The ISL 2007 English Speech Transcription System for European Parliament Speeches

Sebastian Stüker, Christian Fügen, Florian Kraft, and Matthias Wölfel

Institut für Theoretische Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany
{stueker,fuegen,fkraft,wolfel}@ira.uka.de

Abstract

The project *Technology and Corpora for Speech to Speech Translation* (TC-STAR) aims at making a break-through in speech-to-speech translation research, significantly reducing the gap between the performance of machines and humans at this task. Technological and scientific progress is driven by periodic, competitive evaluations within the project. In this paper we describe the ISL speech transcription system for English European Parliament speeches with which we participated in the third TC-STAR evaluation campaing in the spring of 2007. The improvements over last year's system originate from a recognition hypotheses based segmentation, the utilization of unsupervised in-domain training material, a modified cross-system adaptation and combination scheme, and the enhancement of the language model through the use of web based training material.

Index Terms: speech recognition, EPPS, TC-STAR, system description

1. Introduction

TC-STAR - Technology and Corpora for Speech to Speech Translation is a three year integrated project financed by the European Commission within the Sixth Framework Programme. The aim of TC-STAR is to advance research in all core technologies for speech-to-speech translation (SST) in order to reduce the gap in performance between machines and human translators. To foster significant advances in all SST technologies, periodic, competitive evaluations are conducted within TC-STAR for all components involved, including automatic speech recognition (ASR) research, as well as end-to-end systems. The main task for speech-to-speech translation within TC-STAR is European Parliament Plenary Sessions. This tasks consists of speeches delivered in the European Parliament as they are being broadcast through Europe by Sattelite (EbS), the European Union's TV station.

In spring of 2007 the third and final TC-STAR evaluation campaing took place. For the EPPS ASR task the evaluation offered three different conditions: the restricted condition which only allows for the use of EPPS training material, the public condition which allows for all publicly available training material, and the open condition which allows for any training material available to a specific participant. For the training material in all three conditions May 31st was the cut-off date prior to which the training material had to originate.

For the EPPS task we participated for English in the public condition. In this paper we describe the system we used to participate. Our transcription system is centered around a multi-pass decoding strategy[1] that incorporates acoustic models trained on two different phoneme sets and two acoustic-front ends that are combined with the help of conufison network combination (CNC) [2]. The main improvements and modifications to last year's system [3] are:

- Hypotheses based segmentation of the input audio
- Utilization of unsupervised acoustic training material
- Modified cross-system adaptation and combination scheme
- Incorporation of training material obtained by automatized web searches into the language model

Our system was mainly developed with the help of our inhouse Janus Recognition Toolkit (JRTk) which features the Ibis single pass decoder [4].

The rest of the paper is organized as follows. Section 2 introduces the acoustic front-ends used in our system, while Section 3 describes the acoustic model, and Section 4 the segmentation and clustering procedure of this year's system. Section 5 describes the language model used and Section 6 the adaptation and decoding strategy applied. Finally, Section 7 highlights the imrovements in terms of WER that we obtained through the new techniques used in this year's system over last year's.

2. Preprocessing

We trained systems for two different kinds of acoustic front-ends. One is based on the traditional Mel-frequency Cepstral Coefficients (MFCC) obtained from a fast Fourier Transform and the other on the warped minimum variance distortionless response (MVDR). The second front-end replaces the Fourier transformation by a warped MVDR spectral envelope [5], which is a time domain technique to estimate an all-pole model using a warped short time frequency axis such as the Mel scale. The use of the MVDR eliminates the overemphasis of harmonic peaks typically seen in medium and high pitched voiced speech when spectral estimation is based on linear prediction.

For training, both front-ends have provided features every 10 ms. During decoding this was changed to 8 ms after the first stage. In training and decoding, the features were obtained either by the Fourier transformation followed by a Mel-filterbank or the warped MVDR spectral envelope.

For the MVDR front-end we used a model order of 22 without any filter-bank since the warped MVDR already provides the properties of the Mel-filterbank, namely warping to the Mel-frequency and smoothing. The advantage of this approach over the use of a higher model order and a linear-filter bank for dimensionality reduction is an increase in resolution in low frequency regions which cannot be attained with traditionally used Mel-filterbanks. Furthermore, with the MVDR we apply an unequal modeling of spectral peaks and valleys that improves noise robustness, due to the fact that noise is mainly present in low energy regions.

Both frond ends apply vocal tract length normalization (VTLN) [6]. For MFCC this is done in the linear domain, for MVDR in the warped frequency domain. The MFCC front-end uses 13 cepstral coefficients, the MVDR front-end uses 15. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. For both front-ends seven adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 42 dimensions using *linear discriminant analysis* (LDA).

3. Acoustic Model Training

We trained a variety of phoneme based acoustic models for the final evaluation system. All of them are left-right Hidden Markov Models (HMMs) without state skipping with three HMM states per phoneme.

We trained acoustic models for two different kinds of phoneme sets *P1* and *P2*. P1 is a version of the Pronlex phoneme set which consists of 44 phonemes and allophones while P2 is a version of the phoneme set used by the CMU dictionary that consists of 45 phonemes and allophones. We trained models for all four combinations of the two phoneme sets and the two acoustic front-ends described in Section 2.

We trained the models on approx. 80h of English EPPS data provided by RWTH Aachen within the TC-STAR project [7], 9.8h of TED data [8], and 167h of unsupervised EPPS training material that had been collected within TC-STAR by RWTH Aachen but had not been manually transcribed. Transcriptions for the unsupervised training material were obtained by adapting an acoustic model of last year's system on automatic transcriptions provided by RWTH Aachen on that data. We then decoded the data, using the segmentation provided by RWTH Aachen.

All models are semi-continuous quinphone systems that use 16000 distributions over 4000 codebooks. They were trained using incremental splitting of Gaussians training, followed by 2 iterations of Viterbi training. For all models we used one global semi-tied covariance (STC) matrix after LDA [9] as well as Vocal Tract Length Normalization. In addition to that feature space constraint MLLR (cMLLR) speaker adaptive training [10] was applied on top.

We improved the acoustic models further with the help of Maximum Mutual Information Estimation (MMIE) training [11]. We applied MMIE training firstly to the models after the 2 viterbi iterations, and secondly to the models after the FSA-SAT training, taking the adaptation matrices from the last iteration of the maximum likelihood FSA-training and keeping them unchanged during the MMIE training.

This all resulted in eight different acoustic models: for each combination of front-end, MVDR and MFCC, and phoneme set, P1 and P2, one set of models trained with VTLN plus MMIE, and one with FSA-SAT plus MMIE. From now on we refer to these models as *P1-MFCC-VTLN*, *P2-MFCC-VTLN*, *P1-MVDR-VTLN*, *P2-MVDR-VTLN*, *P1-MFCC-SAT*, *P2-MFCC-SAT*, *P1-MVDR-SAT*, and *P2-MVDR-SAT*.

4. Segmentation and Clustering

For this year's system we used a different approach than in last year's system to segment the input data into smaller, sentence-like chunks used for recognition. For the purpose of segmentation we performed a fast decoding pass on the unsegmented input data in order to determine speech and non-speech regions. Segmentation was then done by consecutively splitting segments at the longest non-speech region that was at least 0.3 seconds long. The resulting segments had to contain at least eight speech words and had to have a minimum duration of six seconds.

In order to group the resulting segments into several clusters, with each cluster, in the ideal case, corresponding to one individual speaker we used the same hierarchical, agglomerative clustering technique as last year which is based on TGMM-GLR distance measurement and the Bayesian Information Criterion (BIC) stopping criteria [12]. The resulting speaker labels were used to perform acoustic model adaptation in the multipass decoding strategy described in Section 6.

corpus	#words	weights LM1	weights LM2
EPPS transcr	750k	0.12	0.12
EPPS FT	35M	0.42	0.35
Hub4 BN	130M	0.07	0.04
UN	41M	0.02	0.02
EPPS unsup	1.4M	0.08	0.04
Hansard	48M	0.19	0.16
Gigaword	167M	0.10	0.07
Web	643M	-	0.20

Table 1: The LM training corpora and interpolation weights

5. Language Model and Test Dictionary

5.1. Language Model

For language model (LM) trainig we used data from the following corpora: the manual transcriptions of the EPPS acoustic training data (EPPS transcr), the EPPS final text editions (EPPS FT), Hub4 Broadcast News data (Hub4 BN), the English part of the UN Parallel Text Corpus v1.0 (UN), the automatic transcriptions of the unsupervised EPPS training data (EPPS unsup), the ELDA Hansard corpus consisting of U. K. parliament debates (Hansard), data extracts from the LDC Gigaword corpus (Gigaword), and texts from an inhouse web data collection (Web). Table 1 gives the sizes of the individual corpora.

For the web data corpus we collected training material from the World Wide Web. In order to obey the cut-off date of May 31st the collection was done in the last week of May. The collection was performed with the help of web-collection and filter scripts provided by the University of Washington[13]. For that purpose we built 10k queries which were random shuffles of 784 topic ngrams and 400 general n-gram phrases. The topic n-grams were extracted from the EPPS final text editions by keeping the most frequent 4-grams and 3-grams (not in the 4-grams) not containing any stopwords, hesitations, numbers, or dates. The purpose of the general n-grams is to cover speaking style and were therefore extracted from the EPPS verbatim transcriptions by keeping the most frequent 4-grams and 3-grams (not in the 4-grams) excluding topic words. After reducing redundant webpages and applying a standard text postprocessing filter an amount of 643M words remained.

With the help of the SRI Language Modeling Toolkit [14] we trained separate 4-gram LMs for each corpus. The LMs use modified Kneser-Ney smoothing and interpolation of discounted n-gram probability estimates with lower-order estimates. By interpolating the invididual models we created two

stage/pass	technique	AM/pass	adapted on	dev06	eval06	eval07
1a	Decoding	P1-MFCC-VTLN	incr.	13.4%	11.2%	12.5%
1b	Decoding	P1-MVDR-VTLN	incr.	13.2%	11.3%	12.6%
1c	Decoding	P2-MFCC-VTLN	incr.	13.0%	11.0%	12.3%
1d	Decoding	P2-MVDR-VTLN	incr.	12.6%	11.3%	13.1%
1-1	CNC	1a,1c	–	11.8%	10.2%	11.4%
1-2	CNC	1b,1d	–	11.8%	10.2%	11.5%
2a	Decoding	P1-MFCC-SAT	1-2	11.4%	9.7%	11.0%
2b	Decoding	P1-MVDR-SAT	1-1	11.8%	9.5%	10.5%
2c	Decoding	P2-MFCC-SAT	1-2	11.1%	9.0%	10.5%
2d	Decoding	P2-MVDR-SAT	1-1	11.2%	9.2%	10.8%
2	CNC	2a,2b,2c,2d	–	10.4%	8.5%	9.5%
3a	Decoding	P1-MFCC-SAT	2	10.9%	8.6%	10.2%
3b	Decoding	P1-MVDR-SAT	2	11.1%	8.9%	10.1%
3c	Decoding	P2-MFCC-SAT	2	10.8%	8.4%	10.0%
3d	Decoding	P2-MVDR-SAT	2	11.0%	8.6%	9.8%
3	CNC	3a,3b,3c,3d	–	10.0%	8.0%	9.2%

Table 2: The individual stages and passes in the decoding scheme, giving the acoustic model (AM) used, or which outputs were combined respectively, the output adapted on, and WERs on the 2006 development, 2006 evaluation, and 2007 evaluation sets

language models that we used in our system.

Our first language model (LM1) was the result of the interpolation of all corpora mentioned above, except the web data LM. The interpolation weights were tuned on the 2006 EPPS eval data by minimizing the perplexity of the model. The resulting LM reached a perplexity of 83 on the 2006 EPPS development set, 89 on the 2006 evaluation set, and 85 on the 2007 evaluation set.

Our second language model (LM2) used the web data language model in addition. The interpolation weights were chosen manually as a trade off between tuning on the 2005 EPPS development set, the 2006 EPPS development set, and the 2006 evaluation set. The resulting model has a perplexity of 78 on the 2006 EPPS development set, 83 on the 2006 evaluation set, and 82 on the 2007 evaluation set.

5.2. Test dictionary

The test dictionary remained unchanged to last year's dictionary[3]. It was built by using all words from the EPPS transcripts and all words with more than three occurrences from the EPPS final text editions. Pronuciations missing from the intial dictionaries were created either manually or automatically with the help of Bill Fisher's tool [15] for P1 and Festival [16] for P2 respectively. This resulted in a case sensitive OOV rate of 0.43% on the 2006 EPPS development set, 0.47% on the 2006 evaluation set, and 0.45% on the 2007 evaluation set. All words were mapped to British English spelling.

6. Decoding Strategy and Results

Decoding within our recognition system was performed in three stages. For the first two stages we used the language model LM1, that is the language model without training material collected from the web, and for the third stage we switched to language model LM2 that includes the web data. The acoustic models of the second and third stage were adapted on the output(s) from the previous stage using Maximum Likelihood Linear Regression (MLLR) [17], Vocal Tract Length Normalization (VTLN) [6], and feature-space constrained MLLR (cMLLR) [10]. For the estimation of the cMLLR transformation matrices we used the SAT models before the application of the MMIE training. For the estimation of the other transformations we used the models after the MMIE training and kept the cMLLR matrices fix. For the second and third stage the frame shift during recognition was changed to 8ms as mentioned before.

In the first stage we used the acoustic models P1-MFCC-VTLN, P2-MFCC-VTLN, P1-MVDR-VTLN, P2-MVDR-VTLN. The resulting word lattices of P1-MFCC-VTLN and P2-MFCC-VTLN were then combined via confusion network combination to the output 1-1, and the lattices from P1-MVDR-VTLN and P2-MVDR-VTLN were combined to output 1-2. In this first stage we adapted the acoustic models using incremental VTLN and incrmental fMLLR on a per speaker basis.

For the second stage the MVDR SAT systems for the phoneme sets P1 and P2 were adapted on 1-1, while the MFCC SAT systems for P1 and P2 were adapted on 1-2. The resulting lattices from the second stage were then combined to a single output using CNC.

In the third stage the same acoustic models as in the second stage were then adapted on that output and the results of the recognition runs were combined to the final out, again using CNC. This final sytem yielded a word error rate of 10.2% on the 2006 EPPS development set (dev06), of 8.3% on the 2006 evaluation set (eval06), and of 9.2% on the 2007 evaluation set (eval07). Tabel 2 shows the word error rates of the individual recognition systems on these sets in the different stages in detail.

AM	dev2006		eval2006		eval2007	
	LM1	LM2	LM1	LM2	LM1	LM2
3a	11.1%	10.9%	9.4%	8.6%	10.7%	10.2%
3b	11.4%	11.1%	9.2%	8.9%	10.3%	10.1%
3c	10.9%	10.8%	8.7%	8.4%	10.2%	10.0%
3d	10.9%	11.0%	8.8%	8.6%	10.1%	9.8%
CNC	10.2%	10.0%	8.3%	8.0%	9.4%	9.2%

Table 3: Comparison of the language model w/o (LM1) and with web data (LM2) on the third stage of the decoding

7. Comparative Results

7.1. Segmentation

We examined the improvements from the new segmentation by taking the P1-MFCC-VTLN from the first stage of the evaluation system and comparing its performance on the old and the new segmentation. With the new segmentation the WER drops by 0.5% abs. from 13.9% to 13.4% on dev06, and by 0.6% from 13.1% to 12.5% on eval07.

7.2. Unsupervised Training Material

For measuring the performance gain from the unsuperviced training material we utilized a system from an intermediate stage in our system development. The system consists of 4000 models over 4000 codebooks and was trained with phonemeset P1 and the MFCC front-end as described above, but without MMIE training. The system was trained once without and once with the unsupervised training material. Without the unsupervised material the system reaches a WER of 14.9% on the 2006 development set, of 12.5% on the 2006 evaluation set, and of 14.3% on the 2007 evaluation set. When we include the unsupervised training material the word error rates drops considerably to 14.1% on dev06, 11.9% on eval06, and 13.7% on eval07.

7.3. Web Data

In order to determine the influence of the Web training material on the performance of the system we repeated the third stage of our evaluation system with the same language model than in the second stage, i.e. the language model without the web material. After the confusion network combination of the recognizer outputs the word error rate on eval07 increases by 0.2% absolute to 9.4%. Tabel 3 shows WERs in detail.

8. Conclusions

In this paper we have described our English speech recognition system with which we participated in the third TC-STAR evaluation campaign in spring of 2007 for transcribing European Parliamentary Plenary Sessions. On the 2007 evaluation set our system yields a WER of 9.2%. The main modification's over last year's system are the application of a hypotheses based segmentation to the input signal, the utilization of unsupervised training material, a modification of the cross-system adaptation and combination scheme, and the incorporation of web based training material into the language model.

9. Acknowledgements

This work has been funded by the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

10. References

[1] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, "Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end," in *Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2007.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, October 2000.

[3] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y.-C. Tam, and M. Wölfel, "The isl tc-star spring 2006 asr evaluation systems," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006.

[4] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one pass-decoder based on polymorphic linguistic context assignment," in *ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[5] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectralestimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, September 2005.

[6] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *ICASSP*, Munich, Germany, April 1997.

[7] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross domain automatic transcription on the tc-star epps corpus," in *International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USE, March 2005.

[8] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the ted corpus lectures," in *International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, March 2003.

[9] M. Gales, "Semi-tied covariance matrices for hidden markov models," Cambridge University, Engineering Department, Tech. Rep., February 1998.

[10] ——, "Maximum likelihood linear transformations for hmm-based speech recognition," Cambridge University, Engineering Department, Tech. Rep., May 1997.

[11] D. Povey and P. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, May 2001.

[12] Q. Jin and T. Schultz, "Speaker segmentation and clustering in meetings," in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, October 2004.

[13] M. O. I. Bulyko and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *HLT-NAACL*, Edmonton, Canada, May 2003.

[14] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," Denver, Colorado, USA, 2002.

[15] W. Fisher, "A statistical text-to-phone function using ngrams and rules," in *International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, USA, March 1999.

[16] A. Black and P. Taylor, "The festival speech synthesis system: System docmunation," Human Communication Research Centre, University of Edingburgh, Edingburgh, Scotland, United Kingdom, Tech. Rep., 1997.

[17] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.