The ISL RT-07 Speech-to-Text System

Matthias Wölfel, Sebastian Stüker, and Florian Kraft

Interactive Systems Laboratories
Institut für Theoretische Informatik
Universität Karlsruhe (TH)
Am Fasanengarten 5
76131 Karlsruhe, Germany
{wolfel|stueker|fkraft}@ira.uka.de

Abstract. This paper describes the 2007 meeting speech-to-text system for *lecture rooms* developed at the Interactive Systems Laboratories (ISL), for the multiple distant microphone condition, which has been evaluated in the RT-07 Rich Transcription Meeting Evaluation sponsored by the US National Institute of Standards and Technologies (NIST). We describe the principal differences between our current system and those submitted in previous years, namely the use of a signal adaptive front-end (realized by warped-twice warped minimum variance distortionless response spectral estimation), improved acoustic (including maximum mutual information estimation) and language models, cross adaptation between systems which differ in the front-end as well as the phoneme set, the use of a discriminative criteria instead of the signal-to-noise ratio for the selection of the channel to be used and the use of decoder based speech segmentation.

1 Introduction

In this paper, we present the ISL's most recent lecture meeting speech-to-text system for *lecture rooms*, which has evolved significantly over previous versions [1–4] and which were evaluated in the NIST RT-07 Rich Transcription Meeting Evaluation. The system described in this paper shares many common elements, e.g. the two phoneme sets and the cluster tree, with last years evaluation system as described in [1]. However, it differs from it in several important ways which will be described in this paper.

Notable improvements in the system architecture are:

- besides our standard *warped minimum variance distortionless response* (WMVDR) [5] front-end we have used a signal adaptive front-end provided by *warped-twice warped minimum variance distortionless response* (W2MVDR) [6] spectral estimation
- to exploit benefits from cross system adaptation and system combination we have varied both the front-end and phoneme set [7] which gives an additional accuracy improvement of 0.5% over the usage of a single phoneme-set

- replacement of the *signal-to-noise ratio* (SNR) by a class separability measure for channel selection [8]
- decoder based speech segmentation [9]
- improved acoustic models due to *maximum mutual information estimation* (MMIE) training [10] where the speaker dependent adaptation matrices are unchanged during the MMIE training

We also improved our language models by incorporating additional data collected from the world wide web. We used only acoustic models which have been trained with *vocal tract length normalization* (VTLN) [11], however incremental speaker adaptation in the first pass, as in last year system, was not used. Last but not least, we used different additional acoustic training material.

Most of the decoding experiments described in this paper were either conducted on the lecture meeting portion of the RT-05S development and evaluation set or the current RT-07 development set.

2 Automatic Segmentation

Automatic segmentation for the various conditions of the lecture subtasks is provided by different systems.

For the *individual head-mounted* (IHM) condition, which is particularly difficult due to cross talk from background speakers, we have relied on the segmentation and speaker clusters provided by ICSI [12].

For the *single distant microphone* (SDM) and *multi distant microphone* (MDM) condition we used a different approach than in previous years: We have used a multi-microphone extended version of the single-microphone system which we used in this years English European Parliament Plenary Sessions transcription system developed and evaluated under the TC-STAR project [9]. First, from every session, the channel of the unsegmented recording with the highest SNR is selected. In order to determine speech and non-speech regions a decoding pass is performed on the unsegmented audio. Segmentation is then done by consecutively splitting segments at the longest non-speech region that is at least 0.3 seconds long. The resulting segments had to contain at least three speech words and had to have a minimum duration of three seconds. The maximum duration was set to sixty 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 [13]. The resulting speaker labels were used to perform feature and acoustic model adaptation in the multi-pass decoding strategy as described in Section 4.1.

3 System Training and Development

All speech recognition experiments described in this paper were performed with the help of the *Janus Recognition Toolkit* (JRTk) and the *Ibis* single pass decoder [14].

3.1 Front-End and Phoneme Set

To increase accuracy via cross system adaptation we used two front-ends and phoneme sets. One front-end, identical to last years system, replaces the Fourier transformation by a WMVDR spectral envelope of model order 30. In contrast to our RT-06S system, we have replaced the mel frequency cepstral coefficient front-end by a signal adaptive front-end provided by W2MVDR spectral estimation. Due to the properties of the WMVDR, the mel filterbank has to be replaced by a linear filterbank (in the case of model order 60) or dropped completely (done so for model order 30). The advantages of the WMVDR approach are an increase in resolution in low frequency regions relative to the traditionally used mel filterbanks, and the dissimilar modeling of spectral peaks and valleys to improve noise robustness as noise is present mainly in low energy regions. The advantage of a signal adaptive front-end is that classification relevant characteristics are emphasized while classification irrelevant characteristics are alleviated according to the characteristics of the signal to be analyzed, e.g. vowels and fricatives have different characteristics and therefore should be treated differently.

Both front-ends use a 42-dimensional feature space based on 20 cepstral coefficients with linear discriminant analysis and a global *semi-tied covariance* (STC) transform [15] with utterance-based cepstral mean and variance normalization.

Table 1 and Table 2 compare different front-ends for close and distant data on RT-05S development and evaluation data (lecture meeting). A detailed description about the W2MVDR spectral estimation, the signal adaptive front-end and the training setup can be found in [6]. For close talking the proposed signal adaptive front-end is superior to all investigated front-ends. On distant speech the proposed signal adaptive front-end is superior to most of the investigated front-ends.

The *first phoneme set* (p1) used is an adapted version of the phoneme set used by the *Carnegie Mellon University* (CMU) dictionary that consists of 45 phonemes and allophones. The *second phoneme set* (p2) used is an adapted version of the Pronlex phoneme set which consists of 44 phonemes and allophones. Pronunciations of unknown words were either generated automatically by Festival [16] for the CMU dictionary or by Fisher's grapheme-to-phoneme conversion tool [17] for the Pronlex system.

3.2 Acoustic Model Training

The training setup was based on last years evaluation system. However, this year we selected the training data that performs best on distant talking audio. Therefore, we have used the following training material: CMU (11 hours),

Spectrum	Order	Cepstra	Class Separability			Word Error Rate %					
Test Set			Train	Develop	Eval	Develop			Eval		
Pass						1	2	3	1	2	3
Fourier	–	13	11.007	16.470	16.088	36.1	30.3	28.0	35.3	29.7	27.7
Fourier	–	20	11.620	17.929	16.299	36.0	29.7	27.7	37.2	31.3	28.4
WMVDR	60	13	10.768	16.813	16.261	35.0	30.0	28.2	35.5	29.9	27.6
WMVDR	60	20	11.337	18.022	16.614	34.5	29.1	27.3	35.3	29.6	27.3
WMVDR	30	13	10.900	17.675	16.702	34.6	29.8	27.8	34.7	29.6	27.2
WMVDR	30	20	11.386	18.630	17.318	33.9	29.1	27.4	34.9	29.2	26.9
W2MVDR	60	13	10.893	17.673	16.456	34.5	29.5	27.5	34.1	29.2	27.0
W2MVDR	60	20	11.473	18.510	16.818	34.1	28.8	26.8	35.4	29.0	26.3

Table 1. Class separability and word error rates for different front-end types and settings on close recordings

(note that in the WMVDR front-end with model order 30 applies no smoothing and dimension reduction by a filterbank)

Spectrum	Order	Cepstra	Class Separability		Word Error Rate %					
Test Set			Develop	Eval	Develop			Eval		
Pass					1	2	3	1	2	3
Fourier	–	20	14.786	13.470	61.9	52.0	51.1	61.0	55.0	51.7
WMVDR	60	20	14.487	14.161	60.9	51.2	49.7	59.6	51.7	49.5
WMVDR	30	20	15.111	14.155	59.0	50.5	48.9	59.3	52.1	49.9
W2MVDR	60	20	15.380	14.116	60.3	51.1	49.8	59.9	50.4	47.9

Table 2. Class separability and word error rates for different front-end types on distant recordings

ICSI (72 hours), NIST (13 hours) plus Phase 2 Part 1 which are recordings of meetings, TED (13 hours), and CHIL (10 hours) plus last year's *lecture meeting* development and evaluation data (6 hours) which are recordings of lectures. All the acoustic data is in 16 kHz, 16 bit quality and recorded with head-mounted microphones. Far-field data is available for ICSI, NIST and CHIL. Due to channel mismatch between ICSI and NIST data to the lecture meeting data we have used only the far-field data provided by CHIL for supervised adaptation of the close talking acoustic models to derive distant speech acoustic models.

The model set used this year was unaltered to the one used in the RT-06S evaluation. In comparison to previous systems, e.g. the RT-04S evaluation system [4], it has slight modifications by additional noise models for laughter and other human noises to augment the existing breath and general noise models, and a split of the filler model into a monosyllabic and a disyllabic fillers model.

Acoustic model training was performed with fixed state alignments, which were written by a small system (2000 codebooks) using a mel frequency cepstral coefficient front-end trained on ICSI, NIST (without Phase 2 Part 1) and TED only. We trained four different acoustic models (varying the two front-ends and

the two phoneme sets) for the final evaluation system. All of them are left-right *hidden Markov models* (HMM)s without state skipping with three HMM states per phoneme.

All acoustic models were trained in the same way, resulting in semi-continuous quint phone systems that use 16000 distributions over 4000 codebooks, with a maximum of 64 Gaussians per model.

The adapted gender independent acoustic model training (given the vocal tract normalization values for each speaker by a previous system) can be outlined as follows:

1. Training of the linear discriminant analysis matrix
2. Extraction of samples
3. Incremental growing of Gaussians
4. Training of one global STC matrix
5. Second extraction of samples
6. Second incremental growing of Gaussians
7. Two iterations of Viterbi training to train the distributions for the semi-continuous system and to compensate for the occasionally erroneous fixed-state alignments
8. Four iterations of FSA-SAT speaker adaptive training [18]
9. Decoding of the training data with the previous model and a unigram language model
10. Five iterations of MMIE training [10], leaving the speaker dependent adaptation matrices from the last iteration of the maximum-likelihood speaker adaptive training unchanged during the MMIE training [19]

To adapt to the distant data we adapted the models (after step 7) by

1. Four Viterbi training iterations using the available far-field CHIL data
2. To reduce the impact of distant data on the models we combine the distant adapted models with the clean speech models of step 7 with a weight four times higher than the clean speech models of step 7.

3.3 Language Model Training

We used a 4-gram language model trained on the following corpora: A subset of CHIL transcriptions (ISL_20031028, ISL_20031216_A, ISL_20031125_B, ISL_20040614, ISL_20040616, ISL_20040621, ISL_20040721, ISL_20040830), rt04s-dev and rt04s-eval transcripts, meeting transcripts (ICSI, CMU, NIST, AMI), TED transcripts, Hub4 broadcast news, recent proceedings data ranging from 2002 - 2005, web data from University of Washington (150M words related to CMU, ICSI, NIST meetings), two subsets of inhouse Web data collections and a subset of the RT-06S evaluation data.

Subsets of the following pool from an inhouse web data collection were used. Therefore general phrase 3- and 4-grams were combined with topic phrases.

The general phrases in the queries for the corpora webI-III are based on the most frequent n-grams in CHIL transcriptions and for the corpora webIV-V on most frequent n-grams in the meeting transcripts.

The topic phrases were generated by computing bi-gram based tf-idfs for each proceeding paper. After merging them together and skipping bi-grams including stop-words the top 1,400 topic phrases were mixed randomly with the general phrases until the necessary number of queries were generated. For collecting the data we used the scripts provided by the University of Washington [20]. Table 3 gives an overview of the web data collections and the queries they were based on.

Corpus	General Phrases	Topic Phrases	Queries	Words
webI	CHIL transcripts		1k	146M
webII	CHIL transcripts	recent proceedings	4k	102M
webIII	CHIL transcripts	recent proceedings	10k	311M
webIV	meeting transcripts	recent proceedings	4k	398M
webV	meeting transcripts	recent proceedings	10k	674M

Table 3. Data that the web collection query generation was based on and sizes of collected web data components.

We trained one language model component for a subset of webI-III (318M words) and one component for webIV-webV (613M words). The subset selection was performed by skipping data from irrelevant queries, based on their perplexity with an in-domain LM build on the CHIL data used for query generation and the proceedings data.

Initially all mentioned language model components except the RT-06S evaluation data were interpolated (LM-A) according to an initial held-out set consisting of the CHIL transcriptions ISL_20031111, ISL_20031118, ISL_20031125_A, ISL_20031216_B, ISL_20041111_A, ISL_20041111_B, ISL_20041111_C and ISL_20041112_A. We used the resulting language model to update the held-out set with respect to the RT07 development data. When incrementally adding the RT-06S development set and the RT-06S evaluation set to the held-out set, perplexity on the RT07 development data decreased, while adding the NIST phase 2 part 1 (NIST07) set hurt as shown in Table 4. The motivation for this procedure is to get a tuning set biased to the RT07 development set, which is not selected too narrow.

The extension of the LM-A with the sets NIST07, RT-06S development set and RT-06S evaluation set revealed neither to use the new NIST set nor the RT-06S development set for component modeling purpose, but to use the RT-06S evaluation set. Since experiments for held-out set selection also showed an improvement when adding the RT-06S evaluation set, we split it to use for both. Consequently we used the initial held-out set plus the RT-06S development set plus part of the RT-06S evaluation set as tuning set. The final language model consists of all already mentioned language model components without the NIST07 set and only the part of the RT-06S evaluation set not used for tuning.

LM Components	Tuning Set	PPL
LM-A	initial set	132
LM-A	+ RT-06S-dev	130
LM-A	+ RT-06S-dev + RT-06S-eval	128
LM-A	+ RT-06S-dev + RT-06S-eval + NIST07	132
+ NIST07	initial set	132
+ RT-06S-dev	initial set	132
+ RT-06S-eval	initial set	130
+ subset(RT-06S-eval)	+ RT-06S-dev + subset(RT-06S-eval)	127
+ subset(RT-06S-eval) (i)	+ RT-06S-dev + subset(RT-06S-eval)	120
+ subset(RT-06S-eval) (ip)	+ RT-06S-dev + subset(RT-06S-eval)	123

Table 4. Tuning set selection and test of new corpora on the RT-07 development set.

During the system development we also considered to adapt the language model using a web data collection based on an automated query generation by extraction of topic and style from the hypotheses of previous recognition passes. Unfortunately the methods used lead to no further gain in recognition performance on top of the web data already included.

The final LM was build using the SRILM-toolkit [21]. For discounting we applied the Chen and Goodman's modified Kneser-Ney approach [22] and interpolation of discounted n-gram probability estimates with lower-order estimates was used (marked as (i) in Table 4). Pruning was performed after combining the LM-components (marked as (p) in Table 4) while the threshold was set also with respect to a reasonable decoding time. The perplexities are 123 on the RT-07 development set and 101 on the RT-07 evaluation set.

3.4 Recognition Dictionary and Lexicon

The dictionary contained 58.7k pronunciation variants over a vocabulary of 51.7k. The vocabulary was automatically derived by analysis of BN, Switchboard, meetings (ICSI, CMU, NIST, AMI), TED and CHIL corpora. After applying individual word-frequency thresholds to the corpora, we filtered the resulting list with `ispell` to remove spelling errors and added a few manually checked topic words from the set of topic bigrams used in web data collection. The OOV-rate on *lecture meeting* development and evaluation was 0.7% and 0.6% respectively.

4 Experiments and Results

In this section we present experiments and results on the RT-07 development and evaluation set.

4.1 Decoding Strategy

In order to find the best decoding and cross system adaptation strategy, we performed several different experiments on the lecture meeting development set. We found that the best setup in terms of *word error rate* (WER) and complexity for all conditions uses already vocal tract normalized acoustic models in the first pass while following passes use vocal tract normalized and speaker-adapted models (FSA-SAT). For distant speech data, similar to last year's system, we used close talking models which have been adapted to distant speech data. However, this year, we switched to the close talking model already in the second pass (which gave significant improvements on the development set). Last year, we switched to the close talking models in the third pass.

In another set of experiments, we followed results presented in [23, 24] and our own experience obtained during the development of a system for transcribing English European Parliament Plenary Sessions [25]. There a significant gain (approximately 1.5% absolute) from cross adaptation between systems with different front-ends (WMVDR, Fourier) is seen, and that, when cross adaptation between WMVDR and Fourier leads to no further gains, cross adapting with the Pronlex system improves the WER after *confusion network combination* (CNC) [26] by 0.7% absolute [7].

On development and evaluation data of the lecture meeting RT-07 data we saw improvements by system adaptation of approximately 2% absolut by the combination of the two front-ends with different phoneme sets for all passes.

The processing steps for decoding can be summarized as follows:

1. Decoder based segmentation
2. Speaker clustering
3. estimation of VTLN
4. Calculate SNR for each channel, segment, and utterance
5. decode first pass on combined acoustic channels for WMVDR.p1 and W2MVDR.p2
6. combine runs with confusion networks
7. select channel by class separability measures (uses VTLN) for each individual front-end over each segment and utterance, more detail in Section 4.2
8. adapt VTLN, constrained MLLR and MLLR
9. decode second pass on best channel for WMVDR.p2 and W2MVDR.p1
10. combine runs with confusion networks
11. adapt VTLN, constrained MLLR and MLLR
12. decode third pass on best channel for WMVDR.p1 and W2MVDR.p2
13. combine runs with confusion networks

Using an 8 ms instead of a 10 ms frame-shift for the second and third passes, improves the final WER by about 1% absolute [1].

4.2 Channel Combination and Selection for MDM

In RT-04S, channel combination was performed by decoding all channels and the combination by CNC on the resulting lattices over all channels. No selection

Channel Selection	WER %		
Pass	1	2	3
Signal to Noise Ratio	73.0	62.3	59.5
Class Separability Measure	67.4	57.8	55.1

Table 5. Influence of different channel selection techniques, signal to noise and class separability measure, on the *word error rates* (WER)s on development 2007.

was used, leading to a relatively high computational load for each pass. In the RT-06S system we were able to reduce the computational load by 70% without an increase in WER by performing both channel combination and selection [2].

This years scenario is different as such as we can't assume one dominant speaker and that the best possible microphone is changing for each individual speaker due to head orientation. Therefore, in this years task channel selection is even more important than last year as the signal quality of one channel might be significantly better than those of the other channels. In those cases microphone array or blind source separation techniques might not lead to improvements over the best single microphone.

We have presented a novel channel selection method [8], based on class separability, to improve multi-source far distance speech-to-text transcriptions. Class separability measures have the advantage, compared to other methods such as the SNR, that they are able to evaluate the channel quality on the actual features of the recognition system. Note that for different front-ends different channels might be selected.

A direct comparison between delay-and-sum channel combination and the proposed channel selection technique on the second pass of the RT-07 evaluation system including both front-ends and phone sets combined by CNC shows a relative improvement of 3.6%, from 52.4% to 50.5% WER.

4.3 Overall System Performance

Table 4.3 lists the overall system results for the development and evaluation RT-07S lecure meeting task. The given WERs per pass are after CNC of the lattices of the WMVDR and W2MVDR front-ends with different phoneme sets. The final pass on the IHM and MDM evaluation set give the official numbers as scored by NIST. All other numbers where scored in our laboratory.

On the MDM task, it can be seen that there is a huge gap between the adapted development and evaluation results. On the development set we gain 7.1% by adaptation, while in the evaluation case we gain only 4.4%. The selection of channels is able to improve the system performance by 2.3%.

5 Acknowledgments

This work was partly funded by the European Union (EU) under the integrated project CHIL [27] (IST-506909).

condition	IHM		SDM	MDM		
pass	dev	eval	eval	dev	*compare*	eval
1	36.5	43.1	57.9	56.7	60.2*	56.5
2	29.5	36.3	54.9	50.5	56.8	52.4
3	28.6	36.7	54.4	49.4	54.4	52.1
RT		91	113			114

Table 6. Overall results and real-time factors on RT-05S Eval and RT-06S Eval. In contrast to previous sections, results for the conference meeting part of RT-05S Eval include meeting NIST_20050412-1303. SDM and MDM results were scored with an overlap of one.

(*compare* give the numbers of last years evaluation system on the current evaluation set, note that the fist pass marked with * has been adapted incrementally)

References

1. C. Fügen, M. Wölfel, J. W. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani, "Advances in Lecture Recognition: The ISL RT-06S Evaluation System," in *Interspeech*, 2006.
2. M. Wölfel, C. Fügen, S. Ikbal, and J. W. McDonough, "Multi-Source Far-Distance Microphone Selection and Combination for Automatic Transcription of Lectures," in *Interspeech*, 2006.
3. M. Wölfel and J. McDonough, "Combining Multi-Source Far Distance Speech Recognition Strategies: Beamforming, Blind Channel and Confusion Network Combination," in *Interspeech*, 2005.
4. F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz, "Issues in Meeting Transcription – The ISL Meeting Transcription System," in *Proc. of the Intl. Conf. on Speech and Language Processing (ICSLP)*, 2004.
5. M. Wölfel and J. McDonough, "Minimum Variance Distortionless Response Spectral Estimation Review and Refinements," *IEEE Signal Processing Magazine*, September 2005.
6. M. Wölfel, "Warped-twice minimum variance distortionless response spectral estimation," *Proc. of EUSIPCO*, 2006.
7. 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 *Interspeech*, 2006.
8. M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Interspeech*, 2007.
9. S. Stüker, Fügen, F. Kraft, and M. Wölfel, "The ISL 2007 english speech transcription system for european parliament speeches," in *Interspeech*, 2007.
10. 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.
11. P. Zhan and M. Westphal, "Speaker Normalization Based on Frequency Warping," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997.
12. K. Boakye and A. Stolcke, "Improved speech activity detection using cross-channel features for recognition of multiparty meetings," in *Proc. Interspeech*, 2006.

13. Q. Jin and T. Schultz, "Speaker Segmentation and Clustering in Meetings," in *Proc. of the Intl. Conf. on Speech and Language Processing (ICSLP)*, 2004.
14. H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2001.
15. M. J. F. Gales, "Semi-tied covariance matrices," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.
16. A. W. Black and P. A. Taylor, "The Festival Speech Synthesis System: System documentation," Human Communciation Research Centre, University of Edinburgh, Edinburgh, Scotland, United Kongdom, Tech. Rep. HCRC/TR-83, 1997.
17. W. M. Fisher, "A Statistical Text-to-Phone Function Using Ngrams and Rules," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.
18. M. J. F. Gales, "Adaptive training schemes for robust asr." in *Proc. of ASRU*.
19. J. McDonough, T. Schaaf, and A. Waibel, "On Maximum Mutual Information Speaker-Adapted Training," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.
20. "Scripts for web data collection provided by University of Washington," http://ssli.ee.washington.edu/projects/ears/WebData/web_data_collection.html.
21. A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. of the Intl. Conf. on Speech and Language Processing (ICSLP)*, 2002.
22. S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Computer Science Group, Harvard University, Tech. Rep. TR-10-98, 1998.
23. H. Yu, Y.-C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The ISL RT04 Mandarin Broadcast News Evaluation System," in *EARS Rich Transcription Workshop*, 2004.
24. L. Lamel and J.-L. Gauvain, "Alternate Phone Models for Conversational Speech," in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.
25. S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, and M. W. M. Raab, Y.-C. Tam, "The ISL TC-STAR Spring 2006 ASR Evaluation Systems," in *TC-Star Workshop on Speech-to-Speech Translation*, 2006.
26. L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus among Words: Lattice-based Word Error Minimization," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1999.
27. "CHIL – Computers in the Human Interaction Loop," http://chil.server.de.