Advances in Lecture Recognition: The ISL RT-06S Evaluation System

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

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

Abstract

This paper describes the 2006 lecture recognition system developed at the Interactive Systems Laboratories (ISL), for individual head-microphone (IHM), single distant microphone (SDM), and multiple distant microphones (MDM) conditions. It was evaluated in 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. Our system achieved word error rates of 38.5% (53.4%) and 22.9% (32.2%), respectively, on the MDM and IHM conditions of the RT-05S (RT-06S) lecture evaluation set.

Index Terms: speech recognition, lectures, distant speech, CHIL, RT-06S.

1. Introduction

In this paper, we present the ISL's most recent speech-to-text system for lectures, which has evolved significantly over previous versions [1, 2, 3] and which was evaluated in the NIST RT-06S Rich Transcription Meeting Evaluation.

In [1] we described our improvements over the CHIL [4] evaluation system of January 2005 [2]. The main improvements came from a better selection of training material, new vocabularies and better language models. Our system development experiments were initially performed on the NIST RT-05S development set, but subsequently repeated on the RT-05S evaluation data, as the latter set was richer in that it contained more unique speakers.

The systems described in [1] and [3] shared many common elements; e.g., front-end, phoneme set, training strategy, etc. The system described in this paper differs in several important ways. Notably, we used only speaker-adapted acoustic models. Even in the first pass, we used models already 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) [5] 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 the hope that it would improve the overall result [6, 7], by using the system for cross-adaptation. Last but not least, we used a different speech segmentation algorithm compared to the one used in the RT-04S evaluation system [8].

CMU	ICSI	NIST	TED	CHIL	Hub4-BN
11h	72h	13h	13h	10h	180h

Table 1: *Duration of used acoustic model training data.*

2. Data

2.1. Acoustic Model Training Data

Table 1 contains an overview of the acoustic model training data. CMU, ICSI, and NIST are audio recordings of meetings [9], TED (Translingual English Database) [9], and CHIL are audio recordings of lectures and Hub4-BN [9] contains recordings of broadcast news. All the acoustic data is in 16 kHz, 16 bit quality and recorded with close talking microphones, except for the training data contributed by CMU, which was collected with lapel microphones. For ICSI and NIST, far distant channels were also available.

2.2. Development and Test Data

The decoding experiments described in this paper were conducted on the following data sets:

DEV The lectmtg portion of the official RT-06S development set, which is identical to the RT-05S evaluation set (150min).

EVAL The lectmtg portion of the RT-06S evaluation set (190min).

This year's RT-06S primary condition was MDM and was scored with overlap, i.e. overlap regions were labeled in the reference and were scored. The results on the DEV data presented in this paper were scored without taking regions with overlapping speech into account.

3. Automatic Segmentation

As already mentioned, the current system uses a different automatic speech segmentation approach from previous years [8]. For the IHM condition, automatic segmentation is especially difficult in that significant cross-talk from other speakers is present in the recordings, and this cross-talk should be ignored during automatic transcription. The following speech activity features are extracted on a per frame basis, with a frame size of 32 msec and a frame shift of 10 msec: frame energy in decibels (dB), mean and variance normalized energy passed through a sigmoid function, energy-normalized linear prediction error [10], slope along the frequency axis of a mel-warped filter-bank spectrum , and speech class posterior computed from a multi-layer perceptron (MLP) trained with standard MFCC features to classify speech and non-speech. Using these features, the segmentation is performed in three steps:

1. Background speech activity rejection: For each frame, out of all microphones available for a particular meeting, the microphone with the highest energy is chosen as the current active microphone. From that, unreliable estimates of microphone switches are pruned out, (a) by checking for the presence of minimal percentage of voiced speech using the normalized energy, energy-normalized linear prediction error, and speech class posterior, and (b) by checking for a constraint of minimum duration of 1 sec. Finally, regions where the current microphone is active are tagged as regions of foreground speech.
2. Foreground speech activity detection: Within the regions of foreground speech tagged by the first step, frames with negative spectral slope, high normalized energy, and low energy-normalized linear prediction error are further tagged as foreground speech. These estimates are further smoothed out with a median filter of 0.5 sec duration.
3. Sentence breaking: Regions of foreground speech at the output of the second step are further cut into shorter segments as follows: The point of high confidence non-speech is searched for in the region between times 0.5 sec and 15.0 sec, and a break is made at that point. Then, with that break as the new starting point, the above procedure is repeated to find more break points until the end is reached. Confidences for non-speech regions are measured based on their duration and average energy level.

For IHM, it is assumed that after the segmentation all speech from a single microphone correspond to a single speaker. For SDM, only the sentence breaking step is performed, as all speech in the SDM channel should be recognized and the recognizer is assumed to be the best available system for discarding non-speech. The resulting segments are further tagged with speaker labels using a hierarchical agglomerative speaker clustering technique [8]. For the MDM condition, a single best channel is first determined on the basis of average SNR. Thereafter, the same processing used in the SDM condition is applied to this single best channel.

4. System Training and Development

All experiments described in this paper were run using the Janus Recognition Toolkit (JRTk) and the Ibis single pass decoder [11].

4.1. 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 STC transform [12] 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.

Expt.	System	WER
A	ICSI+NIST+TED	34.8%
	+ CMU	35.1%
	+ BN97	36.0%
B	standard	32.3%
	second incr. growing	32.0%
C	w/o CHIL	32.0%
	with CHIL	31.5%
Overall	ICSI+NIST+CMU+TED+BN, 6000	32.6%
	ICSI+NIST+TED, 4000	31.5%
second pass	ICSI+NIST+CMU+TED+BN, 6000	28.4%
	ICSI+NIST+TED, 4000	27.0%

Table 2: *Training experiments: WERs computed with a first pass FFT systems with incremental VTLN and FSA estimation and a frame shift of 10 msec on the IHM condition of DEV.*

4.2. 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 to perform best on close talking audio, thereby skipping the CMU meeting corpus and the Hub4-BN training material, and yielding a gain of approx. 1% absolute (Table 2 A). We also changed the model set used in RT-04S slightly by adding noise models for laugh and other human noises to the existing breath and general noise models, and splitting the filler model into one for monosyllabic and another one for disyllabic fillers.

Acoustic model training was performed with fixed state alignments, which were written by a small system (2,000 codebooks) trained on the above mentioned corpora. Both the MVDR and FFT systems were trained in the same way, resulting in a size of 16,000 distributions over 4,000 models, with a maximum of 64 Gaussians per model. The training was similar to the one used in [1], with one modification. A second pass for incremental growing of Gaussians was performed after the STC training, so that the complete training procedure is now as follows: (1) a first incremental growing of Gaussians, (2) estimation of the global STC matrix, (3) a second incremental growing of Gaussians; this leads to an additional gain of 0.3% (see Table 2 B). To train the distributions for the semi-continuous system and to compensate for the occasionally worse fixed-state alignments, 2 iterations of Viterbi training were performed.

Since the 10hrs of CHIL training data were released relatively late, we used MAP with a weight of 0.8 for the CHIL data to adapt our current models and gained another 0.6% (Table 2 C). For the ML-SAT models, three additional iterations of ML-SAT [13] were run, wherein feature space adaptation and MLLR parameters were estimated for all speakers in the training set; for these iterations, a weight of 4.0 was applied to the CHIL training data. Comparing the resulting system to the system used in [1], we improved our second pass result by 1.4% absolute (see overall results in Table 2, the second row corresponds to the new system).

In addition to the FTT 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 [14]. For the systems based on this phoneme set, context independent acoustic models were initialized by taking the global mean over all train-

ing data. Several iterations of Viterbi training were then applied in order to train the models. From these context-independent models, forced alignments were obtained and 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 24,000 distributions over 3,000 models with a maximum of 64 Gaussians per model were used and only feature space adaptation parameters were estimated during ML-SAT.

For the far distant channels, we adapted the models by appending two Viterbi training iterations using the far distant meeting (ICSI, NIST) data to the close talking models.

4.3. Language Model Training

A 4-gram mixture language model (LM) was used, with components trained on transcripts from the following corpora: CHIL lectures (45k words), AMI meetings (203k), non-AMI (ICSI, CMU, NIST, LDC) meetings (1.1M), TED (98k), Hub4 Broadcast News (131M), recent speech/language proceedings text (2002-2005) (23M), web data from UW (150M words related to ICSI meetings), and UKA web data collections described further below.

We used the strategy in [15] for web text collection, both with and without changes to the query generation process. In the first collection (web-L), the most frequent 3-grams and 4-grams (1k) from the CHIL lecture data were combined to form queries, as in [15]. In the next collection (web-LP), queries were formed by combining the frequent CHIL n-grams with topic phrases selected from the proceedings data. Finally, a last collection (web-MP) used queries that combined frequent n-grams from the non-AMI meeting data and the topic phrases. The topic phrases were generated by: computing bigram tf-idf (term frequency – inverse document frequency) weights for each paper in the proceedings data, zeroing all but the top 10%, averaging these vectors over the collection, and taking the top 1,400 bigrams excluding any with stopwords. The topic bigrams were mixed randomly with the general phrases until the desired number of queries (14k) were generated. Each collection was perplexity filtered to roughly 150M words so that size would be comparable to the UW meeting-based web data, but the raw collections ranged in size from 559M-1.1B words.

LMs were built using the SRILM-toolkit [16] with modified K-N discounting [17]. The mixture weights were tuned on a held-out set of CHIL lectures, with pruning after interpolation. Results that follow are for our development set (RT05eval), on which the LM with no web data gave a perplexity (PPL) of 142 and 31.1% WER. Adding any web data gives a significant improvement, and both web-L and the UW web data alone yielded similar performance (PPL=132, WER=30.2), though the web-L queries were better matched to the lecture task. A small gain was obtained by using both components (LM-I: PPL=130, WER=30.0), and replacing web-L with two components based on web-LP and web-LM again led to a small improvement (LM-II: PPL=128, WER=29.9). LM-I was used in the MDM and SDM conditions, and LM-II was used in the IHM condition. Compared to the old 4g LM used in [1], we gained 1.6% absolute.

With use of the UKA web data, the mixture weights for both TED and BN were very small, and subsequent experiments showed a slight gain in performance when these were removed. In addition, when UKA larger collections are used (less perplexity filtering), the weight for the UW meeting-based web collection becomes small.

	1st (FFT)	2nd (MVDR)	3rd (FFT)	4th (MVDR)
A	34.2%	30.0%	27.9%	25.5%
B	34.2%	27.0%	25.4%	
C	34.2%	26.8%	25.3%	
D	31.5%	26.5%	25.4%	25.0%

Table 3: *Adaptation experiments, with different acoustic models on the IHM condition of DEV.*

4.4. Recognition Lexicon

The dictionary contains 58,695 pronunciation variants over a vocabulary of 51,731. For the MVDR and FFT system, pronunciations which were unknown in our base dictionary were generated using Festival [18]. For the PRONLEX system, missing words were generated automatically with the help of a grapheme-to-phoneme conversion tool [14]. The vocabulary (same for all systems) was derived by using the corpora: BN, Switchboard PhaseI+II, meetings (ICSI, CMU, NIST, AMI), TED and CHILa. After applying individual word-frequency thresholds to the corpora, we checked the resulting vocabularies with ispell to filter out spelling errors. The OOV-rate on DEV was 0.65%.

5. Experiments and Results

5.1. Decoding Strategy

We performed several experiments to find the best decoding and cross-system adaptation strategy. First we compared different adaptation schemes using different acoustic models from different training stages (Table 3). The following acoustic models were available: speaker-independent models (SI), VTLN-trained models (VTLN), speaker-adapted models (ML-SAT) and were used on a single chain of adaptation passes with alternated front-end (MVDR, FFT). VTLN (V) [19], constrained MLLR (FSA, F) [20] and MLLR (M) [21] adaptation was always done on the confidence weighted hypotheses of the previous pass, whereas the parameters were kept fixed during the subsequent decoding pass. All passes after the first pass were decoded with a frame shift of 8msec instead of 10msec, which gives a gain of about 1% absolute.

In A, we performed the adaptation strictly step-by-step and used only matching models: SI decoding, V estimation and VTLN decoding, V+F estimation and ML-SAT decoding, V+F+M adaptation and ML-SAT decoding. In B, we always applied V+F+M adaptation: SI decoding, V+F+M adaptation and VTLN decoding, V+F+M adaptation and ML-SAT decoding. In C, we modified the second step of B to use V+M adaptation and VTLN decoding. Finally, in D, we used only the speaker-adapted VTLN and ML-SAT models: VTLN decoding with incremental speaker-based VTLN and FSA estimation, V+F+M adaptation and VTLN decoding, V+F+M adaptation and ML-SAT decoding and in the fourth pass the same again. As can be seen, there is no significant difference between B, C, and D; we chose to use D for our RT-06S evaluation systems because it is then not necessary to train another speaker-independent model. Instead, we can easily use models from an earlier training step of the ML-SAT models as a first pass.

In another set of experiments, we followed results presented in [6, 7] and experiences collected during the development of a system for Transcribing English European Parliament speeches. 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 doing confusion network

Pass	IHM	SDM	MDM
DEV first pass	30.3	50.9	46.9
second pass	25.0	45.9	42.0
third pass	23.9	43.4	38.5
fourth pass	23.2		
fifth pass	22.9		
EVAL final pass	32.2	54.7	53.4
RTx	190	110	120

Table 4: *Overall results and real-time factors on DEV and EVAL.*

combination (CNC) [22] with the PRONLEX system in addition by 0.7% absolute.

5.2. Channel Combination and Selection for MDM

In RT-04S, the channel combination was done, by simply decoding all channels and doing a confusion network combination on the resulting lattices over all channels. No selection was done, which means that the computational load for one pass was relatively high. This year, we were able to reduce the computational load by 70% with no increase in WER by doing both channel combination and selection. Therefore we built a single channel at the waveform level, by selecting only those channels for an utterance with a high signal to noise ratio (SNR), which also leads to an improvement in SNR of 2 dB on the DEV set. In addition to the speed-up on MDM 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 4). By further adding selected utterances/ channels by their SNR ratio to the confusion network combination of the BCC channel we saw an additional gain of 0.5%. A detailed explanation is given in [23].

5.3. Overall System Performance

Table 4 lists the overall system results on DEV and EVAL. 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. As described above (Section 5.2), for the first and second pass on MDM blind channel combination was used. For the third pass we added also additional selected utterances/ channels to the confusion network combination step. As for IHM we used in each pass an MVDR and FFT system, but in difference to IHM, the MVDR system was adapted on the CNC result and the FFT system of the MVDR result of the subsequent pass. first and second pass were decoded with far distance acoustic models, but in the third pass, we used the close talking acoustic models.

6. Acknowledgments

This work was partly funded by the *European Union* (EU) under the integrated project CHIL [4] (IST-506909).

7. 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] "CHIL – Computers in the Human Interaction Loop," http://chil.server.de.

[5] M. Wölfel and J. McDonough, "Minimum Variance Distortionless Response Spectral Estimation Review and Refinements," *IEEE Signal Processing Magazine*, September 2005.

[6] 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*, November 2004.

[7] L. Lamel and J.-L. Gauvain, "Alternate Phone Models for Conversational Speech," in *ICASSP*, 2005.

[8] Q. Jin and T. Schultz, "Speaker Segmentation and Clustering in Meetings," in *ICSLP*, 2004.

[9] "Linguistic data consortium," http://www.ldc.upenn.edu.

[10] J. Makhoul, "Linear prediction: A tutorial review," in *Proc. of the IEEE*, 1975, vol. 63(4), pp. 561–580.

[11] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *ASRU*, 2001.

[12] M. J. F. Gales, "Semi-tied covariance matrices," in *ICASSP*, 1998.

[13] J. McDonough, T. Schaaf, and A. Waibel, "On Maximum Mutual Information Speaker-Adapted Training," in *ICASSP*, 2002.

[14] W. M. Fisher, "A Statistical Text-to-Phone Function Using Ngrams and Rules," in *ICASSP*, 1999.

[15] 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, vol. Comp., pp. 7–9.

[16] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *ICSLP*, 2002.

[17] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Tech. Rep. TR-10-98, Computer Science Group, Harvard University, 1998.

[18] A. W. Black and P. A. Taylor, "The Festival Speech Synthesis System: System documentation," Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Edinburgh, Scotland, United Kongdom, 1997.

[19] P. Zhan and M. Westphal, "Speaker Normalization Based on Frequency Warping," in *ICASSP*, 1997.

[20] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," Tech. Rep., Cambridge University, Cambridge, United Kingdom, 1997.

[21] 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.

[22] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus among Words: Lattice-based Word Error Minimization," in *EUROSPEECH*, 1999.

[23] 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.