The RWTH 2009 Quaero ASR Evaluation System for English and German

Markus Nußbaum-Thom, Simon Wiesler, Martin Sundermeyer, Christian Plahl, Stefan Hahn, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{nussbaum,wiesler,sundermeyer,plahl,hahn,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this work, the RWTH automatic speech recognition systems for English and German for the second Quaero evaluation campaign 2009 are presented. The systems are designed to transcribe web data, European parliament plenary sessions and broadcast news data. Another challenge in the 2009 evaluation is that almost no in-domain training data is provided and the test data contains a large variety of speech types. The RWTH participates for the English and German languages with the best results for German and competitive results for the English. Contributing to the enhancements are the systematic use of hierarchical neural network based posterior features, system combination, speaker adaptation, cross speaker adaptation, domain dependent modeling and the usage of additional training data.

Index Terms: speech recognition, probablistic features, system combination, LVCSR

1. Introduction

This paper describes in detail the English and German RWTH automatic speech recognition systems developed for the second Quaero evaluation campaign in 2009. Quaero is a large vocabulary task, with focus on transcribing web data. The data includes speech types like comedy, news, cooking sessions, interviews and talk-shows. Recognition on the data is challenging because of a huge variability in the acoustic conditions and a large portion includes spontaneous speech.

A challenge in the 2009 evaluation is that almost no in-domain training data is given. Only 19 hours of transcribed in-domain audio data is available for training the acoustic models for German. Moreover, the German development and evaluation data sum up to 12 hours. In contrast, no in-domain training data exists for the English task.

The major enhancements to the present system compared to the systems used in last year's evaluation are achieved by using domain specific acoustic and language models, system combination techniques like confusion network combination [4] and cross-adaptation, additional training data and neural network based posterior features [1, 2, 3].

As described in [5, 6], both the German and the English systems consist of several subsystems, that differ in the features or the language models used. For each language a four pass strategy with a 4-gram decoder is performed. We apply a fast vocal tract length normalization (VTLN) in the first pass, and constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR) in the second pass. Depending on the system either language model rescoring or cross-adaptation is applied as third pass. Finally, all subsystems are combined by confusion network system combination.

The paper is organized as follows: In Section 2 we describe our training data, the creation of the lexica and the language modeling. Next, in Section 3, the acoustic models, speaker normalization and adaptation techniques are presented. Section 4 describes the development of the German and English systems and in Section 5 we describe the multi-pass decoding. Finally we present our experiments in Section 6 and the conclusions in Section 7.

2. Language Resources

The development and evaluation data of the Quaero 2009 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. The evaluation is carried out in an open condition, where all the training data before 2008 is allowed. Furthermore, in-domain training data is provided only for German.

2.1. English Training Data

Since there is no in-domain training data available for English other transcribed audio data containing BN and EPPS data are used for training. Table 1 shows the amount of audio data for different corpora. Overall, 500 hours of acoustic training data could be used. The *HUB4* and the *TDT4* corpora contain American English BN [5] only, whereas the *TC-STAR* corpus consists of EPPS [6].

Table 1: *Transcribed data for acoustic modeling for English.*

corpus	duration [h]	# segments	# running words
HUB4	206	119,658	1,617,099
TDT4	186	110,266	1,715,445
TC-STAR	102	66,670	761,234

2.2. German Training Data

In contrast to the English system, in-domain training data for German has been provided. As shown in Table 2 the WEB08 corpus covers podcast data from various domains like comedy, seminar, report and interview, whereas the EPPS08 corpus consists of speeches from the European parliament only. Overall, 19 hours of in-domain data is used.

Table 2: *In-domain transcribed audio data for German.*

corpus	duration [h]	# segments	# running words
WEB08	14	3452	127,086
EPPS08	5	1109	45,796

Due to the small amount of in-domain training data, we also use common training data which is summarized in Table 3 and considers the cut-off date.

Table 3: *Additional transcribed audio data for the German system available for acoustic modeling.*

corpus	duration [h]	# segments	# running words
Verbmobil	63	36440	736,058
WDR	79	42011	625,018
Report Mainz	11	8928	100,641
Zeit	171.5	329384	2,500,866

The *Verbmobil* corpus recorded in the *Verbmobil* project consists of dialogues for travel appointments [7]. In contrast, the audio material of the *WDR* and *Report Mainz* corpora is from the BN domain [8]. Furthermore, RWTH has downloaded read articles from the newspaper *Zeit* for which almost correct transcripts are available. A reasonable segmentation is generated by aligning the transcripts to the audio data. Segment boundaries are introduced on silence chunks not shorter than 35 seconds.

2.3. Lexicon Modelling

For each language the recognition vocabulary is derived from the text data described in Table 4. The text data is cleaned up and normalized by a manually defined set of rules and semi-automatic methods. The lexicon consists of most frequent 65k words for English and 100k words for German respectively. For words, where no pronunciation has been available, the pronunciations are generated by the statistical grapheme-to-phoneme (g2p) conversion toolkit [12].

Table 4: *Text resources for the English (EN) and the German (DE) systems used for language modeling.*

	corpus	# running words	type
EN	HUB4,TDT2-4	10M	BN
	Gigaword	2,600M	newswire
	EPPS verbatim	0.8M	EPPS
	EPPS FTE	34M	EPPS
	total	2,645M	-
DE	TAZ	151M	BN
	German-news	155M	BN
	total	306M	-

In order to cope with American and British English, lexica for both types are generated. The pronunciations of the British English lexicon are based on the British English Example Pronunciation Dictionary, whereas the pronunciations of the American lexicon are based on the American English PRONLEX lexicon. In contrast, the source pronunciations for German are obtained from the German LC-STAR lexicon only.

2.4. Language Modelling

Table 4 shows the text resources available for language model (LM) training for English and German. The text sources consist basically of BN and EPPS data. After training domain specific LMs, the LMs for all languages are linear interpolated using the SRI Language Modelling Toolkit [13], where the interpolation weights are optimized on a holdout data set.

3. Acoustic Modeling

3.1. Baseline Acoustic Modeling

Both, the English and the German systems are composed of several subsystems which use either MFCC or PLP as base features. For each feature type a segment-wise mean and variance normalization is applied and fed into a sliding window of length nine. All feature vectors in the window are concatenated and projected to a 45 dimensional feature space by applying linear discriminant analysis (LDA). The feature vector is augmented with a voicedness feature and phone posterior features estimated using a multilayer perceptron. The hierachical neural network (HMRASTA) is trained using the phonemes of the given language based on MRASTA features [1, 3], as estimated on a phone alignment. The dimensionality of the phone posterior features is reduced using a principal component analysis.

For the sake of simplicity we will refer to the MFCC augmented by a voicedness and MLP features as MFCC+voiced+MLPs and PLP as PLP+voiced+MLP features later on.

Acoustic models for all systems are across word triphone left-to-right hidden markov models (HMMs) based on Gaussian mixtures with globally pooled diagonal covariance matrix. For the English system, 6-state HMMs are used, while for German, 3-state HMMs seemed more adequate. A number of 4500 generalized triphone states defines our HMM states. The baseline acoustic models (AMs) are trained using maximum likelihood (ML)/*Viterbi* on the available training data. The resulting models comprise about 1M gaussians with a globally pooled covariance matrix.

3.2. Speaker Normalization and Adaptation

All systems use the same approach for speaker normalization and adaptation. Vocal Tract Length Normalization (VTLN) is applied to the filterbank within the MFCC or PLP extraction both in training and testing. In recognition, a fast one pass VTLN approach is used, where the warping factor are estimated using a Gaussian mixture classifier, trained on the acoustic training corpora.

Speaker adaptive training (SAT) based on Constrained Maximum Likelihood Linear Regression (CMLLR) [9] is applied to compensate for speaker variation in both training and testing. For the German system Maximum Likelihood Linear Regression (MLLR) is applied to the means of the Gaussian mixtures during recognition.

Both CMLLR and MLLR are text dependent and need a two pass setup. They are carried out in a speaker dependent manner and no speaker identities are provided in the evaluation, so an automatic speaker labeling is performed. To provide a speaker labeling for SAT, a generalized likelihood ratio based segment clustering with a Bayesian information criterion based stopping condition is applied to the segmented training and recognition corpus [10]. In the third pass the German subsystems are cross-adapted to each other.

4. System Development

4.1. Development of the English System

The English system has six different subsystems. The main differences of these subsystems rely on different acoustic features and domain dependent AMs as well as domain dependent LMs.

In the acoustic training more data is available for BN than for EPPS and since the domain BN may be closer to the domain web than parliament 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 language model we apply a similar approach. Since domain dependent language model data is available for EPPS a language model is trained on the TC-STAR data. For the same reason a BN language model is trained on the *HUB4*, *TDT2-4* and *Gigaword* corpora. In order to obtain a faster recognition likelihood pruning is applied to the LM such that smaller LMs are created for use in recognition and larger LMs are applied in lattice-rescoring. The next enumeration summarizes the available AMs and LMs:

- Acoustic Models:
  - a1: Trained on *HUB4* and *TDT4* with MFCC+voiced+MLP features.
  - a2: Trained on *HUB4* and *TDT4* with PLP+voiced+MLP features.
  - a3: Trained on *TC-STAR* with MFCC+voiced+MLP features.
  - a4: Trained on *TC-STAR* with PLP+voiced+MLP features.
- Language Models:
  - ℓ1: Trained on *HUB4*, *TDT2-4* and *Gigaword.*
  - ℓ2: Trained on *TC-STAR*.

Table 5 shows the different systems which arise from the combination of the various domain dependent acoustic and language models. Later we refer to these systems as e1, e2, . . . , e6.

Table 5: *Models used for the English system.*

	BN system				EPPS system	
ac model	a1	a2	a3	a4	a3	a4
lm model	ℓ1				ℓ2	
system	e1	e2	e3	e4	e5	e6

Table 6 gives the amount of training data, number of running words, vocabulary size, perplexities of the final LMs and OOV rates on the English dev09 and eval09 data sets.

Table 6: *Statistics for dev09 and eval09 corpora for English.*

corpus	English					
	dev09			eval09		
domain	web	BN	EPPS	web	BN	EPPS
dur. [h]	10.5	0.18	0.58	2.23	0.5	0.5
run. wrds	123k			35k		
vocab	10k			5k		
PP ℓ1	223			196		
PP ℓ2	353			298		
OOV [%]	2.3			1.75		

4.2. Development of the German System

All German AMs are trained on the whole audio data mentioned in Tables 2 and 3. The German system is subdivided into two subsystems. The main difference of these two subsystems originate from the features used to train the systems. The German subsystem g1 utilizes a MFCC+voiced+MLP whereas the subsystem g2 uses a PLP+voiced+MLP front-end. Both subsystems use the same language model which was estimated on the text data described in Table 4.

Table 7 gives the amount of training data, number of running words, vocabulary size, perplexities of the final LMs and OOV rates on the German dev09 and eval09 data sets.

5. Recognition Process

5.1. Multi-Pass Recognition

For all languages and AMs the first pass is realized by a 4-gram Viterbi decoder using a fast-VTLN normalization. In the second pass a SAT/CMLLR recognition is applied where the statistics for adaptation are collected from the first pass output, the German system uses MLLR in addition.

Table 7: *Statistics for the dev09 and eval09 corpora for German.*

corpus	German					
domain	dev09			eval09		
domain	web	BN	EPPS	web	BN	EPPS
dur. [h]	7.4	0.0	0.38	2.95	0.46	0.42
run. wrds	68k			36k		
vocab	10k			6k		
PP	394			353		
OOV [%]	5.0			4.79		

Figure 1 shows the domain dependent decoding framework for the English system and Figure 2 for German respectively.

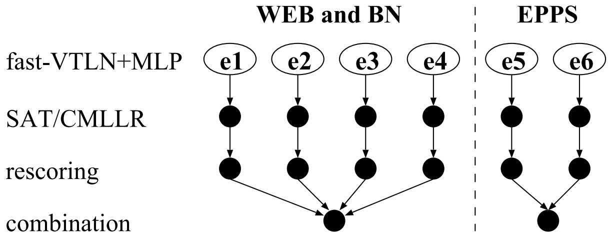

Figure 1: *Schematic diagram of the decoding framework for English.*

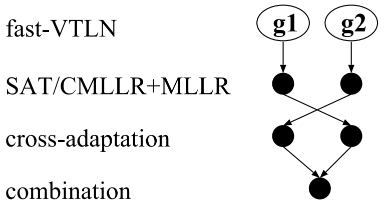

Figure 2: *Schematic diagram of the decoding framework for German.*

In the third pass the English system is lattice-rescored with a full language model. Finally in the fourth pass a system combination is applied for the four subsystems on the WEB and BN domain and on the two subsystems on the EPPS domain.

In contrast, the German subsystems are cross adapted to each other in the third pass. In the last pass the German subsystems are combined using system combination.

6. Experiments

For parameter optimization, the 2009 development sets are used. Statistics are given in Table 6 and Table 7. Tables 8 and 9 summarize the recognition results for the methods applied in the 2009 evaluation for English and German on dev09 and eval09 sets and for comparison the recognition results of the previous year's baseline system are given also.

English The current English system shows a significant improvement compared to the baseline system. Under consideration that the baseline system is the third pass of the subsystem e1 without MLP features, the incorporation of the MLP features leads to an improvement of 41.9% WER relative on dev09 and 34.9% WER relative on eval09. It is clear that these improvements are too promising and probably originate from a not well tuned baseline system. The system combination achieves a reduction of 6.4% WER relative on dev09 and a reduction of 11.9% WER relative on eval09 as expected, compared to the best recognition result of the third pass. Furthermore parts of the improvement are due to the additional training data and domain dependent modeling.

Table 8: *Results on the English dev09 and eval09 corpus.*

corpus	dev09		eval09	
baseline	47.7	-	40.4	-
subsystem	e1	e3	e1	e3
fast-VTLN+MLP	41.8	35.7	39.9	38.6
+SAT/CMLLR	35.4	35.4	33.9	28.7
+rescoring	34.9	34.8	33.5	28.2
+combination	32.7		25.2	

German The current German system achieves a significant improvement compared to the baseline system. The baseline system is the second pass of system g1 without the use of MLP features and the acoustic training data from the *WEB08*, *EPPS08* and the *Zeit* corpora. The application of MLP features and additional audio data leads to a improvement of 26.1% WER relative on dev09 and 31.1% WER relative on eval09, compared to the baseline system. But due to the use of additional training data the gain of the MLP features is not separable and again it seems that the improvements are too promising because of a not well tuned baseline system. Cross-adaptation and system combination give only in a small improvement probably due to the small number of subsystem which are cross adapted and combined.

Table 9: *Results on the German dev09 and eval09 corpus.*

corpus	dev09		eval09	
baseline	44.0	-	40.9	-
subsystem	g1	g2	g1	g2
fast-VTLN+MLP	37.7	37.7	34.2	34.3
+SAT/CMLLR+MLLR	34.9	35.0	31.0	31.2
+cross-adaptation	34.6	34.9	30.5	30.5
+combination	33.3		30.0	

7. Conclusions

In this work the RWTH automatic speech recognition systems developed for the second Quaero evaluation campaign 2009 were presented for the languages English and German. In comparison to the 2008 system, significant improvements were obtained by using MLP based phone posterior features, additional training data, domain dependent modeling, cross-adaptation and system combination of several subsystems. A major contribution of the improvements were achieved because of the use of MLP features. It is notable that all improvements were achieved using only a small amount of in-domain training data from the test domain. The RWTH produced the best results for German and competitive results for English in the 2009 Quaero evaluation.

8. Acknowledgements

This work was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

9. References

[1] F. Valente and H. Hermansky, “Hierarchical and parallel processing of modulation spectrum for asr applications”, In Proc. IEEE Int. Conf. on Acoustics, 2008.

[2] H. Hermansky, D. Ellis, and S. Sharma, “Connectionist feature extraction for conventional HMM systems”, In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2000.

[3] C. Plahl, R. Schlüter, H. Ney, “Hierachical Bottle Neck Features for LVSCR”, submitted at the Interspeech 2010 in Makuhari.

[4] G. Evermann and P. Woodland, ”Posterior probability decoding, confidence estimation and system combination”, In NIST Speech Transcription Workshop, 2000.

[5] B. Hoffmeister, C. Plahl, P. Fritz, G. Heigold, J. Lööf, R. Schlüter, and H. Ney, ”Development of the 2007 RWTH Mandarin LVCSR System“, In IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 2007.

[6] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney, ”The RWTH 2007 TC-STAR Evaluation System for European English and Spanish“, In Interspeech, pages 2145-2148, August 2007.

[7] S. Kanthak, A. Sixtus, S. Molau, R. Schlüter, and H. Ney, “Fast Search for Large Vocabulary Speech Recognition”, In Wolfgang Wahlster: Verbmobil: Foundations of Speech-to-Speech Translation Chp. ”From Speech Input to Augmented Word Lattices”“, pages 63-78, Springer Verlag, July 2000.

[8] W. Macherey, and H. Ney, ”Towards Automatic Corpus Preparation for a German Broadcast News Transcription System“, In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 733-736, May 2002.

[9] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition”, In Computer Speech and Language, vol. 12, no. 2, pp 75-98, Apr. 1998.

[10] 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, vol. 2, pp. 645-648.

[11] F. Guiliani and F. Brugnare, ”Acoustic Model Adaptation with multiple Subervisions”, In Proc. TC-Star Workshop on Speech-To-Speech Translation, June 2006, pp.151-154.

[12] M. Bisani, and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion”, In Speech Communication, May 2008, volume 50, number 5, pages 434-451.

[13] A. Stolcke, ”SRILM - An extensible language modeling toolkit”, In Proc. Int. Conf. on Spoken Language Processing. 2002, vol. 2, pp. 901-904.