Lightly Supervised Acoustic Model Training*

Lori Lamel, Jean-Luc Gauvain, Gilles Adda

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,gauvain,gadda}@limsi.fr

ABSTRACT

Although tremendous progress has been made in speech recognition technology, with the capability of todays state-of-the-art systems to transcribe unrestricted continuous speech from broadcast data, these systems rely on the availability of large amounts of manually transcribed acoustic training data. Obtaining such data is both time-consuming and expensive, requiring trained human annotators with substantial amounts of supervision. In this paper we describe some recent experiments using lightly supervised techniques for acoustic model training in order to reduce the system development cost. The strategy we investigate uses a speech recognizer to transcribe unannotated broadcast news data, and optionally combines the hypothesized transcription with associated, but unaligned closed captions or transcripts to create labeled training. We show that this approach can dramatically reduces the cost of building acoustic models.

1. INTRODUCTION

The last decade has witnessed substantial progress in large vocabulary continuous speech recognition. A number of sites have state-of-the-art systems which can transcribe unrestricted continuous speech from unknown speakers taken from American English television and radio broadcasts with word errors around 20%. However, with today's technology, the adaptation of a recognition system to a new task or another language requires large amounts of transcribed training data. Generating this transcribed data is an expensive process in terms of both manpower and time. There are certain 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 [3].[1]

In this paper we describe recent experiments with lightly supervised acoustic model training. The basic idea is to use a speech recognizer to automatically transcribe unannotated data, thus generating labeled training data. By iteratively increasing the amount of training data, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. We compare the straightforward method of training on all the automatically annotated data with one in which the closed-captions or transcripts are used to filter the hypothesized transcriptions, removing words that are "incorrect". The use of untranscribed data to train acoustic models has been reported recently. BBN describes experiments using completely unsupervised training for conversational speech (Switchboard and Callhome corpora) and reports small improvements by using such data in addition to 3 hours of annotated data, compared to training only on the original 3 hours [13]. Based on their results they conjecture that an order of magnitude more untranscribed data is needed to achieve comparable levels of performance with transcribed data. In [9], 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 authors give little information about the data used to train the language models, and thus it is difficult to assess the level of supervision.

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 Section 5.

2. LIGHTLY SUPERVISED TRAINING

HMM training requires an alignment between the audio signal and the phone models, which usually relies on a perfect orthographic transcription of the speech data and a good phonetic lexicon. In general it is easier to deal with relatively short speech segments so that transcription errors will not propagate and jeopardize the alignment. The orthographic transcription is usually considered as ground truth and training is done in a closely supervised manner. For each speech segment the training algorithm is provided with the exact orthographic transcription of what was spoken, i.e.,

*This work was partially financed by the European Commission under the Language Engineering project LE-5 Coretex.

[1] To avoid confusion, in this paper we group together both of these types of transcripts and refer to them as closed-captions.

the word sequence that the speech recognizer should hypothesize when confronted with the same speech segment.

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:

- Normalize the transcriptions to a common format (some adjustment is always needed as different text sources make use of different conventions).
- Produce a word list from the transcriptions and correct blatant errors (these include typographical errors and inconsistencies).
- Produce a phonemic transcription for all words not in our master lexicon (these are manually verified).
- Align the orthographic transcriptions with the signal using existing models and the pronunciation lexicon (or bootstrap models from another task or language). This procedure often rejects a substantial portion of the data, particularly for long segments.
- Eventually correct transcription errors (or just ignore these if enough audio data is available)
- Run the standard EM training procedure.

This procedure is usually 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, by using an iterative procedure where instead of using manual transcriptions for alignment, at each iteration the most likely word transcription given the current models and all the information available about the audio sample is used. This approach still fits within the EM training framework, which is well-suited for missing data training problems. A completely unsupervised training procedure is to use the current best models to produce an orthographic transcription of the training data, keeping only words that have a high confidence measure. Such an approach, while very enticing, is limited since the only supervision is provided by the confidence measure estimator. This estimator must in turn be trained on development data, which needs to be small to keep the approach interesting.

Between using carefully annotated data such as the detailed transcriptions provided by the LDC and no transcription at all, there is a wide spectrum of possibilities. What is really important is the cost of producing the associated annotations. 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 [2]. Orthographic transcriptions such as closed-captions can be done in a few times real-time, and therefore are quite a bit less costly. These transcriptions have the advantage that they are already available for some television channels, and therefore do not have to be produced specifically for training speech recognizers.

Another approach is to make use of other possible sources of comtemporaneous texts from newpapers, newswires, summaries and the internet. However, since these sources have only an indirect correspondence with the audio data, they provide less supervision.

There are several problems that must be faced when dealing with closed captions instead of speech transcriptions. In addition to providing an exact word-level transcription of what was said, the detailed speech transcriptions 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 are also not a true orthographic transcription of the speech. 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 used for the SDR evaluation to be on the order of 12% [7].

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. Gaussian mixture models for sex and bandwidth identification can be trained on a very small amount of data, so the required labeling is not very costly. 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).

The following training procedure is used in this work:

- Train a language model on all texts and closed captions after normalization
- Partition each show into homogeneous segments and label the acoustic attributes (speaker, gender, bandwidth) [4]
- Train acoustic models on a very small amount of manually annotated data (1h)
- Automatically transcribe a large amount of training data
- Align the closed-captions and the automatic transcriptions (using a standard dynamic programming algorithm)
- Run the standard acoustic model training procedure on the speech segments where the two transcripts are in agreement
- 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 we no longer need to deal with new words and word fragments in the data and we do not need to correct transcription errors. The same basic idea was used to align the automatically generated word transcriptions of the 500 hours of audio broadcasts used in the spoken document retreival task (NIST SDR'99).

3. CORPORA

The unannotated audio data used in these experiments are taken from the DARPA TDT-2 corpus (used in the SDR'99 and SDR'00 evaluations) [3]. The audio corpus used for SDR'99 contains 500 hours of data in (902 shows) from 6 sources: CNN Headline News (550 30-minute shows), ABC World News Tonight (139 30-minute shows), Public Radio International The World (122 1-hour shows), Voice of America VOA Today and World Report (111 1-hour shows). These data were broadcast between January and June 1998. This data comes with associated closed-captions and commercial transcripts. These are divided in about 22k stories with timecodes identifying the beginning and end of each story, and with an average duration of 1min 20secs per story.

The Hub4 acoustic training data (1996 and 1997 releases from the LDC, http://www.ldc.upenn.edu/) contain a total of almost 200 hours of carefully annotated data 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) [8]. 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.

The language model training data are those used for the Hub4 task, with the exception that 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 bought 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 [11].

4. SYSTEM DESCRIPTION

The LIMSI broadcast news transcription system has two main components, the audio partitioner and the 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 procedure [4] 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 [10] 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 audio data from the DARPA Hub4 Broadcast News corpus (the LDC 1996 and 1997 Broadcast News Speech collections) [8]. We used the August 1997 and February 1998 releases of the LDC transcriptions. 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. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wideband and telephone band speech [6]. 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 [5].

Baseline language models were obtained by interpolation of backoff n-gram language models trained on 3 different data sets: BN transcriptions, NAB newspapers and AP Wordstream texts excluding the test epochs, and the transcriptions of the BN acoustic data.

The baseline recognition vocabulary contains 65120 words and 76644 phone transcriptions, and 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. 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 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 [4].

The LIMSI 10x system obtained a word error of 17.1% on the 1999 DARPA/NIST evaluation set (the combined scores in the fourth row in Table 1), and can transcribe unrestricted broadcast data with a word error of about 20% [5]. The word error can be reduced to 15.6% for a system running at 50xRT (last entry in Table 1).

Training	Conditions	*bn99_1*	*bn99_2*	*Average*
1h	1S, LMc	35.2	31.9	33.3
69h	1S, LMc	20.2	18.0	18.9
123h	1S, LMc	19.3	17.1	18.0
123h	4S, LMc	18.5	16.1	17.1
123h	4S, LMa	18.3	16.3	17.1
123h	4S, LMa, 50x	17.1	14.5	15.6

Table 1: Word error rate 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 the last run in column 6. "1S" designates one set of gender-independant acoustic models, whereas "4S" designates four sets of gender and bandwidth dependent acoustic models. The "LMa" language model results from an interpolation of a LM trained on the detailed acoustic transcriptions with one trained on the other text sources excluding the TDT2 closed-captions. "LMc" is trained on the same texts and the TDT2 closed-captions, but no detailed acoustic transcriptions.

5. EXPERIMENTAL RESULTS

In this section we summarize a series of experiments to assess recognition performance as a function of the available acoustic and language model training data. All recognition runs are carried out in under 10xRT unless stated otherwise.

As mentioned above, our usual procedure to build language models for BN data is to interpolate n-gram LMs built on 3 sources of texts[1]: 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 our aim is to investigate the use of acoustic model training data without detailed transcriptions, we built language models for these experiments replacing the detailed transcriptions by the commercial transcriptions (closed captions and radio transcripts) from the TDT2 data. In doing so, a new word list was selected based on the word frequencies in the training data after excluding the detailed transcriptions. Including the TDT2 closed captions in the language model training data provides some supervision in the decoding process when transcribing the TDT2 audio data to produce the reference transcriptions for training purposes. The language models result from an interpolation of individual LMs built on each text source. The language model interpolation coefficients were chosen in order to minimize the perplexity on a development set composed of the second set of the Nov98 evaluation data (3h) and a 2h portion of the TDT2 data from Jun'98 (not included in the LM training data). The resulting interpolation coefficients are 0.45 for the commercial transcript LM, 0.35 for the newspaper LM and 0.20 for the TDT2 closed caption LM. As can be seen in rows 4 and 5 of Table 1, the word error rates with our original language model (LMa) and the new one (LMc) give comparable results on the eval99 test data using our 1999 acoustic models trained on 123 hours of manually annotated data. All the following experiments were run with the LMc language model and with one set of gender and bandwidth independent acoustic models.

In order to bootstrap the training procedure, an initial set of acoustic models were trained on 57 minutes of manually transcribed data from the LDC 1998 Hub4 corpus. The data consist of three shows: ABC Nightline (a960521), CNN Early Prime (e960510a) and NPR All Things Considered (j960510). These acoustic models are quite small compared to our standard Hub4 models. 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 manually transcribed data was only used to bootstrap the process and was not used in building the successive model sets.

These small models were used to first transcribe 208 broadcasts (about 140 hours of data). Two methods were investigated to use the automatically transcribed data for acoustic model training. In the first method, the hypothesized transcriptions were aligned with the closed captions story by story, and only regions where the automatic transcripts agreed with the closed captions were kept for training purposes. After alignment, about 57 hours of speech data were available for training. The second method consists of simply training on all of the aligned data, without trying to filter out recognition errors. In this case about 76 hours of data were available. [2] In both cases the closed-caption story boundaries are used to delimit the audio segments after automatic transcription.

The labeled data was used to train substantially larger acoustic models. These models were then used to transcribe an additional 216 shows. In all, 424 shows were processed (about 287 hours of data), resulting in 140 hours of aligned acoustic data prior to filtering and 108 hours after filtering. With this data models sets close in size to the baseline system were built. The first pass models cover about 5000 triphones (5100 tied states, 80k Gaussians) and the third pass models cover 25000 triphones sharing 11k states and 360k Gaussians.

Several acoustic model sets were trained on subsets of the automatically transcribed data to assess recognition performance as a function of the available data. 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. Recognition results for the two sets of the 1999 Hub4 evaluation test are shown in Table 2. These results can be compared to the first 3 rows of Table 1, which report results using only the detailed manual transcriptions of the training data. 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 based training. The word error reduction does not seem to saturate as the amount of training data increases, so we can still hope to lower the

[2] The difference in the amounts of data transcribed and actually used for training is 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.

Amount of training data		*%werr on bn99_1*		*%werr on bn99_2*		*Average*	
unfiltered	*filtered*	*unfiltered*	*filtered*	*unfiltered*	*filtered*	*unfiltered*	*filtered*
8h	6h	28.9	28.1	24.7	24.0	26.4	25.7
17h	13h	27.6	26.1	23.5	22.1	25.2	23.7
28h	21h	26.5	24.7	22.8	21.0	24.3	22.5
76h	57h	24.4	23.3	21.0	19.6	22.4	21.1
140h	108h	22.8	21.7	19.8	18.7	21.0	19.9

Table 2: Word error rate for increasing quantities of automatically label training data on the 1999 evaluation test sets using (1S) gender and bandwidth independent acoustic models with the language model LMc. All runs were done in less that 10xRT.

error rate by continuing the procedure further. Filtering the automatic transcripts with the closed captions reduces the word error by only 5% relative compared to the error rate obtained by simply training on all the available data. 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 (19.9% versus 18.0% with 1S models). Although part of this difference may be due to the fact that we use different corpora for the training conditions, we believe that this is essentially due to the difference in transcription qualities. These differences can arise from errors in the alignement procedure, word boundary problems, and incorrect labeling of non speech events such as hesitations and breath noises for which no supervision is available.

6. SUMMARY & DISCUSSION

We have investigated the use of low cost data to train acoustic models for broadcast news transcription, with supervision provided by closed captions. We show that recognition results obtained with acoustic models trained on large quantities of automatically annotated data are comparable (under a 10% relative increase in word error) to results with acoustic models trained on large quantities of data with detailed manual annotations. Given the significantly higher cost of detailed manual transcription (substantially more time consuming that producing commercial transcripts, and much more expensive if money is considered because the closed captions and commercial transcripts are produced for other purposes), it is of interest to further explore such methods requiring substantial computation time, but little manual effort. Another 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.

It appears that using the closed captions to provide supervision via the language model is sufficient and that there is only a small advantage in using them to filter the system hypotheses. However, we believe that there should be more effective ways to use the closed captions to improve the models. The procedure is bootstrapped by training acoustic models with 1 hour of manually transcribed data. Since the recognition error rate does not seem to have reached a lower limit, we are continuing the procedure until we process all the TDT-2 data.

REFERENCES

[1] G. Adda, M. Jardino, J.L. Gauvain, "Language Modeling for Broadcast News Transcription," *Proc. ESCA Eurospeech'99*, Budapest, Hungary, **4**, pp. 1759-1760, September 1999.

[2] C. Barras, E. Geoffrois, Z. Wu, Mark Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," to appear in *Speech Communication*.

[3] 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).

[4] 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.

[5] J.L. Gauvain and L. Lamel, "Fast Decoding for Indexation of Broadcast Data," to appear in *Proc. ICSLP'2000*, Beijing, October 2000.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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, 2000.

[13] 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.