RWTH LVCSR Systems for Quaero and EU-Bridge: German, Polish, Spanish and Portuguese

M. Ali Basha Shaik[1], *Zoltan Tüske*[1], *M. Ali Tahir*[1],
Markus Nußbaum-Thom[1], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

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

Abstract

In this paper, German, Polish, Spanish, and Portuguese large vocabulary continuous speech recognition (LVCSR) systems developed by the RWTH Aachen University are presented. All the above mentioned systems for the aforementioned languages are used for the Quaero and EU-Bridge project evaluations. The LVCSR systems developed for these competitive evaluations focus on various domains like broadcast news, podcasts and lecture domain. Transcription of the speech for these tasks is challenging due to huge variability in the acoustic conditions and a significant portion of audio data includes spontaneous speech. Good improvements are obtained using state-of-the-art multilingual bottleneck features, minimum phone error trained acoustic models, language model (LM) adaptation and confusion-network based system combination. In addition, an open vocabulary approach using morphemic units is investigated along with the LM adaptation for the German LVCSR.

Index Terms: LVCSR, European, Quaero, EU-Bridge

1. Introduction

This paper describes various details of the LVCSR systems developed by the RWTH Aachen University. All these systems for various languages are used for the Quaero-2013* and EU-Bridge-2014** evaluation campaign. The LVCSR systems developed for these projects focus on transcribing the speech mainly from Broadcast News, Podcasts and lecture data. All tasks involve large vocabularies for speech transcriptions. Transcription of the speech in these evaluations is challenging because of a huge variability in the acoustic conditions and also a large portion of audio includes spontaneous speech. Most of the described systems used for these projects are built upon the data accumulated from previous Quaero evaluations [1]. The major improvements obtained in the present systems compared to our earlier Quaero systems are achieved by using multilingual bottleneck features, open-vocabulary language models, language model adaptation, and system-combination techniques like confusion network combination [2, 3, 4, 5, 6, 7].

For the development of LVCSR systems, using manually transcribed speech data involves significant cost factor. Therefore, methods which are able to reuse out-of-domain or multilingual resources to ease the model training, have growing interest. Neural networks (NN) have become a major component in the state-of-the-art ASR system, and are used to extract features (probabilistic [8] or bottleneck (BN) TANDEM approach [9]) and/or to model the emission probability in the HMM framework directly (hybrid approach) [10].

In [11, 12] it was observed that Multi Layer Perceptron (MLP) based NN posterior features possess language independent properties to a certain degree: the cross-lingual porting of NNs could lead to significant improvement in a different language. In order to exploit resources of multiple languages in acoustic model training, there is usually a need to unify similar sounds across different languages e.g. by IPA or SAMPA. However, as was shown by [13] the training of NNs on multiple languages is possible without such a mapping if language dependent output layers are used and only the hidden layer parameters are shared between the languages. Combining the multilingual learning with the bottleneck approach [14, 2] demonstrated that the multilingual BN features could benefit from the additional non-target language data and outperformed the unilingual BN. Through better generalization the multilingual BN features can offer improved portability on an new language, and acoustical mismatch between the training and testing can be reduced in the target language by exploiting matched data from other languages [15]. Since transcribed lecture data is not provided for the evaluation for most of the languages, in our systems the BN features are trained on large amount of Broadcast News and conversations data of multiple languages. Covering wide variety of acoustic conditions through the multilingual resources, we aimed at improving the robustness of the acoustic model to recognize acoustically less matched lecture data.

On the other hand, word morphology is an important factor to be considered beforehand for a robust language modeling. In contrast to the European languages like Polish, Spanish and Portuguese, German is a morphologically rich language having a high degree of word inflections, derivations and compounding. Therefore, it is typical to observe high out-of-vocabulary (OOV) rates and poor LM probabilities for a German LVCSR system even when large vocabularies are used. Thus, sub-lexical language modeling is used to decrease the OOV rate and reduce the data sparsity [16, 17, 18]. In this work, we also investigate the use of the state-of-the-art LMs like Maximum Entropy (ME) LMs, which incorporate various knowledge sources as features in the sub-lexical LMs. Furthermore, we also experiment the use of Maximum-A-Posteriori (MAP) adaptation on top of the ME LMs for German, Polish and Spanish systems. Thus, the benefits of both the ME LMs and the traditional N-gram backoff LMs are effectively combined using interpolation. For Portuguese LVCSR system, two separate systems are developed us-

* http://www.quaero.org
** http://www.eu-bridge.eu

ing MFCC and PLP features augmented with multilingual bottleneck features. The advantages of both the systems are combined using confusion network based system combination.

2. Acoustic Model (AM)

The acoustic data supplied by the Quaero project (2009-2012) is used for acoustic modeling. The data could be broadly classified into three categories, namely: web data, broadcast news and the European parliament plenary sessions (EPPS) data.

2.1. Resources

Table 1 lists the amount of audio data used for German LVCSR system [1]. Overall, 140 hours of across-domain acoustic training data is used.

Table 1: *Acoustic Training data (Lng.: Language, dur.: duration (hours), seg.:segments, DE: German, PL: Polish, ES: Spanish and PR: Portuguese)*

Lng.	**Corpus**	**#Dur.**	**#Segs**	**# words**
DE	EPPS08 + WEB08 + Quaero 2010+2011+2012	142	29 K	1.5 M
PL	Quaero +Broadcast News	110	29 K	1.0 M
ES	Quaero 2010+2011+2012	390	214 K	4 M
PR	Broadcast News	110	20 K	1.1 M

2.1.1. Cepstral features

16 Mel-cepstral coefficients (MFCCs) are extracted every 10 ms from the audio files. 20 logarithmic critical band energies (CRBE) are computed over a Hanning window of 25 ms. For the piecewise linear vocal tract length normalization (VTLN) text-independent Gaussian mixture classifier was trained to estimate the warping factor (fast-VTLN). After the segment-wise mean and variance normalization, 9 consecutive frames of MFCC are mapped by linear discriminant analysis (LDA) to a 45-dimensional subspace.

2.1.2. Multilingual bottleneck MRASTA features

Multilingual MRASTA features are applied for German, Polish and Spanish tasks. The temporal trajectories of the CRBEs are smoothed by two-dimensional band pass filters to cover the relevant modulation frequency range (MRASTA) [19]. 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. Our final BN features are extracted from hierarchical, MLP based processing of the modulation spectrum [20, 21]. The input of the first MLP contains the fast modulation part of the MRASTA filtering (228 dim.), whereas the second MLP is trained on the slow modulation components (228 dim.) and on 9-frame context of the PCA transformed BN output of the first MLP (9*43). The modulation features fed to the MLPs are always augmented by the CRBE.

Furthermore, robust MLP features are generated using a multilingual training method as described in [13]. The MLP training data uses four languages: German, English, French, and Polish. The final multilingual BN features are trained on $\sim$ 800 hours of speech data from the Quaero project as shown in Table 2. The feature vectors extracted from the joint corpus of the aforementioned four languages are randomized and fed to the MLPs. Using language specific softmax outputs, back propagation is initiated only from the language specific subset of the output depending on the language-ID of the feature vector. The MLPs are trained according to cross-entropy criterion, and approximate 1500 tied-triphone state posterior probabilities per each language [22]. The BN features were reduced by PCA acounting for 95% of the total variability (38 dim.) To prevent over-fitting and for adjusting the learning rate parameter, 10% of the training corpus is used for cross-validation. The BN features of the German, Polish, Spanish and Portuguese evaluation systems are based on deep MLP. The size of the 6 non-BN hidden layers was set to 2000, the bottleneck layers consisted of 60 nodes, before the last hidden layer.

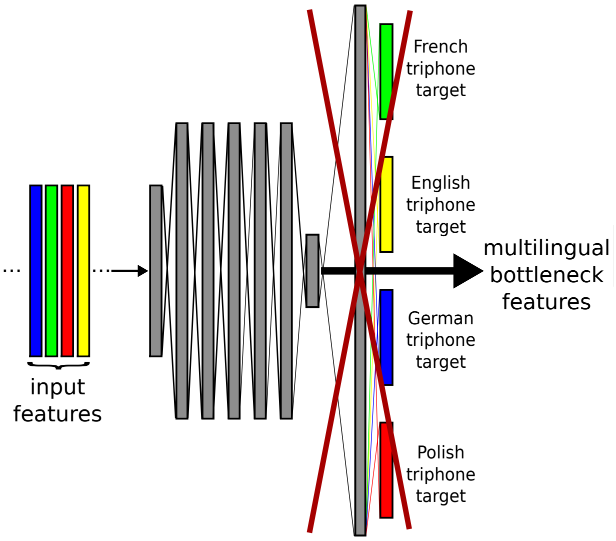

Figure 1: *The joint training of deep context-dependent bottleneck MLP features on multiple languages (DE, FR, EN, PL). The different colors indicate different languages, and language dependent back-propagation from the output layer. The other parts of the network including the bottleneck layer are shared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

Table 2: *Multilingual broadcast news and conversation resources used for BN feature training.*

language	German	English	French	Polish
Duration of speech [h]	142	232	317	110

2.2. AM Training

In our systems, the outputs of a neural network are used as input features for a Gaussian mixture model (GMM). The final 83-dimensional feature vectors are obtained by concatenating the spectral features with the multi-layer-perceptron (MLP) features described in 2.1.1. The acoustic models (AM) training followed similar recipes, the GMMs have been trained according to the maximum likelihood (ML) criterion with the expectation maximization algorithm (EM) with Viterbi approximation and a splitting procedure. The GMMs have a globally pooled, diagonal covariance matrix. $4,500$ generalized triphones determined by a decision-tree-based clustering (CART) are modeled in both languages.

Table 3: *Text Resources (# words: Running words)*

ln.	corpus	# words
DE	TAZ	151M
	German News	155 M
	Blogs+Web	797 M
	Call-Home	5.9 M
	Multilingual Parallel data	104 M
	Lectures	5.0 M
	News + acoustic trans.	971 M
	EU-Bridge in-domain data	8 K
PL	Kurier Lubelski- News	77 M
	Nowosci - News	460 M
	Blogs+Web	710 M
	News	120 M
	Official EU-Bridge data	350 K
	Indomain+ Quaero+ EU-Bridge acoustic trans.	31 M
ES	Gigaword	1.18 B
	Quaero	600 M
	Quaero+Translectures* +EPPS acoustic transcriptions	14 M
PR	Euro-News+ acoustic trans. +In-domain	14 M

*http://www.translectures.eu

Table 4: *Recognition corpus statistics (***ln:** *language,* **crp:** *corpus,* **dur:** *duration in hours,* **vocab:** *vocabulary size,* **OOV:** *effective out-of-vocabulary rate,* **prj:** *project Quaero (QRO) or EU-Bridge (EUB***))*

ln.	prj	domain	crp	dur. (hrs)	vocab	OOV [%]
DE†	QRO	News +podcasts	dev	3.1	300k	0.20
			eval	3.4		0.13
		Lectures (OMTP‡)	dev	3.3	375k	0.25
			eval	3.2		0.46
	EUB	Euro-News	dev	0.5	300k	0.18
			eval	0.6		–
PL	QRO	News +podcasts	dev	3.1	600k	1.10
			eval	3.3		1.30
	EUB	Euro-News	dev	0.5		1.83
			eval	0.5		–
ES	QRO	News +podcasts	dev	3.6	320k	0.45
			eval	3.5		0.42
PR	EUB	News +podcasts	dev	0.36	171k	0.0
			eval	0.49		–

†: Sub-lexical systems are used

‡ OMTP: Online Multimedia Translation Platform

* 2014 dryrun evaluation campaign

2.3. Speaker Adaptation

Several speaker adaptation techniques are used in our tasks. First, mean and variance normalization has been applied to the spectral features. We also applied a vocal tract length normalization (VTLN) to the MFCC features. As an additional pass, speaker adaptation using constrained maximum likelihood linear regression (CMLLR) [23] with a simple target model approach is used [24]. Speaker Adaptive Training (SAT) is performed ie., by applying the CMLLR transformation to the training data to generate new GMMs. We applied speaker adaptation for all the tasks in this paper.

Table 5: *Recognition results (***Ln.***: Project-Language ID,* **FW***: full-word system,* **SW***: sub-lexical system,* **Crp***: corpus,* **BN***: Broadcast News + podcasts corpus,* **OMTP***: Multimedia corpus (eg.: lectures),* **PPL** *: perplexity,* **CNC***: confusion network combination,* **Ord***: LM Order,* **AM***: Acoustic Model,* **ML***: Maximum Liklihood AMs (2^{nd} pass),* **BO***: backoff LM,* **ME:** *BO Interpolated Maximum Entropy LM,* **sp:** *supervised adapted MELM,* **usp:** *unsupervised adapted MELM)*

Prj.	Expt.	Crp	AM	LM	LM Adap	Ord	PPL [%]	WER [%]
QUAERO - DE*	SW (BN)	dev	ML	BO	no	5	255.5	16.1
				ME		4	252.6	
		eval		BO	no	5	324.0	14.5
				ME		4	321.0	
					usp		308.1	**14.4**
QUAERO - DE*	SW (OMTP)	dev	ML	BO	no	5	435.8	21.3
				ME		4	435.4	
		eval		BO	no	5	523.9	25.7
				ME		4	515.7	25.6
					sp		491.8	25.3
					usp		477.5	**25.2**
QUAERO - PL*	FW (BN)	dev	ML	BO	no	5	656.1	12.5
			MPE					11.8
				ME		4	645.3	11.7
		eval	MPE	BO	no	5	658.8	13.7
				ME		4	640.3	13.5
					sp		601.9	**13.2**
					usp		618.5	13.3
QUAERO - ES*	FW (BN)	dev	ML	BO	no	4	189.3	15.0
			MPE					14.1
				ME			182.3	14.0
		eval	MPE	BO	no		187.5	**13.1**
				ME	usp		187.1	
EUB- DE*	**project******-baseline**	dev	–	–	–	–	–	23.6
		eval						20.8
	RWTH SW (BN)	dev	ML	BO	no	5	334.8	9.3
		eval					–	**11.8**
EUB- PL*	**project******-baseline**	dev	–	–	–	–	–	28.3
		eval						18.4
	RWTH FW (BN)	dev	MPE	BO	no	5	431.2	11.9
		eval					–	**9.4**
EUB - PR*	**project******-baseline**	dev	–	–	–	–	–	35.8
		eval						28.0
	MFCC	dev	ML	BO	no	5	348.9	25.3
	PLP							25.8
	CNC (BN)							24.6
		eval					–	**18.7**

*Best system in-terms of the WER in the evaluation.

**Project baseline

2.4. Minimum Phone Error (MPE) Models

In general, MPE based discriminative training provides additional gain in terms of WER compared to conventional Maximum Liklihood (ML) training [25, 26]. The initial Gaussian acoustic model has been further trained discriminatively by MPE training. There are $r = 1, ..., R$ training utterances each with transcription W_r and feature sequence $X_r = x_1, ..., x_{T_r}$ of length T_r. The sentence-level minimum phone error criterion incorporates an accuracy score $A(W, W_r)$, which is the phone transcription accuracy of hypothesis sentence W given the reference sentence W_r [27]. This is nearly equal to the number of reference phones minus the number of errors. Therefore :

$$\mathcal{F}_{MPE}(\Lambda) = -\tau_\Lambda ||\Lambda - \Lambda_0||^2 + \sum_{r=1}^{R} \sum_{W \in \mathcal{M}_r} P_\Lambda(W|X_r) A(W, W_r) \tag{1}$$

$$P_\Lambda(W_r|X_r) = \frac{\left(p(W_r)^{\frac{1}{\eta}} \cdot p_\Lambda(X_r|W_r)\right)^\beta}{\sum_{W \in \mathcal{M}_r} \left(p(W)^{\frac{1}{\eta}} \cdot p_\Lambda(X_r|W)\right)^\beta} \tag{2}$$

$$p_\Lambda(X_r|W) = \max_{s_1^{T_r}|W} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) p(x_t|s_t) \right\} \tag{3}$$

Center regularization is used in MPE training, which loosely binds Λ to their initial values Λ_0. $\mathcal{M}_r$ is the set of all possible word sequences, η is a language model scale, and β is a posterior scale. $p(W)$ are word sequence prior probabilities obtained from language model. $p(x|s)$ is the emission probability of feature x given acoustic model HMM state s, and $p(s_t|s_{t-1})$ is the transition probability from state s_{t-1} to state s_t.

3. Corpus statistics

The development and evaluation corpus statistics for all the experimented languages are shown in Table 4. OOV rates and perplexities are not shown in Tables 4 and 5 for EU-Bridge eval corpora, as they are not released by the project committee.

4. Language Model (LM)

The LM text is collected from various sources like Broadcast News, Podcasts, Blogs, Web and Audio-transcriptions. The noisy raw text is normalized using language dependent set of rules and semi-automatic methods. Different types of data used to generate a LM are shown in Table 3. Vocabulary is generated based on the frequency of the words and then domain specific backoff N-gram LMs are created. Heldout corpus is used to generate the interpolation weights and domain adapted LMs are created using using linear interpolation [28].

For German LVCSR sub-lexical experiments, Morfessor is used to decompose the words [29]. Low frequency words are excluded while generating the Morfessor model. The decomposed words are post-processed to produce a cleaner set of sub-lexical units and boundary markers are added to regenerate full-words later after recognition. Very short units are avoided as they are usually difficult to recognize and also could harm the overall WER with more insertion errors [30, 31]. To generate N-gram backoff sub-lexical LMs, different hybrid vocabularies are selected, where top-most 5k full-word forms are preserved. For OMTP system, two different domain specific language models are selected for interpolation. The first language model consists of largely BN domain, where as the second language model consists of lectures domain. It is observed that interpolated language model performed better than standalone lectures domain language model in terms of perplexity and thus, is used during recognition. 5-gram backoff LMs are created for all the systems, but for Spanish. Alternatively 4-gram Maximum Entropy (ME) language models are created along with language model adaptation. The language model adaptation uses MAP principle [32]. In adaptation, the parameters estimated from the across domain data are used as the prior means to learn the parameters from the in-domain data [33]. Here, the development data is used as an in-domain data in a supervised adaptation. Similarly, the automatic transcriptions generated from the initial pass are used as an in-domain data in an un-supervised adaptation [3]. Both the backoff N-gram LMs and adapted/non-adapted ME language models are linearly interpolated for robust probability estimates.

5. Recognition Results

The recognition systems have a multi-pass setup. After an initial non-adapted pass, transcriptions are obtained which are used for the CMLLR-adapted recognition pass. After speaker adaptation pass, ME language models are used during N-best list rescoring. N-best (N=5000) list is selected based on the optimal WER on dev corpus. Alternatively for Portuguese system, confusion network based system combination is used to combine the results of both MFCC and PLP system. As a baseline EU-Bridge systems, initial baseline WERs are provided by the Fondazione Bruno Kessler (FBK) research group, Italy. The description of these baselines are not released by the project committee.

Recognition results are shown in Table 5. For German BN system, it is observed that the WERs are better than the OMTP system. This is mainly due to the presence of spontaneous speech in OMTP corpus. For all the systems where ME language models are used, limited reductions are observed in terms of perplexity and the same effect is also reflected in WER. MAP based language model adaptation provided significant gains for German and Polish tasks in-terms of the WER. MPE trained acoustic models provided noticeable improvements for Polish and Spanish tasks. Confusion network combination helped to achieve significant gains for Portuguese task. The following reductions in WER are achieved in comparison with the project baselines for the EU-Bridge systems, summarized as : German LVCSR [eval: ≈ abs: 9.0 %, rel: 43.3%], Polish LVCSR [eval: ≈ abs: 9.0%, rel: 48.9%] and Portuguese LVCSR [eval: ≈ abs: 9.3 %, rel: 33.2 %].

6. Conclusions

Multiple LVCSR systems developed for European languages across BN and lecture domains for the evaluations of Quaero and EU-Bridge projects are presented. Acoustic level multilingual features using neural networks, domain dependent language modeling, supervised and unsupervised adaptation and system combination of subsystems were investigated. On the whole, noticeable improvements are obtained mainly due to the use of multilingual features, MPE trained acoustic models, open-vocabulary language models, language model adaptation and confusion network based system combination. For morphologically rich languages like German and Polish, significant reductions in WER are achieved using open vocabulary language modeling along with the language model adaptation. The RWTH LVCSR systems for the languages described in this work all ranked first in both the Quaero 2013 and EU-Bridge 2014 evaluation campaigns.

7. Acknowledgements

This work was partly funded by the European Community's 7^{th} Framework Programme under the project EU-Bridge (FP7-287658) and partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation. Hermann Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

8. References

[1] 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 *Interspeech*, Makuhari, Chiba, Japan, Sep. 2010, pp. 1517 – 1520.

[2] Z. Tüske, R. Schlüter, and H. Ney, “Multilingual Hierarchical MRASTA Features for ASR,” in *Interspeech*, Lyon, France, Aug. 2013, pp. 2222–2226.

[3] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, “Investigation of Maximum Entropy Hybrid Language Models for Open Vocabulary German and Polish LVCSR,” in *Interspeech*, Portland, OR, USA, Sep. 2012.

[4] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, “Using Morpheme and Syllable Based Sub-words for Polish LVCSR,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4680 – 4683.

[5] B. Hoffmeister, “Bayes risk decoding and its application to system combination,” Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, july 2011.

[6] M. Shaik, Z. Tüske, S. Wiesler, M. Nußbaum-Thom, S. Peitz, R. Schlüter, and H. Ney, “The RWTH Aachen German and English LVCSR systems for IWSLT-2013,” in *The International Workshop on Spoken Language Translation*, Heidelberg, Dec. 2013.

[7] G. Evermann and P. Woodland, “Posterior Probability Decoding, Confidence Estimation And System Combination,” in *Proceedings of the NIST and NSA Speech Transcription Workshop*, College Park, MD, USA, 2000.

[8] H. Hermansky, D. P. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Jun. 2000, pp. 1635 – 1638.

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

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

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

[12] 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. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 321–324.

[13] 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*, Brisbane, Australia, Sep. 2008, pp. 2711–2714.

[14] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, “The language-independent bottleneck features,” in *IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, Dec. 2012, pp. 336–341.

[15] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, “Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 7349–7353.

[16] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, “Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR,” in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1441 – 1444.

[17] M. Bisani and H. Ney, “Open Vocabulary Speech Recognition with Flat Hybrid Models,” in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 725 – 728.

[18] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, “Sub-lexical language models for German LVCSR,” in *IEEE Workshop on Spoken Language Technology*, Berkeley, CA, USA, Dec. 2010, pp. 159 – 164.

[19] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR,” in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 361–364.

[20] F. Valente and H. Hermansky, “Hierarchical and parallel processing of modulation spectrum for ASR applications,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Mar. 2008, pp. 4165–4168.

[21] C. Plahl, R. Schlüter, and H. Ney, “Hierarchical Bottle Neck Features for LVCSR,” in *Interspeech*, Makuhari, japan, Sep. 2010, pp. 1197–1200.

[22] Z. Tüske, R. Schlüter, and H. Ney, “Deep hierarchical bottleneck MRASTA features for LVCSR,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 6970–6974.

[23] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[24] G. Stemmer, F. Brugnara, and D. Giuliani, “Adaptive training using simple target models,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, Pennsylvania, USA, Mar. 2005, pp. 997–1000.

[25] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, United States ed. Prentice Hall, 1993.

[26] M. A. Tahir, M. Nußbaum-Thom, R. Schlüter, and H. Ney, “Simultaneous Discriminative Training and Mixture Splitting of HMMs for Speech Recognition,” in *Interspeech*, Portland, OR, USA, Sep. 2012.

[27] D. Povey, “Discriminative training for large vocabulary speech recognition,” Ph.D. dissertation, Cambridge, England, 2004.

[28] A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.

[29] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0,” Computer and Information Science Helsinki University of Technology, Finland, Tech. Rep., Mar. 2005.

[30] K. L. de Ipiña, I. Torres, L. Oñederra, A. Varona, and L. J. Rodríguez, “Selection of sublexical units for continuous speech recognition of Basque,” in *Interspeech*, Beijing, China, oct 2000, pp. 544 – 547.

[31] K. L. de Ipiña, N. Ezeiza, G. Bordel, and M. Graña, “Morphological segmentation for speech processing in Basque,” in *Proceedings of IEEE Workshop on Speech Synthesis*, Santa Monica, USA, sept 2002, pp. 187 – 190.

[32] C. Chelba and A. Acero, “Adaptation of maximum entropy capitalizer: Little data can help a lot,” *Computer Speech and Language*, vol. 20, no. 4, pp. 382 – 399, 2006.

[33] T. Alumäe and M. Kurimo, “Efficient Estimation of Maximum Entropy Language Models with N-gram features: an SRILM extension,” in *Interspeech*, Chiba, Japan, September 2010.