The 2006 RWTH Parliamentary Speeches Transcription System

J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney

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

Abstract

In this work, the RWTH automatic speech recognition systems developed for the second TC-STAR evaluation campaign 2006 are presented. The systems were designed to transcribe parliamentary speeches taken from the European Parliament Plenary Sessions (EPPS) in European English and Spanish, as well as speeches from the Spanish Parliament. The RWTH systems apply a two pass search strategy with a fourgram one-pass decoder including a fast vocal tract length normalization variant as first pass. The systems further include several adaptation and normalization methods, minimum classification error trained models, and bayes risk minimization. For all relevant individual components contrastive results are presented on the EPPS Spanish and English data.

Index Terms: speech recognition, speaker normalisation, VTLN

1. Introduction

The TC-STAR (Technology and Corpora for Speech to Speech Translation) project [1] is envisioned as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST), including Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS) (speech synthesis). The project targets a selection of unconstrained conversational speech domains (speeches and broadcast news) and three languages (British English, European Spanish, and Mandarin Chinese). For the TC-STAR project, language resources (LR) for English and Spanish parliamentary speeches were collected for training and system development, as well as the TC-STAR evaluation campaigns. Within the restricted conditions of the TC-STAR evaluations the training data is restricted to these LR. This paper describes in detail the English and Spanish RWTH ASR system which were developed for the restricted condition of the TC-STAR Second Evaluation Campaign 2006. The systems comprises a one-pass fourgram decoder including fast vocal tract length normalization (VTLN), constrained maximum likelihood linear regression (CMLLR) including speaker adaptive training (SAT), maximum likelihood linear regression (MLLR), discriminative training including minimum classification error (MCE) training, as well as *Bayes* risk minimization (MBR). Further internal system combination, including ROVER or confusion network combination (CNC), did not result in further improvements.

2. Language Resources

2.1. Data

The English and Spanish LR both contain recordings from the European Parliament Plenary Sessions (EPPS), whereas the Spanish LR additionally include speeches from the Spanish Parliament and Congress (SPC). Approximately 100h of speech recordings per language were manually transcribed. These verbatim transcriptions (VT) include a segmentation into sentence like units, speaker labels, and topic headings.

The web site of the European Parliament [2] provides all EPPS reports since April 1996 translated in all official languages of the EU. These documents are known as the final text edition (FTE) and differ notably from the VT as the FTE aims for high readability. Table 1 specifies the data used for language modelling. The recordings, the corresponding manual transcripts, and the text LR were produced by Universitat Politècnica de Catalunya (UPC) and RWTH Aachen University. Table 2 gives the statistics of the acoustic training data used in the RWTH system. Development and evaluation sets were provided by ELDA. The English and Spanish EPPS development and evaluation sets each consisted of about three hours of speech, plus 4h of Spanish parliament data for evaluation. Table 3 gives an overview of the corpora.

Table 1: *Text resources available for language modelling.*

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

Table 2: *Transcribed recordings from the EPPS (both) and SPC (Spanish) domain available for acoustic modelling.*

	English	Spanish
Acoustic Data [h]	87.5	91.3
# Segments	66,670	101,608
# Running Words	704,883	743734

2.2. Lexicon Modeling

The recognition word lists were derived from the restricted domain data as described in Sec. 2.1. The available textual data was cleaned up and normalized, using a manually defined set of rules and semi-automatic methods. The word lists were produced as follows. All words from the verbatim transcriptions occuring at least twice were chosen. For the additional textual data a cut-off value was calculated requiring an out-of-vocabulary (OOV) rate below one percent on the development and a final lexicon of at least 50k words.

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 [3]. Using the dictionaries statistical grapheme-to-phoneme conversion models were trained [4] for Spanish and English. The models were used to produce pronunciations for words not covered by the original lexica. In Table 3 the lexicon statistics on the development and evaluation data are presented.

Table 3: *Development and evaluation data from the EPPS domain, and from the SPC domain (for evaluation Spanish only).*

	English		Spanish	
	Dev	Eval	Dev	Eval (+SPC)
Audio [h]	3.2	3.2	2.4	6.9
# Run. wrd	27,029	29,829	20,982	60,039
# Speakers	41	41	31	63
Vocab size	52,429		60,156	
4-gram PP	99.7	108.7	78.2	88.9
OOV [%]	0.81	0.58	0.61	1.22

2.3. Handling of OOV Words

We had originally assumed that the EPPS task would exhibit a very high lexical diversity leading to inevitably high out-of-vocabulary (OOV) rates. To address this problem an open vocabulary recognition approach was examined. In this so-called flat hybrid approach we augment the recognition vocabulary by a set of word fragments each consisting of a short sequence of phonemes with associated spelling information. The set of fragments is derived from the baseline pronunciation dictionary using a maximum likelihood criterion. The language model used by the recognizer is estimated from a modified version of the training corpus where each OOV word is replaced by its most likely sequence of fragments. This technique has been applied quite successfully on the *Wall Street Journal* database [4]. We have tried the identical techniques on the EPPS data, however without success: OOV words were recognised only very rarely, while spurious insertions of small fragments increased the overall error rate. We attribute this failure to the surprisingly low lexical diversity of the EPPS task. A low frequency of OOV words in training causes the estimation of the ”OOV part” of the model to be unreliable, due to lack of data. At the same time a low OOV rate in testing means that false alarms may easily exceed the potential improvement from OOV detection.

3. Acoustic Modeling

3.1. Baseline Acoustic Modeling

The acoustic front end comprises Mel-Frequency Cepstral Coefficient (MFCC) features derived from a bank of 20 filters. 16 cepstral coeficients including the zeroth coefficient were used, and cepstral mean normalization was applied. The MFCC features were augmented with a *voicedness feature* [5]. 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 dimensions.

Acoustic models were triphone based Gaussian mixture models (GMMs) with a globally pooled diagonal covariance matrix. The triphones were top down clustered using CART, rendering 4501 generalized triphone states.

The baseline acoustic models were maximum likelihood (ML)/*Viterbi* trained using the manually transcribed training data provided for the restricted condition, cf. Sec. 2.1 and Table 2.

3.2. Speaker Normalization and Adaptation

Three different approaches were used in combination to compensate for the acoustical variations due to speaker differences. First, a fast one-pass variant of Vocal Tract Length Normalization (VTLN) was applied to the filterbank within the MFCC extraction both in training and testing. The fast VTLN performs warping factor estimation using GMMs trained on a subset of the training corpus, for wich warping factors were estimated using the usual grid search.

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* (STM) 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 VTLN features was used. As a contrast experiment, when no SAT was used in acoustic model training, standard CMLLR was performed in recognition.

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 CMLLR is carried out in a speaker dependent manner, and since no speaker identities were provieded in the evaluation, an automatic speaker labeling was done. For SAT the speaker labels provided in the training data was used. The details of the two-pass system is described in Sec. 5.2.

3.3. Discriminative Training

To refine the ML trained acoustic model discriminative training was performed. Here the Maximum Mutual Information (MMI) criterion and the Minimum Classification Error (MCE) criterion were used as they have proven to perform best in our system. For the experiments the lattice based MCE was taken, which was originally presented in [8] for a large vocabulary speech recognition task. As in ML training, only the manually transcribed training data was used. 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. For MCE the spoken word sequence needs to be contained in the lattice. To guarantee this the best alignment of the spoken word sequence was merged into the training lattices. 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 a recognition on the development corpus and was about 10. The resulting models comprise about 800–900k Gaussians.

4. Language Modeling

4.1. Baseline Language Modeling

The language model (LM) training also was done using the restricted task data. For English, the data includes the transcriptions of the acoustic training data and the FTE data. From both data sets we trained seperate case sensitive fourgram LMs. The applied smoothing was modified Kneser-Ney discounting with interpolation. The final LM was the result of a linear interpolation of the two preliminary models, where the interpolation weights were optimized on the English development set. We used the SRI Language Modeling Toolkit to build and interpolate the LMs [9]. For English the optimal weights were 0.71 for VT and 0.29 for FTE. For Spanish additional restricted data from the Spanish Parliament (SPC) was used. Thus, three preliminary LMs were built. The linear interpolation weights were optimized in a grid search on the Spanish development data. The optimal weights were: VT: 0.53, FTE: 0.15, and SPC: 0.32. Table 3 gives the perplexities of the final LMs on the development and evaluation data.

4.2. Punctuation Modeling

For punctuation a sentence segmentation algorithm, also applied by the RWTH SLT system was used. For each estimated sentence break, a full stop is inserted; no further punctuation marks were produced. The segmentation approach originates from [10]. A decision for placing a segment boundary is made based on a log-linear combination of language model and prosodic features. In contrast to existing approaches, an explicit optimization over the number of words in the segment is performed by adding a length model feature. For a more detailed presentation of this method, see the presentation of the RWTH spoken language translation system [11].

5. Search Issues

5.1. Baseline One-Pass Recognizer

The RWTH baseline system realizes a one-pass fourgram Viterbi decoder using 6-state left-to-right HMM cross-word generalized triphone models. HMM states are tied pairwise such that each 6-state HMM is modeled by three separate Gaussian mixture distributions. A phonetic decision tree is used for tying the triphone models. The baseline system uses voicedness features (cf. Sec. 3.1) and fast VTLN (cf. Sec. 3.2).

5.2. Two-Pass Speaker Adapted System

As described in Sec. 3.2, a two-pass search strategy is used to facilitate speaker adaptation. The first pass was performed using the baseline VTLN/voicedness system, with the ML estimated acoustic model. Since no fine-grained segmentation of the data was provided in the evaluation, the complete recordings were used as input to the system. The recordings varied in length between a couple of minutes and half an hour. The silence information from the first recognition pass is used to segment the audio data for the second pass. The segment breaks are 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 [12]. The segmented and clustered corpus was used to estimate the CMLLR and MLLR matrices needed by the adaptation. The second pass finally was performed using the best acoustic models, discriminatively trained on the CMLLR-SAT transformed features, and adapted using the estimated CMLLR and MLLR matrices.

5.3. Bayes Risk Minimization

The quality of a speech recognition system is typically assessed by its word error rate (WER). However, the standard decision rule is based on minimizing the Bayes risk using the *sentence* instead of the *word* error count as cost function. As a consequence, a rescoring pass using the Minimum Bayes Risk (MBR) criterion with a WER based cost function was applied. The experiments reported here were carried out with the algorithm proposed in [13] which is applied on N-best lists.

6. Experiments

The experiments described in this paper were done in the context of the second TC-STAR ASR evaluation campaign. To monitor the progress of the system development several recognition experiments were performed comparing the effectiveness of different methods applied. Due to the large number of available methods, not all possible combinations were investigated.

Since the evaluation data certainly was not available beforehand, not all contrast experiments carried out on the development data were performed on the evaluation data. For Spanish, the development was mainly carried out on the EPPS part of the development corpus.

6.1. Baseline System

As described in Sec. 5.1 the baseline system already included VTLN and voicedness features. As a contrast, and for use in system combination, experiments were also performed with a plain baseline without VTLN and voicedness. Table 4 summarizes the results comparing the two baseline systems, both for English and Spanish.

Table 4: *Baseline WER [%] on EPPS development data.*

	English	Spanish
Baseline	18.5	13.2
VTLN+voice	17.2	11.9

6.2. Speaker Adaptation

On top of the ML baseline system already including VTLN, four different adapted systems were used, differing w.r.t. SAT model, and MLLR usage. Table 5 show the performance of the different systems for English and Spanish on the development corpora used. While SAT gives a clear improvement in the case without MLLR, SAT with MLLR was not observed to lead to further improvements. On the other hand, when SAT is used the improvement of MLLR is somewhat inconclusive: for English the improvement is substantial, but for Spanish it is neglectable. Note that for Spanish, the baseline already contains an improved language model.

Table 5: *Adaptation WER[%] on EPPS development data.*

	English	Spanish
Baseline	17.2	10.7
CMLLR	15.7	9.2
SAT	15.2	8.6
CMLLR+MLLR	14.0	8.6
SAT+MLLR	14.0	8.6

6.3. Discriminative Training

Table 6 summarizes the improvements resulting from discriminative training. Note that discriminative training in combination with MLLR did not perform consistently: for English discriminative training is beneficial whereas for Spanish the word error rate even increases. However, CMLLR-SAT combined with discriminative training and MLLR yields improvements on both corpora (see also discussion in Sec. 6.5). Furthermore, MCE slightly outperformed MMI.

Table 6: *Discriminative training performance (WER[%]) on EPPS development data.*

	English	Spanish
MLLR	14.0	8.6
MLLR+MMI	13.6	8.8
MLLR+MMI+SAT	13.3	-
MLLR+MCE+SAT	13.1	8.0

6.4. Bayes Risk Minimization

Table 7 compares the results for the best two-pass, SAT-based, discriminatively trained systems with and without using MBR. In English a marginal improvement was obtained whereas in Spanish no improvement could be observed. The observed reduced performance may be due to the relatively low error rates obtained for these tasks.

Table 7: *Performance of Bayes risk minimization (WER[%]) on EPPS development data.*

	English	Spanish
No MBR	12.9	7.8
MBR	12.8	7.8

6.5. Summary of Results

Tables 8 and 9 show the chronological progression of the results during the preparation for the evaluation campaign[1], as well as the corresponding results for the evaluation corpus, where available. Note that while the separate improvements of STM-SAT and discriminative training were small, the combined improvement was larger than the sum of the separate improvements, when compared to a ML trained system with both CMLLR and MLLR. A similar effect has been described in [14], where discriminative training was reported to give larger improvements when the system is using SAT and MLLR, as compared to only using MLLR.

Table 8: *Overview of English system performance (WER[%]).*

	Dev	Eval
Baseline	18.5	-
+VTLN+voice	17.2	14.4
+CMLLR	15.7	-
+MLLR	14.0	11.8
+MMI	13.6	11.7
+SAT	13.3	10.8
+New LM	12.9	10.3
+MBR	12.8	10.2

Table 9: *Overview of Spanish system performance (WER[%]). Note that the evaluation data contains EPPS and STC data.*

	Dev	Eval
Baseline	13.2	-
+VTLN+voice	11.9	-
+New LM	10.7	16.1
+MLLR	8.6	11.3
+MCE	8.8	11.1
+SAT	8.0	-
+Tuning	7.8	10.2

7. Conclusions & Outlook

In this work, the RWTH automatic speech recognition systems developed for the second TC-STAR evaluation campaign 2006 were presented. The systems were designed to transcribe parliamentary speeches taken from the European Parliament Plenary Sessions (EPPS) in European English and Spanish, as well as speeches from the Spanish Parliament. Using a two-pass decoding strategy a number of improvements could be obtained. Using several speaker adaptation and normalization schemes, speaker adaptive training, MCE and MMI training, and *Bayes* risk minimization, the overall improvement obtained on top of the baseline system ranged up to about 2/3. For all relevant system components, contrastive results are presented on the EPPS Spanish and English data. In addition experiments on system combination were performed but not used in the final evaluation. These experiments are described in [15].

Acknowledgements

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

8. References

[1] "TC-STAR, Technology and Corpora for Speech-to-Speech Translation Components," http://www.tc-star.com.

[2] "European Parliament," http://www.europarl.eu.int.

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

[4] Maximilian Bisani and Hermann Ney, "Joint-models for grapheme-to-phoneme conversion," *Speech Communication*, In press.

[5] 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.

[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] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sept. 2005, pp. 2133 – 2136.

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

[10] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proc. Int. Conf. on Spoken Language Processing*, Sidney, Australia, 1998.

[11] E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popović, S. Hasan, and H. Ney, "The RWTH machine translation system," in *Proc.* TC-STAR *Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 31–36.

[12] 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.

[13] A. Stolcke, Y. König, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," in *Proc. European Conf. on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997, pp. 373 – 400.

[14] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge, England, 2004.

[15] 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.

[1] The entry *Tuning* refers to language model scale tuning