Transcribing Lectures and Seminars*

L. Lamel, G. Adda, E. Bilinski, J.L. Gauvain
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,gadda,bilinski,gauvain}@limsi.fr

ABSTRACT

This paper describes recent research carried out in the context of the FP6 Integrated Project CHIL in developing a system to automatically transcribe lectures and seminars. We made use of widely available corpora to train both the acoustic and language models, since only a small amount of CHIL data were available for system development. For 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. Word error rates of about 25% are obtained on test data extracted 12 seminars.

1. INTRODUCTION

The EC FP6 Integrated Project *Computers in the Human Communication Loop* (CHIL) is exploring new paradigms for human-computer interaction. The idea is to develop services that are provided in an unobtrusive manner so as to suit human needs. To do so the partners are developing robust, multi-modal perceptual user interfaces which can track and identify people, recognize what they are doing and take appropriate actions based on the context. To use the project terms, the goal is to models the "Who, Where, What, Why and How of Human activities and communication" (http://isl.ira.uka.de/chil). At LIMSI we are developing technologies for audio based speech activity detection, speaker recognition and tracking (Who and Where), automatic speech recognition and the extraction of linguistic meta-data (What), topic detection and emotion recognition (How and Why).

One of the CHIL services aims to provide support for lecture situations. The services can be either on-line and off-line, each with different technological constraints. For on-line services, the lecture must be transcribed and annotated in close to real time, while the lecture is happening. Such an interactive application would allow latecomers to catch up on what was already presented earlier in the talk, by either reading the transcript or an automatically created summary. If someone needs to step out of the lecture for a few minutes, the service would allow the person to scan the missing portion. Many possible off-line applications can also be envisioned that would benefit from automatic transcription, annotation, indexing and retrieval. These technologies could be used to archive all public presentations (conferences, workshops, lectures) 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. Given the large number of parallel oral sessions at most major conferences, such services could allow attendees to interactively access talks they were unable to attend. In this paper we report on experiments in developing a first transcription system for lectures and seminars for off-line applications.

2. CORPORA

Although within the CHIL project there is a multi-site effort to collect seminar data, at the time of the dry-run and first year evaluation no CHIL data were available for training purposes. Therefore one of the problems addressed was to locate appropriate audio and textual resources with which to develop the recognizer models. Of the publicly available corpora, the most closely related audio data are the TED recordings of presentations at the *Eurospeech* conference in Berlin 1993 [2]. The majority of presentations are made by non-native speakers of English. Although there are 188 speeches (about 50 hours) of audio recordings, transcriptions are only available for 39 lectures [3]. 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 [4, 5, 6]. Using a single microphone channel per speaker for the data from all four sources (distributed by LDC), a total about 97 hours of audio training data is available. The amount of data per corpus is summarized in Table 1.

According to the CHIL evaluation protocol, the development data segments from the Jun'04 and Jan'05 seminars were allowed to be used for training or supervised adaptation purposes. Therefore the development segments from the seminars were included in the acoustic

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

Source	*Microphone*	*Type*	*Amount*
TED	lapel	39 lectures	9.3h
ISL	lapel	18 meetings	10.3h
ICSI	head mounted	75 meetings	59.9h
NIST	head mounted	19 meetings	17.2h

Table 1: Summary of audio data sources.

model training data for the primary system. The primary models are therefore adapted to the most of the evaluation speakers. A contrast system was also run where the 1h of development segments from the seminars was not used.

The language model training data 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, but in addition we used transcriptions of conversational telephone speech from the CallHome, SwitchBoard and Fisher collections. We also tried using assorted transcriptions from Broadcast News (BN) data, but since these did not reduce the perplexity they were not used to estimate the language models. The amount of words in the each audio transcript source are given in Table 2.

TED: 75k words
NIST: 150k words
ISL: 115k words
ICSI: 756k words
CTS: 3M words
CHIL jun04 dev: 7604 words
jan05 dev: 5894 words

Table 2: Summary of audio transcripts.

In addition to the audio transcripts, a large number of texts on audio, speech and language processing can be obtained from conference and workshop proceedings (see Table 3). The 18972 proceedings texts were processed using scripts derived from ones shared by ITC-IRST to convert postscript and pdf files to ascii texts. Further processing removed unwanted materials (email, websites, telephone numbers, addresses, mathematical formulas and symbols, figures, tables, references) as well as special formatting characters and ill-formed lines.

3. 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 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 by means of a decision tree.

TED texts:	426 papers	935k words
ASRU'99-03:	359 papers	930k words
DARPA'97-99:	129 papers	310k words
Eurospeech'97-03:	2941 papers	5749k words
ICASSP'95-04:	8258 papers	12448k words
ICME'00,03:	1004 papers	2060k words
ICSLP'96-02:	2910 papers	5407k words
LREC'02,04:	898 papers	2570k words
ISCA+other workshops:	2047 papers	5050k words

Table 3: Summary of proceedings texts (19k articles, 35.4M words).

Acoustic models

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. The cepstral coefficients are normalized on a segment-cluster basis so that 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 (32 Gaussians per state). The triphone-based context-dependent phone models are word-independent but word position-dependent and gender-independent. The first decoding pass uses a small set of acoustic models with about 5000 contexts and tied states, and a total of 165k Gaussians. A larger set of acoustic models, used in the second and third passes, cover about 19000 phone contexts represented with a total of 11k states, for a total of about 360k Gaussians.

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.

Language models

The n-gram language models were obtained by interpolation [11] of backoff n-gram language models trained on the sources listed in Tables 2 and 3. For language modeling, the 18972 articles (35.4M words) from speech-related workshops and conferences (ICASSP, Eurospeech, ICSLP, ISCA workshops, LREC, ...) were processed using scripts similar to ones provided by IRST to convert postscript and pdf files to text. Further text processing was carried out to remove undesired items such as email, addresses, mathematical formulas and symbols, figures, tables, references, special formatting characters and ill-formed lines. The cleaned texts were then transformed to be closer to a natural speaking style, by applying our standard normalizations for numbers and dates, and transformations for acronyms and compound words.

Language modeling also made use of transcriptions of the audio data from TED (75k words), NIST meetings (150k words), ISL meetings (115k words) and ICSI meetings (756k words). The EARS conversational speech (CTS) transcripts were used as an additional source of spontaneous speech data for language modeling.

The above text sources were grouped into four subsets:

1. All the proceedings texts
2. Chil dev data
3. TED, meeting audio transcriptions
4. Conversational Telephone Speech transcription

Bigram, trigram and fourgram language models were estimated on each of the four text sources and interpolated, with interpolation weights of about 0.3 for the texts and 0.1 for CTS. The resulting perplexities are:

4-gram	jun04	px=97.6	jan05	px=107.1
3-gram	jun04	px=99.2	jan05	px=109.9
2-gram	jun04	px=110.5	jan05	px=132.6

We also investigated other combinations of the four text sources. The most important contribution comes from the CHIL development transcriptions, which give a large drop in perplexity particularly for the Jun04 data (Without these transcripts, the 4-gram perplexities of the jun04 data is 127.6 and of the jan05 data is 122.1).

Lexicon

The CHIL word list was selected from the audio transcripts and the proceedings texts. The original 65k BN wordlist had an OOV rate of 8.0% on the jun04 data and of 6.2% on the jan05 data. A 20k wordlist comprised of only the words in the audio transcripts had an OOV rate of about 1.3%. By combining the audio and textual sources, the resulting 35k wordlist has OOV rates of 0.23% and 0.17% on the jun04 and jan05 data sets respectively.

Pronunciations for several thousand words were added to the LIMSI American English dictionary in order to be able to train acoustic models on the audio corpora. Many of the words were compound words formed by concatenating pronunciations from components words, inflected forms and spelled or spoken acronyms. We are also investigating the use of pronunciation rules as a means of better modeling foreign accented speech.

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 lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [10]. 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 trigram language model (15M trigrams and 4M bigrams). The trigram lattices are rescored with a 4-gram language model (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 [8] with only two regression class. The lattice is generated for each segment using a bigram LM and position-dependent triphones with 19k contexts and 11k tied states (32 Gaussians per state). As in the first pass, the lattices are rescored with a 4-gram language model.

4. EXPERIMENTS AND RESULTS

Experimental results are reported on two sets of ISL seminars. The seminars were recorded with both near and far-field microphones, including a microphone array. In addition to the audio data, simultaneous calibrated video recordings were made. The first set, comprised of recordings from seven seminars (7 different speakers, all with German accents) was used in the June 2004 early evaluation, which aimed to assess existing technology on CHIL seminar data. Each seminar was split into four 5-minute segments, two for development and two for test. In total the development and test data each contain 1.2 hours of speech.

The second set, used for the Jan 2005 technology benchmark, is comprised of five seminars (5 different speakers, with German, American, Italian and Indian accents). Two of the five seminars were split into development and test portions, and the remaining three were only used for testing purposes. For this test there was about 0.75h of development data and 2.1hours of test.

All data were manually transcribed and segmented by ELDA with corrections provided by the CHIL ASR partners (IBM, UKA, LIMSI). The Jun'04 data has 3971 dev segments and 3077 eval segments. The Jan'05 data has 1395 and 1764 dev and eval segments respectively. Speech recognition tests were carried out on both the close-talking microphone (CTM) data and far-field microphone data. For the far-field task, the data from the individual microphone channels could be used, as well as the result of a delay-and-sum beam-forming performed at UKA [9].

Tables 4 and 5 give the word error rates of our primary system on the CTM data for two test sets. For the Jun'04 data each seminar has two test portions (denoted s1 and s2). The word error rates range from about 17% (for Jun'04 seminar ctm_2003-11-25_A to over 40% for seminars ctm_2003-11-25_B and ctm_2003-12-16_A).

Our first recognition results was obtained by running the LIMSI RT03 BN system on the Jun04 seminar data.

Segment	WER
ctm_2003-10-28_s1	21.9
ctm_2003-10-28_s2	22.2
ctm_2003-11-11_s1	25.0
ctm_2003-11-11_s2	22.3
ctm_2003-11-18_s1	21.8
ctm_2003-11-18_s2	22.3
ctm_2003-11-25_A_s1	18.1
ctm_2003-11-25_A_s2	15.8
ctm_2003-11-25_B_s1	34.3
ctm_2003-11-25_B_s2	42.6
ctm_2003-12-16_A_s1	37.9
ctm_2003-12-16_A_s2	44.1
ctm_2003-12-16_B_s1	24.3
ctm_2003-12-16_B_s2	23.7

Table 4: Word error rates of the primary system on the Jun04 test seminars per seminar segment.

Seminar	WER
ctm_20041111_1100	26.7
ctm_20041111_1400	20.7
ctm_20041111_1545	23.1
ctm_20041112_1030	26.8
ctm_20041112_1400	22.0

Table 5: Word error rates of the primary system on the Jan05 test data by seminar.

The initial word error rates were quite high, being over 50% on the CTM data and over 80% for chan00 of the microphone array. We focused all of our development work at improving performance on the CTM data. This is due to our belief that improvements for this data will also apply to far-field data, and also to the lack of available far-field training data representative of the CHIL rooms. The overall results are summarized in Table 6. On the Jan05 data, the CHIL primary system obtained a word error rate of 23.6%, compared with the 42.2% obtained with the LIMSI RT04 BN transcription system. The effect of adding just a small amount of speech (1 hour total) from the test speakers to the almost 100 hours of other data can be seen by comparing the CTM primary and nodev systems. There is a 13% relative gain on the Jun04 data where all seminars had specified development segments, and 9% on the Jan05 data where only two of the five seminars had development portions. The last entry gives with word error rate on the beam-formed microphone array data, which is about twice that obtained on the close-talking microphone data.

5. CONCLUSIONS AND FUTURE WORK

The general task of transcribing lectures and seminars is a challenging one, combining the difficulties encountered in the processing spontaneous speech and the difficulties of far-field speech recognition. Most of the results reported here are for close-talking microphone data

System	*Jun04*	*Jan05*
RT04 BN	-	42.2
CTM, primary	26.2	23.6
CTM, no dev	30.2	26.0
Beam, primary	57.6	51.9

Table 6: Overall word error rates on the Jun04 and Jan05 data.

since there was no far-field microphone data available for system development. It is our belief that most of techniques which improve recognition of CTM data will also improve far-field speech recognition. Since no audio or textual training data were available, other data sources were used for acoustic and language model training. Including about 1 hour of development data in the training data resulted in a 10% reduction in word error rate. Future work will investigate automatic partitioning of the data into speaker turns and multi-microphone training to improve the far-field recognition.

REFERENCES

[1] A. Waibel, H. Steusloff, R. Stiefelhagen, "CHIL - Computers in the Human Interaction Loop," *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, April 2004. (http://isl.ira.uka.de/chil)

[2] L.F. Lamel, F. Schiel et al., "The Translanguage English Database TED," *ICSLP'94*, Yokohama, Sep 1994. (LDC2002S04)

[3] The Translanguage English Database (TED) Transcripts, LDC catalog number LDC2002T03, isbn 1-58563-202-3.

[4] S. Burger, V. MacLaran, H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style, *ICSLP'02*, Denver, Sep 2002. (LDC2004S05, LDC2004E04, LDC2004E05)

[5] A. Janin, D. Baron et al, "The ICSI Meeting Corpus," *ICASSP'03*, Hong Kong, Apr 2003. (LDC2004S02, LDC2004T04)

[6] J.S. Garofolo, C.D. Laprun et al., "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] 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.

[9] D. Macho, J. Padrell et al., "First experiments of automatic speech activity detection, source localization and speech recognition in the CHIL project," *Workshop on Hands-Free Speech Communication and Microphone Arrays*, Rutgers University, Piscataway, NJ, 2005.

[10] L. Mangu, E. Brill, A. Stolcke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeech'99*, 495-498, Budapest, Sep 1999.

[11] P.C. Woodland, T. Niesler, E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR," presented at the 1998 Hub5E Workshop, Sep 1998.

[12] L. Nguyen et al., "The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System," *DARPA RT04 workshop*, Palisades, NY, Nov 2004.