INVESTIGATING 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

The last decade has witnessed substantial progress in speech recognition technology, with todays state-of-the-art systems being able to transcribe broadcast audio data with a word error of about 20%. However, acoustic model development for the recognizers requires large corpora of manually transcribed 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 different levels of supervision for acoustic model training in order to reduce the system development cost. The experiments have been carried out using the DARPA TDT-2 corpus (also used in the SDR99 and SDR00 evaluations). Our experiments demonstrate that light supervision is sufficient for acoustic model development, drastically reducing the development cost.

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 cost, both in terms of human effort and financial needs, required to adapt a recognition system to a new task or another language. One of the most often cited costs is that of obtaining the 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 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].

In a recent paper [10] some preliminary experiments with lightly supervised acoustic model training were described, where the basic idea is to use a speech recognizer to automatically transcribe unannotated data, thus generating "approximately" 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. A straightforward approach of training on all the automatically annotated data was compared with one in which the closed-captions are used to filter the hypothesized transcriptions, removing words that are "incorrect". To our surprise, somewhat comparable recognition results were obtained both with and without filtering, suggesting that inclusion of the closed-captions in the language model training material provided sufficient supervision. Although the idea of using untranscribed data to train acoustic models has been proposed before (see [13] and [9]), we are not aware of any other large scale experiments with this technique on a publicly available corpora.

In this paper we investigate the effects of using different levels of supervision, as provided by the language model training texts, on the accuracy of the acoustic models constructed using automatically generated word transcriptions. 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 Section 4.

2. ACOUSTIC MODEL 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. Training acoustic models usually entails carrying out a sequence of operations once the audio data and transcription files have been loaded [10]. First the transcriptions need to be converted to a common format (some adjustment is always needed as different corpora make use of different conventions), and a pronunciation lexicon de-

*This work was partially financed by the European Commission under the Human Language Technologies project Coretex.

rived. The orthographic transcriptions are then aligned with the signal using existing models (bootstrap models from another task or language). This procedure often rejects a substantial portion of the data, particularly for long segments. If enough audio data is available these errors can simply be ignored, but often the principal transcription errors are manually corrected. Once the alignments are available, the standard EM training procedure is carried out. This procedure is usually iterated several times to refine the acoustic models, where in general, each iteration recovers a portion of the rejected data.

One can imagine training acoustic models in a less supervised manner, via an iterative procedure where instead of using manual transcriptions for alignment, at each iteration the most likely word transcription given the current models and any known information about the audio sample is used. This approach still fits within the EM training framework, which is well-suited for missing data training problems. Compared with commonly used training procedures [10], the manual work is considerably reduced, both in generating the annotated corpus and 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.

3. SYSTEM DESCRIPTION

The LIMSI broadcast news transcription system has two main components, the audio partitioner and the word recognizer. Data partitioning [4] serves to divide the continuous audio stream into homegenous segments, associating appropriate labels for cluster, gender and bandwidth with the segments. 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 for cluster-based acoustic model adaptation using the MLLR technique [11] prior to word graph generation. A 3-gram LM is used in the first two decoding steps. The final hypotheses are generated with a 4-gram LM and acoustic models adapted with the hypotheses of step 2.

In the 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]. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wideband and telephone band speech [6]. The models contain 28000 position-dependent, cross-word triphone models with 11700 tied states and approximately 360k Gaussians [5].

The baseline language models are obtained by interpolation of models trained on 3 different data sets (excluding the test epochs): about 790M words of newspaper and newswire texts; 240M word of commercial broadcast news transcripts; and the transcriptions of the Hub4 acoustic data. The recognition vocabulary contains 65120 words.

4. EXPERIMENTAL RESULTS

In this section a series of experiments assessing recognition performance as a function of the available acoustic and language model training data are summarized. All recognition runs were carried out in under 10xRT unless stated otherwise. In particular we investigate the accuracy of the acoustic models obtained after recognizing the audio data using different levels of supervision via the language model. With the exception of the baseline Hub4 language models, none of the language models include a component estimated on the transcriptions of the Hub4 acoustic training data. The language model training texts come from contemporaneous sources such as newpapers and newswires, and commercial summaries and transcripts, and closed-captions. The former sources have only an indirect correspondence with the audio data and provide less supervision than the closed captions.

For each set of LM training texts, a new word list was selected based on the word frequencies in the training data. 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 a development set composed of the second set of the Nov98 evaluation data (3h) and a 2h portion of the TDT2 data from Jun98 (not included in the LM training data). The following combinations were investigated:

- **LMa** (baseline Hub4 LM): newspaper+newswire (NEWS), commercial transcripts (COM) predating Jun98, acoustic transcripts
- **LMn+t+c:** NEWS, COM, closed-captions through May98
- **LMn+t:** NEWS, COM through May98
- **LMn+c:** NEWS, closed-captions through May98
- **LMn:** NEWS through May98
- **LMn+to:** NEWS through May98, COM through Dec97

It should be noted that 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 include the closed captions in the LM training data provide additional supervision in the decoding process when transcribing audio data from the same epoch.

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 [12]. The LIMSI 10x system obtained a word error of 17.1% on the evaluation set (the combined scores in the penultimate row in Table 1 4S, LMa) [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, LMn+t+c	35.2	31.9	33.3
69h	1S, LMn+t+c	20.2	18.0	18.9
123h	1S, LMn+t+c	19.3	17.1	18.0
123h	1S, LMn+t	19.8	17.7	18.6
123h	1S, LMn+c	20.7	17.9	19.1
123h	1S, LMn	22.4	19.4	20.6
123h	1S, LMn+to	19.8	18.0	18.7
123h	4S, LMn+t+c	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 row. "1S" designates one set of gender-independant acoustic models, whereas "4S" designates four sets of gender and bandwidth dependent acoustic models.

Amount of training data			*%werr*	
raw	*unfiltered*	*filtered*	*unfiltered*	*filtered*
14h	8h	6h	26.4	25.7
28h	7h	13h	25.2	23.7
58h	28h	21h	24.3	22.5
140	76h	57h	22.4	21.1
287	140h	108h	21.0	19.9
503	238h	188h	20.2	19.4

Table 2: 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 language model LMn+t+c. All runs were done in less that 10xRT.

As can be seen in Table 1, the word error rates with our original Hub4 language model (LMa) and the one without the transcriptions of the acoustic data (LMn+t+c) give comparable results using the 1999 acoustic models trained on 123 hours of manually annotated data (123h, 4S). The quality of the different language models listed above are compared in Table 1 using speaker-independent (1S) acoustic models trained on the same Hub4 data. As can be observed, removing any text source leads to a degradation in recognition performance. It appears it is more important to include commercial transcripts (LMn+t), even if they are old (LMn+to) than the closed captions (LMn+c). 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 (LMn+t+c).

This basic idea is used to align the automatically generated word transcriptions of the 500 hours of audio broadcasts used in the spoken document retrieval task (NIST SDR99) [3]. The audio corpus is comprised of 902 shows from difference sources (CNN, ABC, PRI, VOA), broadcast between January and June 1998.

The lightly supervised training procedure is as follows. In order to bootstrap the training procedure, an initial set of acoustic models were trained on 57 minutes (3 shows) of manually transcribed data from the LDC 1998 Hub4 corpus. These acoustic models have significantly fewer parameters than the standard Hub4 models. The manually transcribed data was only used to bootstrap the process and was not used in building the successive model sets. The recognition performance using the bootstrap models is given in the first entry of Table 1.

These small models were used to transcribe 208 broadcasts (about 140 hours of data). Table 2 compares two methods investigated in [10] 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. The second method consists of simply training on all of the aligned data, without trying to filter out recognition errors.[1] In both cases the closed-caption story boundaries are used to delimit the audio segments after automatic transcription.

The automatically labeled data was used to train substantially larger acoustic models, which in turn were used to transcribe an additional 216 shows. In all, 902 shows were processed (about 500 hours of data), resulting in about 200 hours of aligned acoustic data. With this data models sets close in size to the baseline system were built.

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 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

[1] 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*				
raw	*unfiltered*	*LMn+t+c*	*LMn+t*	*LMn+c*	*LMn*	*LMn+to*
14h	8h	26.4	27.6	27.4	29.0	27.6
28h	7h	25.2	25.7	25.6	28.1	25.7
58h	28h	24.3	25.2	25.7	27.4	25.1

Table 3: Word error rate for different language models and increasing quantities of automatically labeled training data on the 1999 evaluation test sets using (1S) gender and bandwidth independent acoustic models.

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% filtered/20.2% unfiltered 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.

In Table 3 word error rates are given for the different language models, with increasing quantities of automatically labeled training data using gender and bandwidth independent acoustic models. Performance is seen to improve with increasing amounts of training data, with the best LM trained on all text sources. The commercial transcripts (LMn+t and LMn+to), even if predating the data epoch, are seen to be more important than the closed-captions (LMn+c), supporting the earlier observation that they are closer to spoken language. Even if only news texts from the same period (LMn) are available, these provide adequate supervision for lightly supervised acoustic model training.

5. 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 than producing commercial transcripts, and more expensive since closed captions and commercial transcripts are produced for other purposes), such approaches are very promising as they require 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.

One possible way to improve this method is to take advantage of a priori knowledge of the broadcast show type. Although the amount of available data will be reduced, training models only on a subset of the shows is likely to better match the unannotated data and thus result in a better approximate transcription.

REFERENCES

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

[2] 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, Jan. 2001.

[3] C. Cieri, D. Graff, M. Liberman, "The TDT-2 Text and Speech Corpus," *DARPA Broadcast News Workshop*, Herndon. (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," *ARPA Speech Recognition Workshop*, Chantilly, VA, pp. 56-63, Feb. 1997.

[5] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *ICSLP'2000*, **3**, pp. 794-798, Beijing, Oct. 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," *8th Text Retrieval Conference TREC-8*, Nov. 1999.

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

[9] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *ESCA Eurospeech'99*, Budapest, **6**, pp. 2725-2728, Sept. 1999.

[10] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised Acoustic Model Training," *ISCA ITRW ASR2000*, pp. 150-155, Paris, 18-20 Sept. 2000.

[11] 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.

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

[13] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, pp. 301-305, Feb. 1998.