The RWTH Aachen German and English LVCSR systems for IWSLT-2013

M. Ali Basha Shaik[1], *Zoltan Tüske*[1], *Simon Wiesler*[1],
Markus Nußbaum-Thom, Stephan Peitz, Ralf Schlüter[1] *and 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, wiesler, nussbaum, peitz, schlueter, ney}@cs.rwth-aachen.de

Abstract

In this paper, German and English large vocabulary continuous speech recognition (LVCSR) systems developed by the RWTH Aachen University for the IWSLT-2013 evaluation campaign are presented. Good improvements are obtained with state-of-the-art monolingual and multilingual bottleneck features. In addition, an open vocabulary approach using morphemic sub-lexical units is investigated along with the language model adaptation for the German LVCSR. For both the languages, competitive WERs are achieved using system combination.

1. Introduction

This paper describes in detail the German and English RWTH large vocabulary continuous speech recognition recognition systems developed for the IWSLT-2013 evaluation campaign. Automatic speech recognition track in IWSLT-2013 evaluation campaign focuses on transcribing lecture data. One of the major challenge in the IWSLT-2013 evaluation is that no acoustic modeling training data is provided for the aforementioned languages, but the development data. The data includes speech types like lectures, talks and conversations. Recognition on the data is challenging because of a huge variability in the acoustic conditions and a large portion includes spontaneous speech.

In the development of ASR systems transcribed speech data is still a significant cost factor. Therefore, methods which are able to reuse out-of-domain or multilingual resources to ease the model training, have growing interest, and this demand exists not only for under-resourced languages. The neural networks (NN) have become a major component in the state-of-the-art ASR system, and are used to extract features (probabilistic [1] or bottleneck (BN) TANDEM approach [2]) and/or to model the emission probability in the HMM framework directly (hybrid approach) [3]. In [4, 5] 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 [6] 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 [7, 8] 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 [9]. Since transcribed lecture data were not provided for the evaluation, 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, German is a morphologically rich language having a high degree of word inflections, derivations and compounding. For a morphologically rich language like German, high out-of-vocabulary (OOV) rates and poor LM probabilities are generally observed. Thus, sub-lexical language modeling is used to decrease the OOV rate and reduce the data sparsity [10, 11, 12]. In this work, we also investigate the use of the state-of-the-art LMs like Maximum Entropy (MaxEnt) LMs, which provide modular structure to incorporate various knowledge sources as features in the sub-lexical LMs. Furthermore, we experiment the use of Maximum a-posteriori (MAP) adaptation over the MaxEnt LMs. Thus, the benefits of both the MaxEnt LMs and the traditional N-gram backoff LMs are effectively combined using interpolation, followed by confusion network based system combination.

The rest of the paper is organized as follows: In Section 2 speaker independent and dependent acoustic models are described along with the investigated features. In Section 3, the use of various full-word and sub-lexical language models are investigated. In Section 3.7, the generation of the lexicon is described. In Section 4, various recognition setups are described. Results are discussed in Section 5, followed by conclusions.

2. Acoustic Model (AM)

In this work, the data from the Quaero project is used for acoustic modeling. The training data for the IWSLT-2013 evaluation campaign consist of data from three domains. While the majority of the data is from the web (WEB), data from broadcast news (BN) and European parliament plenary sessions (EPPS) is also covered.

2.1. Resources

2.1.1. German

Table 1 lists the amount of audio data used from different domains [13] for German LVCSR . Overall, 140 hours of across-domain acoustic training data is used. The data includes the audio from BN, EPPS and the web domains.

Table 1: *Acoustic Training data (dur.: duration (hours), seg.:segments)*

Corpus	#Dur.	#Segs	# Running words
EPPS08	5	1109	45,796
WEB08	14	3452	127,086
Quaero 2010+2011+2012	123	25061	1,391,468

2.1.2. English

Similarly, Table 2 lists the amount of audio data, which is collected from different domains. Overall, 142 hours of acoustic training data is used [13]. The *HUB4* and the *TDT4* corpora contain only American English Broadcast News, whereas the *TC-STAR* corpus consists of European Planery Parliamentary Speech data.

Table 2: *Acoustic Training data (dur.: duration (hours), seg.:segments)*

Corpus	# Dur.	#Segs	# Running words
Quaero	268	57,629	1,666,733
HUB4	206	119.658	1,617,099
TDT4	186	110.266	1,715,445
EPPS	102	66,670	761,234
TED	200	21,614	1,857,660

Table 2 lists the amount of audio data used for acoustic model training. The largest database is the English Quaero corpus[1], which consists of 268 hours transcribes web podcasts. HUB4 and TDT4 are American English broadcast news corpora. EPPS consists of 102 hours of English European Parliament speeches.

All this data has in common that it is out-of-domain for a lectures recognition system. Therefore, we downloaded 200 hours videos from the TED website[2]. All videos have been uploaded to the TED website before the IWSLT cut-off date December31 2010. We used the video subtitles as transcriptions. We used a low pruning threshold for aligning the data and discarded the segments which could not be aligned. In total, we used 962 hours audio training data with a mix of British and American English and from various domains.

2.2. Feature Extraction

2.2.1. Cepstral features

From the audio files 16 Mel-cepstral coefficients (MFCC) were extracted every 10 ms. The 20 logarithmic critical band energies (CRBE) were 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 were mapped by linear discriminant analysis (LDA) to a 45-dimensional subspace.

2.2.2. Multilingual bottleneck MRASTA features

For both evaluation systems the same multilingual MRASTA features are applied. The original RASTA filters were introduced to extract features which are less sensitive to linear distortion [14]. According to [15], the temporal trajectories of the CRBEs were smoothed by two-dimensional bandpass filters to cover the relevant modulation frequency range (MRASTA). 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 [16, 17]. 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 PCA transformed BN output of the first MLP. The modulation features fed to the MLPs were always augmented by the CRBE.

Furthermore, in order to extract robust MLP features a multilingual training method proposed by [6] is applied. The MLP training data covered four languages — English, French, German, and Polish —, and the final multilingual BN features are trained on $\sim$ 800 hours of speech data collected within the Quaero project as shown in Table 3. The multilingual corpus incorporates the complete German and part of the English resources described in Subsection 2.1.1 and 2.1.2. The feature vectors extracted from the joint corpus of the four languages were 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 [18]. 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 evaluation systems were based on deep MLP. The size of the 6 non-BN hidden layers was set to 2000, the bottleneck layers consisted of 60 nodes and was always placed before the last hidden layer.

[1] http://www.quaero.org/

[2] www.ted.com

Table 3: *Multilingual broadcast news and conversation resources used for* BN *feature training.*

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

In addition, four additional experiments were carried out to select the best MLP features for German LVCSR : In the classical (shallow) 5-layer uni- and multilingual BN networks the hidden layers had 7000 nodes. In deep BN, making the last hidden layer language dependent (4x2000) increased the number of trainable parameters and did not increase the MLP training time. On the contrary, testing a single large hidden layer (8000 nodes) after the BN increased the number of parameters even further, and resulted in longer training time. The final submissions are based on this later BN structure, one level of the hierarchy is also shown in Fig. 1.

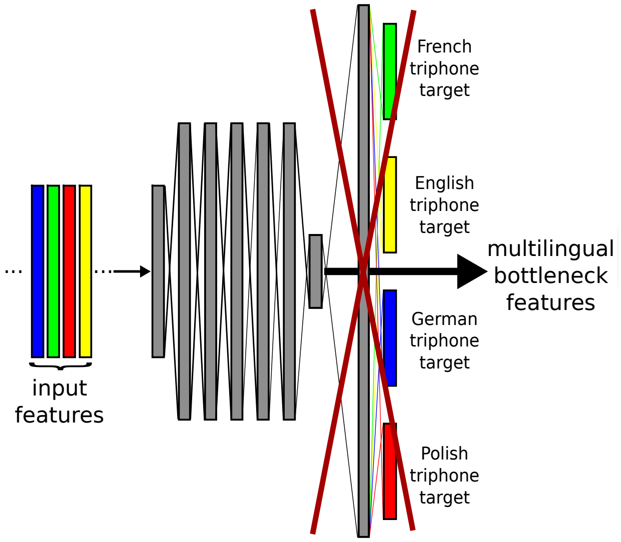

Figure 1: *The joint training of deep context-dependent bottleneck MLP features on multiple languages (FR, EN, DE, PL). The different colors indicate different languages, and language dependent back-propagation from the output layer. The other parts of the network including the bottleneck layer are shared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

2.3. AM Training with Speaker Adaptation

The English acoustic models have been trained on the complete data as described in Subsection 2.1.2, whereas the German acoustic models are built using mostly Quaero data as described in Subsection 2.1.1.

All our systems are based on a bottleneck tandem approach, i.e., the outputs of a neural network are used as input features for a Gaussian mixture model (GMM). The final 83-dimensional feature vectors were obtained by concatenating the spectral features with the multi-layer-perceptron (MLP) features described in 2.2.1. 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.

Speaker adaptation is of crucial importance for the performance of a lecture recognition system. If significant amount of audio data is available along with the speaker related information, this helps to capture the speaker variabilities and helps in reduction of the WER. Several speaker adaptation techniques are used in our system. First, mean and variance normalization has been applied to the spectral features. Furthermore, we applied a vocal tract length normalization (VTLN) to the MFCC features. The VTLN warping factors were obtained by performing a grid search on the audio training data. A Gaussian classifier has been trained on the results and applied to the training and recognition data to obtain the VTLN-transformed features. In addition, speaker adaptation using constrained maximum likelihood linear regression (CMLLR) [19] with the simple target model approach [20] is applied. The CMLLR transformation is applied to the training data and a new GMM is trained (speaker adaptive training). In recognition, the CMLLR transforms are estimated from a first recognition pass and then, a second recognition pass with the GMM from speaker adaptive training (SAT) is performed. The speaker labels required for for CMLLR adaptation were obtained by clustering speech segments optimizing the Bayesian information criterion [21]. Both the speaker independent and adaptive GMM models ended up over 1M densities. This is referred as common system for both English (system-1) and German LVCSRs.

In addition to the system described above, for English a second system (system-2) is trained which uses the MLP features of our IWSLT-12 submission [22]. These MLPs were only trained on the English Quaero data and have less layers. In order to improve system variability, we also performed an additional recognition pass with maximum likelihood linear regression (MLLR) [19]. In our experience, MLLR does not improve performance of Tandem systems, but it may be advantageous to have an MLLR system in the system combination.

3. Language Model

3.1. Resources

The distribution of words in any spoken language is captured by the LM text. The LM text is collected from various domains. Relatively as more amount of acoustic training data is available for BN than for EPPS and since the BN domain could be closer to the web domain than parliamentary speeches, we decide to build an American English BN AM and a British English EPPS AM in order to get better domain dependent modeling. For the training of the LM we apply a similar approach, as domain dependent LM data is used. The text is normalized using language dependent predefined set of rules and semi-automatic methods. For example, Dates and Roman numerals are converted into text format. Punctuation's are discarded. In this paper, LM text is used for both the German and English LVCSR task as recommended by the IWSLT evaluation committee[3], as shown in Table 4

Table 4: *Text Resources for German and English LVCSR*

Lang	Corpus	# Running words
DE	Podcast	46k
	IWSLT LM data	2.5M
	Lecture Talks	2.5M
	CALL HOME - speech	5.9M
	Multilingual Parallel data	104M
	Web	384M
	News + acoustic trans.	971M
EN	IWSLT LM data	3M
	WMT 2012 news-commentary	5M
	Acoustic transcriptions	8M
	WMT 2012 news-crawl	2.8B
	Gigaword corpus	3B

3.2. Backoff LM

As described in Table 4, the LM text is collected from multiple sources. The top N most frequent words are selected as a vocabulary from the full-word text. For English, 150k most frequent words are used to generate modified Kneser-Ney smoothed 4-gram and 5-gram full-word LMs . Similarly for German, 150k and 200k full-word vocabularies are selected to generate 5-gram LMs.

3.3. Sub-lexical LMs

For an open vocabulary speech recognition, sub-lexical units are used in the language modeling for German LVCSR [11]. In general, a LM comprising sub-lexical units with or without a fraction of full-words is called a sub-lexical LM. In general, morphemes could be extracted using linguistic or data-driven morphological decomposition. When sub-lexical LMs are used, the data sparsity problem is relatively reduced compared to the full-word LMs, leading to lower OOV rates and higher lexical coverage. Furthermore, as the count based statistics are improved, the LM probability estimates are relatively better estimated compared to a full-word LM [10, 11, 12].

In this work, words are decomposed using a Morfessor [23]. Word decomposition model is trained using unique words that occur more than 5 times in the LM text. Low frequency words are excluded to avoid noise that are harmful during training. This model is also used to decompose new words. The decomposed words are processed so as to produce a cleaner set of sub-lexical units and to avoid very short units which are usually difficult to recognize. This is found to be helpful to improve the final WER. To generate sub-lexical LMs, 200k hybrid vocabulary is selected, where topmost 5k full-word forms are preserved. Standard N-gram backoff models are created using SRILM toolkit [24].

3.4. Maximum Entropy LMs

Alternatively, for German LVCSR, state-of-the-art MaxEnt LM is generated to capture the long range dependencies [25]. In principle, MaxEnt LM uses the information obtained from multiple knowledge sources as feature constraints. The knowledge sources could be different types of features having different constraints (i.e., probability distribution functions). MaxEnt LM estimates a unified model in a feature space by selecting the distribution function of the highest entropy satisfying all the constraints from an intersection of all the imposed feature constraints. If w is a word/morpheme taken from a vocabulary W, $f(.)$ is the feature function, λ is an optimal weight, h is the context, $Z(h)$ is the normalization factor for all the seen contexts, MaxEnt model can be computed using Eq. 1.

$$p_{me}(w|h) = \frac{e^{\sum_i \lambda_i f_i(w,h)}}{Z(h)} \qquad (1)$$

$$\text{Where, } Z(h) = \sum_{w_i \epsilon W} e^{\sum_j \lambda_j f_j(w_i,h)}$$

3.5. Adaptation

In general, adapted LMs are known to perform better than non-adapted LMs in cases of domain mis-match or if the LM corpus is diverse. In this paper, the LM data is obtained from multiple domains for LVCSR. It is often unrealistic to significantly reduce the WER without adapting the LM to in-domain data [26]. For this reason, we apply LM adaptation over MaxEnt LMs. Here, Maximum a-posteriori (*MAP*) adaptation is performed, using Gaussian priors over the generated MaxEnt models (cf. Section 3.4). The MaxEnt model is trained on background data including the N-gram features of the in-domain data. The prior parameters computed from the background data are used to learn the parameters from the in-domain data. During MaxEnt training, the prior has zero mean during Gaussian prior smoothing. But during adaptation, the prior distribution is centered at the background data parameters. The regularized log-likelihood of the adaptation training data is maximized during adaptation.

As an in-domain data, two different types of adaptation, namely supervised and unsupervised are investigated [25]. In supervised adaptation, the development data is used as an in-domain data. Whereas, for an unsupervised adaptation, the automatic transcriptions are used from the first pass recognition. Here, the adaptation is performed over both morpheme and feature based MaxEnt models. The 5-gram MaxEnt and adapted models are created using SRILM-extension [27].

In general, N-gram backoff LMs are known to perform better in capturing the short range context dependencies.

[3] http://www.iwslt2013.org/59.php

When the data is sufficiently available, the likelihood estimates of the frequently occurring N-grams are generally better estimated and reliable. In this work, morphemic MaxEnt LMs are linearly interpolated with N-gram LMs [28].

3.6. Perplexity

Perplexity is a entropy related metric which measures the average branching factor for the LM, during search. On the other hand, perplexities across various systems can only be compared when the (same) finite vocabulary is used. The word level standard equation of the perplexity (PP_w) in log domain is :

$$PP_w(w_1^k) = log \left[\prod_{l=1}^{K} p(w_l|w_h)\right]^{-\frac{1}{K}} \tag{2}$$

Thus, Eq. 2 is renormalized using at character level as:

$$PP_c(w_1^k) = log \left[\prod_{l=1}^{K} p(w_l|w_h)\right]^{-\frac{1}{K}\frac{K}{K_c}} \tag{3}$$

Where, K is the total number of words observed in the recognition corpus. K_c represents the actual number of characters including word boundaries and a representative character per sentence-end token. Thus, using Eq. 3, full-word LM and the sub-lexical LM could be easily compared.

3.7. Lexical Modeling

The full-word lexicon consists of 150k words for English LVCSR. Similarly, lexicons consisting of 150k and 200k full-words are generated for German LVCSR. For most of the full-words as the pronunciations are not available, statistical grapheme-to-phoneme (G2P) conversion toolkit is used for both the languages [29]. The full-word pronunciations are aligned to its corresponding sequence of morphemic sub-lexical units using the expectation-maximization (EM) algorithm as described in [12]. Thereby, lexicon is generated using the sub-lexical entries of size 200k.

3.8. Word Reconstruction

For sub-lexical experiments, full-words are needed to be reconstructed from the morphemes. An identifier '+' is marked at the end of each non-boundary morpheme. After recognition, the recognized morphemes are combined using the predefined marker to regenerate the full-words. For example: *wasch+ masch+ ine* → *waschmaschine* (washing machine in English). Alternatively, the effective OOV rate of any corpus is computed in such a way that a word is considered an OOV if and only if it is not found in the vocabulary and it is not possible to compose it using in-vocabulary sub-lexical units.

4. Recognition Setup

The evaluation systems have a multi-pass recognition setup. In an initial non-adapted pass, a first transcription is obtained, which is used for the CMLLR-adapted recognition pass. The development and evaluation corpus statistics for both the languages are shown in Table 5.

Table 5: *Details of the IWSLT-13 Recognition Corpus*

Language	Corpus	#Duration (hrs.)
English	dev2012	2.0
	tst2011	1.3
	tst2012	2.2
	tst2013	4.8
German	dev2012	3.3
	tst2013	3.2

For the English LVCSR system, CMU segmentation is used [30]. A 4-gram domain adapted backoff LM is created to construct the search space and 5-gram LM is used for rescoring word lattices. For our alternative system (system 2), an non-adapted and a CMLLR-pass are performed as in system-1. In addition, a third recognition pass with MLLR adaptation is performed. Finally, the word lattices are rescored. Confusion network based system combination is used to combine the results of both systems.

For the German LVCSR system, two different systems are experimented with LIUM [31] and RWTH audio segmentation [32]. 5-gram domain adapted backoff LM is created to construct the search space. This recognition setup is similar to the system-1 of the English LVCSR. After the speaker adaptation, N-best (N=5000) lists are generated from the lattices for LM rescoring. The N-best lists are rescored using the interpolated LMs as described in Section 3.5. Similarly, the advantages of both the full-word and the sub-lexical systems are combined using confusion network decoding.

5. Results

In this Section, detailed results for the various systems are described in terms of the WER and the OOV rates. For both the languages, WERs for the development corpus are generated using the unofficial scoring script, where as the WERs for the evaluation corpus are obtained using official scoring script. For English LVCSR system, the recognition results are shown in Table 6. The WER of system-1 is better than system-2. Significant improvements are obtained using speaker adapted acoustic models over the speaker independent models. Further improvements are obtained using confusion network decoding. In addition, noticeable WERs are reported on the tst2011 and tst2012 corpora for IWSLT-2013 evaluation, compared to our previous IWSLT-2012 WERs for English LVCSR as shown in Table 7. Test transcriptions are not released by the IWSLT-13 evaluation committee, yet.

For the German LVCSR system, the first set of experiments are shown in Table 8. The BN features used in the evaluation system were optimized using the 200k sub-lexical LM, as it is better than the 150k or 200k full-word system interms of the WER. Thus, the recognition results using 150k vocabulary are not shown in this paper. The experiments were carried out with RWTH segmentation and sub-lexical language models containing 200k sub-lexical units. As can

Table 6: *WERs[%] of the English LVCSR system (OOV Rate:0.7, dev2012 PPL:129, Vocabulary size:150k).*

Corpus	Pass	System-1	System-2
dev2012	VTLN	17.6	21.9
	CMLLR	15.2	18.5
	MLLR	-	18.8
	LM-rescoring	14.8	17.9
	CN decoding	**14.4**	
tst2011		**10.2**	
tst2012		**11.3**	
tst2013		**16.0**	

Table 7: *Progressive WER [%] improvements : IWSLT-12 Vs. IWSLT-13 English LVCSR Systems*

Corpus	IWSLT 2012	IWSLT 2013	Rel. gain
tst2011	13.4	10.2	23.9
tst2012	13.6	11.3	20.3
tst2013	-	16.0	-

be seen in Table 8, the deep unilingual BN features trained on out-of-domain BN/BC data did not result in better WER compared to the shallow ones (1st and 3rd rows). Including multiple languages in the BN training improved the results significantly, and the performance gap increased further after the speaker adaptation step (3rd and 4th rows), similar to our observation in [8]. Furthermore, the results also show that the deep structure is more beneficial for multilingual training and outperforms the shallow multilingual BN (2nd, 4th rows). Different types of last hidden layers described in Subsection 2.2.1 were also investigated. Applying language dependent hidden layers between the bottleneck and output layer did not resulted in lower error rate (5th row). On the contrary, if the number of parameters were increased by larger language independent hidden layer further reduction in WER (6th row) is observed.

Table 8: *WER[%] comparison of speaker independent (SI) and speaker adapted (SA) uni- and multilingual BN features with different structures - German LVCSR with* **no** *word compounding (***Seg***: audio segmentation,* **SI***: speaker independent models,* **SA***: speaker adapted models)*

	AM	Seg	Dev2012		Eval2013	
			SI	SA	SI	SA
BN features	Shallow	RWTH	22.3	20.1	30.0	27.5
	+multilingual		21.7	19.1	29.4	26.1
	Deep		22.1	20.5	30.1	28.1
	+multilingual		20.9	19.0	28.0	25.8
	+lang.dep.hidden		20.8	19.1	27.9	26.1
	+large hidden		**20.6**	**18.8**	**27.7**	**25.7**
		LIUM	20.8	19.0	27.9	25.9

Table 9: *Recognition results for 200k German LVCSR with* **no** *word compounding (***FW***: full-word system,* **Crp***: corpus,* **MW***: sub-lexical system, PP_w: word-level perplexity, PP_c: character level perplexity,* **unsp***: unsupervised adapted LM ,* **CN***: confusion network decoding,* **CER***: character error rate,* **Effective OOV rate** *:- Dev:0, eval:0.9)*

Expt.	Crp	LM	Adap	PP_w/PP_c	WER [%]	CER [%]
FW	dev	backoff	no	314/2.1	19.6	7.6
	eval			226/2.2	26.0	15.6
MW	dev	backoff	no	284/2.2	18.8	7.5
		+ME		282/2.2	18.8	7.5
	eval	backoff	no	240/2.3	25.4	15.4
		+ME+unsp	yes	239/2.3	25.4	15.4
CN dec. MW+FW	dev	backoff	no	–	**18.4**	7.5
	eval			–	**25.2**	15.4

Alternatively, as shown in Table 9, non-adapted and adapted MaxEnt models are applied on the morpheme systems interpolated with the backoff LM. Character-level perplexities are shown for fair comparison between full-word and morpheme based systems. Applying LM adaptation did not affect either the perplexity or the WER for both development and evaluation corpus. To capture the advantages of both the sub-lexical and full-word systems, system combination is used. Using confusion network decoding based system combination, further improvements are achieved compared to the stand-alone sub-lexical based system.

6. Conclusions

In this paper, the descriptions of the German and English LVCSR systems developed by the RWTH Aachen for the IWSLT 2013 evaluation are presented. Here, state-of-the-art acoustic level multilingual features, domain dependent language modeling, supervised and unsupervised adaptation and system combination of subsystems are experimented. Noticeable contribution of the improvements were achieved because of the use of multilingual features. Language model adaptation did not affect the WER. Although sub-lexical systems performed significantly better than the full-word systems, system combination outperformed all other systems. The RWTH produced competitive results for German and English LVCSRs in the IWSLT 2013 evaluation campaign.

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] 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, June 2000, pp. 1635 – 1638.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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, Sept. 2008, pp. 2711–2714.

[7] 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.

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

[9] 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.

[10] M. Bisani and H. Ney, "Open Vocabulary Speech Recognition with Flat Hybrid Models," in *Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 725 – 728.

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

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

[13] 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, Sept. 2010, pp. 1517 – 1520.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

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

[16] 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.

[17] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical Bottle Neck Features for LVCSR," in *Interspeech*, Makuhari, japan, Sept. 2010, pp. 1197–1200.

[18] 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.

[19] 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.

[20] 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.

[21] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1998, pp. 645–648.

[22] S. Peitz, S. Mansour, M. Freitag, M. Feng, M. Huck, J. Wuebker, M. Nuhn, M. Nußbaum-Thom, and H. Ney, "The RWTH Aachen Speech Recognition and Machine Translation System for IWSLT 2012," in *The International Workshop on Spoken Language Translation*, Hongkong, Dec. 2012, pp. 69–76.

[23] 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.

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

[25] 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, Sept. 2012.

[26] 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.

[27] 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.

[28] C. Samuelsson and W. Reichl, “A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Phoenix, AZ, USA, Mar. 1999, pp. 537 – 540.

[29] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, May 2008.

[30] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, “Automatic segmentation, classification and clustering of broadcast news audio,” in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, USA, Feb. 1997, pp. 97–99.

[31] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, “An Open-source State-of-the-art Toolbox for Broadcast News Diarization,” in *Interspeech*, Lyon, France, aug 2013, pp. 1477 – 1481.

[32] D. Rybach, C. Gollan, R. Schlüter, and H. Ney, “Audio Segmentation for Speech Recognition using Segment Features,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 4197 – 4200.