Article Submitted to Computer Speech and Language, published in Jan. 2002

Lightly Supervised and Unsupervised Acoustic Model Training

Lori Lamel, Jean-Luc Gauvain and Gilles Adda

Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay Cedex, FRANCE

Abstract

The last decade has witnessed substantial progress in speech recognition technology, with todays state-of-the-art systems being able to transcribe unrestricted broadcast news audio data with a word error of about 20%. However, acoustic model development for these recognizers relies on the availability of large amounts of manually transcribed training data. Obtaining such data is both time-consuming and expensive, requiring trained human annotators and substantial amounts of supervision.

This paper describes some recent experiments using lightly supervised and unsupervised techniques for acoustic model training in order to reduce the system development cost. The approach uses a speech recognizer to transcribe unannotated broadcast news data from the Darpa TDT-2 corpus. The hypothesized transcription is optionally aligned with closed captions or transcripts to create labels for the training data. Experiments providing supervision only via the language model training materials show that including texts which are contemporaneous with the audio data is not crucial for success of the approach, and that the acoustic models can be initialized with as little as 10 minutes of manually annotated data. These experiments demonstrate that lightly or unsupervised supervision can dramatically reduce the cost of building acoustic models.

1. Introduction

Despite the rapid progress made in large vocabulary continuous speech recognition, there remain many outstanding challenges. One of the main challenges is to reduce the development costs required to adapt a recognition system to a new task or another language. With today's technology, the adaptation of a recognition system to a new task or another language requires large amounts of transcribed training data. One of the most often cited costs in development is that of obtaining this necessary transcribed acoustic training data, which is an expensive process in terms of both manpower and time. There are certain audio sources such as radio and television news broadcasts, that can provide an essentially unlimited supply of acoustic training data. However, for the vast majority of audio data sources there are no corresponding accurate word transcriptions. Some of these sources, in particular, the main American television channels also broadcast manually derived closed-captions. The closed-captions are a close, but not exact transcription of what is being spoken, and these are only coarsely time-aligned with the audio signal. Manual transcripts are also available for certain radio broadcasts (Cieri et al., 1999). [*] There may also exist other sources of information with different levels of completeness such as approximate transcriptions, summaries or keywords, which can be used to provide some supervision.

This paper describes a series of recent experiments aimed at reducing the level of supervision required for acoustic model training. The basic idea is to use a speech recognizer to automatically transcribe raw audio data, thus generating approximate transcriptions for the training data. Training on all of the automatically annotated data is compared with using the closed-captions/transcripts to filter the hypothesized transcriptions, thus removing words that are potentially incorrect and training only on the words which agree. The effects of using different levels of supervision, via selection of the language model training texts is also assessed.

Although the idea of using untranscribed data to train acoustic models has been proposed before (see Zavaliagkos and Colthurst (1998); Kemp and Waibel (1999)), we are not aware of any large-scale, thorough experiments with this technique on publicly available corpora. In the experiments carried out at BBN completely unsupervised acoustic training from a conversational speech corpus (Switchboard and Callhome Spanish corpora) were combined with 3 hours of manually annotated data (Zavaliagkos and Colthurst, 1998). Small improvements (0.4 to 3% relative) were reported which led to their conjecture that an order of magnitude more untranscribed data is needed to achieve comparable levels of performance with transcribed data. Kemp and Waibel report significant word error reductions using untranscribed data for German broadcast news transcription from one source. They show that comparable levels of performance can be obtained by using twice as much untranscribed data as transcribed data (30 hours versus 15 hours). [†]

The remainder of this paper is as follows. The next section presents the basic ideas of lightly supervised training, followed by a description of the corpora used in this work and an overview of the LIMSI broadcast news transcription system. The experimental results are given in Sections 5 through 7 varying the amount

[*] In this paper we group together both of these types of transcripts and refer to them as closed-captions.

[†] The authors give little information about the data used to train the language models, and thus it is difficult to assess the level of supervision.

of manually transcribed data used to estimate the initial acoustic models and the quantity and sources of language model training texts. Finally some possible extensions of this work are discussed.

2. Lightly supervised acoustic model training

HMM training requires an alignment between the audio signal and the phone models, which usually relies on an orthographic transcription of the speech data and a good phonemic lexicon. The orthographic transcription is usually considered as ground truth, and assumed to be the word sequence that the speech recognizer should hypothesize when confronted with the same speech segment. In general it is easier to deal with relatively short speech segments so that transcription errors will not propagate and jeopardize the alignment.

Training acoustic models for a new corpus (which could also reflect a change of task and/or language), usually entails the following sequence of operations once the audio data and transcription files have been loaded:

1. Normalize the transcriptions to a common format (some adjustment is always needed as different text sources make use of different conventions).
2. Produce a word list from the transcriptions and correct blatant errors (these include typographical errors and inconsistencies).
3. Produce a phonemic transcription for all words not in the lexicon (these are manually verified).
4. Viterbi align the orthographic transcriptions with the signal using existing models (which can be bootstrapped from another task or language) and the pronunciation lexicon to produce a time-aligned phone transcription. Since the reference transcriptions and the phonemic lexicon are not really perfect, this alignment procedure may not succeed. Failure occurs when there is no complete Viterbi alignment due to beam-pruning or when some duration criteria are not respected such as a maximum allowable phone duration. [‡]
5. Correct transcription errors in the unaligned data or simply ignore these segments, discarding the corresponding data, if enough audio data is available.
6. Run the standard EM training procedure.

These operations may be iterated several times to refine the acoustic models. In general each iteration recovers a portion of the rejected data.

One can imagine training acoustic models in a less supervised manner. In fact, any related linguistic information available about the audio sample can be used in place of the manual transcriptions required for alignment. This information can be used in training a language model, which can be used to produce the most

[‡] A phone duration longer than 500ms is likely to be indicative of an error, for phones other than silence or breath noise.

likely word transcription given the current models. This language model can range from a simple left-to-right word graph corresponding to the orthographic transcription to a very open N-gram model encoding the available linguistic content of the training data. An iterative procedure can successively refine the models and the transcription. This approach still fits within the EM training framework, which is well-suited for missing data training problems. The automatically produced orthographic transcriptions of the training data can eventually be filtered using confidence measures (Zavaliagkos and Colthurst, 1998; Kemp and Waibel, 1999) or approximate manual transcriptions (such as closed captions) keeping only words that are likely to have been correctly recognized.

Detailed annotation requires on the order of 20-40 times real-time of manual effort, and even after manual verification the final transcriptions are not exempt from errors (Barras et al., 2000). Orthographic transcriptions such as closed-captions can be produced in a few times real-time, and therefore are quite a bit less costly. These transcriptions have the added advantage that they are already available for some television channels. However, there are several problems that must be faced when dealing with closed captions instead of accurate speech transcriptions. In addition to providing an exact word-level transcription of what was said, the detailed speech transcriptions often provide a wealth of additional information that is not available in the closed-captions. This includes the marking of non-speech events such as respiration, coughing, throat clearing; indication of speaker turns, as well as the speaker identities and gender; indication of the acoustic conditions, such as the presence of background music or noise, and the transmission channel; and the annotation of non-speech segments such as music. The closed-captions, while accurately reflecting the meaning, are much less precise. Hesitations and repetitions are not marked and there may be word insertions, delections and changes in the word order. NIST found the disagreement between the closed-captions and manual transcripts on a 10-hour subset of the TDT-2 data to be on the order of 12% (Garofolo et al., 1999).

In order to use the closed-captions for training we need to automatically produce some of the missing information such as an audio segmentation into speaker turns, with (intra-show) speaker identifiers, and identifying nonspeech segments and acoustic conditions. [§] Also, each word in the closed-caption needs to be aligned to the audio signal, which must allow for the transcription errors (such as insertions, deletions and substitions).

In addition to the closed-captions, some other possible electronic sources of text come from newpapers and newswires, and the internet. The text data may be contemporaneous with the audio data or may predate, or in the case of archives, postdate, the period. However, since these sources have only an indirect correspondence with the audio data, they evidently provide less supervision than orthographic transcriptions or the closed-captions.

[§] The models used in the partitioning process are Gaussian mixture which can be trained on a very small amount of data, so the required labeling is not very costly.

The following training procedure is used in this work which can be used with all of the different levels of supervision schemes described above:

1. Normalize the available text materials (e.g., newspaper and newswire, commercially produced transcripts, closed-captions, detailed transcripts of acoustic training data) and train an n-gram language model
2. Partition each show into homogeneous segments, labeling the acoustic attributes (speaker, gender, bandwidth) (Gauvain et al., 1997)
3. Train acoustic models on a small amount of manually annotated data (1 hour or less)
4. Automatically transcribe a large amount of raw training data
5. Optionally align the closed-captions with the automatic transcriptions (using a dynamic programming algorithm) removing speech segments where the two transcripts disagree.
6. Run the standard acoustic model training procedure on the speech segments using the automatic transcripts
7. Reiterate from step 4.

It is easy to see that the manual work is considerably reduced, not only in generating the annotated corpus but also during the training procedure, since there is no longer a need to deal with new words and word fragments in the data and errors in the detailed manual transcriptions do not need to be corrected.

One way to iterate the procedure is to retranscribe all of the available training data with the new models. This being costly, we choose to process the data by chunks, where each new chunk is transcribed using models trained on all of the previously transcribed chunks. In this way the training data is only transcribed once, with the quality of the transcription improving at each step. (An exception to the chunked processing is an experiment is reported in Section 7 where the models from a given iteration are used to retranscribe the same data.) This approach was used to produce automatic word transcriptions of the 500 hours of audio broadcasts used in the TREC spoken document retrieval task (NIST SDR'99).

Three sets of experiments exploring lightly supervised and unsupervised acoustic model training were carried out. The first set of experiments, reported in Section 5, aim to assess recognition performance as a function of the available acoustic training data, in terms of both the amount of raw data and the quality of the transcription. Results are given for supervised training and for light supervision with and without the use of closed-caption filtering (Tables 1 and 2).

The second set of experiments, described in Section 6, investigates the impact of different levels of supervision via the language model training materials. Seven language models (see Table 3) were estimated using various combinations of the text sources, from the same epoch as the TDT-2 audio data or predating the period. As mentioned earlier, newpaper and newswires sources have only an indirect correspondence with the audio data and therefore provide less supervision than the closed captions and commercially generated transcripts.

Section 7 explores unsupervised acoustic model training. In this third set of experiments, the amount of available acoustic and language model training data are severely reduced in order to better understand the tradeoff between cost and performance. Except for the initial bootstrap models estimated on 10 minutes of manually transcribed data, all acoustic model training is unsupervised.

With the exception of the baseline Hub4 language models used for comparative purposes, none of the language models include a component estimated on the transcriptions of the Hub4 acoustic training data. All recognition runs are carried out in under 10xRT unless stated otherwise.

3. Corpora

The unannotated audio data used in these experiments are taken from the Darpa TDT-2 corpus (Cieri et al., 1999). The corpus used in this work consists of over 550 hours of data from 6 sources: CNN Headline News (550 30-minute shows), ABC World News Tonight (139 30-minute shows), Public Radio International The World (103 1-hour shows), Voice of America Today and World Report (111 1-hour shows). These TDT-2 collection contains data broadcast between January and June 1998, and have associated closed-captions for the TV shows and commercially produced transcripts for the radio shows (The data from January were not used in these experiments.). The data is divided in about 22k stories with timecodes identifying the beginning and end of each story, and with an average duration of 1 minute and 20 seconds per story.

The language model training data are those used for the Darpa Broadcast News task, with the exception that in this work none of the manual transcriptions of the acoustic training data were used for either word list selection or language model estimation. [¶] These data include: about 790M words of newspaper and newswire texts distributed by LDC (Jan 1994 - May 1998) from the Hub4 and TDT corpora; 240M words of commercial broadcast news transcripts distributed by the LDC (years 92-95) and directly from PSMedia (years 96-97); and the closed captions (predating June 98) distributed as part of the TDT-2 corpus.

For testing purposes we use the 1999 Hub4 evaluation data, which is comprised of two 90 minute data sets selected by NIST. The first set was extracted from 10 hours of data broadcast in June 1998, and the second set from a set of broadcasts recorded in August-September 1998 (Pallett et al., 2000).

[¶] Since we have worked on the broadcast news transcription task for several years, and have therefore acquired a fair amount of knowledge about this task, we paid extra attention to avoid inadvertently including information in the unsupervised training experiments. One source of knowledge that could not be avoided concerns the lexical pronunciations which were not modified for this work.

4. System Description

The LIMSI broadcast news transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning serves to divide the continuous stream of acoustic data into homegenous segments, associating appropriate labels with the segments. The segmentation and labeling process (Gauvain et al., 1997) first detects and rejects non-speech segments, and then applies an iterative maximum likelihood segmentation/clustering procedure to the speech segments. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

The speech recognizer uses 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. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypotheses are used in cluster-based acoustic model adaptation using the MLLR technique (Leggetter and Woodland, 2000) prior to word graph generation. A 3-gram language model is used for the first two decoding passes. The final hypotheses are generated with a 4-gram language model and acoustic models adapted with the hypotheses of step 2.

In our baseline system used in DARPA evaluation tests, the acoustic models were trained on about 150 hours of manually transcribed audio data from the DARPA Hub4 Broadcast News corpus (the LDC 1996 and 1997 Broadcast News Speech collections) (Graff, 1997), predating the TDT-2 data used for unsupervised training. These carefully annotated data comes from a variety of sources: ABC (Nightline, World News Now, World News Tonight), CNN (Early Prime, Headline News, Prime News, The World Today, Early Edition, Prime Time Live), CSPAN (Washington Journal, Public Policy), and NPR (All Things Considered, Marketplace) (Graff, 1997). We used the August 1997 and February 1998 releases of the LDC transcriptions. In addition to the word transcriptions, the annotations include speech fragments and non-speech events, speaker turns and identities, and markers for overlapping portions and non-English speech. Overlapping speech portions were detected in the transcriptions and removed from the training data.

The acoustic feature vector has 39-components comprised of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives Gauvain et al. (2002). Gender-dependent acoustic models were built using MAP adaptation (Gauvain and Lee, 1994) of SI seed models for wideband and telephone band speech. For computational reasons, smaller sets of acoustic models are used in the first decoding pass. These position-dependent, cross-word triphone models cover 5500 contexts, with 6300 tied states and 16 Gaussians per state. For the second and third decoding passes, a larger set of 28000 position-dependent, cross-word triphone models with 11700 tied states are used, with approximately 180k and 360k Gaussians (Gauvain and Lamel, 2000).

Baseline language models were obtained by interpolation of backoff n-gram language models trained on 3 different data sets: commercial broadcast news transcripts (240M word), newspapers (North American Business News) and Associated Press (AP) Wordstream texts (790M words) excluding the test data epochs, and the transcriptions of the broadcast news acoustic data.

The baseline recognition vocabulary contains 65120 words and 76644 phone transcriptions, and has a lexical coverage of over 99% on all evaluation test sets from the years 1996-1999. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. About 15% of the lexical entries have more than one pronunciation, with 90% of these having two alternatives. The pronunciations make use of a set of 48 phones set, where 3 phone units represent silence, filler words, and breath noises. The filler and breath phones model only these two events and are not used in transcribing other lexical entries. The lexicon contains compound words for about 300 frequent word sequences, as well as word entries for common acronyms, providing an easy way to allow for reduced pronunciations (Gauvain et al., 1997).

The LIMSI 10x system[‖] obtained a word error of 17.1% on the 1999 DARPA/-NIST evaluation set, and can transcribe unrestricted broadcast data with a word error of about 20% (Gauvain and Lamel, 2000). The word error can be reduced to 15.6% for a system running at 50xRT.

5. Impact of the amount of acoustic training data

As mentioned above, in our standard broadcast news system the language models are built by interpolating n-gram LMs built on 3 sources of texts (Adda et al., 1999): large amounts of newspaper and newswire texts, large amounts of commercial BN transcriptions, and much smaller amounts (what ever is available) of detailed BN transcriptions. Since this work investigates the use of unannotated acoustic model training data, the component language model estimated on the detailed BN transcriptions was replaced with one trained on closed-captions from the TDT-2 corpus used in these experiments. A 65k word list was selected based on the word frequencies in the training texts, i.e., excluding the detailed transcriptions. The language model interpolation coefficients were chosen in order to minimize the perplexity on a 38,000 word development set composed of the second set of the Nov'98 evaluation data (3h) and a 2h portion of the TDT-2 data from Jun'98 (not included in the language model training data). The resulting language model News.Com.Cap is the interpolation of language models trained on newspaper and newswires (News), commercially produced transcripts (Com), and closed-captions (Cap) through May98. The interpolation coefficients

[‖] The notation 10x means that the entire transcription process including partitioning, word recognition and adaptation requires at most 10 hours of computation time to process one hour of data. This was the targeted processing speed in the 1999 evaluation.

are 0.45 for the commercial transcript language model, 0.35 for the newspaper language model and 0.20 for the TDT-2 closed caption language model.

Amount of training data		*Conditions*	*Language Model*
Raw	*Usable*		*News.Com.Cap*
1.5h	1h	1S	33.3
50h	33h	1S	20.7
104h	67h	1S	19.1
200h	123h	1S	18.0
200h	123h	4S	17.1

Table 1: Supervised acoustic model training: Word error rate (%) on the 1999 evaluation test data for various conditions using acoustic models trained on the HUB4 training data with detailed manual transcriptions. All runs were done in less than 10xRT, except for the final row. "1S" designates one set of gender-independant acoustic models, whereas "4S" designates four sets of gender and bandwidth dependent acoustic models. The "News.Com.Cap" language model is trained on the available text sources including TDT-2 closed-captions, but without the detailed transcriptions of the acoustic training data.

As a baseline, Table 1 gives the word error rate with the News.Com.Cap language model as a function of the amount of manually annotated acoustic training data. The raw data reflects the size of the audio data before partitioning, and the usable data the amount of data used in training the acoustic models. As expected, increasing the amount of training data has a large impact when the total amount is small. With only 1 hour of data the word error rate is 33.3%, whereas with 33 hours the word error is 20.7%, a relative error rate reduction of 38%. Successive doubling of the training data reduces the error by less the 10% and 5% relative. It should be noted that initially there is also a large increase in model size, accounting for the improved model accuracy. However, for our system, once a model size of about 11500 tied states with 360k Gaussians is obtained, larger models do not give any significant improvement in the desired operating range of about 10xRT decoding. The word error rate with the News.Com.Cap language model is the same as that of the original Hub4 language model on the eval99 test data using the 1999 acoustic models trained on 123 hours of manually annotated data. Using 4 sets of gender and bandwidth dependent acoustic models reduces the relative word error rate by 5% to 17.1%. All of the remaining experiments in this section were run with the News.Com.Cap language model and with one set of gender and bandwidth independent acoustic models.

In the following, the use of automatically transcribed audio data for acoustic model training is investigated. In order to bootstrap the training procedure, an initial set of acoustic models were trained on 1 hour of manually transcribed data from the LDC 1996 Hub4 corpus (these are the same models as used for the second entry in Table 1). The data consist of three shows: ABC Nightline (a960521), CNN Early Prime (e960510a) and NPR All Things Considered (j960510). These resulting acoustic models have significantly fewer parameters than the standard Hub4 models.[**] The 1 hour of manually transcribed data was only used to bootstrap the process and was not used in building the successive model sets.

Amount of training data			*Average %werr*	
raw	*unfiltered*	*filtered*	*unfiltered*	*filtered*
1.5h	1h manual		33.3	
14h	8h	6h	26.4	25.7
28h	17h	13h	25.2	23.7
58h	28h	21h	24.3	22.5
140h	76h	57h	22.4	21.1
270h	140h	108h	21.0	19.9
558h	238h	188h	20.2	19.4

Table 2: Lightly supervised acoustic model training: Word error rate (%) for increasing quantities of automatically labeled training data on the 1999 evaluation test sets using (1S) gender and bandwidth independent acoustic models with the News.Com.Cap language model.

Recognition results with increasing amounts of automatically transcribed audio data are shown in Table 2 on the 1999 Hub4 evaluation test. The bootstrap models were used to successively transcribe 14, 28, 58, and 140 hours of raw data. For each of these data sets, two sets of acoustic models were built, with and without filtering the transcripts by the closed-captions (see Table 2). [††] Closed-caption filtering consists of using dynamic programming to align the hypothesized transcription with the closed captions on a story-by-story basis, and retaining only regions where the automatic transcripts agree with the closed captions. All words that disagree are considered “incorrect” and the corresponding audio segments are discarded. The filtering is seen to remove about one-third of the audio data (compare the unfiltered and filtered durations). The second model set is training using all of the transcribed data, without trying to filter out recognition errors. The amount of data actually used to train the acoustic models are shown both with and without closed-caption filtering. As a consequence, the unfiltered model sets are about 25% larger in terms of the number of triphone contexts covered and the total number of Gaussians than those built with the filtered data. It should be noted that in both cases the closed-caption story

[**] The first pass models cover only 1737 triphone contexts (893 tied states and 21k Gaussians), and the second and third pass models cover 3416 triphone contexts (899 tied states, 14k and 22k Gaussians, respectively).

[††] The difference in the amounts of raw data transcribed and actually used for training are due to three factors. The first is that the total duration includes non-speech segments which are eliminated prior to recognition during partitioning. Secondly, the story boundaries in the closed captions are used to eliminate irrelevant portions, such as commercials. Thirdly, since there are many remaining silence frames, only a portion of these are retained for training.

boundaries are used to determine the limits of the retained audio segments after automatic transcription (i.e., this information is not used by the recognizer).

Acoustic models trained on the 140 hours of automatically transcribed data were in turn used to transcribe an additional 130 hours of raw data (yielding a total of 270 hours). Acoustic models were trained on this data, and used to process the remaining 288 hours of audio data. In total, 558 hours of raw data were processed with which acoustic models were trained. The first pass models cover about 5000 triphones (5100 tied states, 80k Gaussians), the second pass models cover about 15000 triphone contexts with 10k tied states and 160k Gaussians, and the third pass models cover 25000 triphones sharing 11k states and 360k Gaussians.

Several observations can be made about these results. As expected, when more training data is used, the word error rate decreases. This is true for both the filtered and unfiltered training data. The word error reduction does not seem to saturate as the amount of training data increases, so we can still hope to lower the error rate by continuing the procedure further. Filtering the automatic transcripts with the closed captions reduces the word error by only about 5% relative compared to the error rate obtained by simply training on all the available data. This implies that including the closed captions in the language model training data seems to provide enough supervision to ensure proper convergence of the training procedure. The best word error rate obtained with this procedure is about 10% higher than what can be obtained by training with the 123 hours of detailed annotated transcriptions (19.4% versus 18.0% with 1S models in Table 1). Although part of this difference may be due to the fact that different corpora are used for training, we believe that it is essentially due to differences in transcription quality. These differences can arise from errors in the alignment procedure, word boundary problems, and incorrect labeling of non-speech events such as hesitations and breath noises for which no supervision is available (recall that these are not present in the closed-captions).

6. Impact of the language model training material

The above results with unfiltered data indicate that the language model used in the experiments provided sufficient supervision for the lightly supervised approach to be successful. However the conditions are quite advantageous in that the language model contains a lot of information about the unannotated acoustic training data. Given the current state-of-the-art in broadcast news transcription, the first requirements in developing a system for a different language are acquiring the necessary audio and textual resources. The costs of obtaining these resources evidently depends upon the target language, and only some of the resources may be accessible.

In an effort to understand the contribution of the different text sources and the importance of the epoch of the text data, a series of experiments were carried out in which different combinations of text materials were used to train the language models. The training text sources (see Section 3) include newpapers and newswires, commercially produced summaries and transcripts, and closed-captions. The following combinations were investigated:

- **LMa** (baseline Hub4 LM): newspaper and newswire (News), commercially produced transcripts (Com) predating Jun98, and acoustic transcripts
- **News.Com.Cap:** newspaper and newswire, commercially produced transcripts, and closed-captions (Cap) through May98
- **News.Com:** newspaper and newswire, and commercially produced transcripts through May98
- **News.Cap:** newspaper and newswire and closed-captions through May98
- **News:** newspaper and newswire through May98
- **News.Com97:** newspaper and newswire through May98, commercially produced transcripts through Dec97
- **News.Com97.Cap:** newspaper and newswire and closed-captions through May98, commercially produced transcripts through Dec97
- **News97:** newspaper and newswire through Dec97

It should be noted that with the exception of the last language model (News97), all of the conditions include newspaper and newswire texts from the same epoch as the audio data. These provide an important source of knowledge particularly with respect to the vocabulary items. Conditions which also include the closed captions in the language model training data evidently provide closer supervision in the decoding process.

For each combination of LM training texts, a word list was obtained by including the most frequent words in the training texts. All language models are formed by interpolating individual LMs built on each text source. The interpolation coefficients were chosen in order to minimize the perplexity on the same development described above (see Section 5).

LM Training	*Word Error Rate (%)*
LMa	18.0
News.Com.Cap	18.0
News.Com	18.6
News.Cap	19.1
News	20.6
News.Com97	18.7
News.Com97.Cap	18.1
News97	20.9

Table 3: Supervised acoustic model training: Word error rate (%) on the 1999 evaluation test sets for various language model training conditions using one set of gender-independant acoustic models trained on the HUB4 training data (123 hours) with detailed manual transcriptions.

As a reference, Table 3 compares the value of the different language models listed above using speaker-independent (1S) acoustic models trained on the baseline Hub4 data. The first two entries are a reminder that the original Hub4 language model (LMa) and the light supervision language model (News.Com.Cap) had the same word error (18.0%) with these acoustic models. It can be observed that removing any text source leads to a degradation in recognition performance. It appears it is more important to include commercially produced transcripts (Com), even if they are old (Com97) than the closed captions (Cap). This suggests that the commercial transcripts more accurately represent spoken language than closed-captioning. Even if only newspaper and newswire texts are available, the word error increases by only 14% over the best configuration (News.Com.Cap) and even using older newspaper and newswire texts (News97) does not substantially increase the word error rate.

Amount of raw data	1.5h	14h	28h	58h	140h	270h
News.Com	33.7	27.6	25.7	25.2	22.6	21.4
News.Cap	34.4	27.4	25.6	25.7	22.9	22.0
News	35.9	29.0	28.1	27.4	25.2	23.7
News.Com97	33.9	27.6	25.7	25.1	22.5	21.4
News.Com97.Cap	33.3	26.4	25.0	24.2	21.9	-
News97	36.1	30.6	28.9	27.9	25.2	24.4

Table 4: Lightly supervised acoustic model training training: Word error rate (%) for different language models and increasing quantities of automatically labeled training data on the 1999 evaluation test sets using one set of gender and bandwidth independent acoustic models.

For each language model, the approach used in the previous section was taken, that is the bootstrap acoustic models were used to successively transcribe 14, 28, 58 hours of raw data. Then all of the automatically annotated data was used to build acoustic models, which were in turn used to transcribe the next chunk of data. In these experiments there is no filtering with the closed-captions, only the closed-caption story boundaries are used to delimit the audio segments. Table 4 gives the word error rates for the different language models, with increasing quantities of automatically labeled training data. The first column shows for each language model the word error rate with acoustic models trained on only 1 hour of manually transcribed data. These word error rates range from 33% to 36% across the language models, indicative of the anticipated word error rate for the raw data to be transcribed with the various configurations. With 14 hours (raw) of approximately labeled training data, the word error is reduced by about 20% for all LMs compared with training on 1h of data which has carefully manual transcriptions. As expected the best performance is obtained with a language model which is trained on the most data, but all language models behave in a similar manner. The models including the commercial BN transcripts (News.Com and News.Com97), even if predating the data epoch, are seen to perform slighly better when the commercially produced transcripts are replaced with closed-captions (News.Cap), supporting the earlier observation that the commercial BN transcripts are closer to spoken language. Even if only news texts (News and News97) are available, these provide adequate supervision, with only slightly better results when the texts are from the same period as the audio data. The same relative improvements are observed for all language models, with the largest gain in the early iterations. This may imply that the additional data is less useful, or can be linked to the method used here where the transcription quality is improving from one iteration to the next, but the portions of the data transcribed with the early models have significantly higher word error rates than the latter data. One possible solution is to retranscribe all the data with the best available system.

7. Unsupervised acoustic model training

The preceding experiments indicated that given sufficient language model training texts, the exact source and epoch were not critical for the success of the approach. They also showed that although filtering the hypotheses resulted in slightly better acoustic models, the filtering is also not required (see Table 2). The experiments reported in this section look at drastically reducing the amount of acoustic training data and/or the quantity of language model training texts in order to find the minimal requirements for bootstrapping the procedure. Three conditions are investigated for unsupervised acoustic model training:

- Removing the story boundary filtering.
- Training the acoustic models on only a very small amount of manually annotated audio data, in this case 10 minutes taken from the beginning of a single show (a960521.sph)[‡‡]
- Training the language models on substantially less data from a short time period predating the epoch of the acoustic training data: down to 1.8 M words.

To evaluate how much data is needed to train the bootstrap models, the amount of annotated acoustic training was reduced from 1 hour to 10 minutes. The resulting acoustic models are very small, covering only a few hundred phone contexts, sharing 300 tied states with 4500 Gaussians. For this first experiment the News.Com.Cap language model was used (see Table 5, column News.Com.Cap). The initial word error with this configuration is 53.1%, high enough that we may question whether or not this approach can possibly work. Given this high initial word error we decided to carry out more iterations, processing smaller amounts of data in each chunk. First only 6 shows were processed and used to train acoustic models. Since the difference in performance with and

[‡‡] We also looked at using an entire show for training, but since the initial model performances were about the same as only using the first 10 minutes we decided to use the smaller amount of training data for the remaining experiments.

without story boundary (SBF) filtering is relatively small, this procedure was eliminated in the remaining steps. On each successive iteration the amount of data processed is roughly doubled, with relative error reductions on the order of 10-15%. After the 5th iteration, the word error is 23.4%, which is close to that obtained previously with seed models trained on 1 hour of manually annotated data (22.4%, see line 5 in Table 2). The remaining difference in performance may be due to the removal of the story boundary filtering procedure. This confirms the earlier hypothesis that the language model provides sufficient supervision for the training procedure to converge rapidly.

		WER (%)	
Iteration	*Raw Acoustic training data*	News.Cap.Com	News
bootstrap models	10 min manual	53.1	55.6
1 (6 shows), with SBF	4 h	35.6	-
1 (6 shows)	4 h	37.3	41.9
2 (+12 shows)	12 h	31.7	35.6
3 (+24 shows)	28 h	27.9	31.0
4 (+48 shows)	58 h	26.0	28.7
5 (+118 shows)	140 h	23.4	-
Retranscribe data with 1st iteration models			
retranscribe 90 shows (4)	58 h (2x)	24.9	28.4
retranscribe 90 shows (4)	58 h (3x)	24.4	-

Table 5: Unsupervised acoustic model training: Word error rate (%) on the 1999 evaluation test sets using one set of gender and bandwidth independent acoustic models. The initial acoustic models were trained on only 10 minutes of manually annotated data. The News.Com.Cap language model was trained on newspaper and newswire texts, commercially produced transcripts and closed-captions through May98. The News language model was trained on newspaper and newswire texts through May98. SBF: story boundary filtering applied.

Since the acoustic data is transcribed in chunks, each set of acoustic models is built on data with a range of word errors. To assess the potential improvement that could be obtained by iterating over the same subset of training data, the models from iteration 4 (trained on 90 shows) were used to retranscribe the same data. As seen in the lower part of Table 5 the word error is reduced from 26.0% to 24.9% by reprocessing the same 58 hours of data once and to 24.4% processing the data a third time. The word error rate is quite close to the 24.3% word error rate with light supervision reported in Table 2 with 58 hours of data, and about 15% higher than the word error of 20.7% obtained with 50h of supervised training data (see Table 1). Since the News.Com.Cap language model includes components estimated on closely related, manually produced text data, a similar experiment was carried out using only newspaper and newswire sources. The results given in Table 5 under the column News show a similar trend, but have a 10% higher overall error rate. It can also be noted that retranscribing the

58 hours of data gives a much smaller improvement than that obtained with the News.Com.Cap language model.

This experiment shows that the amount of manually annotated data used to train the bootstrap models is not crucial, so we next questioned the effects of dramatically reducing the language model training data. Recall that the News.Com.Cap models are trained on a billion words of text. In the next experiment, language models were estimated on only 1.8 million words of newspaper and newswire texts from December 26-31, 1997 i.e., predating the audio data. The corresponding lexicon contains only 40k words, including the most frequent words in the text corpus already in our American English master lexicon. For reference, these language models were tested using acoustic models trained on the standard Hub4 training data (200 hours) and on the 1.5-hour and 10-minute training sets. The results are summarized in Table 6, along with the word error rates for the News.Com.Cap language model. The word error with only 10 minutes of data is 65.3%. This condition was chosen as the starting point for further exploration of the unsupervised acoustic model training.

	Raw Acoustic training data		
Language model	*200 hours*	*1.5 hours*	*10 minutes*
News.Com.Cap, 65k	18.0	33.3	53.1
News, 65k	20.9	36.1	55.6
1.8 M words, 40k	28.8	46.9	65.3

Table 6: Supervised acoustic model training: Reference word error rates (%) on the 1999 evaluation test data with varying amounts of manually annotated acoustic training data and a language model trained on 1.8 M words of text.

Raw Acoustic training data		*WER (%)*
bootstrap models	10 min manual	65.3
1 (6 shows)	4 h	54.1
2 (+12 shows)	12 h	47.7
3 (+24 shows)	28 h	43.7
4 (+48 shows)	58 h	41.4
5 (+60 shows)	108 h	39.2
6 (+58 shows)	140 h	37.4

Table 7: Unsupervised acoustic model training: Word error rate (%) on the 1999 evaluation test data with varying amounts of automatically transcribed acoustic training data and a language model trained on 1.8 M words of text.

The acoustic training data was chunked in the same manner as in the preceding experiment, processing exactly the same files in each iteration. The first observation that can be made, is that even using a recognizer with a word error of 65% the procedure is converging properly by training acoustic models on automatically labeled data. This is even more surprising since the only supervision is via a language model trained on a small amount of text data predating the raw acoustic audio data. These conditions are substantially improverised compared to the initial experiments carried out in Section 5. As the amount of automatically transcribed acoustic data is successively doubled, there are consistent reductions in the word error rate. While these error rates are quite a bit higher than reported in the previous section, we may expect that retranscribing the same shows should reduce the word error further, as observed in Table 5. As a reminder, the word error with the Hub4 acoustic models trained on 200 hours of data is 28.8% with this language model, substantially higher than the 18.0% word error obtained with the News.Com.Cap language model.

8. Conclusions

In this work we have investigated the use of low cost data to train acoustic models for broadcast news transcription. We have shown that detailed manual transcriptions are not a requirement for acoustic model training.

We first explored a scheme using approximate transcriptions, such as closed captions to provide light supervision. When closed captions are available, the recognition results obtained with acoustic models trained on a large quantity of automatically annotated data is comparable (under a 10% relative increase in word error) to results with acoustic models trained on a large amount of manually annotated data.

Different levels of supervision provided by the language model training data were investigated and it was found that the procedure converges for all of the tested configurations, and that the differences across language models are relatively small. This implies that the technique can be applied even if closely related texts are not available.

We have further shown that the level of supervision can be considerably reduced, i.e., that the training can be done essentially without manual transcripts (only 10 minutes of data used to construct bootstrap models), and that there is no need to manually locate story boundaries. Finally, we have show that even though the language model is the only source of supervision in the training process, the procedure converges even using a poor language model.

This method requires substantial computation time, but little manual effort. An advantage offered by this approach is that there is no need to extend the pronunciation lexicon to cover all words and word fragments occurring in the training data. By eliminating the need for manual transcription, automated training can be applied to essentially unlimited quantities of task-specific training data. A question that remains unanswered is can better performance be obtained using large amounts of automatically annotated data than with a large, but lesser amount of manually annotated data? and if so, how much data is needed?

Recent work has shown that the same basic idea proposed here can be used as a transparent method for adapting the generic models to a specific task, thus achieving a higher degree of genericity, and to improve acoustic models for portability across tasks (Lefevre et al., 2001). This approach can also reduce the cost of porting to another language. Once some reasonable amount of language model training texts and a pronunciation lexicon are available, bootstrap models can be used to initialize the transcription process of a large quantity of audio data at a low cost.

Acknowledgements

This work was partially financed by the European Commission under the Language Engineering project IST-1999 11876 Coretex.

References

G. Adda, M. Jardino and J.L. Gauvain, "Language Modeling for Broadcast News Transcription," *Proc. ESCA EuroSpeech'99*, Budapest, Hungary, **4**, pp. 1759-1762, September 1999.

C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2), pp. 5-22, January 2001.

C. Cieri, D. Graff, M. Liberman, "The TDT-2 Text and Speech Corpus," *Proc. DARPA Broadcast News Workshop*, Herndon, VA. (see also http://morph.ldc.upenn.edu/TDT).

P. Clarkson, R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge Toolkit," *EuroSpeech'97*, Rhodes, Greece, pp. 2707-2710, September 1997.

J. Garofolo, C. Auzanne, E. Voorhees, W. Fisher, "1999 TREC-8 Spoken Document Retrieval Track Overview and Results," *Proc. 8th Text Retrieval Conference TREC-8*, November 1999.

J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, Chantilly, VA, pp. 56-63, February 1997.

J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**, pp. 1335-1338, Sydney, December 1998.

J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," to appear in *Speech Communication*, 2002.

J.L. Gauvain and L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'2000*, **3**, pp. 794-798, Beijing, October 2000.

J.L. Gauvain, C.H. Lee, “Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains,” *IEEE Trans. on SAP*, **2**(2), pp. 291-298, April 1994.

D. Graff, “The 1996 Broadcast News Speech and Language-Model Corpus,” *Proc. ARPA Speech Recognition Workshop*, Chantilly, VA, pp. 11-14, February 1997.

T. Kemp and A. Waibel, “Unsupervised Training of a Speech Recognizer: Recent Experiments,” *Proc. ESCA Eurospeech'99*, Budapest, Hungary, **6**, pp. 2725-2728, September 1999.

F. Kubala, J. Cohen *et al.*, “The Hub and Spoke Paradigm for CSR Evaluation,” *Proc. ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, pp. 9-14, 1994.

L. Lamel, J.L. Gauvain, G. Adda, “Lightly Supervised Acoustic Model Training,” *Proc. ISCA ITRW ASR2000*, pp. 150-154, Paris, September 2000.

F. Lefevre, J.L. Gauvain, L. Lamel, “Genericity and Adaptability Issues for Task-independent Speech Recognition,” *Proc. ISCA ITRW on Adaptation Methods in Speech Recognition*, Sophia Antipolis, France, August 2001.

C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

D.S. Pallett, J.G. Fiscus, *et al.* “1998 Broadcast News Benchmark Test Results,” *Proc. DARPA Broadcast News Workshop*, pp. 5-12, Herndon, VA, February 1999.

D. Pallett, J. Fiscus, M. Przybocki, “Broadcast News 1999 Test Results,” *Proc. NIST/NSA Speech Transcription Workshop*, College Park, Maryland, May 2000.

A. Waibel, P. Geutner, L. Mayfield Tomokiyo, T. Schultz, M. Woszczcyna, “Multilinguality in Speech and Spoken Language Systems,” *Procceedings of the IEEE*, special issue on Spoken Language Processing, **88**(8), pp. 1297-1313, August 2000.

G. Zavaliagkos and T. Colthurst, “Utilizing Untranscribed Training Data to Improve Performance,” *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, pp. 301-305, February 1998.