The ISL RT-06S Speech-to-Text System

Christian Fügen[1], Shajith Ikbal[1], Florian Kraft[1], Kenichi Kumatani[1], Kornel Laskowski[1], John W. McDonough[1], Mari Ostendorf[1,2], Sebastian Stüker[1], and Matthias Wölfel[1]

[1] Interactive Systems Laboratories, Universität Karlsruhe (TH), Karlsruhe, Germany
[2] Dept. of Electrical Engineering, University of Washington, Seattle, WA, USA
{fuegen|shajith|fkraft|kumatani|kornel|jmcd|mo|stueker|wolfel}@ira.uka.de

Abstract. This paper describes the 2006 lecture and conference meeting speech-to-text system developed at the Interactive Systems Laboratories (ISL), for the individual head-mounted microphone (IHM), single distant microphone (SDM), and multiple distant microphone (MDM) conditions, which was evaluated in the RT-06S 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 improved acoustic and language models, cross adaptation between systems with different front-ends and phoneme sets, and the use of various automatic speech segmentation algorithms.

1 Introduction

In this paper, we present the ISL's most recent speech-to-text systems for lectures and conference meetings, which have evolved significantly over previous versions [1–3] and which were evaluated in the NIST RT-06S Rich Transcription Meeting Evaluation.

The systems described in [1] and [3] shared many common elements, e.g. front-end, phoneme set, and training strategy. The systems described in this paper differ from them in several important ways. Notably, we used only speaker-adapted acoustic models. Even in the first pass, we used models trained with vocal tract length normalization (VTLN), and employed speaker-based incremental adaptation during decoding. Several acoustic models with different front-ends were trained: besides our standard FFT MFCC front-end, we also trained a system with a minimum variance distortionless response (MVDR) [4] front-end. Furthermore, in addition to our standard phoneme set, which was used in RT-04S [3], we also trained a system based on the PRONLEX phoneme set in order to exploit benefits from cross-adaptation and system combination [5]. We also improved our language models by incorporating data collected from the world wide web. Last but not least, we used different speech segmentation algorithms compared to the one used in our RT-04S evaluation system [6].

Most of the decoding experiments described in this paper were conducted on the lecture meeting portion of the official RT-06S development set, which is

identical to the RT-05S evaluation set further referred to as *lectDEV*, and only a small portion of experiments were done on the conference meeting portion of this set, *confDEV*. The corresponding evaluation sets are named *lectEVAL* and *confEVAL*. The *confDEV* results in this paper exclude one NIST meeting (`NIST_20050412-1303`), as it contained a participant on speakerphone, a condition which was guaranteed not to appear in *confEVAL*.

2 Automatic Segmentation

Automatic segmentation for the various conditions of the lecture and conference subtasks is provided by different systems. For the IHM condition, which is particularly difficult due to cross-talk from background speakers, we developed separate systems for the lecture and conference meeting subtasks. The lecture segmenter is an improved version of a single-microphone system which we used in the TC-STAR project [7]; the conference segmenter is a further evolution of our multi-microphone RT-04S IHM segmentation system. During RT-06S development, the two IHM systems further diverged due to subdomain-specific challenges: participants without microphones in the lecture task, and overlap and participant interaction in the conference task. We describe the two IHM segmenters below. The MDM and SDM segmentation was the same for both the lecture and conference meeting subtasks, and brief mention is included in the lecture segmentation description.

Lecture Meeting Segmentation

Our IHM lecture segmentation approach uses the following speech activity features extracted with a frame size of 32 ms and a frame shift of 10 ms: frame energy in decibels (E), mean and variance normalized E passed through a sigmoid function (E_n), energy-normalized linear prediction error (L) [8], slope along the frequency axis of a mel-warped filter-bank spectrum (S), and SPEECH/NON-SPEECH posteriors (P) computed using a multi-layer perceptron (MLP) trained with standard MFCC features. Segmentation for the IHM condition is performed in three steps [9]:

1. *Background speech activity rejection* discards regions of prominent cross-talk. This step uses E from all available microphones for a particular meeting and additional constraints such as presence of a minimal percentage of voiced speech and minimum duration.
2. *Foreground speech activity detection* identifies regions of reasonably prominent foreground speech activity. This step uses S, E_n, and L and a median filter of length 0.5 s.
3. *Sentence breaking* further cuts down the segments into shorter segments at points of high confidence NON-SPEECH, assuming those would correspond to actual sentence breaks; NON-SPEECH confidences were estimated using duration and average E.

For the IHM condition, after these steps, all segments from a single microphone are assumed to be produced by a single speaker. For SDM, only the sentence breaking step above is performed, assuming that the recognizer is the best system for discarding NON-SPEECH. The resulting segments are further tagged with speaker labels using a hierarchical agglomerative speaker clustering technique [6]. For the MDM condition, a single best channel is first chosen based on average SNR to perform further processing similar to the SDM condition.

Conference Meeting Segmentation

IHM segmentation in this subdomain is performed in 3 steps:

1. Initial label assignment is performed using our RT-04S IHM conference meeting segmenter.
2. Rather than decoding the 2-state SPEECH/NON-SPEECH activity of each participant independently [10, 11], we find the best Viterbi path through a 2^K-state vocal interaction space, where K is the number of participants in the test meeting [12]. This allows us to impose constraints on the degree of overlap in each meeting [13]. Single-Gaussian, multivariate acoustic models are trained on the test data using the initial labels from (1) above, and algorithms published in [12]. Our transition model is trained on the multichannel, manual turn segmentation available from the orthographic transcription of meetings collected at the ISL. It has the form:

$$P\left(q_{t+1} = S_j \mid q_t = S_i \right) \doteq P\left(\|S_j\| \, , \, \|S_j \cap S_i\| \mid \|S_i\| \right) \tag{1}$$

   where $\|S_i\|$ and $\|S_j\|$ are the numbers of participants in SPEECH in the interaction states S_i and S_j, respectively, and $\|S_j \cap S_i\|$ is the number of participants in SPEECH in **both** S_i and S_j. A best single-participant SPEECH/NON-SPEECH path ψ_k^* is then extracted for each participant from the best multi-participant interaction path q^*.
3. Each single-participant path ψ_k^* is independently smoothed, by eliminating short intervals of speech activity and short speech activity gaps.

This algorithm significantly outperforms the segmenter used in our RT-04S evaluation system. In Table 1 we show 8th pass WER results using our RT-04S meeting recognizer on the RT-04S eval data, together with our automatic RT-04S segmentation, our automatic RT-06S segmentation, and manual segmentation. Similarly, the table gives 1st pass WER results using our RT-06S meeting recognizer on *confDEV* and *confEVAL*, with the same three segmentation systems. As this table shows, for different passes, different recognizers, and different data sets, the word error rate using our RT-06S conference meeting segmentation is 50%-80% lower than that using our RT-04S segmenter, relative to manual segmentation.

Table 1. ASR errors committed by the last pass of our RT-04S STT system and the first pass of our RT-06S STT system, using our RT-04S meeting segmentation, our RT-06S meeting segmentation, and manual segmentation, on the IHM condition for conference meeting data.

Segmentation	RT-04S, last pass				RT-06S, first pass							
	RT-04S eval data				*confDEV*				*confEVAL*			
	del	ins	sub	WER	del	ins	sub	WER	del	ins	sub	WER
RT-04S (V43c)	19.3	2.5	13.9	35.7	17.8	3.4	18.4	39.5	22.1	10.0	20.4	52.4
RT-06S (01)	14.2	1.8	14.1	30.1	13.9	3.1	20.0	37.0	15.1	5.3	21.5	42.0
manual	11.5	2.8	14.7	28.9	10.3	2.8	21.3	34.4	9.5	5.0	23.0	37.6

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].

The following acoustic model training data was used: CMU (11hrs), ICSI (72hrs), NIST (13hrs) and AMI (16hrs) which are recordings of meetings, TED (13hrs), and CHIL (10hrs) which are recordings of lectures, and Hub4-BN (180hrs) which contains recordings of news broadcasts . All the acoustic data is in 16 kHz, 16 bit quality and recorded with head-mounted microphones, except for the CMU and Hub4-BN training data, which were recorded with either lapel or other microphones. For ICSI, NIST and AMI, farfield channels were also available.

3.1 Signal Processing

In contrast to our RT-04S system, we used two different front-ends to increase performance via cross-adaptation. The first front-end uses a 42-dimensional feature space based on MFCC with linear discriminant analysis (LDA) and a global semi-tied covariance (STC) transform [15] with utterance-based cepstral mean subtraction (CMS). It is identical to the one used in RT-04S. The second front-end replaces the Fourier transformation by a warped minimum variance distortionless response (MVDR) spectral envelope of model order 30. Due to the properties of the warped MVDR, neither the mel-filterbank nor any other filterbank was used. The advantages of the MVDR 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. Furthermore, the number of cepstral coefficients has been increased from 13 to 20. As before, a 42-dimensional feature space after LDA and a global STC transform with utterance based CMS was used.

3.2 Acoustic Model Training

The training setup was based on experiments performed during the development of the lecture translation system [1]. We selected the training data that

performs best on close talking audio, using only the ICSI, NIST and TED data and skipping the CMU and the Hub4-BN training material. This reduced the WER from 36.0% to 34.8% on *lectDEV*. We also changed the model set used in RT-04S slightly by adding noise models for laughter and other human noises to the existing breath and general noise models, and splitting the filler model into one for monosyllabic and another for disyllabic fillers.

Table 2. IHM improvements over the system developed in [1] on *lectDEV*. First pass with incremental VTLN and feature-space constrained MLLR (FSA) estimation and a frame shift of 10 ms, second pass with static VTLN, FSA and MLLR and 8 ms frame shift.

Pass	Acoustic Model Training Data	#codebooks	WER
1	ICSI+NIST+CMU+TED+BN	6000	32.6
	ICSI+NIST+TED	4000	31.5
2	ICSI+NIST+CMU+TED+BN	6000	28.4
	ICSI+NIST+TED	4000	27.0

For both subdomains, lecture and conference meetings, acoustic model training was performed with fixed state alignments, which were written by a small system (2k codebooks) trained on the corpora mentioned above. Both the MVDR and the FFT system were trained in the same way, resulting in a size of 16k distributions over 4k models, with a maximum of 64 Gaussians per model. The training was similar to that used in [1], with one modification. A second pass for incremental growing of Gaussians was performed after the STC training, which leads to an additional gain of 0.3% resulting in a WER of 32.0% on the IHM data of *lectDEV*. To train the distributions for the semi-continuous system and to compensate for the occasionally erroneous fixed-state alignments, 2 iterations of Viterbi training were performed. For the ML-SAT models, three additional iterations of maximum-likelihood speaker adaptive training (ML-SAT) [16] were run, wherein feature space adaptation and MLLR parameters were estimated for all speakers in the training set.

In addition to the FFT and MVDR systems, we trained another system using the PRONLEX phoneme set. The initial versions of the training and recognition lexica were a merger of the `callhome_english_lexicon_97061` dictionary and the LIMSI SI-284 training dictionary. Frequently missing words were added manually, and all other missing words were generated automatically with the help of a grapheme-to-phoneme conversion tool [17]. For the systems based on this phoneme set, context-independent acoustic models were trained from flat models. From them, fully context-dependent models were clustered in the same way as for the other phoneme set. The training of the context-dependent models followed the same scheme as for the other phoneme set, with the difference that 24k distributions over 3k models with a maximum of 64 Gaussians per model were used and only feature space adaptation parameters were estimated during ML-SAT.

For the lecture meeting system we used maximum a posteriori (MAP) adaptation with a weight of 0.8 for the CHIL data to adapt our semi-continuous models and gained 0.6% on top of the 32.0% WER on *lectDEV*. In a post-eval experiment, we added that data to our initial training set instead of using MAP and obtained a slight gain on *lectDEV*. During ML-SAT training, these models were applied to the CHIL training data with a weight of 4.0. Comparing the resulting system to the system used in [1], we improved our second pass result by 1.4% absolute (see Table 2, second row).

For the conference meeting system, we used exactly the same acoustic models, except for one difference: the PRONLEX system was additionally adapted using MAP with a weight of 0.8 for the AMI training data.

For the farfield channels, we adapted the models by appending two Viterbi training iterations using the farfield ICSI and NIST meeting data to the close-talking models. Using the AMI farfield data gave no further gains on *lectDEV*.

3.3 Language Model Training

All systems use 4-gram mixture language models (LMs). Three separate LMs were trained – for lectures, non-AMI conference-style meetings, and AMI conference meetings – since the speaking style and topics were qualitatively different in these subsets. The meeting transcripts, web text, and other sources used in training were subdivided so that component LMs might be weighted differently according to style (see Table 3). Mixture weights for each LM were optimized on a held out set of data: 30k for lectures, and 53k for the AMI and non-AMI conference meetings. All the LMs were built using the SRILM-toolkit [18], with modified Kneser-Ney discounting [19]. Pruning was performed after the interpolation of the LM-components, using a fixed threshold 10^{-9}.

For web text collection, we employed two different web query strategies. For the web-L and web-M-A collections, we followed the same web text collection framework as proposed in [20], where frequently spoken 3-grams and 4-grams from the target task training data are combined to form queries. For the other collections, the frequent n-grams from different lecture or conference meeting transcripts were combined with topic bigrams to form queries: web-MP with frequent n-grams from the conference meeting transcripts and web-LP with frequent n-grams from the lecture transcripts, respectively.[3] The goal was to obtain text reflecting a broad variety of topics, some of which are not represented in the training set. All UKA web data was perplexity filtered to 60% of the original collection sizes, with the exception of the query-based filtering where size was chosen to roughly match the UW meeting-based web collection (UW web-M).

The topic phrase generation consisted of: computing bigram tf-idf (term frequency – inverse document frequency) weights for each document in the proceedings data, zeroing all but the top 10%, averaging these weight vectors over the collection, and taking the top 1,400 bigrams excluding any with stop-words

[3] The web-LP corpus includes as a subset the web-L corpus, with redundancy between the collections removed.

Table 3. Corpora and size used in training the LM components. Data that the web collection query generation was based on is given in square brackets.

For all LMs:	
non-AMI meetings (ICSI, CMU, NIST, LDC)	1095 K
AMI meetings (RT-05S Dev: AMI-draft, AMI-final)	203 K
CHIL lectures	74 K
UW web-M [non-AMI meetings]	150M
UKA web-MP [non-AMI meetings, proceedings]	613M
w/ query-based filtering	124M
For the lecture meeting LM only:	
Translingual English Database (TED)	98 K
Hub4 Broadcast News	131M
recent speech/language proceedings (2002-2005)	130M
UKA web-L [CHIL]	146M
UKA web-LP [CHIL, proceedings]	318M
w/ query-based filtering	130M
For conference meeting LMs only:	
Switchboard CTS	4M
Fisher CTS	22M
UKA web-M-A [AMI meetings]	458M
UW web-F [Fisher CTS]	525M

or numbers (e.g. "Section 1"). The topic bigrams were mixed randomly with the general phrases until the desired number of queries (14k) was generated.

Table 4. Perplexity (PPL) and word error rate (WER) on *lectDEV* using language models with different data source mixture components.

LM	Components	PPL	WER
0	No web data	142	31.1
A	+ UW web-M	131	30.2
B	+ UKA web-L	132	30.2
C	+ UW web-M + UKA web-L	130	30.0
D	+ all web (query filtered)	128	29.9
E	+ all web (doc filtered)	126	29.8
F	(E) – UW web-M	126	29.6
G	(E) – BN,TED	126	29.7

We ran a series of experiments with different sets of web data as shown in Table 4. Not surprisingly, the biggest impact is associated with incorporating any web data, regardless of type. Using more web data gives further improvement (LMs A-B vs. C-G). Both web-L and the UW web data alone yielded similar performance (LMs A vs. B), though the web-L queries were better matched to the lecture task. However, the two are somewhat complementary and give a small gain when combined. We compared query-based vs. document-based perplexity filtering (LMs D vs. E), since some of the queries generated by randomly com-

bining topic words and lecture n-grams effectively mixed topics. Size differences in the collections make it difficult to compare methods, but since the difference was small and document-based filtering is more flexible, we used the latter in subsequent experiments. Examining the weights used for different components of LM-E (see Table 5), we noted that a small weight was given to the UW web-M data once the other (more topic-oriented) collections were available, and we observed a small gain in performance when it was removed (LM-F). We separately explored removing the low weight text sources (LM-G) and again observed a small but not significant gain. Overall, the best case model reduced perplexity by roughly 10% and WER by roughly 5% relative, compared to using no web data at all. Due to time constraints, the lecture system as applied in the evalua-

Table 5. Weights learned for the different component LMs for the lecture task associated with LM-E in Table 4.

Speech Transcripts		Web Text		Other	
CHIL lectures	0.25	UKA web-LP	0.20	proceedings	0.19
non-AMI meetings	0.14	UKA web-MP	0.10	TED	0.004
AMI meetings	0.08	UW web-M	0.05	BN	0.004

tion used LM-C for most conditions, though LM-D was used in later passes for the IHM condition. Compared to the old 4-gram LM used in [1], we gain 1.6% absolute from using LM-C, or 1.9% absolute if we use the best model obtained with subsequent development.

We did no further development for the conference meeting language models other than to introduce new web data. Based on the results of prior ICSI work [21], we did not use the BN or TED data but included CTS transcripts. In Table 6, the weights for the different language model components confirm the different nature of the AMI meetings. In addition to the expected differences of matched vs. mismatched collections, the AMI meetings do not leverage the Fisher data nearly as much as the non-AMI meetings. Interestingly, the combined weights of the different meeting-related web corpora are the same (.28) for both LMs. The overall perplexity on the two data sets is quite different (70 vs. 98 for AMI vs. non-AMI subsets of *confDEV*), though both have a WER of 31.1%.

3.4 Recognition Lexicon

For the lecture system, the dictionary contained 58.7k pronunciation variants over a vocabulary of 51.7k. The vocabulary was derived by using the corpora: BN, Switchboard, meetings (ICSI, CMU, NIST, AMI), TED and CHIL. 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 *lectDEV* was 0.65%. The conference meeting system used a dictionary of 56k pronunciation variants over a vocabulary of 48k entries from Switchboard, Fisher, meetings, and CHIL corpora. In this case, we used the

Table 6. Weights learned for the different component LMs for the conference meeting task, with separate tuning for the non-AMI and AMI meetings.

Component	LM weight	
	non-AMI	AMI
non-AMI speech	**0.31**	0.08
AMI speech	0.01	**0.42**
CHIL speech	0.002	0.005
Switchboard CTS	0.03	0.03
Fisher CTS	**0.30**	0.12
UW web-M	0.11	0.03
UW web-F	0.06	0.06
UKA web-MP	0.10	0.09
UKA web-M-A	0.07	0.16

SRI vocabulary selection technique [22] available in the SRILM toolkit, again followed by `ispell` filtering and the inclusion of topic words as well as skipping those vocabulary entries not available in the lecture system dictionary.

Pronunciations for new words for most systems were generated using Festival [23]. For the PRONLEX system, pronunciations were generated automatically using Fisher's grapheme-to-phoneme conversion tool [17].

4 Experiments and Results

4.1 Decoding Strategy

In order to find the best decoding and cross-system adaptation strategy, we performed several different experiments on *lectDEV*. The best setup in terms of word error rate and complexity for all conditions uses only VTLN-trained models (VTLN) or speaker-adapted models (ML-SAT) and no speaker-independent models, even in the first decoding pass:

1. VTLN decoding using incremental, speaker-based VTLN [24] and feature-space constrained MLLR (FSA) [25] adaptation.
2. VTLN, FSA and MLLR [26] adaptation on the confidence-weighted hypothesis of the first pass and VTLN decoding with fixed adaptation parameters.
3. VTLN, FSA and MLLR adaptation on the output of second pass and ML-SAT decoding.
4. Same as in the third pass.

Using an 8 ms instead of a 10 ms frame-shift for passes 2–4, improves the final WER by about 1% absolute [9] on *lectDEV*.

In another set of experiments, we followed results presented in [27, 28] and our own experience obtained during the development of a system for transcribing English European Parliament Plenary Sessions [7]. It was seen that we gain significantly (approx. 1.5% absolute) from cross-adaptation between systems with different front-ends (MVDR, FFT), and that, when cross-adaptation between

MVDR and FFT leads to no further gains, cross-adapting with the PRONLEX system improves the WER after confusion network combination (CNC) [29] by 0.7% absolute [5].

4.2 Channel Combination and Selection for MDM

In RT-04S, channel combination was performed by decoding all channels and doing a confusion network combination on the resulting lattices over all channels. No selection was used, leading to a relatively high computational load for one pass. This year, we were able to reduce the computational load by 70% with no increase in WER by performing both channel combination and selection. We constructed a single channel at the waveform level by selecting only those channels for an utterance with a high signal-to-noise ratio (SNR); this leads to an improvement in SNR of 2 dB on *lectDEV*. In addition to the speed-up on the MDM condition, we gained 4% in WER with this blind channel combination (BCC) approach compared to the SDM condition (see first and second pass overall results in Table 7). Including additional utterances and/or channels based on their SNR ratio to the confusion network combination of the BCC channel yields a further gain of 0.5% absolute. A detailed explanation is given in [30].

4.3 Overall System Performance

Table 7 lists the overall system results with automatic segmentation for RT-06S. The WERs per pass are after CNC of the lattices of the MVDR, FFT, and/or PRONLEX system used in that pass. In each pass of the IHM system, both an MVDR and an FFT system were used and cross-adapted on the previous pass. In the fourth pass, we only used the PRONLEX system and adapted the fifth pass systems (FFT, PRONLEX) on the CNC result of lattices from the third and fourth pass.

Table 7. 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.

Pass	IHM				SDM			MDM		
	lect		conf		lect		conf	lect		conf
	dev	eval	dev	eval	dev	eval	eval	dev	eval	eval
1	30.3	39.6	41.7	37.6	50.9	65.9		46.9	61.2	
2	25.0	34.7	35.2	31.9	45.9	59.0	60.1	42.0	57.0	
3	23.9	33.6	33.7	30.8	43.4	55.5	58.3	38.5	53.9	53.8
4	23.2	32.7	32.6	30.2		54.7			53.4	
5	22.9	32.2	31.9	30.2						
RTx	190				110			120		

As described above (Section 4.2), the first and second pass for the MDM condition used blind channel combination. In the third pass we added additional

utterances and/or channels to the confusion network combination step. As for IHM we used both an MVDR and an FFT front-end in each pass, but in contrast to IHM, the MVDR system was adapted on the CNC result and the FFT system on the MVDR result of the subsequent pass. The first and second passes were decoded with farfield acoustic models, but in the third pass we used the close-talking acoustic models.

On the lecture meeting task, it can be seen that there is a huge gap between the development and the evaluation data results. This comes from the additional data collected by sites other than UKA. While the IHM error rates for UKA (23.9%) and IBM (27.3%) are similar to those on the development data, which were collected by UKA only, the error rates on data from AIT (35.3%), ITC (31.8%) and UPC (54.0%) are much worse. The reason for that is likely the more interactive style of the non-UKA lecture meetings, e.g. coffee breaks (UPC), and the higher proportion of non-native speakers.

5 Acknowledgments

This work was partly funded by the European Union (EU) under the integrated project CHIL [31] (IST-506909).

References

1. C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel, "Open Domain Speech Recognition & Translation: Lectures and Speeches," in *ICASSP*, 2006.
2. M. Wölfel and J. McDonough, "Combining Multi-Source Far Distance Speech Recognition Strategies: Beamforming, Blind Channel and Confusion Network Combination," in *INTERSPEECH*, 2005.
3. F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz, "Issues in Meeting Transcription – The ISL Meeting Transcription System," in *ICSLP*, 2004.
4. M. Wölfel and J. McDonough, "Minimum Variance Distortionless Response Spectral Estimation Review and Refinements," *IEEE Signal Processing Magazine*, September 2005.
5. 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.
6. Q. Jin and T. Schultz, "Speaker Segmentation and Clustering in Meetings," in *ICSLP*, 2004.
7. 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.
8. J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
9. 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.

10. T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," in *Proc. ASRU*, 2001.
11. S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multichannel Audio," *IEEE Trans on Speech and Audio Processing*, vol. 13, pp. 84–91, 2005.
12. K. Laskowski and T. Schultz, "Unsupervised Learning of Overlapped Speech Model Parameters for Multichannel Speech Activity Detection in Meetings," in *Proc. ICASSP*, 2006.
13. Ö. Çetin and E. Shriberg, "Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap," in *Proc. ICASSP*, 2006.
14. H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *ASRU*, 2001.
15. M. J. F. Gales, "Semi-tied covariance matrices," in *ICASSP*, 1998.
16. J. McDonough, T. Schaaf, and A. Waibel, "On Maximum Mutual Information Speaker-Adapted Training," in *ICASSP*, 2002.
17. W. M. Fisher, "A Statistical Text-to-Phone Function Using Ngrams and Rules," in *ICASSP*, 1999.
18. A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *ICSLP*, 2002.
19. 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.
20. I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures," in *Proc. HLT-NAACL*, 2003.
21. Ö. Çetin and A. Stolcke, "Language Modeling in the ICSI-SRI Spring 2005 Meeting Speech Recognition Evaluation System," International Computer Science Institute, Berkeley, CA, USA, Tech. Rep. TR-05-006, 2005.
22. A. Venkataraman and W. Wang, "Techniques for Effective Vocabulary Selection," in *Proc. Eurospeech*, 2003.
23. 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.
24. P. Zhan and M. Westphal, "Speaker Normalization Based on Frequency Warping," in *ICASSP*, 1997.
25. M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," Cambridge University, Cambridge, United Kingdom, Tech. Rep., 1997.
26. C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
27. 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.
28. L. Lamel and J.-L. Gauvain, "Alternate Phone Models for Conversational Speech," in *ICASSP*, 2005.
29. L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus among Words: Lattice-based Word Error Minimization," in *EUROSPEECH*, 1999.
30. 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.
31. "CHIL – Computers in the Human Interaction Loop," http://chil.server.de.