of one language should provide rather homogeneous recording conditions.

In this work we study the behavior of NN used for feature extraction trained on a large database corpus – English Fisher. Although the performance of ported monolingual system would be worse in comparison with the multilingual one (due to the limited acoustic space coverage and phonemic variability), it still should reveal the trends.

The focus of this study is to find out, how the coverage and partitioning of acoustic space bounded by a single language phonology will affect the performance in the target language. The limitation to a single language phoneme set makes it possible to alter the phonetic resolution of the acoustic space by means of triphone clustering. Such change should reveal if finer resolution of otherwise the same acoustic space will lead to better performance in the target language. An alternative to the phonetic resolution is acoustic coverage of the units. By changing the amount of training data – by changing the number of speakers as well as the number of utterances per speaker, the acoustic variability of given unit will also change.

It would be also interesting to see how the language with large data can be used together with a multilingual set. The databases used for multilingual model training are usually more or less balanced. If the aim of multilingual processing is to use any transcribed data, large differences in the amounts of data per language may appear. A case study can reveal if this might be a problem or if the multilingual training procedure can deal with it.

2. Experimental setup

In this study, we observe the WER obtained from a tandem[13] system where the features for the final GMM-HMM classifier are the Bottle-Neck (BN)[14] features obtained from Stacked Bottle-Neck (SBN) Neural Network (NN) hierarchy[15]. A simple maximum-likelihood trained model without any speaker adaptation is used.

The GMM-HMM model is trained on the target language which is represented by the limited language pack of the following data sets release:

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, 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.

Table 1. Statistics of the data for target languages.

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
# 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

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, 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.

Statistics for target languages are given in Tab. 1. The amounts of data refer to the speech segments after chopping out long portions of silence – only 150 ms of silence were left at the beginning and end of each utterance; when pause longer than 300 ms was detected, the utterance was split into two. The vocabulary and language model (LM) training data consist of speech word transcriptions of the training data. 3-gram LM was used for the decoding.

The features for GMM-HMM are transformed BN outputs without any additional features. Thus the system performance will directly reflect the changes made in neural network training. The transformation used is the Maximum Likelihood Linear Transform (MLLT), which considers HMM states as classes. The GMM models are trained by single-pass retraining from an HLDA-PLP initial system (described below). 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 initial system is based on PLP coefficients which are together with their first, second and third order derivatives transformed using HLDA. The HLDA treats every Gaussian component as a class and its output dimensionality is 39. The conversation side mean and variance normalization is used on top of the transformed features. The HMM states correspond to cross-word tied-states triphones, each state consists of 18 Gaussian mixture components. The model is trained from scratch using mix-up maximum likelihood training. This model is used for forced alignment of the training data and for seeding the final BN features based HMMs.

2.1. SBN neural network hierarchy

The SBN is a two-stage structure of 6-layer NNs as described in [15]. 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 applied on squared FFT magnitudes. The fundamental frequency features consist of F0 and probability of voicing estimated according to [16] and smoothed

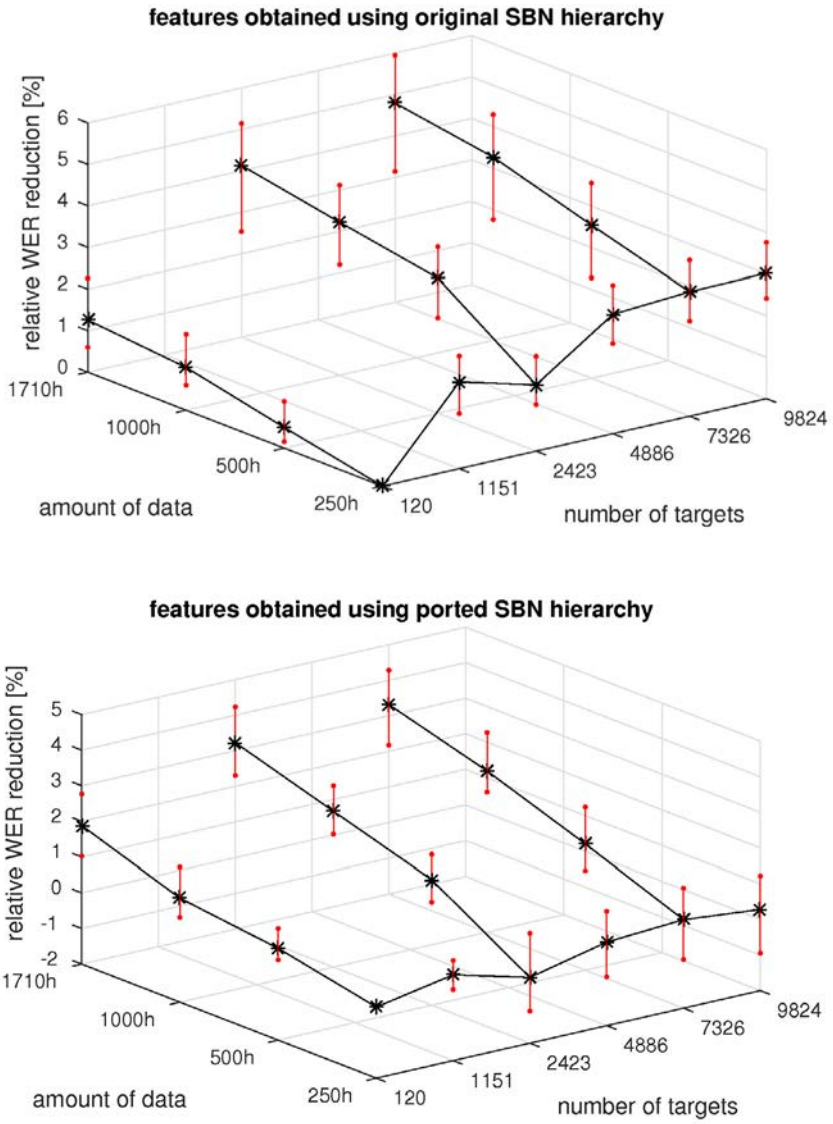

Fig. 1. Average relative WER reduction over five target languages as a function of training data amount and number of NN targets. The reference point is the NN trained on 250 hours with phoneme state targets. Red bars connect minimum and maximum values for given setting.

by dynamic programming, F0 estimates obtained by Snack tool[1] function *getf0* and seven coefficients of Fundamental Frequency Variations spectrum[17,18]. Together, there are 10 F0 related coefficients.

The 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 $34 \times 6 = 204$ coefficients on the first stage NN input. The whole data set is mean and variance normalized.

[1] `www.speech.kth.se/snack/`

3. SBN trained on Fisher database

The purpose of these experiments is to address the sensitivity of target language WER on the coverage and partitioning of source language acoustic space. By coverage, we mean meant inter- and intra- speaker variability presented by the data. Different coverage is simply achieved by using only selection of the data for NN training. The partitioning is presented by the NN training targets. Generally, the more targets we use, the finer acoustic event can be classified and the higher is the partitioning.

The source language here is English taken from the Fisher database[2]. The initial HLDA-PLP models have 9824 tied triphone states. The generated decision tree was climbed up to create clusterings with different number 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 original triphone tied states, 7326, 4886, 2423, and 1151 for the reduced triphone states and 120 for phoneme states. The amount of data used for NN training was either 250, 500, 1000 or 1710 hours of data. The 1710 hours set represents whole training data after removing long portions of silence (see above). The selection of data used for smaller training sets was done randomly on a segment level. For the 250 hours training set, 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.

Two sets of experiments were done:

- The SBN hierarchies trained on Fisher data were directly used to extract BN features for target language.
- The SBN hierarchies were ported to target language and then the BN features were extracted. The porting was done as follows: The last layer of NN trained on Fisher was trimmed, and new layer with appropriate number of outputs was initialized randomly. Then only the new layer was trained for six epochs on target language. Finally, whole network was trained (fine tuned) on target language data for another six to eight epochs. The learning rate for this fine tuning training is set to $1/10$ of the normal training.

To present results in a compact way, the relative WER reduction with respect to system based on features obtained from SBN trained on 250 hours with phoneme state targets was computed for each target language. Then the relative WER reductions were averaged over the target languages. The results obtained using original and ported SBN NNs were kept apart. The minimal and maximal relative WER reduction was found for each training condition. The results are shown in Fig. 1 where the points connected by black line corresponds to the average WER reduction and the red bars around it shows the minimal and maximal values.

It can be seen that for features obtained from the original SBN trained on Fisher data only, the gains are achieved for increased partitioning as well as coverage. Finer clustering is more important to achieve better performance, but increasing the coverage and partitioning together is far more efficient.

When porting to the target language is done, the partitioning does not play so important role. Moreover, having bigger partitioning without sufficient coverage may lead to the performance degradation. This is interesting behavior revealing that fine-tuning of NN is not able to utilize the higher clustering. One could think that porting from NNs which distinguish between more acoustic units would be easier but for some languages it just might be the opposite. The only safe way to introduce higher clustering is to increase the coverage too.

4. Fisher as part of multilingual data set

This part of the study focuses on the effect of additional large data for multilingual training. The Fisher data is added to 11 IARPA BABEL[3] languages shown in Tab 2. The speech was force-aligned using our BABEL ASR system[19] and long portions of silence were removed (see Sec.2). The phoneme states are used as targets for the multilingual NN training as it is more practical and was also found more efficient[20]. More details about the characteristics of the languages can be found in[21].

[2] Fisher 1,2; LDC2004S13, LDC2005S13 for speech data; LDC2004T19, LDC2005T19 for transcripts
[3] http://www.iarpa.gov/index.php/research-programs/babel

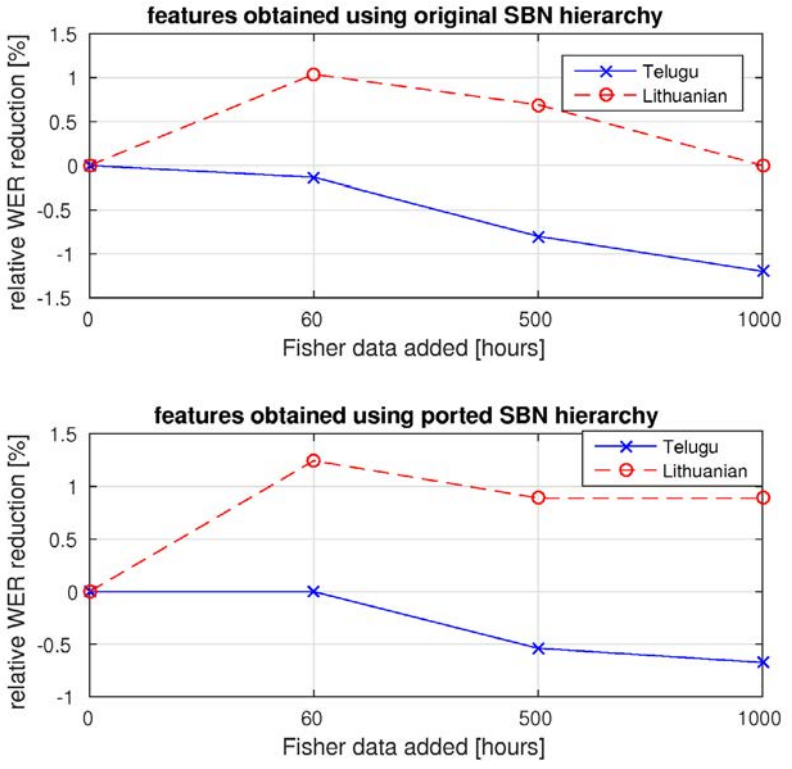

Fig. 2. Relative WER reduction for Telugu and Lithuanian as a function of additional Fisher data used in multilingual training. The reference point is for no Fisher data added.

Table 2. Data used for multilingual training of SBN networks.

Language	Dataset	hours	# phn. states
Cantonese	IARPA-babel101-v0.4c	65.0	471
Pashto	IARPA-babel104b-v0.4aY	64.7	216
Turkish	IARPA-babel105-v0.6	56.6	126
Tagalog	IARPA-babel106-v0.2g	44.1	252
Vietnamese	IARPA-babel107b-v0.7	53.2	303
Assamese	IARPA-babel102b-v0.5a	46.7	141
Bengali	IARPA-babel103b-v0.4b	53.6	147
Haitian Creole	IARPA-babel201b-v0.2b	55.0	99
Lao	IARPA-babel203b-v3.1a	71.6	411
Tamil	IARPA-babel204b-v1.1b	72.7	102
Zulu	IARPA-babel206b-v0.1e	57.8	219
Total		641.0	2487

Since the Haitian Creole, Lao and Zulu are presented in the multilingual training set, the evaluation is done on the remaining two languages – Telugu and Lithuanian.

First, the multilingual NN is trained without the Fisher data to have a reference point. Then, 60, 500 and 1000 hours from Fisher are added as another language to the training set. In the first case, the amount of data roughly corresponds to the size of language pack. This results in a balanced training set in terms of language representation. In the second case, the size of Fisher data is comparable to the whole multilingual set. The last case, when 1000 hours

of Fisher are added, the English data forms large majority of all the data. Note, that phoneme states were used as targets for the Fisher database (120 targets).

Again, two sets of experiments were done: In the first set, the multilingual SBN NNs are used directly for the feature extraction. In the second set, the NNs are ported to target language prior to feature generation.

The results are shown in Fig. 2. The relative WER reduction for each language is shown. The reference system is trained on features obtained by multilingual NNs without Fisher data. As can be seen from the curves, the effect of adding 60 hours of Fisher data is either none (for Telugu) of positive (for Lithuanian) which is the expected result after adding one more language to the multilingual training set. But with increasing amount of data coming from this new language, the performance on target languages decreases. The degradation is more distinctive on NNs without porting as the fine-tuning on target data largely improves performance of the system [5]. This results show the potential danger of over-presenting a language (or a group of languages) in the multilingual training set.

5. Conclusions

In this study we have shown how to use large data from single language in order to create a system well performing on different target language. It was show, that not only the acoustic space coverage – amount of data, but also acoustic space partitioning – level of detail, is important for good performance. The results show that these properties are tied and joint effect of increasing them both is bigger than what would be expected by summing up improvements in each direction.

The results also reveal possible danger which can be encountered while porting the NN to target language. The possibly danger setting is to train NN on acoustic space with high partitioning but without sufficient coverage. In such case, the resulting system can perform worse than one ported from NNs with only basic partitioning – phoneme state targets.

An interesting phenomenon is the improvement obtained by Fisher SBN with ~1100 targets. This increase in performance appears across all the languages for purely Fisher trained NNs but disappears after porting. One explanation might be that the number of targets is close to the number of states used in the target language. The smooth curve after porting would be then caused by the porting process, which actually sets the number of targets to be the same for all the SBNs.

The second sets of experiments employed the large data in the multilingual training. The results reveal that having largely unbalanced amount of data per language in the training set might cause degradation of the final system. For both of our test languages, the final performance goes down as the amount of used Fisher data increases over the balanced set. The porting procedure can reduce the negative effect if the target language can benefit from the largely presented language.

Although this study was based on tandem system, where features generated by NNs are used for GMM-HMM model, we believe that the findings apply also to hybrid system as the NNs used are virtually the same. But for hybrid, the porting step is obligatory.

In the future, we would like to investigate what really matters in the acoustic space partitioning. The question we would like to answer is: Is there an optimal clustering of source language acoustic space? Or is there a strategy to create well performing clustering? The second track is the challenge of better combination of highly imbalanced training sets. One way is to scale down the back-propagation error for the over-presented language. This might be better than simply selecting only a portion of the data but still can be far from optimal.

Acknowledgments

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. This work was also supported by the European Union's Horizon 2020 project No. 645523 BISON, and by Technology Agency of the Czech Republic project No. TA04011311 "MINT".

References

1. Scanzio, S., Laface, P., Fissore, L., Gemello, R., Mana, F.. On the use of a multilingual neural network front-end. In: *Proceedings of INTERSPEECH-2008*. ISBN 978-1-4244-2353-8; 2008, p. 2711–2714.
2. Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., et al. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In: *Proc. International Conference on Acoustictics, Speech, and Signal Processing*; vol. 2010. IEEE Signal Processing Society. ISBN 978-1-4244-4296-6; 2010, p. 4334–4337.
3. Swietojanski, P., Ghoshal, A., Renals, S.. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In: *Spoken Language Technology Workshop (SLT), 2012 IEEE*. ISBN 978-1-4673-5125-6; 2012, p. 246–251.
4. Veselý, K., Karafiát, M., Grézl, F.. Convolutive bottleneck network features for LVCSR. In: *Proceedings of ASRU 2011*. ISBN 978-1-4673-0366-8; 2011, p. 42–47.
5. Grézl, F., Karafiát, M.. 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; 2014.
6. Tüske, Z., Golik, P., Nolden, D., Schlüter, R., Ney, H.. Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages. In: *Interspeech*. Singapore; 2014, p. 1420–1424.
7. Yu Zhang Ekapol Chuangsuwanich, J.G.. Language ID-based training of multilingual stacked bottleneck features. In: *Interspeech*. Singapore; 2014.
8. Tüske, Z., Schlüter, R., Ney, H.. Multilingual hierarchical MRASTA features for ASR. In: *Interspeech*. Lyon, France; 2013, p. 2222–2226.
9. Vu, N.T., Schultz, T.. Multilingual multilayer perceptron for rapid language adaptation between and across language families. In: *Proceedings of Interspeech 2013*; 8. 2013.
10. Knill, K., M.J.F.Gales, , Rath, S., Zhang, P.W.C., Zhang, S.X.. Investigation of multilingual deep neural networks for spoken term detection. In: *Proc. of ASRU 2013*. 2013.
11. Grézl, F., Karafiát, M., Veselý, K.. 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: IEEE; 2014.
12. Vu, N.T., Weiner, J., Schultz, T.. Investigating the learning effect of multilingual bottle-neck features for ASR. In: *Interspeech*. Singapore; 2014.
13. Hermansky, H., Ellis, D.P.W., Sharma, S.. Tandem connectionist feature extraction for conventional HMM systems. In: *Proc. ICASSP 2000*. Turkey; 2000.
14. Grézl, F., Karafiát, M., Kontár, S., Černocký, J.. Probabilistic and Bottle-Neck features for LVCSR of meetings. In: *Proc. ICASSP 2007*. Honolulu, Hawaii, USA. ISBN 1-4244-0728-1; 2007, p. 757–760.
15. Grézl, F., Karafiát, M., Burget, L.. Investigation into Bottle-Neck features for meeting speech recognition. In: *Proc. Interspeech 2009*. 2009, p. 294–2950.
16. Talkin, D.. A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Paliwal, K., editors. *Speech Coding and Synthesis*. New York: Elseviever; 1995.
17. Laskowski, K., Heldner, M., Edlund, J.. The fundamental frequency variation spectrum. *in Proceedings of FONETIK 2008* 2008;:29–32.
18. Laskowski, K., Edlund, J.. 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. ISBN 2-9517408-6-7; 2010.
19. Karafiát, M., Grézl, F., Hannemann, M., Veselý, K., Černocký, J.H.. BUT BABEL System for Spontaneous Cantonese. In: *Proceedings of Interspeech 2013*; 8. 2013, p. 2589–2593.
20. Grézl, F., Egorova, E., Karafiát, M.. Further investigation into multilingual training and adaptation of stacked Bottle-Neck neural network structure. In: *Proceedings of 2014 Spoken Language Technology Workshop*. IEEE Signal Processing Society. ISBN 978-1-4799-7129-9; 2014, p. 48–53.
21. Harper, M.. The BABEL program and low resource speech technology. In: *Proc. of ASRU 2013*. 2013.