UNSUPERVISED ACOUSTIC MODEL TRAINING[*]

Lori Lamel, Jean-Luc Gauvain and Gilles Adda

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

ABSTRACT

This paper describes some recent experiments using unsupervised techniques for acoustic model training in order to reduce the system development cost. The approach uses a speech recognizer to transcribe unannotated raw broadcast news data. The hypothesized transcription is used 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 unsupervised training is a viable training scheme and can dramatically reduce the cost of building acoustic models.

1. INTRODUCTION

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. Detailed annotation requires on the order of 20-40 times real-time, and even after manual verification the final transcriptions are not exempt from errors [1]. 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. While for some of the main American television channels closed-captions are manually produced, these are not available for most other languages. 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 some recent experiments with unsupervised acoustic model training. The basic idea is to use a speech recognizer (with bootstrap models trained on a very small corpus) to automatically transcribe raw audio data, generating approximate transcriptions for the training data. We extend the work reported in [7] in which closely associated text data where used to provide indirect or "light" supervision for acoustic model training. These experiments demonstrated that comparable acoustic models could be estimated on automatically annotated data with and without the use of closed-caption filtering to remove potentially incorrect words in the hypothesized transcriptions. In [8] the effects of using different levels of supervision, via selection of the language model training texts was assessed, and it was shown that the exact source and epoch of the texts were not critical for the success of the approach.

Other work has been reported using untranscribed data to train acoustic models. The experiments reported in [11] with unsupervised acoustic training 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 [6] reported significant word error reductions using untranscribed data for German broadcast news transcription from one source. They showed 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 describes the American English corpus used in this work, followed by a short overview of the LIMSI broadcast news transcription system. Section 4 explores unsupervised acoustic model training. Except for the initial bootstrap models estimated on 10 minutes of manually transcribed data, all acoustic model training is unsupervised. No manual transcriptions of the Hub4 acoustic training were used for language model estimation, and all recognition runs are carried out in under 10xRT.

2. AMERICAN ENGLISH CORPUS

The unannotated audio data used in these experiments are taken from the DARPA TDT-2 corpus [2]. The corpus used in this work consists of over 430 hours of data from 6 sources: CNN (440 30-minute shows), ABC (109 30-minute shows), PRI (82 1-hour shows), VOA (75 1-hour shows). This 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 and June 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 texts are from 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. The text data consist of about 790M words of newspaper and newswire texts (Jan94-May98) from the Hub4 and TDT corpora distributed by LDC; 240M words of commercial broadcast news transcripts distributed by the LDC (years 92-95) and directly from PSMedia

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

(years 96-97); and the texts (prior to Jun98) from the TDT-2 corpus.

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 knowledge coming from the manual annotations 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.

The 1999 Hub4 evaluation data, comprised of two 90 minute data sets (from June98 and Aug-Sep98) selected by NIST are used for testing [10].

3. 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 homogeneous segments, associating appropriate labels with the segments. The segmentation and labeling process [3] first detects and rejects non-speech segments, and then applies an iterative maximum likelihood segmentation/clustering procedure on 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. The acoustic feature vector has 39-components comprised of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives [4]. 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 [9] 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. 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].

The LIMSI 10x system can transcribe unrestricted broadcast data with a word error of about 20% [5].[1] For reference, the word error on the 1999 evaluation test data is given in Table 1 for different amounts of manually annotated training data using one set of gender-independent acoustic models and a language model News.Com.Cap resulting from the interpolation of individual language models trained on newspaper and newswires (News), commercially produced transcripts (Com), and closed-captions (Cap) through May98.

The interpolation coefficients 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		*Language Model*
Raw	*Usable*	News.Com.Cap
10min	10min	53.1
1.5h	1h	33.3
50h	33h	20.7
104h	67h	19.1
200h	123h	18.0

Table 1: Supervised acoustic model training: Word error rate (%) on the 1999 evaluation test data for various conditions using one set of gender-independent acoustic models trained on subsets of the HUB4 training data with detailed manual transcriptions. The language model is trained on the available text sources, without any detailed transcriptions of the acoustic training data.

4. UNSUPERVISED TRAINING

Our earlier work [8] indicated that given sufficient language model training texts, the exact source and epoch were not critical for the success of the approach. The experiments also showed that although aligning the hypotheses with the closed-captions and keeping only the speech portions where the words agreed resulted in slightly better acoustic models, but the difference is not large. Therefore the step of closed-caption filtering is not required [7]. The experiments reported here 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 that was used in [7].
- 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)[2]
- 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 2, 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. In our previous experiments the story boundary time codes provided with the TDT-2 data, which serve mainly to remove advertisements were used to delimit the speech data. In Table 2 the word error rates with and without story boundary filtering are compared for the first iteration. Since the difference in performance with and without story boundary (SBF) filtering is relatively small, filtering was not used 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

[1] Using 4 sets of gender and bandwidth dependent models, the word error reported in the 1999 DARPA/NIST evaluation was 17.1% with the 10x systems and 15.6% for a system running at 50xRT.

[2] 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.

		WER (%) / (relative reduction)	
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 / (-33%)	-
1 (6 shows)	4 h	37.3 / (-30%)	41.9 / (-25%)
2 (+12 shows)	12 h	31.7 / (-15%)	35.6 / (-15%)
3 (+23 shows)	27 h	27.9 / (-12%)	31.0 / (-13%)
4 (+44 shows)	53 h	26.0 / (-8%)	28.7 / (-7%)
5 (+118 shows)	135 h	23.4 / (-10%)	-
Retranscribe data with 1st iteration models			
retranscribe 85 shows (4)	53 h (2x)	24.9	28.4
retranscribe 85 shows (4)	53 h (3x)	24.4	-

Table 2: 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.

	Raw Acoustic training data		
Language model	*200 hours*	*1.5 hours*	*10 min*
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 3: 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 news texts from 1997.

word error is 23.4%, which is close to the 22.4% reported in [8] using seed models trained on 1 hour of manually annotated data. The remaining difference in performance may be due to the removal of the story boundary filtering procedure. This confirms the earlier assertion that the language model provides sufficient supervision for the training procedure to converge rapidly.

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 85 shows) were used to retranscribe the same data. As seen in the lower part of Table 2 the word error is reduced from 26.0% to 24.9% by reprocessing the same 53 hours of data once and to 24.4% processing the data a third time. This word error rate is quite close to that reported in [8] with 53 hours of raw data and using seed models trained on 1 hour of manually annotated data. From these results we can conclude that the amount of manually annotated data used to train the bootstrap models is not crucial.

Since the goal is to avoid manual transcription, can be noted that the word error of 24.4% obtained by transcribing 53 hours of data three times is only about 15% higher the word error of 20.7% obtained with 50h of supervised training data.

Since the News.Com.Cap language model includes components estimated on closely related, manually produced transcriptions of audio data, a similar experiment was carried out using only newspaper and newswire sources. This condition is more realistic than the preceding one in that a newspaper and newswire texts can be located on the Internet for a number of languages, whereas commercially produced transcriptions are harder to obtain. The results given in Table 2 under the column News show a similar trend, but have a 10% higher overall error rate. The relative word error reductions are seen to be roughly the same for the two conditions. It can also be noted that retranscribing the 53 hours of data gives a much smaller improvement than that obtained with the News.Com.Cap language model. This may be because the reiteration with the News.Com.Cap LM was actually carried out in two steps: in the first step the 41 shows in iterations 1-3 were transcribed, and these new models, which had a relative gain of 6% compared to iteration 4 were used to transcribe the remaining 44 shows. For the News LM, the reiteration was done in one step: the 4th iteration models were used to retranscribe all 85 shows. This suggests that the two step approach is more effective than a single step one.

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 (+23 shows)	27 h	43.7
4 (+44 shows)	53 h	41.4
5 (+60 shows)	103 h	39.2
6 (+58 shows)	135 h	37.4

Table 4: 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 news texts from 1997.

We now question 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 3, 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 the unsupervised acoustic model training.

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 lan-

Acoustic models	Training data	*Word Error*
bootstrap models	3.5 h manual	42.6%
unsupervised	30.0 h automatic	39.1%

Table 5: Supervised versus unsupervised acoustic model training for Portuguese.

guage model trained on a small amount of text data predating the raw acoustic audio data. These conditions are substantially impoverished compared to the previous experiments reported in [7, 8]. 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 (underway) the same shows should reduce the word error further, as observed in Table 2. 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 (see Table 3).

5. LANGUAGE PORTABILITY

A preliminary set of experiments were carried out applying this approach to another language. We make use of a corpus of Portuguese broadcast news data for which substantially less manually transcribed data are available. RTP and INESC, partners in the Alert project (http:alert.uni-duisburg.de) provided 5 hours of manually annotated data from 11 different news programs. Two of the programs (82 minutes) were reserved for testing purposes (JornalTarde_20_04_00 and 24Horas_19_07_00). The remaining 3.5 hours of data were used for acoustic model training. The language model texts were obtained from the following sources: the Portuguese Newswire Text Corpus distributed by LDC (23M words from 1994-1998); Correio da Manha (1.6M words), Expresso (1.9M words from 2000-2001), and Jornal de Noticias (46M words, from 1996-2001), The recognition lexicon contains 64488 words. The pronunciations are generated by grapheme-to-phoneme rules, and use 39 phones.

Initial acoustic model sets trained on the 3.5 hours of available data were used to transcribe 30 hours of Portuguese TV broadcasts. Acoustic models were built on the automatically transcribed data. The recognition results given in Table 5 for the two conditions show that better results are obtained with 30 hours of automatic transcripts (39.1%) than with 3.5 hours of manual transcripts (42.6%).

This preliminary experiment, which was carried out in a different manner than the experiments reported in the previous section, supports the feasibility of lightly supervised and unsupervised acoustic model training.

6. 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, and that the training can be done essentially without manual transcripts (only 10 minutes of data used to construct bootstrap models). Although the language model is the only source of supervision in the training process, the procedure converges even using a poor language model. There is no need to remove advertisements by manually locating story boundaries, this is successfully done by the partitioner.

This method requires substantial computation time,[3] 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. Recent experiments on Portuguese suggest that the unsupervised technique can be use to reduce the development costs in porting to another language. A question that remains unanswered is can better performance be obtained using a very large amount of automatically annotated data than the performance obtained by a state-of-the-art broadcast news transcription system trained on a substantial amount (200 hours) of manually annotated data? and if so, how much data is needed?

REFERENCES

[1] 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.

[2] 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).

[3] 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.

[4] J.L. Gauvain, L. Lamel, G. Adda, “The LIMSI Broadcast News Transcription System,” to appear in *Speech Communication*, 2002.

[5] J.L. Gauvain and L. Lamel, “Fast Decoding for Indexation of Broadcast Data,” *Proc. ICSLP’2000*, **3**, pp. 794-798, Beijing, October 2000.

[6] 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.

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

[8] L. Lamel, J.L. Gauvain, G. Adda, “Investigating Lightly Supervised Acoustic Model Training,” *Proc. ICASSP-01*, Salt Lake City, May 2001.

[9] 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.

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

[11] 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.

[3] Evaluating one condition (combination of acoustic seed models and language model) requires 10 hours of processing time for each hour of raw audio data. Processing 135 hours of data therefore requires 8 weeks of computation on a single CPU. The 2 days of CPU time to train and test the models is negligible compared to this decoding time.