The LIMSI RT06s Lecture Transcription System

L. Lamel, E. Bilinski, G. Adda, J.L. Gauvain, and H. Schwenk[⋆]

LIMSI-CNRS, BP 133
91403 Orsay Cedex, France

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 presentations. Widely available corpora were used to train both the acoustic and language models, since only a small amount of CHIL data was available for system development. 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

One of the CHIL services is to provide on-line and off-line support for lecture situations. 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 and seminars) 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. At LIMSI our focus is on developing a lecture and seminar transcription system for off-line applications.

The speech recognizer for CHIL has been developed from the LIMSI Broadcast News transcription system for American English [6]. Since only a small amount of CHIL data was available for system development widely available corpora were used to train both the acoustic and the language models. Acoustic model training made use of the transcribed portion of the TED corpus of Eurospeech recordings, the ICSI, ISL, and NIST meeting corpora, and a few CHIL seminars. For language model training,

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

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 recognizer used in the January 2005 evaluation is described in [8, 10]. In the remainder of this paper the 2006 speech recognizer is described, highlighting differences from the 2005 system and development results are provided.

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 [6]. The transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning is based on an audio stream mixture model [6], 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 better deal with the farfield microphone data [17]. 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 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 2006 word list is case-sensitive and contains 58k words, and several thousand compound words and acronyms. (The 2005 word list had 35k words, and both the word list and language models were case-insensitive). Pronunciations for several thousand words were added to the LIMSI American English dictionary. Many of the additional words were compound words formed by concatenating pronunciations from existing words, inflected forms and spelled or spoken acronyms.

Word recognition is performed in multiple decoding passes, where each 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 [13]. The words with the highest posterior in each confusion set are hypothesized. The final decoding pass makes use of a connectionist language model interpolated with a 4-gram model.

3 Training Corpora

Although multi-site data collection is ongoing, only a limited number of transcribed seminars were available for speech recognizer training since priority was given to selection of the evaluation test data. Therefore one of the problems was to locate appro-

Table 1. Summary of audio data sources.

Source	*Microphone*	*Type*	*Amount*
TED	lapel	39 lectures	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

Table 2. Summary of additional CHIL audio data sources used in 2006. s1 and s2 correspond to different segments from the same seminar.

CHIL_ctm_2003-10-28_[s1,s2], CHIL_ctm_2003-11-11_[s1,s2], CHIL_ctm_2003-11-18_[s1,s2]
CHIL_ctm_2003-11-25_A_[s1,s2], CHIL_ctm_2003-11-25_B_[s1,s2],
CHIL_ctm_2003-12-16_A_[s1,s2], CHIL_ctm_2003-12-16_B_[s1,s2],
CHIL_ctm_20041111_1100, CHIL_ctm_20041111_1400, CHIL_ctm_20041111_1545,
CHIL_ctm_20041112_1030, CHIL_ctm_20041112_1400, ISL_20040614_ctm
ISL_20040616_ctm, ISL_20040621_ctm, ISL_20040721_ctm, ISL_20040830_ctm

priate 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 [9]. 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 [1]. 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 [3, 5, 7]. The amount of data per corpus is summarized in Table 1. The first four corpora were used in training the 2005 system, and the last two entries are new in the 2006 system. Using a single microphone channel per speaker for the data from all four sources (distributed by the LDC), a total about 97 hours of audio training data were available in 2005 and an additional 76 hours of data were used in 2006. As can be seen, in the 2005 system only close-talking microphone data were used for acoustic model training, whereas some distant microphone channels were used in 2006.

Since one of the aims in CHIL is speech recognition of farfield data, and the primary RT06s task being the multiple distant microphone condition, in the 2006 system farfield data were also used in training. To this end a selection of the farfield data in the ICSI corpus were used. Since for the ISCI data there are a varying number of channels, for each of the speakers, a single farfield channel was selected as being the most appropriate for that speaker. The microphone channel was chosen as that having the highest likelihood during forced alignment on a subset of data for each speaker. During training these data were pooled with the close-talking microphone data. The CHIL seminars included in this year's training are listed in Table 2.

Table 3. Summary of audio transcripts.

TED 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

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. In addition transcriptions of conversational telephone speech from the CallHome, SwitchBoard and Fisher collections (distributed by the LDC) were used. 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 3. Compared with last year's system we made use of some additional data from the AMI/IDIAP meeting corpus, the NIST RT04 and RT05 development data, and the transcripts of the additional CHIL seminars (24k words more than last year).

Table 4. Summary of proceedings texts (20k articles, 42M 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

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 as shown in Table 4. The almost 20k papers in the proceedings texts were processed using scripts derived from ones shared by ITC-IRST to convert postscript and pdf files to ascii texts. The texts were extracted from the PDF files, after removing corrupted documents. 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. A stricter filtering was applied this year than last year in order to have cleaner texts for language model training.

4 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 are 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 [6]. A set of 152 questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

The baseline acoustic models were those used in the LIMSI 2005 CHIL system. This year several enhancements were added to the system. Speaker adaptive (SAT) training was used in the system which gave about a 1% absolute improvement for the ihm condition. The acoustic models were trained on multi-style data, including close-talking and farfield microphones. Different training strategies were investigated, and the best for us was to pool the close-talking head-mounted data with the table-top audio data for the models used in the ihm condition. For the farfield conditions (sdm, mdm, and mm3a) the best results were obtained by adapting the ihm models with the farfield data. We also explored adapting broadcast news models since this was advantageous last year, but better results were obtained on the development data using lecture models.

5 Language modeling

The 58k recognizer word list was determined using the following process: The 75k most probable words were selected by linear interpolation of the unigram language models obtained from the 8 different seminar and meeting data sources listed in Table 3 and a component trained on 46M words of proceedings texts. While the CTS data were used for language model training, they were not used for vocabulary selection. Compared with last year, about 300k words of additional audio transcripts were available. This 75k word list was then filtered using our master pronunciation dictionary for American English in order to eliminate errors. (Pronunciations for several hundred frequently occurring words in the transcripts had already been added to the master dictionary for AM training.) The final word list contains 57768 words, and has an OOV rate of 0.46% on the development data (RT05s eval). Last year’s 35k word list had an OOV rate of 0.61% on the same data.

For language model estimation the available corpora were grouped into 3 sources:

- Seminar and meeting transcriptions (1.42M words)
- Proceedings texts (46M words)
- Transcriptions of Conversational Telephone Speech databases available from LDC (29M words)

The proceeding texts are comprised of the proceedings from 54 conferences and workshops in speech and language, which represent about 20,000 PDF documents.

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 perplexity of the 4-gram LM is 130 on the development data, which can be compared to 140 with last years’ model.

Our text normalization process has been drastically changed since last year. Our previous normalization was case insensitive (all words were in upper case), no compound words were allowed, and all acronyms were split in a sequence of letters. In the new normalization, some compound words are kept (words which contain an internal hyphen such as *‘air-conditioning’*) according to a reference list. The reference list of compound word was derived from compound words in the American Heritage Dictionary, completed with a list of first and last names in different languages, and some place names found in geographical and historical encyclopedias. The texts were also processed with a primitive named entity detector, in order to preserve the proper names.

Thus, for instance, for a compound word ’A-B’ present in the text, 2 cases are distinguished:

1. If the word ‘A-B’ is present in one of the reference lists or was tagged as a ‘proper name’, the word ‘A-B’ is kept as a lexical entry.
2. If the first case is not true, the word ‘A-B’ is split into the two words ‘A B’.

The new text processing is case sensitive, which means that to have correct texts, a decision must be taken as to what is the true case for the first word of each sentence. Moreover, in some texts word case may be vague either for stylistic reasons (signifying emphasis) or due to segmentation errors in the proceedings texts. For these texts the case of all words needs to be reconsidered. This process is done by adding all the possible cases encountered in the texts for all words for which case is potentially ambiguous. For this process the available 472M words of broadcast news texts were also used. To attribute the correct case for the sentence-initial word an interpolated language model was constructed with a set of texts after removing the first word of each sentence. Case is then added to the original sentence by creating a graph with all possible cases for all words with multiple forms, and parsing the graph using the interpolated language model. Finally, all acronyms are considered as words and have not been split.

The aim of this new normalization is not to optimize the lexical coverage or to decrease the perplexity of the language model, but to be able to deliver a transcription containing more information, and thus facilitate the further downstream processing (named entity detection, summarization, ...).

Connectioninst language models [2, 14] have been shown to be performant when LM training data is limited. The basic idea is to project the word indices onto a continuous space and to use a probability estimator operating on this space, as illustrated in Figure 1. Both tasks are performed by a neural network. This is still a n-gram approach, but the n-gram LM probabilities are “interpolated” for any possible context of length n-1 instead of backing-off to shorter contexts. Since the resulting probability densities are continuous functions of the word representation, better generalization to unknown n-grams can be expected.

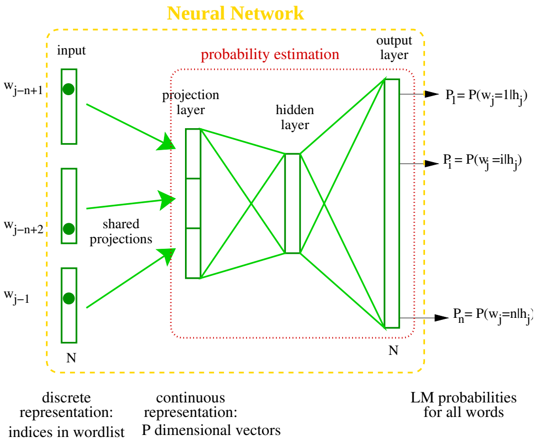

Fig. 1. Neural network language model.

The neural network LM was trained on the transcriptions of the audio data and the proceedings, but not the CTS data. This language model reduces the perplexity on the development data from 135 to 108.

5.1 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 [13]. 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 (7M 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 [11] with only two regression class. The lattice is generated for each segment using a 58k word bigram LM and position-dependent triphones with 24k contexts and 11k tied states (32 Gaussians per state). As in the first pass, the lattices are rescored with a 58k word 4-gram language model (9M fourgrams, 19M trigrams and 5M bigrams).

6 Experiments and results

All the development work at LIMSI was carried out using a portion of the RT05s evaluation data for which audio files and transcripts were shared by UKA. The data consist of the ihm channels of excerpts from 16 seminars shown in Table 5. For the sdm condition and mdm ccondition the subset of seminars used are listed in the lower part of the table.

Table 5. Development set used at LIMSI for ihm and sdm/mdm conditions (from RT05s eval).

ihm:
CHIL_20041123-0900-[E1,E2]_h01_001, CHIL_20041123-1000-[E1,E2]_h01_001,
CHIL_20041123-1100-[E1,E2,E3]_h01_001, CHIL_20041123-1500-[E1,E2]_h01_001,
CHIL_20041123-1600-[E1,E2]_h01_001, CHIL_20041124-1000-[E1,E2]_h01_001,
CHIL_20041124-1100-[E1,E2]_h01_001, CHIL_20050112-0000-[E1,E2]_h01_001,
CHIL_20050126-0000-E1_h01_001, CHIL_20050127-0000-E1_h01_001,
CHIL_20050128-0000-[E1,E2]_h01_001, CHIL_20050202-0000-[E1,E2]_h01_001,
CHIL_20050214-0000-E1_h01_001, CHIL_20050310-0000-[E1,E2]_h01_001,
CHIL_20050310-0001-E1_h01_001, CHIL_20050314-0000-[E1,E2]_h01_001
sdm/mdm:
CHIL_20041123-1600-E1: d01, d02, d03,d04, d05 CHIL_20050202-0000-E2: d01, d02, d03,d04, d05
CHIL_20050314-0000-E2: d01, d02, d03,d04, d05 CHIL_20050128-0000-E1: d01, d02, d03,d04, d05
CHIL_20050310-0001-E1: d01, d02, d03,d04, d05

The starting point was an updated version of the LIMSI 2005 evaluation system which incorporated an automatic data partitioner (last year's evaluation used manually determined speech segments). As cen be seen in Table 6 the word error rate with the baseline system was 26.1%. This is the same system as was used to automatically transcribe the Q&A development seminars. Using updated acoustic models (AM 1), with the 35k LM from last year gave a small increase in performance. Resgmenting the training data and reestimating acoustic models (AM 2) gave an additional small gain, with a larger gain from SAT training (AM 2 + SAT). The use of the 58k wordlist and language model with the baseline acoustic models gave essentially the same performance as the baseline system (verifying that the new normalization did not degrade performance). Retraining acoustic models with the same text normalization results in a better match of the cross-word context-dependent phone models during training and test, as can be seen in the second entry in the lower part of Table 6. Tuning the system gave a further improvement, as did speaker adaptive training (+SAT), pronunciation probabilities (+ pron probs) and the connectionist language model (+ NNLM). The neural network LM achieves almost a 1% absolute word error reduction on top of the other improvements. Overall on the development data a relative word error reduction of 13% was obtained compared to the 2005 system.

Table 6. Word error rates on the ihm development data (see Table 5).

System	*WER (%)*
Baseline, 35k LM	26.1
Updated AM 1, 35k LM	25.9
Updated AM 2, 35k LM	25.7
Updated AM 2+SAT, 35k LM	25.0
58k wordlist, LM	26.0
58k LM, updated AM	25.3
+ tuning	24.6
+ SAT	24.0
+ pron probs	23.5
+ NNLM	22.6

The close-talking microphone contrast condition for RT06s aimed at transcribing only the speech from the primary talker. This is very different from the previous CHIL evaluations where the segments of speech from other speakers were removed prior to scoring. In order to compare the performance of the LIMSI 2006 system to that of the 2005 system, we decided to score using the same method as was used last year, that is ignoring speech in inter-segment gaps. (Strong arguments can be made for both points of view. At LIMSI we believe that the role of the speech recognizer is to transcribe speech into words, independent of who spoke them. It is the role of the speaker diarization system to associate words with the person who spoke them. This is not the point of view taken in the RT06s evaluation, and the evaluation plan, while clearly stating this for SAD is a bit ambiguous on this point for STT.) The LIMSI system did not make any attempt to exclude speech from other speakers and therefore has a very high insertion rate (over 121%, more than 22k words), giving an overall error rate of 147%. (Note that the development set we used did not contain background speech so it was not possible to do a serious development activity. Therefore we did not try.)

Table 7 gives unofficial, comparative results on the RT06s evaluation data with the baseline system and the RT06s evaluation system on the ihm audio data, ignoring speech in inter-segment gaps. Although these numbers cannot be compared to other sites, they allow us to measure the improvement of our models this year. The overall error rate has been reduced by almost 12% absolute (28% relative). A more appropriate measure would be to score all of the speech recognized, but at the time of this writing reference transcriptions for the inter-segment gap regions are not available.

Development for the farfield data condition was carried out on portion of the available data consisting of 5 seminars listed in the lower portion of Table 5. Initially the

Table 7. Unofficial, comparative results on the RT06s evaluation data with the baseline system and the RT06s evaluation system on the ihm audio data, ignoring speech in inter-segment gaps. There are a total of 19373 reference words.

System	*Cor*	*Subs*	*Del*	*Ins*	*WER*
ihm baseline	64.0	25.7	10.3	6.1	42.1
ihm RT06s	72.9	19.5	7.6	3.2	30.3

standard data partitioner was used, but when the SAD/SPKR systems for RT06s were finalized (see [17]), these were used for further development work and in the final evaluation system. As can be seen in Table 8, the word error on the sdm data was reduced from 64% to 55% when scoring with overlap. The mdm system output was created by combining with ROVER [4] the outputs of all the sdm channels. For this condition the development word error rate was reduced from 55.7% to about 51%.

Table 8. Word error rates on the sdm and mdm development data (see Table 5).

	sdm WER (%)		*mdm WER (%)*
System	*overlap*	*non overlap*	*overlap*
58k LM, update AM 2	64.0	62.9	
58k LM, adapt AM with FF	62.5	61.3	55.7
+ tuning	60.4	58.8	
+ SAT	60.1	58.5	
+ mdm partitioner	56.6	57.1	53.3
+ pron prob	56.1	55.3	
+ NNLM	55.2	54.4	51.1

The first two entries in Table 9 give the official NIST results for the for the mdm and sdm conditions. As for the development results, the mdm hypotheses are obtained by applying ROVER [4] to the hypotheses of the individual microphone channels. The lower part of the table gives unofficial contrastive results on the mm3a beam-formed multiple mark III microphone array data provided by UKA for the baseline RT06s farfield system and with acoustic model adaptation using about 5 hours of beam-formed multiple mark III microphone array data also provided by UKA. These data correspond to the portion of the development data for which manual transcripts were available. Since no data remained for system development, we chose not to submit a system for this condition in the official evaluation. Although all the word error rates are quite high, the best performance is obtained on the mm3a data using the acoustic models adapted with the beam-formed data.

Table 9. NIST official RT06s results on farfield data: mdm and sdm conditions (top). Unofficial, comparative results on the RT06s evaluation beam-formed data distributed by UKA with the baseline farfield models and a post RT06 model set adapted with 5 hours of beam-formed data. There are a total of 17986 reference words. (Results are missing for seminar AIT_20051011_B_Segment1 due to a partitioning error for the baseline system.)

System	*Cor*	*Subs*	*Del*	*Ins*	*WER*
mdm	48.3	39.3	12.4	11.8	63.5
sdm	43.2	38.6	18.2	6.5	63.3
mm3a baseline	32.6	20.8	46.6	1.4	68.8
mm3a post RT06	48.1	31.7	20.2	5.6	57.5

7 Conclusions

This paper has described our research aimed at developing a system 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. Results were reported for both close talking and far-field microphones, for both development and evaluation data. This was LIMSI's first participation to the multiple farfield microphone task. Compared to the LIMSI 2005 system, the overall error rate has been reduced by over 10% on the development data for the ihm condition and about 15% on the RT06s eval mm3a data.

References

1. The Translanguage English Database (TED) Transcripts, LDC catalog number LDC2002T03, isbn 1-58563-202-3.
2. Y. Bengio and R. Ducharme, "A neural probabilistic language model," *Advances in Neural Information Processing Systems (NIPS)*, **13**:933-938, 2001.
3. 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)
4. J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," *Proc. ASRU'97*, 347-354, Santa Barbara, December 1997.
5. 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)
6. J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.
7. 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)
8. L. Lamel, G. Adda, E. Bilinski and J.L. Gauvain, "Transcribing Lectures and Seminars," *Proc. ISCA Eurospeech'05*, Lisbon, Sep 2005.
9. L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani and H. Tillmann, "The Translanguage English Database TED," *ICSLP'94*, Yokohama, Sep 1994. (LDC2002S04)

10. L. Lamel, H. Schwenk, J.L. Gauvain, G. Adda and E. Bilinski, “Improvements in Transcribing Lectures and Seminars,” *Proc. MLMI’05*, Edinburgh, July 2005.
11. 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.
12. D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. W olfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, and S. Chu, “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.
13. L. Mangu, E. Brill and A. Stolcke, “Finding Consensus Among Words: Lattice-Based Word Error Minimization,” *Eurospeech’99*, 495-498, Budapest, Sep 1999.
14. H. Schwenk, “Efficient training of large neural networks for language modeling,” *IJCNN* , pp. 3059–3062, 2004.
15. 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)
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. Gauvain, “Speaker Diarization: from Broadcast News to Lectures,” *Proc. RT06s*, submitted.
18. X. Zhu, C. Barras, S. Meignier and J.L. Gauvain, “Combining speaker identification and BIC for speaker diarization” *Proc. Interspeech’05, pp.2441-2444*, Lisboa, September, 2005
19. X. Zhu, C.C. Leung, C. Barras, L. Lamel, and J.L. Gauvain, “Speech activity detection and speaker identification for CHIL,” *Proc. MLMI’05*, Edinburgh, July 2005.