Multilingual Hierarchical MRASTA Features for ASR

Zoltán Tüske[a], *Ralf Schlüter*[a], *Hermann Ney*[a,b]

[a]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany
[b]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{tuske, schlueter, ney}@cs.rwth-aachen.de

Abstract

Recently, a multilingual Multi Layer Perceptron (MLP) training method was introduced without having to explicitly map the phonetic units of multiple languages to a common set. This paper further investigates this method using bottleneck (BN) tandem connectionist acoustic modeling for four high-resourced languages — English, French, German, and Polish. Aiming at the improvement of already existing high performing automatic speech recognition (ASR) systems, the multilingual training of the BN-MLP is extended from short-term to hierarchical long-term (multi-resolutional RASTA) feature extraction. Furthermore, deeper structures and context-dependent target labels are also examined. We experimentally demonstrate that a single state-of-the-art BN feature set can be trained for multiple languages, which is superior to the monolingual feature set, and results in significant gains in all the four languages. Studying the scalability of the multilingual BN features, a similar gain is observed in small (50 hours) and in larger scale (300 hours) ASR experiments regardless of the distribution of the data amount between the languages. Using deeper structures, context-dependent targets, and speaker adaptation, the multilingual BN reduces the word error rates by 3–7% relative over the target language BN features and 25–30% over the conventional MFCC system.

Index Terms: deep MLP, bottleneck, multilingual, hierarchical, MRASTA, LVCSR

1. Introduction

With the development of ASR systems for an increasing number of languages, methods being able to generalize over languages have particularly growing interest recently. Since manually transcribed speech data is still a significant cost factor in the development of Large Vocabulary Continuous Speech Recognition System (LVCSR), the demand to improve the acoustic models using multilingual resources exists not only for under-resourced languages.

As neural networks (NN) have become a major component of recent Hidden Markov Model (HMM) based ASR techniques, it was observed that MLP based posterior features possess language independent properties to a certain degree [1, 2, 3]. Although, the posterior features trained e.g. on English speech data can significantly improve the pure cepstral based systems in entirely different languages, like Arabic or Mandarin [1], usually the NN features trained on the target language result in better recognition performance [4, 5, 6]. However, the cross-lingual NNs are shown to be a good initialization before training NN based features on a new language [7], especially in a low-resource scenario [8].

The incorporation of MLP based posterior estimation as additional features into the Gaussian Mixture Model (GMM) based HMMs, known as tandem approach, was introduced in [9] and was improved by the bottleneck concept of [10]. The bottleneck approach can be interpreted as a dimension reduction method using non-linear discriminant analysis. The bottleneck features in general outperform the posterior features, and are usually concatenated with classical cepstral features resulting in lower error rates than the NN features alone.

To exploit resources of multiple languages in acoustic model training, there is usually a need to unify similar sounds across different languages. This could be done either in a knowledge based way, e.g. with phonetic alphabets such as IPA or SAMPA [7, 11, 12, 13], or by various data driven approaches [5, 14, 15, 16]. Due to the fact that the available lexicons for ASR are usually simplified (e.g. by phone folding), mapping phones of multiple languages on a common set is often ambiguous or inaccurate. However, the training of neural networks on multiple languages is possible without such a mapping by sharing hidden layers across languages [17]. A similar motivation can be found in [18], where the factored GMM models of different languages share parameters in a subspace.

In [19] and [6] the approach of [17] was applied to bottleneck features. They showed that the multilingual bottleneck features outperform the monolingual features, and that the multilingual features offer more robust cross-lingual generalization for unseen languages. As [19] pointed out, this method also outperformed the IPA based phoneme unification approach. Furthermore, [6] demonstrated that the multilingual BN can be successfully applied to reduce the mismatch between training and testing acoustical conditions by reusing matched data from other languages.

This paper extends our previous investigation [6] to the multilingual BN-MLP approach as follows. First, the multilingual bottleneck MLP training is applied to hierarchical processing of long-term, multi resolutional RASTA filter outputs [20, 21]. Second, the scalability of the language independent features is explored in large and small scale experiments. Third, based on the recent success of deep neural networks in hybrid acoustic modeling [22] and BN features [23, 24], multilingual 7-hidden-layer BN-MLPs are trained on tied-triphone state targets. We also investigate the effect of introducing language dependent hidden layers after the bottleneck.

The paper is organized as follows: after a short description of the training and testing corpora of the four languages in Section 2, Section 3 gives the details of the feature extraction and also summarizes the multilingual training of the BN features. We describe the experimental setup in Section 4 followed by results in Section 5. The final conclusions are drawn in Section 6.

2. Language resources

The QUAERO project aims at developing LVCSR systems for European languages to transcribe podcast and broadcast news data. For our research purpose the following four languages were selected: French (FR), English (EN), German (GE), and Polish (PL). The corpus statistics can be found in Table 1. The acoustic training is mainly based on speech data collected and transcribed within the project, and only the German corpus contains about 14 hours of additional external audio recordings downloaded from the web. Thus, the final multilingual BN features are trained on ∼800 hours of speech data. For the estimation of language models (LM) various training materials were used. Beside the in-domain text data provided for all project partners, the LM data were extended by text resources from the web by crawling RSS news feeds, web archives, etc. For further details we refer to [25, 26].

Table 1: *Training and testing corpora statistics on four different languages, the number of phonemes (phn.) in the corresponding lexicons, and perplexity (PPL) values measured on the development sets*

	Corpus	Total data [h]	# running words	# phn.	lexicon size	PPL
FR	Training	317	3.9M	49	200k	131
	Dev11	2.9	36k			
	Eval11	3.1	38k			
EN	Training	232	2.7M	42	150k	130
	Dev11	3.7	45k			
	Eval11	3.3	35k			
GE	Training	142	1.4M	50	300k	250
	Dev11	3.8	35k			
	Eval11	3.1	29k			
PL	Training	110	1.0M	40	600k	673
	Dev11	3.4	28k			
	Eval11	3.6	27k			

3. Feature extraction

3.1. Cepstral features

From the audio files, vocal tract length normalized (VTLN) Mel cepstral coefficients (MFCC) are extracted. The factors for the piecewise linear warping are estimated by language specific text-independent Gaussian Mixture classifiers (fast-VTLN). Then nine consecutive frames of the 16-dimensional, segmentwise mean-and-variance normalized MFCC vectors are projected by LDA into a 45-dimensional subspace.

3.2. Bottleneck MRASTA features

In this paper the multilingual MLP training (see Section 3.3) is applied to long-term MLP features. The bottleneck feature extraction pipeline is similar to that in [24] and is based on the work of [20]. One second trajectory of each critical band is filtered by first and second derivatives of the Gaussian function, where the standard deviation varies between 8 and 60 ms resulting in 12 temporal filters per band. The bottleneck features are extracted from hierarchical, MLP based processing of the modulation spectrum [27]. The input of the first MLP contains the fast modulation part of the MRASTA filtering, whereas the second MLP is trained on the slow modulation components and the BN output of the first MLP. In both cases, the feature vectors fed into the MLPs are augmented by the logarithm of critical band energies. In our investigation the bottleneck layers consisted of 60 nodes. The final features are obtained by concatenating LDA transformed MFCCs with PCA transformed linear output of the BN layer.

In the experiments where a classical 5-layer BN-MLP is used, the number of nodes in the hidden layers is fixed to 7000. In the MLPs with deeper structures the size of the non-bottleneck layers is set to 2000. Using backpropagation algorithm in the mini-batch mode (512 frames), the randomly initialized, fully connected MLPs are trained according to cross-entropy criterion, and approximate phoneme or tied-triphone state posterior probabilities. In the later case, based on our previous study [24], the output layers are limited to 1500 target labels. To prevent overfitting and for adjusting the learning rate parameter, 10% of the training corpus is used for cross-validation.

3.3. Multilingual training of bottleneck MLPs

In order to extract robust MLP features from multilingual resources, we apply a training method proposed by [17], which avoids the mapping of phonemes of different languages to a common set. The raw feature vectors of the multiple languages are merged, randomized and presented to the MLPs for training along with the phonetic and language labels. In contrast to standard MLPs, the network uses language specific softmax function as output non-linearity, also referred to as interval-based softmax [19]. Depending on the language-ID of each feature vector, backpropagation is initiated only from the language specific subset of the output. Applying the multilingual training to BN-MLP [19, 6], the key idea is that up to the bottleneck layer the network is shared between the languages, and the multilingual training forces the net to extract a more language-independent representation from the input vectors. Fig. 1 shows the multilingual BN-MLP of one level of the hierarchy with deep 9-layer bottleneck MLP structure and context-dependent targets, where language dependent hidden layers after the BN are also introduced. The motivation behind this idea is that each language can have its own non-linearity to generate the posterior estimates from the low-dimensional bottleneck feature space. The language dependent hidden layer structure corresponds to a special case: it can be also interpreted as a single weight matrix between the outputs and the last hidden layers where the language dependent weights show a non-square block diagonal structure. Therefore, this could be learned by the neural network directly. However, having n languages, the block diagonal constraint increases the number of trainable parameters only linearly with n, whereas a fully connected output and last hidden layers of n languages would result in an n^2 increase. Furthermore, due to the language dependent backpropagation the introduction of language dependent hidden layers does not increase the computational cost.

4. Experimental setup

4.1. Acoustic and language modeling

All recognition experiments are carried out with the freely available RASR decoder [28]. Instead of training the acoustic models from scratch, an initial alignment generated by previous best systems are used to estimate the GMM and MLP parameters. Speaker adapted results are based on Constrained Maximum Likelihood Linear Regression (CMLLR) [29] with the simple target model approach [30]. Discriminative training was not performed.

For each language a 4-gram LM was estimated and smoothed by the modified Kneser-Ney method. To handle the

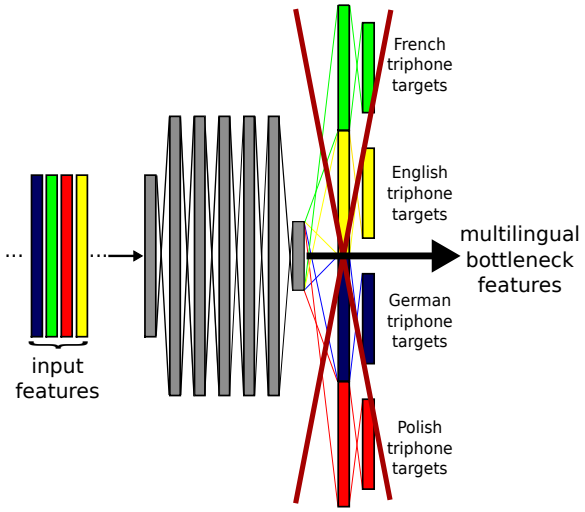

Figure 1: *The joint training of deep context-dependent bottleneck MLP features on multiple languages (FR, EN, GE, PL). The different colors indicate different languages, and language dependent backpropagation from the output layer. The other parts of the network including the bottleneck layer are shared between the languages.*

high lexical variety of the German language (due to compounding and inflection), an alternative LM containing full-word and sublexical units was estimated [31]. The lexicon sizes and perplexity values of the unpruned LMs are listed in Table 1. For further details on the LM estimation we also refer to [26, 32].

5. Experimental results

5.1. Small scale experiments

In the first experiments the effect of multilingual BN features was investigated, assuming similar amount of speech data is available in all languages. Therefore, about 50 hours from each corpus were selected. The multilingual BN features are based on a classical 5-layer MLP structure and trained on the 200 hours of data using phoneme targets. Table 2 compares the recognition results of classical cepstral features based systems with tandem systems using either BN trained only on the target language or BN trained in a multilingual way. As can be seen, all languages benefit from the multilingual BN features, although the target language represents only 25% of the data during the MLP training. The relative improvement compared to the target language BN features exceeds 5% for English, German, French and 2% for Polish.

Table 2: *Small-scale speaker-independent recognition results on Eval11 corpora of four different languages using **target** language or **multi**lingual BN features. The results are in word error rate (WER).*

Language	FR	EN	DE	PL
MFCC	25.5	31.7	25.0	18.9
+BN_{target}	22.2	26.8	21.3	15.7
+BN_{multi}	21.1	24.9	20.1	15.4

In order to have a better understanding how important the presence of other languages during the BN training is, different combinations of the languages are tested. According to our previous investigation [6], the target language data was always presented to the neural networks. Demonstrating only with the French recognition setup, Table 3 clearly shows that the target language BN features can be improved using resources from any of the other three languages. Furthermore, the more languages were used the larger the potential improvement became. Using all languages and training a single bottleneck network, we were able to end up in a single set of BN features for four languages.

Table 3: *Speaker-independent recognition results (in WER) on French Eval11 test set using multilingual BN features trained on different combination of languages*

BN trained on							
FR	FR +EN	FR +DE	FR +PL	FR +EN+DE	FR +EN+PL	FR +DE+PL	FR+EN +DE+PL
22.2	21.8	**21.6**	21.6	21.7	21.7	**21.5**	**21.1**

5.2. Large scale experiments

Based on our observation in small scale experiments, in the following, multilingual BN features were always trained on all four languages. Since the available data are usually not equally distributed between the languages, we also investigate the effect of unbalanced corpus sizes. Using 800 hours of speech data and phoneme targets, the first three rows of Table 4 show that the relative improvement, which originates from training BN features on multiple languages, does not change drastically. Compared to Table 2, the word error rates are still reduced by more than 5% relative for English and German, while the French and Polish systems show 3% relative improvement. The previous experiment demonstrated that despite the unbalanced corpus sizes and larger amount of speech data in all languages, the multilingual BN features are able to extract a more discriminative speech representation than the monolingual BN.

Table 4: *Large-scale speaker-independent recognition results on Eval11 corpora of four different languages using **target** language or **multi**lingual BN features. The results are in WER.*

		Language	FR	EN	DE	PL
		MFCC	23.6	28.6	23.3	18.1
MLP targets	phonemes	+BN_{target}	19.3	23.1	19.0	14.5
		+BN_{multi}	18.7	21.3	17.9	14.0
	1500 tied-triphone states	+BN_{target}	18.3	21.8	18.4	13.8
		+BN_{multi}	18.1	21.3	17.6	13.9
		+deep BN_{target}	17.4	20.3	17.3	13.0
		+deep BN_{multi}	17.1	19.7	16.4	12.6
		+lang.dep. hidden layer	16.8	19.7	16.2	12.4

Switching from phoneme to 1500 tied-triphone states targets resulted in 3–5% relative improvement in the monolingual case (3rd row in Table 4). Training a single BN feature set for the four languages in the described multilingual way improved the results only for French and German. The gain is 1–2% relative over the target BN trained on context-dependent targets. It is interesting to see, that for English the phoneme based multilingual BN features perform better than the target language BN-MLP trained on tied-states (2nd column, 3rd and 4th rows), and that increasing the targets of the multilingual BN does not always result in better performance (3rd and 5th rows). As a summary, the multilingual BN clearly outperforms the monolingual features if phoneme targets are used. Nevertheless the use of context-dependent targets to train BN features mitigates the advantage of the multilingual 5-layer BN in speaker-independent recognition scenario.

Since the BN features can also benefit from deeper structures and context-dependent phonetic targets [24], the multilingual bottleneck features are also investigated using more complex MLPs. In our previous study about deep BN features only a symmetrical MLP structure (three hidden layers before and after the BN) was used. Therefore, before the multilingual training, it was first tested on the French recognition task where to place the bottleneck layer, given a 9-layer MLP with fixed non-bottleneck hidden layer size of 2000. As Table 5 shows, pushing the BN closer to the output layer does not have any significant influence on the recognition error rate, as long as there is at least a single hidden layer between the output and the BN. The best performance was observed on the development set when only a single hidden layer was used after the BN. However, it resulted only in an insignificant gain on Eval11 set. In the following, the structure, which shares five non-bottleneck hidden layers between the four languages before the BN layer, is used for multilingual MLP training. A similar MLP is shown in Figure 1. The results in Table 5 also indicate that taking the principal components account for 95% of the total variability of the BN features is a nearly optimal choice for deep BN features as well.

Training monolingual context-dependent (CD) deep BN features, we observed huge, over 25% relative improvement in all languages (6^{th} row in Table 4). As can be seen, the deep multilingual CD-BN features opposed to 5-layer CD-BN_{multi} improved the results in each language. This could be explained by the increased capacity of the MLP to extract the low-dimensional language-independent speech representation which can then discriminate between 1500 classes of any of the four languages. The relative gain over the deep monolingual BN features is between 2-5%.

The speaker-independent French, Polish, and English systems show 28%, 31%, 30% relative improvement over the classical cepstral system, whereas the German ASR performance is improved by 40% relative. Introducing language dependent hidden layers could only slightly reduce the error rate.

Table 5: *Effect of the place of BN and the PCA size on WER using deep 9-layer MLP structure. The speaker-independent recognition performance is measured on Eval11 corpus of the French task. Bold font indicates the PCA dimension size accounting for the 95% of the variability.*

# hidden layers	PCA dimension							
after the BN	17	23	29	35	41	47	53	59
3	17.7	17.3	17.4	**17.4**	17.6	17.6	17.6	17.7
2	17.5	17.4	17.4	**17.3**	17.4	17.4	17.5	17.8
1	17.8	17.3	17.3	17.4	**17.4**	17.5	17.5	17.6
0	18.0	17.9	17.7	17.8	17.8	17.9	**17.9**	18.1

5.3. Effect of speaker adaptation

In the following experiments, the effect of CMLLR based speaker adaptation on the multilingual MLP features is studied. As can be seen in the 1^{st}, 2^{nd}, and 4^{th} rows of Table 6, the classical 5-layer context-dependent monolingual BN features improved the classical cepstral based system by 20% relative. The deep structures show significant gain over the 5-layer BN, and result in 21-28% relative improvement over MFCC. As the results reported in the 3^{rd} and 5^{th} rows indicate, the speaker adaptation does not eliminate the effect of multilingual BN features, but rather intensifies it. In contrast with Table 4, the speaker adapted 5-layer multilingual CD-BN features always outperformed the monolingual ones (2^{nd} and 3^{th} rows), and reduced the error rate by 2% to 5% relative. For Polish and German the multilingual 5-layer BN resulted in similar or even better performance than the target deep BN. This effect might be explained as follows: the multilingual BN output is a language independent representation, therefore CMLLR performs not only speaker, but also language adaptation. Moreover, the deep structure increased the performance gap between mono- and multilingual BN features (4^{nd} and 5^{th} rows). The relative gain is 3%–7% over target BN, and 25%–30% over the MFCC system. The slight benefit of language-dependent hidden layers vanished after the CMLLR adaptation.

Table 6: *Speaker adapted recognition results on Eval11 corpora of four different languages using **target** language or **multi**lingual BN features. The results are in WER.*

		Language	FR	EN	DE	PL
		MFCC	21.6	26.4	21.4	15.9
MLP targets	1500 tied-triphone states	+BN_{target}	17.3	19.7	17.2	12.3
		+BN_{multi}	17.0	19.2	16.3	12.1
		+deep BN_{target}	16.7	18.8	16.8	12.1
		+deep BN_{multi}	16.2	18.1	15.7	11.7
		+lang.dep. hidden layer	16.3	18.2	15.7	11.7

6. Conclusions

A recently introduced multilingual MLP approach was extensively evaluated to improve LVCSR system for four high-resourced languages. We have shown through a series of experiments that (1) common, multilingual BN features can be trained for a set of languages; (2) the gain of multilingual BN features is superior to monolingual ones even if the available amount of data is unbalanced between the different languages; (3) context-dependent targets and deeper structures can improve the multilingual BN features further; (4) the effect of speaker adaptation and multilingual BN features is additive. With the help of multilingual BN features our French, English, German, and Polish LVCSR systems were improved by 3%, 4%, 7%, 3% relative over monolingual BN features, and 25%, 31%, 27%, 26% relative compared to the MFCC systems.

Although the multilingual BN outperforms the monolingual MLP features, there is a hint that further language specific tunning might be due. As a future work, we also intend to investigate the application of the multilingual BN to under-resourced languages.

Acknowledgement

This work has received funding from the Quaero Programme funded by OSEO, French State agency for innovation, and from the European Union Seventh Framework Programme EU-Bridge (FP7/2007-2013) under grant agreement no.287658. The study was partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. 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.

H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

7. References

[1] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2006, pp. 321–324.

[2] Ö. Çetin, M. Magimai.-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2007, pp. 36–41.

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

[4] L. Tóth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual Portability of MLP-Based Tandem Features–A Case Study for English and Hungarian," in *Proc. of Interspeech*, 2008, pp. 2695–2698.

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

[6] 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 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, p. *accepted for publication*.

[7] 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. of Interspeech*, 2012.

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

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

[10] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 757–760.

[11] T. Schultz and A. Waibel, "Fast Bootstrapping Of LVCSR Systems With Multilingual Phoneme Sets," in *Proc of Eurospeech*, 1997.

[12] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2009, pp. 4333–4336.

[13] D. Imseng, H. Bourlard, and M. Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

[14] T. Schultz and A. Waibel, "Development of Multilingual Acoustic Models in the GlobalPhone Project," in *Proc. of Workshop on Text, Speech, and Dialogue*, 1998.

[15] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 2000, pp. 1029–1032.

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

[17] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the Use of a Multilingual Neural Network Front-End," in *Proc. of Interspeech*, 2008, pp. 2711–2714.

[18] L. Burget, P. Schwarz, M. Agarwal, P. Akayazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2013*, 2010, pp. 4334–4337.

[19] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. of IEEE Workshop on Spoken Language Technology*, 2012, pp. 336–341.

[20] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. of Interspeech*, 2005, pp. 361–364.

[21] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical Bottle Neck Features for LVCSR," in *Proc. of Interspeech*, 2010, pp. 1197–1200.

[22] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in *Proc of Interspeech*, 2011, pp. 437–440.

[23] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 4153–4156.

[24] Z. Tüske, R. Schlüter, and H. Ney, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, p. *accepted for publication*.

[25] M. Nußbaum-Thom, S. Wiesler, M. Sundermeyer, C. Plahl, S. Hahn, R. Schlüter, and H. Ney, "The RWTH 2009 QUAERO ASR evaluation system for English and German," in *Proc. of Interspeech*, 2010, pp. 1517–1520.

[26] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A.-D. Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 QUAERO ASR Evaluation System for English, French, and German," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 2212–2215.

[27] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 4165–4168.

[28] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[29] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[30] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 997–1000.

[31] M. A. B. Shaik, A. E.-D. Mousa, R. Schlüter, and H. Ney, "Hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR," in *Proc. of Interspeech*, 2011, pp. 1441–1444.

[32] M. A. B. Shaik, A. E.-D. Mousa, R. Schlüter, and H. Ney, "Using morpheme and syllable based sub-words for Polish LVCSR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 4680–4683.