The LIMSI RT07 Lecture Transcription System

L. Lamel, E. Bilinski, J.L. Gauvain, G. Adda, C. Barras[1], and X. Zhu[1] ⋆

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
[1]also with Univ Paris-Sud, F-91405, Orsay, France
{lamel,bilinski,gauvain,gadda,barras,xuan}@limsi.fr

Abstract. A system to automatically transcribe lectures and presentations has been developed in the context of the FP6 Integrated Project CHIL. In addition to the seminar data recorded by the CHIL partners, widely available corpora were used to train both the acoustic and language models. Acoustic model training made use of the transcribed portion of the TED corpus of Eurospeech recordings, as well as the ICSI, ISL, and NIST meeting corpora. For language model training, text materials were extracted from a variety of on-line conference proceedings. Experimental results are reported for close-talking and far-field microphones on development and evaluation data.

1 Introduction

In the Computers in the Human Interaction Loop (CHIL) project (http://chil.server.de) services are being developed which use computers to improve human-human communication. One of the CHIL services is to provide support for lecture situations, such as providing transcriptions and summaries in close-to-real time for interactive applications or providing off-line support for archiving, search and retrieval, all of which can benefit from automatic processing. One can imagine a future where all public presentations (classes, lectures, seminars, workshops and conferences) are archived for future viewing and selected access. Automatic techniques can provide a wealth of annotations, enabling users to search the audio data to find talks on specific topics or by certain speakers. At LIMSI a transcription system for off-line processing of lecture and seminar has been developed within the context of the CHIL project..

The speech recognizer for CHIL has been derived from the LIMSI Broadcast News transcription system for American English [7]. In addition to the CHIL data available, acoustic and language model training made use of widely available corpora including the TED corpus of Eurospeech recordings, the ICSI, ISL, and NIST meeting corpora. For language model training, in addition to the transcriptions of the audio data, text materials were extracted from a variety of on-line conference proceedings. The LIMSI CHIL speech recognizers used in previous evaluations are described in [9, 11, 12]. In the remainder of this paper the 2007 speech recognizer is described, and development results are provided.

⋆ This work was partially financed by the European Commission under the FP6 Integrated Project IP 506909 CHIL

Table 1. Summary of audio data sources. The top part of the table lists the audio data available in 2005 (97h of IHM data from 4 sources). The middle lists the additional 76h of data used in training the 2006 system and the bottom lists the additional data used in the 2007 system.

Source	*Microphone*	*Type*	*Amount*
TED	lapel	39 speeches	9.3h
ISL	lapel	18 meetings	10.3h
ICSI	head mounted	75 meetings	60h
NIST	head mounted	19 meetings	17.2h
ICSI	tabletop	75 meetings	70h
CHIL	head mounted	17 seminars	6.2h
TED	lapel mics, lightly supervised	190 speeches	46h
Beamformed lecture data	tabletop	rt05s, rt06s, dev07	$<$7h

2 Recognizer Overview

The speech recognizer uses the same core technology and is built using the same training utilities as the LIMSI Broadcast News transcription system described in [7]. The transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning is based on an audio stream mixture model [7], and serves to divide the continuous stream of acoustic data into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment. This year the data partitioner was adapted to the MDM beamformed data [18]. For each speech segment, the word recognizer determines the sequence of words, associating start and end times and an optional confidence measure with each word. The word recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and n-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained via a decision tree.

The language models (LMs) are interpolated backoff n-gram models estimated on subsets of the available training texts. The recognition word list was selected from the audio transcripts and the proceedings texts so as to minimize the out-of-vocabulary (OOV) rate on a set of development data. The vocabulary contains 58k (case-sensitive) words, including several thousand compound words and acronyms.

Word recognition is performed in multiple decoding passes, where each decoding pass generates a word lattice with cross-word, position-dependent, gender-dependent acoustic models, followed by consensus decoding [14] with 4-gram and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [13] techniques prior to each decoding pass.

3 Training Corpora

One of the challenges of the lecture transcription task is locating appropriate audio and textual resources with which to develop the recognizer models. Although multi-site data collection was carried out in the CHIL project, most of this data was reserved for

Table 2. Summary of audio transcripts from various sources (top) and proceedings texts (bottom) along with the number of words by source.

TED oral presentations	71k words
NIST meetings	156k words
ISL meetings	116k words
ICSI	785k words
CTS	3M words
AMI/IDIAP meeting	143k words
NIST RT04, RT05 data	57k words
CHIL Jun04/Jan05 seminars	55k words
CHIL summer04 seminars	38k words

TED texts:	426 papers	929k words
ASRU'99-05:	427 papers	1140k words
DARPA'97-99,04:	119 papers	317k words
Eurospeech'97-05:	3485 papers	7650k words
ICASSP'95-05:	7831 papers	14318k words
ICME'00,03:	996 papers	2101k words
ICSLP'96-04:	3202 papers	7198k words
LREC'02,04:	891 papers	2553k words
ISCA+other workshops:	2333 papers	6077k words

development and testing purposes with only a limited amount of transcribed data available for speech recognizer training. Therefore a variety of publicly available corpora were used for training. The most closely related audio data are the TED recordings of presentations at the *Eurospeech* conference in Berlin 1993 [10]. The majority of presentations are made by non-native speakers of English. Although there are 188 oral presentations (about 50 hours of audio recordings), transcriptions are only available for 39 lectures [1]. This year a biased-LM version of the LIMSI RT06 close-talking microphone speech system was used to transcribe the remaining 190 speeches (46h) so these could be also used for acoustic model training. Other related data sources are the ISL, ICSI and NIST meeting corpora which contain audio recordings made with multiple microphones of a variety of meetings (3-10 participants) on different topics [5, 6, 8]. The amount of data per corpus is summarized in Table 1. The first four corpora were used in training the 2005 system and contain data recorded with individual head-mounted microphones (IHM); the middle two entries were added in the 2006 system; and the last two were added in the 2007 system. From the available farfield data in the ICSI corpus for which there are a varying number of channels, the farfield microphone channel with highest likelihood during forced alignment was selected as being the most appropriate for each speaker. The 2007 acoustic models were trained on pooled data from all sources, including close-talking microphone data, tabletop distant microphone data and a small amount of beamformed data. The ICSI delay&sum signal enhancement software [2] was used to process all the available lecture training and test data (rt05s, rt06s, and dev07).

The language model training data are the same as were used in the 2006 system and consist of manual transcriptions of related audio data as well as the proceedings texts from a variety of speech and language related conferences and workshops. The audio transcripts come from the same sources as are used for acoustic training. In addition transcriptions of conversational telephone speech (CTS) from the CallHome, Switch-Board and Fisher collections (distributed by the LDC) were used. The amount of words in the each audio transcript source are given in Table 2. In addition to the audio transcripts, almost 20k papers in the proceedings of workshops and conferences in the audio, speech and language processing domain were used for language modeling. These texts shown in the lower part for Table 2 were processed by tools derived from ones shared by ITC-IRST.

4 Audio Partitioner

The LIMSI RT-07S speaker diarization system for the conference and lecture meetings is fully described in [18]. This system builds upon the RT-06S diarization system designed for lecture data. The diarization system combines agglomerative clustering based on Bayesian information criterion (BIC) with a second clustering using state-of-the-art speaker identification (SID) techniques [4, 17]. The system has 5 steps which use a 38-dimensional feature vector consisting of 12 cepstral coefficients, Δ and Δ-Δ coefficients plus the Δ and Δ-Δ log-energy. 1) Speech activity detection (SAD), which locates speech portions in the signal using a Log-Likelihood Ratio (LLR) based speech activity detector [17]. The SAD acoustic models, each with 256 Gaussians, were trained on about 5 hours of conference meeting data from the NIST RT'04 and RT'05 evaluations. 2) Initial segmentation, which is performed by taking the maxima of a local Gaussian divergence measure between two adjacent sliding windows of 5 seconds. 3) Viterbi resegmentation is used to refine the segment boundaries using 8-component GMMs trained from the initial speech segments. 4) BIC clustering which is used to successively merge speech segments, and 5) Speaker clustering is carried out using speaker recognition methods [3, 15].

Since the speech activity detection error of the baseline system was relatively high (about 10%) on lecture data, some of the normalization techniques and acoustic representations that were explored to improve performance are described in [18]. The RT07 diarization system integrating these improvements obtains comparable results on both the RT-07S conference and lecture evaluation data for the multiple distant microphone (MDM) condition.

5 Acoustic modeling

The acoustic feature vector has 39-components comprised of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives. The cepstral parameters are derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis

using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance.

The acoustic models are context-dependent, 3-state left-to-right hidden Markov models with Gaussian mixture. The triphone-based phone models are word-independent and gender-independent, but word position-dependent. The acoustic models are MLLT-SAT trained, with different sets of tied-state models used in successive decoding passes. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state position of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states [7]. A set of 152 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

Two sets of models were estimated on all the available training data, and MAP adapted with the beamformed data. Since only a very small amount of beamformed data was available, for the final system, the RT06s data used for development was also included in the adaptation data. The small set covers 5k phone contexts and has 5.2k tied states with 32 Gaussians per state. The large set covers 25k phone contexts, with 11.5k tied states and 32 Gaussians per state.

6 Language modeling

The LIMSI RT07 system used two language models, a case-insensitive 35k LM from the 2005 system and the 58k case-sensitive LM from the RT06 system. The recognizer word lists were determined by interpolating unigram language models trained on different subsets of the available training texts listed in Table 2. The proceeding texts are comprised of the proceedings from 54 conferences and workshops in speech and language, which represent about 20,000 PDF documents. While not used for vocabulary selection, the CTS data were used for language model training.

For language model estimation the available corpora were grouped into 3 sources: 1) Seminar and meeting transcriptions (1.42M words); 2) Proceedings texts (46M words); 3) Transcriptions of Conversational Telephone Speech databases available from LDC (29M words). Three backoff n-gram language models were estimated, one on each of the data subsets. The component language models were interpolated [16], and the weights were chosen to minimize the perplexity of the development data. The largest weight is for the transcriptions (0.6), with weights of 0.3 and 0.1 for the proceedings texts and CTS transcripts respectively. The perplexities and OOV rates of the 4-gram LMs are shown in Table 3. The 58k LM contains 8.8M fourgrams, 19M trigrams, 5M bigrams, and the 35k LM contains 6.6M fourgrams, 15M trigrams, 4M bigrams. More information concerning the language models can be found in [12].

7 Decoding

Word recognition is performed in two passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. The posterior probabilities of the lattice edges are estimated using the forward-backward algorithm. The 4-gram lattices are converted to a confusion network with posterior probabilities by iteratively merging

Table 3. Perplexities and OOV rates of the 35k and 58k language models on the development and test data.

Data set	*rt06*		*dev07*		*rt07*	
Language Model	OOV	Px	OOV	Px	OOV	Px
35k 4-gram	0.4	157	0.9	165	0.7	136
58k 4-gram	0.4	162	0.8	163	0.7	138

lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [14]. The words with the highest posterior in each confusion set are hypothesized.

Pass 1: Initial Hypothesis Generation - This step generates initial hypotheses which are then used for speaker-based acoustic model adaptation. This is done via one pass (about 1xRT) cross-word trigram decoding with gender-independent sets of position-dependent triphones (5k contexts, 5k tied states) and a 35k word trigram language model (15M trigrams and 4M bigrams). The trigram lattices are rescored with a 4-gram language model (6.6M fourgrams, 15M trigrams and 4M bigrams).

Pass 2: Adapted decode - Unsupervised acoustic model adaptation of speaker-independent models is performed for each speaker using the CMLLR and MLLR techniques [13] with only two regression class. The lattice is generated for each segment using a 58k word bigram LM and position-dependent triphones with 25k contexts and 11.5k tied states (32 Gaussians per state). As in the first pass, the lattices are rescored with a 58k word 4-gram language model (8.8M fourgrams, 19M trigrams and 5M bigrams) and pronunciation probabilities.

8 Experiments and results

Some initial experiments were carried out using the designated RT07 development set comprised of 5 seminars, one from each CHIL data collection site. The seminars have different durations, ranging from 23 to 44 minutes. The baseline results with the LIMSI RT06 farfield system had a word error rate of 64.4% on the beamformed signal. Although results are reported here only for the second decoding pass, the improvement relative to the first pass is in the range of 4-8% depending upon the test set and system configuration. Updating the segmentation gave a slight error reduction (64.0%). Since the development seminars are significantly longer than the test data which consists of 5-min excerpts, these were divided into 5-min chunks. The WER on the chunked data is 63.0% with the RT06 system.

Since this development data is not representative of the test, and in light of the very limited amount of beamformed data that could be used for model adaptation, the RT06s evaluation data was used for all further system development. These data are comprised of 38 5-minute lecture excerpts contributed by 5 of the CHIL partners: AIT, IBM, ITC, UKA and UPC. Table 4 provides some of the development results. The baseline WER with the RT06s MDM acoustic models on the beamformed data was 65.2%. By adding the beamformed data to acoustic training data, the WER is reduced by 0.8%. MAP

Table 4. Recognition error rates on the RT06s evaluation data for the baseline acoustic models (MDM AM); pooling the beamformed training data (+ pool bmf); MAP adaptation with the beamformed training data (+ MAP with bmf); and decoding tuning (+ tuning).

RT06s bmf	*Corr (%)*	*Subs (%)*	*Del (%)*	*Ins (%)*	*WER (%)*
RT06s MDM AM	41.3	37.8	20.9	6.6	65.2
+ pool bmf data	41.2	35.4	23.3	5.6	64.4
+ MAP with bmf data	43.4	34.6	22.0	5.6	62.2
+ tuning	44.4	33.2	22.4	5.4	61.0

Table 5. Official NIST SASTT and STT results on the RT07s evaluation data.

scoring	*Cor (%)*	*Sub (%)*	*SpSub (%)*	*Del (%)*	*Ins (%)*	*WER (%)*	*SER (%)*
SASTT	47.6	29.8	4.2	18.4	5.5	57.9	40.0
STT	51.8	29.7		18.4	5.6	53.7	38.3

adapting this models with the beamformed data, gives a further error reduction of over 2% and after tuning a word error of 61% is obtained. This represents an error reduction of about 6% relative to the baseline models.

Table 5 reports the official NIST SASTT results, along with the STT scoring. For the evaluation system, MAP adaptation was performed with all the available beamformed data, including RT06s. The SASTT word error rate is 57.9%, including the 4.2% of erroneous speaker associations. The equivalent STT WER is 53.7%. No system development was done this year for the sdm or ihm conditions, but a few contrastive post-evaluation runs were done. Using the segmentations provided by SRI resulted in a 42.2% WER with the LIMSI RT06s ihm acoustic models and a 40.6% WER with the RT07s multistyle acoustic models. Using a single distant microphone with the an STT WER of 60.7% and an SASTT WER of 63.9% were obtained with the RT07s multistyle acoustic models.

9 Conclusions

This paper has described the LIMSI RT07 system aiming to automatically transcribe lectures and seminars for off-line applications. Publicly available corpora were used to train both the acoustic and language models, since only a small amount of CHIL data were available for system development. This was LIMSI's second participation to the multiple farfield microphone task. This year the ICSI beamforming software was used to process the lecture training and test data. In addition to including the available beamformed data during acoustic model training, the remaining TED speeches were transcribed and used in a lightly supervised manner. Compared to the LIMSI 2006 system, this year's system also used a revised audio partitioner which significantly reduced the speaker diarization error on the primary MDM test condition.

References

1. The Translanguage English Database (TED) Transcripts, LDC catalog number LDC2002T03, isbn 1-58563-202-3.
2. X. Anguera, and C. Wooters, and J. Hernando, "Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion", in *Automatic Speech Recognition and Understanding (IEEE, ASRU'05)*, San Juan, Puerto Rico, 2005.
3. C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *IEEE ICASSP 2003*, Hong Kong, 2003.
4. C. Barras, X. Zhu, S. Meignier and J-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *The IEEE Transactions on Audio, Speech and Language Processing*, September, 2006.
5. S. Burger, V. MacLaran and H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style, *ICSLP'02*, Denver, Sep 2002. (LDC2004S05, LDC2004E04, LDC2004E05)
6. J.S. Garofolo, C.D. Laprun, M. Michel, V.M. Stanford and E. Tabassi, "The NIST Meeting Room Pilot Corpus," *LREC'04*, Lisbon, May 2004. (LDC2004S09, LDC2004T13)
7. J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.
8. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, "The ICSI Meeting Corpus," *ICASSP'03*, Hong Kong, Apr 2003. (LDC2004S02, LDC2004T04)
9. L. Lamel, G. Adda, E. Bilinski and J.L. Gauvain, "Transcribing Lectures and Seminars," *Proc. ISCA Eurospeech'05*, Lisbon, Sep 2005.
10. L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani and H. Tillmann, "The Translanguage English Database TED," *ICSLP'94*, Yokohama, Sep 1994. (LDC2002S04)
11. L. Lamel, H. Schwenk, J.L. Gauvain, G. Adda and E. Bilinski, "Improvements in Transcribing Lectures and Seminars," *Proc. MLMI'05*, Edinburgh, July 2005.
12. L. Lamel, E. Bilinski, G. Adda, J.L. Gauvain, H. Schwenk. "The LIMSI RT06s Lecture Transcription System" *Proc. RT06s Workshop*, Washington DC, USA, May 2006.
13. C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, **9**(2):171-185, 1995.
14. L. Mangu, E. Brill and A. Stolcke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeech'99*, 495-498, Budapest, Sep 1999.
15. J. Schroeder and J. Campbell, Eds., *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Academic Press, 2000.
16. P.C. Woodland, T. Niesler and E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR," presented at the 1998 Hub5E Workshop, Sep 1998.
17. X. Zhu, C. Barras, L. Lamel and J.L. Gyauvain, "Speaker Diarization: from Broadcast News to Lectures", In *MLMI 2006 Meeting Recognition Workshop*, Washington DC, USA, May 2006.
18. X. Zhu, C. Barras, L. Lamel and J-L. Gauvain, "Multi-Stage Speaker Diarization for Conference and Lecture Meetings," *Proc. NIST RT'07*, Baltimore, May 2007.
19. X. Zhu, C. Barras, S. Meignier and J.L. Gauvain, "Combining speaker identification and BIC for speaker diarization" *Proc. Interspeech'05, pp.2441-2444*, Lisbon, September, 2005