MULTILINGUAL MRASTA FEATURES FOR LOW-RESOURCE KEYWORD SEARCH AND SPEECH RECOGNITION SYSTEMS

Zoltán Tüske[a], David Nolden, 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, nolden, schlueter, ney}@cs.rwth-aachen.de

ABSTRACT

This paper investigates the application of hierarchical MRASTA bottleneck (BN) features for under-resourced languages within the IARPA Babel project. Through multilingual training of Multilayer Perceptron (MLP) BN features on five languages (Cantonese, Pashto, Tagalog, Turkish, and Vietnamese), we could end up in a single feature stream which is more beneficial to all languages than the unilingual features. In the case of balanced corpus sizes, the multilingual BN features improve the automatic speech recognition (ASR) performance by 3-5% and the keyword search (KWS) by 3-10% relative for both limited (LLP) and full language packs (FLP). Borrowing orders of magnitude more data from non-target FLPs, the recognition error rate is reduced by 8-10%, and the spoken term detection is improved by over 40% relative on Vietnamese and Pashto LLP. Aiming at the fast development of acoustic models, cross-lingual transfer of multilingually "pretrained" BN features for a new language is also investigated. Without the need of any MLP training on the new language, the ported BN features performed similarly to the unilingual features on FLP and significantly better on LLP. Results also show that a simple fine-tuning step on the new language is enough to achieve comparable KWS and ASR performance to that system where the target language is also involved in the time-consuming multilingual training.

Index Terms— Babel, MRASTA, MLP, bottleneck, hierarchical, neural network, tandem, ASR, KWS

1. INTRODUCTION

Artificial neural networks (NN) have become an essential part of the state-of-the-art ASR systems. In the Hidden Markov Model framework, they are applied to extract features for Gaussian mixture models (tandem approach [1, 2]) and/or to estimate state posterior probabilities directly (hybrid approach [3]). Nowadays, the application of ASR technologies to an increasing number of languages induces a growing interest for methods which are able to exploit out-of-domain data or even resources of other languages in order to improve the recognition performance. Besides their great success in standard acoustic modeling, NNs turned out to be also capable to benefit significantly from other available resources — e.g. through multi-task [4, 5] and multilingual training [6]. As has been shown, a simple cross-domain/lingual porting of NNs could also be beneficial [7]. Furthermore, this improvement is usually not limited to low-resourced scenarios [8, 9].

The aim of the IARPA funded Babel project is the development of robust speech technologies, especially for spoken term detection, which can be applied to any language with a limited amount of transcription in a limited time [10]. To achieve the main goals of the project, the participants have to develop systems at the end of each program period with more and more limitations on the amount of transcribed speech provided and on the time allowed to create the system for a previously unseen surprise language. In the project, two training database configurations are defined. The full language pack (FLP) contains about 100 hours of speech, whereas the limited language pack (LLP) comprises only about 10 hours of transcribed data. In the first period, five languages have been released — Cantonese, Pashto, Turkish, Tagalog, and Vietnamese as surprise language —, which were completely transcribed.

This paper extends our previous multilingual investigation on European languages [11] to these more diverse and low-resourced languages. A first set of experiments for both FLP and LLP aimed at multilingual training of deep hierarchical MRASTA features using balanced amount of data from each language, including the target language. Moreover, in order to mitigate the lack of large amount of data for the LLP, the use of non-target FLP data for multilingual training was considered, thereby taking advantage of hundreds of hours of transcribed speech from the non-target language FLPs in addition to just ten hours for the target language. The objective of this work is to find an effective way for rapid development of improved BN features for a new language. As will be shown, with hardly any loss in recognition and KWS performance, this latter goal can be achieved by multilingual initialization and target-language specific fine-tuning of the hierarchical BN features.

The paper is organized as follows, Section 2 gives an overview of related work. After the description of the training and testing corpora in Section 3, we give the details on our experimental setups in Section 4. The uni- and multilingual ASR and KWS results are presented in Section 5, followed by a discussions in Section 6. The paper closes with conclusions in Section 7.

2. RELATION TO PRIOR WORK

This work investigates hierarchical tandem MRASTA bottleneck features [12, 13] on five languages. The MRASTA BN features were previously optimized for large vocabulary speech recognition in [14], and were extended by multilingual training in [11]. The multilingual parameter sharing in MLPs can be defined on different levels due to the feed-forward structure of multiple layers. In this paper, the multilingual training method introduced in [6] is applied. Based on our previous investigation in [11], the parameters are shared up to the last hidden layer, and language specific output is used similar to [15]. Multilingual training using universal phone set is another widely used approach [16]. In this case, the whole network is shared between the languages [17, 18]. However, as was shown in [15], a BN-MLP with language dependent outputs achieved usually lower error rates than with a single output based e.g. on IPA [19]. Target language specific fine-tuning of a NN transferred from other language(s) has also been investigated for both

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Table 1. *Statistics of training and testing corpora*

Language	Id	Set	Amount of data [h]		# phn.	lexicon
			Training*	Test	/ton.	size
Cantonese	101	LLP	10	20	145	5.9k
		FLP	69			18k
Pashto	104	LLP	10	20	44	7.0k
		FLP	72			21k
Turkish	105	LLP	10	20	45	12k
		FLP	68			46k
Tagalog	106	LLP	10	20	49	6.6k
		FLP	72			24k
Vietnamese	107	LLP	11	20	92	4.1k
		FLP	77			6.7k

*after segmentation

tandem and hybrid MLPs e.g. in [18, 20, 21] and [8, 22]. Compared to the work in [18], we study a different multilingual method with deep and hierarchical BN features using context-dependent targets. During the fine-tuning step of our hierarchical BN features, the convolutive approach of [23] is applied similar as in [24].

3. CORPUS DESCRIPTION

The Babel corpora of the first five languages consist of only narrow-band telephony data recorded in various environments from different landline and mobile devices, and cover several dialects. Each corpus contains mainly conversational speech and less than 30% scripted phonetically rich sentences (e.g. short responses to questions, phone numbers, dates). Pronunciation lexicons covering only words appearing in transcribed training data are also provided to the participants. Although more than 100 hours of data are recorded in each full language pack (FLP), considerable portion of the data is non-speech. Furthermore, we used only the conversational part of the FLP in our experiments. About 20 hours of recordings in the limited language packs are predefined subsets of the FLPs. The ASR and KWS performance is measured in all experiments on the development sets of the project. Our training corpus is based on the reference segmentation, whereas the segmentation of the test sets was provided by our project partner IBM. Table 1 summarizes the corpus statistics, and shows the amount of speech retained for acoustic model and BN-MLP training after segmentation and silence normalization steps. For Cantonese and Vietnamese the different tones are also considered. Our lexicons contain only the words that appear in the training data of the corresponding language packs.

4. EXPERIMENTAL SETUPS

4.1. Feature extraction

4.1.1. Cepstral features

In preliminary experiments we noticed that Gammatone features (GT) performed slightly better than MFCC for narrow-band speech. Therefore, all the experiments are based on this type of feature extraction. For more details we refer to [25]. For the MRASTA filtering the 15-dimensional Gammatone critical band energies (CRBE) were extracted after the spectral smoothing and 10th root compression steps. After segmentwise mean-and-variance normalization, the GT features were concatenated to fundamental frequency (F0) and voicedness features [26, 27]. The final 45-dimensional features used for Gaussian Mixture Model (GMM) training were extracted after linear discriminant analysis (LDA) of 9 consecutive frames.

4.1.2. Uni- and multilingual MRASTA BN features

According to [28], the temporal trajectories (101 frames) of the Gammatone CRBEs were smoothed by two-dimensional bandpass filters (MRASTA). Following the work of [12, 13], the final BN features are extracted by hierarchical MLPs. 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 windowed and LDA transformed BN output of the first MLP. The modulation features fed into the MLPs were always augmented by the CRBE, 9 frames of fundamental frequency (F0) and voicedness features. Independent of the language, F0 features were always used.

The multilingual BN features were trained on fully randomized feature vector set extracted from the joint corpora of the languages. Although different approaches are available to handle the different targets of the languages [16], in this paper the method proposed in [6] is applied because it avoids the ambiguous mapping to a common set. With language dependent softmax outputs, backpropagation is initiated from the language specific subset of the output depending on the language-ID of the feature vector. All hidden layers — including the BN layer — were shared between the languages. Using non-target language data from the Babel project, we moved from the primary ”BaseLR” condition, where only the target language data provided with the project is allowed, to the ”BabelLR” condition, where all Babel data from non-target languages can be used.

The MLPs are initialized by discriminative pretraining [29] and trained according to the cross-entropy criterion. The targets are 1500 tied-triphone states per language, and are determined by earlier stopping of the clustering algorithm also used for GMM training in Subsection 4.2. For adjusting the learning rate parameter, 10% of the training corpus is held out for cross-validation. In the MLP structures 6 non-BN hidden layers with 2000 neurons and sigmoid function are used, the BN layer consists of 60 nodes and was placed before the last hidden layer. When we performed fine-tuning, i.e. joint training of the two-level NN on the target language, the BN-MLP was transformed to time-convolutive NN due to the windowing function in the middle, similar to [15].

4.2. Acoustic Modeling

The initial GMMs were trained according to the maximum likelihood criterion with the expectation maximization algorithm. The Gaussian components share a globally pooled diagonal covariance matrix. Speaker adaptive training was applied using constrained maximum likelihood linear regression [30, 31]. On each FLP, 4500 generalized triphones tied by a decision-tree-based clustering were modeled. Due to the limited amount of observations, clustering ended up with only 2000-3000 tied states on LLP. Furthermore, Minimum Phone Error (MPE) criterion based discriminative training was performed to sharpen the speaker adapted acoustic models [32].

4.3. Language Modeling

One of the biggest challenges in the Babel project is the sparse language model training data. According to the ”BaseLR” submission condition, no additional resources were used during the 4-gram language model (LM) estimation. To smooth the LMs, the discount parameters were estimated according to [33].

4.4. Speech Recognition and Keyword Search Systems

The speech recognition experiments were carried out with the publicly available RASR toolkit [34]. To perform the keyword search on the pruned word lattices, weighted finite state transducer based tools provided by IBM were used [35]. For in-vocabulary queries the spoken term detection is performed on word-level, whereas out-of-vocabulary terms are searched in the phonetic form of the lattice after grapheme-to-phoneme conversion. In addition, in some languages the phonetic form of the queries was further expanded with a transducer modeling phone confusions [36]. The KWS was carried out on word-graphs containing about 2000 arcs/sec on average for Cantonese, Pashto, Tagalog, Turkish, and 9000 arcs/sec for Vietnamese. In Vietnamese KWS, the lattices were obtained by decoding with a bigram LM. The speech-to-text performance was measured in token error rate (TER). Depending on the language a token corresponds to a character for Cantonese, a word for Pashto, Turkish, and Tagalog, and a syllable for Vietnamese. The spoken term detection efficiency was measured on the development keyword list in terms of Maximum Term Weighted Value (MTWV) (for details see [37]) after sum-to-one score normalization per keyword [36].

5. EXPERIMENTAL RESULTS

5.1. Effect of the number of languages

In the first set of experiments, the multilingual BN features were trained on either FLPs or LLPs corresponding to the training set condition of the target language. Since similar amounts of data are available in the packs, the joint corpus is well balanced between the languages. First, we compared uni- to multilingual BN features trained on the first four development languages of the project, Cantonese (CA), Pashto (PA), Turkish (TU), Tagalog (TA). Then, experiments were also carried out including the Vietnamese pack (VI) in the BN training. As Table 2 shows, like our previous investigation [11], significant TER improvement can be observed on all languages, although these languages are more diverse. Surprisingly, using three non-tonal languages in the multilingual training gave comparable improvement in Canontese as in Tagalog or Turkish. Including a fifth language led to a consistent but modest gain in TER. Again, although the fifth language was tonal, the Tagalog system showed similar improvement as the Canontese. We also experimented with LLPs and observed similar improvements w.r.t. the number of languages involved in the BN training (results not reported here).

Training the BN on five languages had the advantage that a single set of multilingually trained features could be used on all languages. Besides the speech-to-text, KWS performance was also measured after MPE training. Results in Table 3 indicate that 3-5% relative ASR performance improvement can be achieved through multilingual BN features. On average the KWS results show larger gains of 3-10% relative on FLP and over 10% on LLP.

Table 2. *TER [%] comparison of uni- and multilingual BN features based ASR systems after speaker adaptation on FLPs.*

Language	Id	Unilingual	Multilingual	
			CA+PA +TU+TA	CA+PA +TU+TA+VI
Cantonese	101	42.3	40.7	40.2
Pashto	104	52.8	51.6	51.3
Turkish	105	50.5	49.3	49.2
Tagalog	106	50.5	48.5	48.0

Table 3. *ASR (in TER [%]) and KWS (in MTWV) performance comparison of uni- and multilingual BN features based systems after speaker adaptation and MPE on FLPs and LLPs.*

Language	Id	Set	Unilingual		Multilingual	
			TER	MTWV	TER	MTWV
Cantonese	101	FLP	40.6	0.5528	39.4	0.5677
		LLP	55.7	0.3023	52.5	0.3495
Pashto	104	FLP	52.2	0.4194	50.3	0.4537
		LLP	64.9	0.1899	62.4	0.2186
Turkish	105	FLP	49.5	0.6347	48.1	0.6504
		LLP	65.4	0.3503	62.6	0.4115
Tagalog	106	FLP	49.1	0.4885	46.6	0.5365
		LLP	62.3	0.2801	59.9	0.3116
Vietnamese	106	FLP	50.4	0.4569	48.2	0.4763
		LLP	64.4	0.1996	62.0	0.2436

5.2. Improving LLP results with non-target FLPs

In the previous experiments only LLP data was used for multilingual training for the LLP case. Since the amount of provided transcribed data will be reduced in each period of the project, this limited multilingual training could probably be a realistic scenario only close to the end of the project. However, currently orders of magnitude more transcribed data are at hand in non-target language FLPs. Therefore, in the next experiments we investigated how the unbalanced multilingual corpus influences the KWS and ASR performance on the target language (LLP). In order to verify our results, besides the surprise language of the previous evaluation (Vietnamese) the experiments were repeated on another, non-tonal language of the development package (Pashto). In this latter case, data were borrowed from Cantonese, Turkish, Tagalog, and Vietnamese.

As the results in the 1st, 5th, and 9th rows of Table 4 show, by including FLPs instead of LLPs of other languages in the multilingual training, the TER can be significantly reduced. The improvement over the unilingual features is 9% relative for both languages. The spoken term detection shows even larger gain, due to the better acoustic model we measured 0.1237 absolute MTWV improvement on Vietnamese, and 0.0711 on Pashto. However, the training time for such a hierarchical MLP increased also by orders of magnitude with more training data, as shown in the last column.

5.3. Effect of fine-tuning of the hierarchical BN on the target language data

Since the BN features are trained in a hierarchical manner (2 MLPs after each other), and the multilingual training is carried out jointly on all data, we investigated whether the joint fine-tuning of the hierarchy using only the target language could result in additional gain. The 2nd, 6th and 10th rows of Table 4 show, that the ASR performance improved further. The unilingual systems show only modest improvement in TER from fine-tuning. Fine-tuning the multilingual BN trained on non-target FLPs and target LLP resulted in slight reduction in MTWV on Vietnamese, whereas further TER reduction could still be observed (9-10th rows). On Vietnamese, the degradation in MTWV after the fine-tuning might be related to suboptimal pruning parameters to extract the more dense lattices for KWS (sec-

Table 4. *Effect of exploiting non-target FLPs for LLPs, porting multilingual BN, and fine-tuning with target language. Bold red font indicates the amount of "borrowed" data, and blue the data available in the target language. Results are obtained after SAT and MPE.*

Hierarchical BN trained on	Fine tuning	Language				MLP training time [h]*
		Vietnamese		Pashto		
		TER	MTWV	TER	MTWV	
1 LLP	no	64.4	0.1996	64.9	0.1899	3
	yes	64.0	0.1834	64.3	0.1901	5
4 LLP	no	65.7	0.1940	64.7	0.1845	0
	yes	62.6	0.2498	62.0	0.2241	4
4 LLP+1 LLP	no	62.0	0.2436	62.4	0.2186	16
	yes	60.9	0.2541	61.9	0.2342	19
4 FLP	no	59.7	0.2511	60.0	0.2507	0
	yes	57.6	0.2902	58.6	0.2700	4
4 FLP+1 LLP	no	59.0	0.3233	59.4	0.2610	72
	yes	57.1	0.3170	58.4	0.2770	75

*for the target language, measured on Vietnamese ond and last rows). In contrast, if balanced corpora are used, additional 4-7% relative MTWV improvement is observed (5-6th rows).

5.4. Cross-lingual transfer and initialization

As the previous results show, multilingual BN features resulted in a solid improvement in both TER and MTWV. However, the multilingual training is rather time consuming. Due to the time constraints in the evaluation phase, our concern was also to find a fast development way to obtain the improved BN features. First, we tested the cross-lingual porting of the multilingual BN features trained on non-target languages. The third row in Table 4 shows that porting BN trained on non-target LLPs did not improve the Vietnamese results and only a slight gain in TER is observed in Pashto. Nevertheless, the transfer of multilingual BN trained on FLPs (7th row) shows more than 6% relative TER improvement over the best unilingual features (2nd row). In addition, the cross-lingual porting does not need any MLP training time on the new language. It should be mentioned, in general, the unilingual BN outperforms the cross-lingually ported multilingual features if enough data are available in the target language without acoustic mismatch between the training and testing conditions, see [15, 9] and second and third rows of Table 5. In the case of LLP only 10 hours of speech are available, and might not cover the wide range of the acoustical variation of the telephony speech.

In further experiments, the initialization with multilingually trained NN was investigated. Excluding the target language from the multilingual training enables training the multilingual BN beforehand, saving considerable amount of time during the evaluation phase of the surprise language. For the sake of completeness, these experiments were also carried out with multilingual MLPs where the target language was also included during the BN training. Not surprisingly, after fine-tuning of the ported NNs large gains were observed if the multilingual training was performed without the target language data. Using only LLPs 3.1% and 2.7% absolute TER improvement was measured in Vietnamese and Pashto over the ported BN features (3rd, 4th rows). The fine-tuned BN outperformed the best unilingual BN features by 2-4% relative in TER. Similar observations can be made if more data is borrowed from FLPs (8th row). Although this type of ported features already outperformed the unilingual ones, the performance gap increased further after the fine-tuning step. The TER of the fine-tuned ported features is similar but slightly higher than the results achieved by multilingually trained (and fine-tuned) features where the target language is also involved in the multilingual training (10th row). Regarding the KWS performance, on Vietnamese 0.0906, on Pashto 0.0799 absolute improvement was measured over the best unilingual system with only 4 hours of MLP training time. Although there is still a gap to fill between the performance of the best multilingual features (9th and 10th rows) and the fast developed BN features, the former ones can easily become impractical with larger numbers of languages and stronger time constraints.

6. DISCUSSION

The MLP training time accounting for the fine-tuning step consists of two steps. First, the last hidden layer is initialized by a pretraining step since the target language was not present in the multilingual training. The parameter fine-tuning is then performed on the complete MLP. If the target language was included in the multilingual training then there was no need for initialization. In our implementation the windowing in the middle of the hierarchy is pushed to the input to perform the MLP training with fully randomized training data. Using only segmentwise randomization could significantly reduce the duration of the fine-tuning step. The cross-lingual porting and initialization experiments were also repeated on FLPs. As can be seen in Table 5, similar observations can be made regarding the fast-development steps (4th row) for a new language. Experimental results suggest the following multilingual strategies for the development and surprise languages of the project. Since there are no hard time constraints, the multilingual training can be done for development languages on all non-target language FLPs including the target language (FLP or LLP), followed by a fine-tuning. During the evaluation of a surprise language (for both FLP and LLP), the multilingual BN features previously trained on all available development FLPs can be an effective initialization scheme, and only a fine-tuning step would be necessary to achieve improved ASR and KWS performance.

Table 5. *Porting multilingual BN, and its target language specific fine-tuning on FLPs. Bold red font indicates the amount of "borrowed" data, and blue the available data in the target language. Results are after SAT and MPE.*

Hierarchical BN trained on	Fine tuning	Language				MLP training time [h]*
		Vietnamese		Pashto		
		TER	MTWV	TER	MTWV	
1 FLP	no	50.4	0.4569	52.2	0.4194	16
	yes	49.4	0.4681	51.7	0.4270	29
4 FLP	no	50.9	0.4620	52.4	0.4130	0
	yes	47.4	0.4712	50.7	0.4607	25
4 FLP+1 FLP	no	48.2	0.4763	50.3	0.4537	80
	yes	47.2	0.4765	50.0	0.4658	93

*for the target language, measured on Vietnamese

7. CONCLUSIONS

This study has evaluated recently proposed multilingual BN features on linguistically diverse set of low-resourced languages. Keyword spotting and speech recognition results indicate the superiority of the multilingual features over the unilingual ones, in particular for low target language training resources. From the results where multilingually trained BN features were cross-lingually transferred we can conclude: (1) borrowing orders of magnitude more data from other languages, cross-lingually ported BN can outperform the unilingual BN features if only a limited amount of data is available in the target language; (2) a fine-tuning step can significantly improve the results, clearly outperforming the unilingual system in all cases; (3) although the best performance is achieved if the target language is also included in the multilingual training, comparable results can be obtained after fine-tuning of cross-lingually transferred BN features trained only on non-target languages.

As future work, we intend to extend our multilingual experiments on several new and more diverse language packs (e.g. Zulu and Lao) released in the project recently. This might influence the optimal number of languages involved in the multilingual training.

Acknowledgement The study was 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. This effort used the following IARPA Babel Program language collection releases: babel101-v0.4c (Cantonese), babel104b-v0.4aY (Pashto), babel105b-v0.4 (Turkish), babel106-v0.2f (Tagalog), babel107b-v0.7 (Vietnamese). 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.

8. REFERENCES

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

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

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

[4] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *Proc. of the Tenth International Conference on Machine Learning*, 1993, pp. 41–48.

[5] S. Parveen and P. Green, "Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks," in *Proc. of Eurospeech*, 2003, pp. 1813–1816.

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

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

[8] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8619–8623.

[9] Z. Tüske *et al.*, "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, pp. 7349–7353.

[10] "http://www.iarpa.gov/Programs/ia/Babel/babel.html."

[11] Z. Tüske *et al.*, "Multilingual Hierarchical MRASTA Features for ASR," in *Proc. of Interspeech*, 2013, pp. 2222–2226.

[12] 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.

[13] C. Plahl *et al.*, "Hierarchical Bottle Neck Features for LVCSR," in *Proc. of Interspeech*, 2010, pp. 1197–1200.

[14] Z. Tüske *et al.*, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 6970–6974.

[15] K. Veselý *et al.*, "The language-independent bottleneck features," in *Proc. of IEEE Workshop on Spoken Language Technology*, 2012, pp. 336–341.

[16] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001.

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

[18] N. T. Vu and T. Schultz, "Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families," in *Proc. of Interspeech*, 2013, pp. 515–519.

[19] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

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

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

[22] A. Ghoshal *et al.*, "Multilingual training of Deep-Neural networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 7319–7323.

[23] A. Waibel *et al.*, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[24] K. Vesely *et al.*, "Convolutive Bottleneck Network Features for LVCSR," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 42–47.

[25] R. Schlüter *et al.*, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 649–652.

[26] X. Lei *et al.*, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. of Int. Conf. on Spoken Language Processing*, 2006, pp. 1237–1240.

[27] A. Zolnay *et al.*, "Robust Speech Recognition Using a Voiced-Unvoiced Feature," in *Proc. of Int. Conf. on Spoken Language Processing*, 2002, pp. 1065–1068.

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

[29] F. Seide *et al.*, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 24–29.

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

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

[32] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2002, pp. I–105–I–108.

[33] M. Sundermeyer *et al.*, "On the Estimation of Discount Parameters for Language Model Smoothing," in *Proc. of Interspeech*, 2011, pp. 1433–1436.

[34] D. Rybach *et al.*, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[35] B. Kingsbury *et al.*, "A High-Performance Cantonese Keyword Search System," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8277–8281.

[36] L. Mangu *et al.*, "Exploiting diversity for spoken term detection," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8282–8286.

[37] J. G. Fiscus *et al.*, "Results of the 2006 spoken term detection evaluation," in *Proc. of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51–57.