The RWTH 2007 TC-STAR Evaluation System for European English and Spanish

J. Lööf, Ch. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, Ch. Plahl, D. Rybach R. Schlüter, and H. Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,gollan,hahn,heigold,hoffmeister,plahl,rybach,schluter,ney}@cs.rwth-aachen.de

Abstract

In this work, the RWTH automatic speech recognition systems developed for the third TC-STAR evaluation campaign 2007 are presented. The RWTH systems make systematic use of internal system combination, combining systems with differences in feature extraction, adaptation methods, and training data used. To take advantage of this, novel feature extraction methods were employed; this year saw the introduction of Gammatone features and MLP based phone posterior features. Further improvements were achieved using unsupervised training, and it is notable that these improvements were achieved using a fairly low amount of automatically transcribed data. Also contributing to the improvements over last year was the switch to MPE training, and the introduction of projecting SAT transforms.

Index Terms: speech recognition, system combination

1. Introduction

This paper describes in detail the English and Spanish RWTH Automatic Speech recognition systems, which were developed for the 2007 TC-STAR Evaluation Campaign. The TC-STAR EPPS Task, described in detail in [1], is a large vocabulary task with the focus on parliamentary speeches and political debates, in English and Spanish.

The main improvement compared to the system used in the last year's evaluation [1] was the systematic use of system combination techniques, combining systems with differences in feature extraction, adaptation methods, and training data. For system combination both minimum frame WER [2] and ROVER were used.

All the individual systems were based on a one-pass fourgram decoder with optional fast vocal tract length normalization (VTLN). In a second pass, speaker adaptation was applied. All systems included constrained maximum likelihood linear regression (CMLLR) including speaker adaptive training (SAT), maximum likelihood linear regression (MLLR), and discriminative minimum phone error (MPE) training. As an optional step before system combination, the lattices produced by the second recognition step can be rescored with an improved language model.

For each of the languages, the evaluation was carried out in three different conditions, differing in what training data was eligible for use. In the *restricted* condition, only certain training data provided through the TC-STAR consortium was allowed. For the *public* condition, any publicly available training data respecting a cut-off date (i.e before May 31st, 2006) was eligible. In the *open* condition any training data before the cut-off could be used. The RWTH systems participated in the restricted condition for English and Spanish, as well as in the public condition for English.

2. Acoustic Modeling

The RWTH system consisted of four independent subsystems, each differing in the acoustic modelling used. The individual systems were numbered from 1 to 4.

The acoustic models were trained using the data allowed in the restricted condition. The data consisted of approximately 300h of audio data from the European parliament plenary sessions (EPPS) for both English and Spanish, of which 100h were manually transcribed. The transcriptions include a segmentation into sentence like units, speaker labels, and topic headings. In Spanish, this was augmented with about 40h of manually transcribed recordings from the Spanish Parliament and Congress (SPC). See Table 1 for details. About 100h of the untranscribed data was new compared to last year.

Table 1: *Recordings from the EPPS (both) and SPC (Spanish) domain available for acoustic modelling.*

	En EPPS	Es EPPS	Es SPC
Total Data [h]	278.8	249.1	38.4
Transcribed Data [h]	91.6	61.9	38.4
# Segments	66,670	59,490	42,118
# Running Words	660,603	458,917	257,481
Untranscribed Data [h]	187.2	187.2	-

2.1. Baseline Acoustic Modeling

The largest difference between the different systems is in the basic acoustic modeling, and especially in the feature extraction methods used. All systems were trained on the manually transcribed acoustic data allowed in the restricted condition; System 4 additionally used the untranscribed data provided for the restricted condition.

System 1 used an acoustic front end consisting of mel-frequency cepstral coefficient (MFCC) features derived from a bank of 20 filters. 16 cepstral coefficients including the zeroth coefficient were used, and cepstral mean normalization was applied. The MFCC features were augmented with a *voicedness feature* [3]. The MFCCs and voicedness features from nine consecutive frames were concatenated and a linear discriminative analysis (LDA) was used to project the resulting vector to 45 components.

In System 2, gammatone cepstral coefficients were used, as presented in [4]. The gammatone filterbank is reported to give a good approximation of the human auditory filter. Acoustic features derived from a gammatone filterbank were shown to perform comparable to standard features such as MFCC and PLP. Both cepstral mean and variance normalization were applied to the Gammatone cepstral features. In addition a LDA estimated as in System 1 was used.

System 3 used the features from System 1, but augmented them with phone posterior features estimated using a multilayer perceptron (MLP). The neural network was trained using the phonemes of the given language, as estimated by a phone alignment. The input to the neural network was multiple time resolution features (MRasta [5]), which were based on PLP features. The dimensionality of the phone posterior features was

reduced using a KLT-transform, and the result was concatenated to the features from System 1.

System 4 differs from System 1 mainly through the use of the untranscribed data allowed in the restricted condition, using unsupervised training with word posterior confidence selection. The recordings were automatically transcribed using systems optimized for the raw recording conditions, which differ from the evaluation conditions, where noise, music, and foreign speech segments are already manually excluded. Therefore, the automatic segmentation, speaker clustering, and data threshold parameters were optimized on a raw recording development set derived from the TC-STAR development corpus of the first TC-STAR evaluation campaign. Notable here is that even though a fairly low amount of automatically transcribed data was used, significant improvements were still obtained. System 4 used features similar to System 1 but used both cepstral mean and variance normalization.

Acoustic models for all systems were cross-word triphone based 6-state left-to-right Gaussian mixture hidden Markov models with a globally pooled diagonal covariance matrix. A number of 4500 generalized triphone states was used. The baseline acoustic models were maximum likelihood (ML) / *Viterbi* trained using the training data provided for the restricted condition.

2.2. Speaker Normalization and Adaptation

Depending on the individual system, two or three different approaches for speaker normalization / adaptation were applied. First, in all subsystems using MFCC features, vocal tract length normalization (VTLN) was applied to the filterbank within the MFCC extraction both in training and testing. For recognition, a fast one pass VTLN method was used, where the warping factor was estimated using a Gaussian mixture classifier, trained on the acoustic training corpus. Warping factors were estimated using a grid search over 21 factors in the range 0.8 – 1.2. For Systems 1 and 3 the classifier was trained only on a subset of the corpus, and was used to estimate warping factors in acoustic training. In System 4 the warping factors estimated using grid search were used for acoustic model training. System 3, based on gammatone features, used no VTLN.

In Systems 3 and 4, speaker adaptive training (SAT) based on constrained maximum likelihood linear regression (CMLLR) [6] was used to compensate for speaker variation in both training and testing. The *simple target model* approach [7] was used, since results in [7] indicate that it outperforms the standard CMLLR-SAT method [6]. As target model an acoustic model with a single Gaussian per state trained on baseline features was used. Systems 1 and 2 used a similar approach to SAT, but used projecting affine feature transforms instead of CMLLR. The speaker specific transforms were estimated using the so called MMI′ criterion, and replaced the LDA matrix in feature extraction, see [8] for details.

Finally, maximum likelihood linear regression (MLLR) was applied to the means of the acoustic model in recognition. A regression class tree was used to adjust the number of regression classes to the amount of data available.

Since both CMLLR and MLLR are text dependent, a two pass setup is needed. Also, since they are carried out in a speaker dependent manner, and since no speaker identities were provided in the evaluation, an automatic speaker labeling was applied. For SAT, the speaker labels provided in the training data were used. The details of the two-pass system is described in Sec. 4.1.

2.3. Discriminative Training

To refine the ML trained acoustic model, discriminative training was performed. The minimum phone error (MPE) criterion [9] was used as preliminary experiments had shown an improvement compared to the maximum mutual information (MMI) and minimum classification error (MCE) criteria used in last years systems. The discriminative training was initialized with the ML trained acoustic model.

The word-conditioned word lattices used in training were generated with the VTLN/voicedness system in combination with a bigram language model. Since the lattices were dominated by silence and noise arcs, the lattices were filtered. The idea behind this filtering was to correct the posteriors for accumulation of discriminative statistics. This step was (particularly for the English system) essential for good performance. A similar effect was observed if normalizing the pronunciation scores to 1. However, this approach yielded slightly worse results.

For acoustic rescoring during discriminative training iterations the exact match approach was used, i.e. the word boundary times were kept fixed. The optimal number of training iterations was determined by recognition on the development corpus. The resulting models comprise about 800–900k Gaussians.

3. Lexicon and Language Modeling

3.1. Lexicon Modeling

Last year's recognition lexica have been improved using various approaches. For both English and Spanish, we corrected some minor errors and added the names of the politicians who have joined the European Parliament (EP) in the meantime. We also enlarged the vocabulary utilizing publicly available language modeling data (cf. Sec. 3.2): on parts of the Gigaword-corpus, unigrams with their respective counts have been calculated and sorted according to descending rank. For each language, the top 100k words of these unigram-counts have been checked if they already occur within the lexicon. From the missing words, approximately 1,300 words for English and 1,000 words for Spanish have been included. The pronunciations for the names of the additional members of the EP as well as for the LM-derived ones were generated using the grapheme-to-phoneme conversion (g2p) tool [10]. Additionally, we introduced pronunciation weights based on relative frequencies calculated on an alignment of the training data. To prevent overfitting, these frequencies were smoothed with a uniform distribution over the number of pronunciations. A uniform distribution was also assumed if there were no observations of a certain orthography within the training data but more than one possible pronunciation (as e.g. for words added by the g2p tool).

The English pronunciation lexicon was derived from the British English example pronunciation dictionary (BEEP). The Spanish pronunciation lexicon was derived from the lexicon of the LC-STAR project [11]. Using these dictionaries, statistical grapheme-to-phoneme conversion models were trained [10] for Spanish and English. The models were used to produce pronunciations for words not covered by the original lexica.

3.2. Language Modeling

For the restricted condition, we trained standard case sensitive fourgram LMs with all available data. This includes the final text editions (FTE) and verbatim transcriptions (VT) up until May 2006 excluding the previous evaluation time intervals. Thus, there are approximately 2M additional running words for both English and Spanish due to the extended allowed time period compared to last year's system. For Spanish, there are also the debates of the Spanish Parliament and Congress available with the same time constraints (see [1] for an detailed description of the available text sources). For each language, an LM for each data source has been trained using modified Kneser-Ney discounting as the smoothing method. These LMs have been linearly interpolated whereas the interpolation weights have been optimized on the development set. Our final restricted con-

dition language model for English contains approximately 7.5M multi-grams, the one for Spanish about 14M multi-grams. Table 2 gives an overview of the text sources used for the restricted condition.

In contrast to last year's Evaluation, we also submitted results for the public condition data track for English. For this track, we trained an additional fourgram language model on the data of the British Parliament (BP) made available by ELDA (approximately 51M running tokens) and parts of the English Gigaword corpus (GW), which may be purchased by LDC (approximately 174M running tokens). This LM has been interpolated with the language model for the restricted condition. The final public condition LM contains about 26M multi-grams. This LM was not used for recognition but for rescoring on the word graphs of the restricted condition. We used the SRI Language Modeling Toolkit to build and interpolate the LMs [12]. Statistics of the resulting language models are shown in Table 3.

Table 2: *Text resources available for language modelling, restricted condition.*

	running words		
	VT	FTE	Spanish SPC
English	781,649	33,894,405	-
Spanish	516,936	35,190,383	47,181,386

Table 3: *Language model statistics.*

	#multi-grams	PP on Dev06	PP on Eval07
English, restricted	7,472,949	95.9	110.8
Spanish, restricted	14,286,867	76.5	104.8
English, public	25,759,369	110.3	107.1

4. Recognition Process

4.1. Two-Pass Speaker Adapted System

The RWTH baseline system realizes a one-pass fourgram Viterbi decoder. Each of the individual systems used two pass recognition to facilitate speaker adaptation, as described in Sec. 2.2. The first pass was performed using an ML estimated acoustic model. Since no fine-grained segmentation of the data was provided in the evaluation, the complete recordings were recognized using the first pass of system 1. The recordings varied in length between a couple of minutes and half an hour. The silence information from the recognition was used to segment the audio data. The segment boundaries were chosen at the longest silence regions in such a way that no segment is longer than 35s, while keeping the number of segments at a minimum.

To provide a speaker labeling, a generalized likelihood ratio based segment clustering with a *Bayesian* information criterion based stopping condition was applied to the segmented recognition corpus [13]. For System 1, the output of the segmentation recognition was used to estimate SAT and MLLR matrices needed by the adaptation, while for systems 2 – 4 a first pass recognition performed on the segmented recordings was used. The second pass was performed to produce lattices using the best acoustic models, discriminatively trained on the SAT transformed features, and adapted using the MLLR matrices. In the English public systems, the lattices were rescored using the larger language model described in Sec. 3.2.

4.2. System Combination

A common method to improve the recognition performance is to combine the output of several subsystems. Crucial for good combination results are subsystems that are diverse in the errors they make. According to the literature different acoustic features and the usage of (partly) different training data are two techniques to achieve systems that produce different errors.

Table 4: *Recognition corpora statistics*

	English			Spanish		
	dev06	eval06	eval07	dev06	eval06	eval07
Audio [h]	3.2	3.2	2.9	2.4	6.9	6.2
# Run. wrd	27k	30k	27k	21k	60k	57k
# Speakers	41	41	50	31	63	53

Hence, the four systems described in chapter 2 set up a promising base for system combination.

For the 2007 evaluation campaign we investigated two combination approaches: the well-known ROVER (Recognizer Output Voting Error Reduction) [14] approach and the minimum frame WER (min-fWER) approach [2]. All ROVER experiments were done using confidence scores. The confidence scores were calculated from lattices produced by the final decoding step, see [15].

The min-fWER approach is defined within a Minimum Bayes' Risk (MBR) framework. Similar to confusion networks it aims to find consensus among a lattice but utilizes a different approximation. The min-fWER approach can easily be extended to work on a union of lattices and thus find consensus over several subsystems.

The search space used by ROVER consists of all possible paths through the alignment of the first-best hypothesis of all subsystems. In particular, it allows paths that are not present in one of the subsystems' lattices. This can lead to broken language model contexts in the final hypothesis, which is known to harm subsequent machine translation. In contrast, the search space for min-fWER is explicitly restricted to the union over the lattices produced by the different subsystems. Thus, we ensure that our final hypothesis does not break language model context.

5. Experiments

The development and evaluation sets were transcribed by ELDA. For parameter optimization, the 2006 evaluation development and evaluation sets were available. Statistics are given in Table 4, together with the statistics for the 2007 development data. Note that for the Spanish 2006 development set the statistics are for the EPPS part only.

5.1. Individual Systems

For the development of the individual systems, the 2006 development set was used for development and the 2006 evaluation set was used as verification. Results are presented for two of the individual systems, one in English and one in Spanish. The other systems produced similar results. In the case of the English system, results were produced both with and without a final rescoring with the public condition data language model. This rescoring was the only difference between the restricted and public systems.

Table 5 shows the development results of the English System 4, using untranscribed data, while Table 6 summarises the development results for the Spanish System 1. All error rates but the min-fWER results are Viterbi decoding results. The min-fWER results were calculated over the lattices later used for system combination. Note that in all results for the Spanish 2006 development set, the results reported only include the EPPS portion of the set. For the evaluation sets, the results for the combined EPPS plus SPC corpora are reported.

5.2. System Combination Results

In the following section the results from the system combination experiments are presented. The system combination weights were tuned on the development set of 2006. The evaluation 2006 set served as test set and results are presented for this set and also for the 2007 evaluation set. Table 7 shows the English system results, both for the four individual systems, and for the

Table 5: *System 4 English results*

	dev06	eval06
baseline	15.7	13.1
+SAT	14.0	11.5
+Unsupervised	12.9	-
+MPE	12.5	-
+MLLR	11.8	9.8
+ New Lexicon	11.6	9.6
+ Public LM	11.0	8.5
min-fWER decode	10.6	8.4

Table 6: *System 1 Spanish results*

	dev06	eval06
baseline	9.9	13.8
+ SAT	7.9	-
+ MPE	7.3	9.6
+ MLLR	7.1	9.3
+ New Lexicon	7.1	9.3

combination methods. The results for the Spanish (restricted) system are shown in Table 8.

Table 7: *English combination results*

	Restricted		Public	
	eval06	eval07	eval06	eval07
System 1	9.4	10.6	8.7	10.1
System 2	10.1	11.6	9.0	10.9
System 3	10.3	12.4	9.4	11.8
System 4	9.6	10.8	8.5	9.8
ROVER	8.8	9.9	7.9	9.3
Min-fWER	8.8	9.7	7.8	9.0

Table 8: *Spanish combination results*

	eval06	eval07
System 1	9.2	9.3
System 2	10.1	9.8
System 3	9.8	9.8
System 4	9.9	9.9
ROVER	8.7	8.9

5.3. Comparison to the 2006 RWTH Systems

Table 9 summarizes the results from our current system, and compares them to the results obtained using the 2006 system, as well as to the best of the individual systems (in restricted condition). Compared to last year's evaluation, the systems for the final TC-STAR evaluation presented here lead to relative improvements of around 12-13% in word error rate for the restricted condition.

Table 9: *Summary of results*

	English		Spanish	
	eval06	eval07	eval06	eval07
2006 Restricted	10.2	11.3	10.2	10.1
2007 Single sys.	9.4	10.6	9.2	9.3
2007 Restricted	8.8	9.7	8.7	8.9
2007 Public	7.8	9.0	-	-

6. Conclusions

In this work, the RWTH automatic speech recognition systems developed for the second TC-STAR evaluation campaign 2007 were presented. In comparison to the 2006 system, large improvements were obtained by using system combination techniques, which depend on the use of new acoustic features: the Gammatone features and the MLP based phone posterior features. Furthermore, improvements were achieved using unsupervised training, and it is notable that these improvements were reached using a fairly low amount of automatically transcribed data. Another contribution to the improvements was the use of MPE training, and the introduction of projecting SAT transforms. The English as well as the Spanish system achieved improvements of 12–13% relative in word error rate compared to the systems used last year. The RWTH systems produced the best results (w.r.t. WER) of all participating single site systems, in the restricted and public condition of both the English and the Spanish task.

Acknowledgements

This work was partly funded by the European Union under the Human Language Technologies project TC-STAR (FP6-506738).

7. References

[1] J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006, pp. 105 – 108.

[2] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006.

[3] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, USA, Sept. 2002, vol. 2, pp. 1065 – 1068.

[4] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Honolulu, HI, USA, Apr. 2007.

[5] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sept. 2005, pp. 361 – 164.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, Apr. 1998.

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

[8] J. Lööf, R. Schlüter, and H. Ney, "Efficient estimation of speaker-specific projecting feature transforms," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, Submitted.

[9] D. Povey and P. C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, May 2002, vol. 1, pp. 105 – 108.

[10] M. Bisani and H. Ney, "Multigrambased grapheme-to-phoneme conversion for LVCSR," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, vol. 2, pp. 933 – 936.

[11] "LC-STAR, Lexica and Corpora for Speech-to-Speech Translation Components," http://www.lc-star.com.

[12] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, Sept. 2002.

[13] 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.

[14] J. G. Fiskus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 1997.

[15] F. Wessel, R. Schlüter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 33 – 36, 2001.