RECENT ADVANCES IN TRANSCRIBING TELEVISION AND RADIO BROADCASTS

Jean-Luc Gauvain, Lori Lamel, Gilles Adda, Michèle Jardino

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

ABSTRACT

Transcription of broadcast news shows (radio and television) is a major step in developing automatic tools for indexation and retrieval of the vast amounts of information generated on a daily basis. Broadcast shows are challenging to transcribe as they consist of a continuous data stream with segments of different linguistic and acoustic natures. Transcribing such data requires addressing two main problems: those related to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Prior to word transcription, the data is partitioned into homogeneous acoustic segments. Non-speech segments are identified and rejected, and the speech segments are clustered and labeled according to bandwidth and gender. The speaker-independent large vocabulary, continuous speech recognizer makes use of n-gram statistics for language modeling and of continuous density HMMs with Gaussian mixtures for acoustic modeling. The LIMSI system has consistently obtained top-level performance in DARPA evaluations, with an overall word transcription error on the Nov98 evaluation test data of 13.6%. The average word error on unrestricted American English broadcast news data is under 20%.

INTRODUCTION

In this paper we report on recent progress in transcribing the television and radio broadcast news, and describe the LIMSI system used in the Nov98 ARPA benchmark test. This system is an extension of our Nov97 Hub4E system [5], using maximum likelihood partitioning and a 3-step decoding approach with acoustic model adaptation.

Radio and television broadcasts contain signal segments of various linguistic and acoustic natures, with abrupt or gradual transitions between segments. Data partitioning serves to divide the acoustic signal into homegenous segments, and to associate appropriate labels with the segments. The segmentation and labeling procedure[4] first detects and rejects non-speech segments using GMMs. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

The speech recognizer uses context-dependent (word-independent but position-dependent) triphone-based phone models. Each phone model is a tied state left-to-right CD-HMMs with Gaussian mixtures, where the state tying is 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, each with two passes. The initial hypothesis are used for cluster-based acoustic model adaptation using the MLLR technique. The final hypothesis is generated using a 4-gram interpolated with a category trigram model with 270 automatically generated word classes [8].

Our development work aimed at improving the partitioning algorithm[5, 6] and improving the acoustic and language models. The main differences from our Nov97 system are the use of additional acoustic and language model training data, the use of divisive decision tree clustering instead of agglomerative clustering for state-tying, the generation of word graphs using adapted acoustic models as well as acoustic model adaptation prior to successive decoding passes, the use of interpolated LMs trained on different data sets instead of training a single model on weighted texts, and a 4-gram LM interpolated with a category model.

DATA PARTITIONING

While it is evidently possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-foward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities. Prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. By using acoustic models trained on particular acoustic conditions, overall performance can be significantly improved, particularly when cluster-based adaptation is performed. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), reduces the computation time and simplifies decoding.

The segmentation and labeling procedure introduced in [5] is as follows. First, the non-speech segments are detected and rejected using Gaussian mixture models (GMMs). The GMMs, each with 64 Gaussians, serve to detect speech, pure-music and other (background). The acoustic feature vector used for segmentation contains 38 parameters. It is the same as the recognition feature vector except that it does not include the energy, although the delta energy parameters are included. Each GMM was trained on about 1h of acoustic data, extracted from the training data after segmentation with the transcriptions. The speech model was trained on data of all types, with the exception of pure music segments and silence portions of segments transcribed as speech over music. In order to detect speech in noisy conditions, a second speech GMM was trained on only data labeled as speech in noise.The music model was trained only on portions of the data that were labeled as pure music, so as to avoid mistakenly detecting speech over music segments. The background model was trained on the segments labeled as silence during forced alignment, excluding silences in segments labeled as speech in the presence of background music. All test segments labeled as music or silence are rejected prior to further processing.

A maximum likelihood segmentation/clustering iterative procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors corresponding to a show $(x_1, \ldots, x_T)$, the goal is to find the number of sources of homogeneous data (each modeled by a p.d.f. $f(\cdot|\lambda_k)$ with a known number of parameters) and the places of source changes. The result of the procedure is a sequence of non-overlaping segments $(s_1, \ldots, s_N)$ with their associated segment cluster labels $(c_1, \ldots, c_N)$, where $c_i \in [1, K]$ and $K \leq N$. Each segment cluster is assumed to represent one speaker in a particular acoustic environment.

The process is initialized using a simple segmentation algorithm based on the detection of spectral change (similar to the first step used in the CMU'96 system[10]). The procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge, and the segment boundary penalty. When no more merges are possible, the segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary penalty, within a 1s interval. This is done to locate the segment boundaries at silence portions, attempting to avoid cutting words (but sometimes this still occurs).

Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to locate telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

RECOGNIZER OVERVIEW

Acoustic Modeling

The acoustic models were trained on all the available transcribed task-specific training data, amounting to about 150 hours of audio data. This data was used to train the Gaussian mixture models needed for segmentation and the acoustic models for use in word recognition. The acoustic analysis derives cepstral parameters from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms[4]. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization.

Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wideband and telephone band speech[3]. For computational reasons, a smaller set of acoustic models is used in the bigram pass to generate a word graph. These position-dependent, cross-word triphone models cover 5416 contexts, with 11500 tied states and 32 Gaussians per state. For trigram decoding a larger set of 27506 position-dependent, cross-word triphone models with 11500 tied states was used. Acoustic model development aimed to minimize the word error rate on the eval96 test data.

State-tying was done via divisive decision tree clustering using a set of 184 questions about the phone position, the distinctive features (and identities) of the phone and the neighboring phones. This is particularly interesting when there are a very large number of states to cluster since it is at the same time both faster and more robust than a bottom-up greedy algorithm.

Language modeling

All language models used in the different steps were obtained by interpolation of backoff n-gram language models trained on different text sets: 203M words of BN transcriptions, 343M words of NAB newspaper texts and AP Wordstream; and 1.6M words corresponding to the transcriptions of the BN acoustic training data. The interpolation coefficients of the LMs were chosen in order to minimize the perplexity on the Nov96 and Nov97 evaluation test sets. A backoff 4-gram LM was derived from this interpolation by merging the LM components[11]. Interpolating LMs trained on the different data sets resulted in lower perplexities than training a single model on all the texts (weighted) as we have done in the past[5]. This is a better approach, both cleaner and more accurate. The perplexity of the eval97 test set with an interpolated 4-gram LM is 162.0, compared with 179.5 with a 4-gram trained on empirically weighted data. The resulting 4-gram LM was interpolated with a 3-gram class based language model, with 270 automatically determined word classes[8]. The classification procedure uses a Monte-Carlo algorithm to minimize the conditional relative entropy between a word-based bigram distribution and a class-based bigram distribution. Bigram and trigram LMs were built in a similar manner for use in the first two decoding steps.

The broadcast news training texts were cleaned in order to be homogeneous with the previous texts, and filler words such as UH and UHM, were mapped to a unique form. As was done in previous years, the training text were processed to add a proportion of breath markers (4%), and of filler words (0.5%)[4]; some frequent word sequences were mapped to compound words, and the 1000 most frequent acronyms in the training texts treated as words instead of as sequences of independent letters.

Lexical Modeling

The recognition vocabulary contains 65,122 words and 72,788 phone transcriptions, and is comprised of all words occuring at least 15 times in the broadcast news texts (63,954 words) or at least twice in the acoustic training

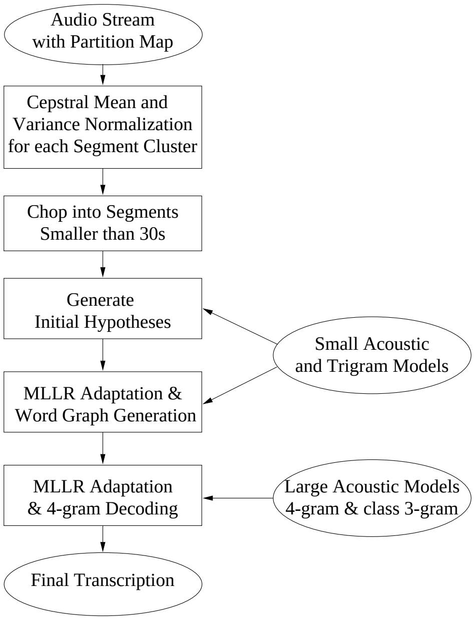

Figure 1: Word decoding.

data (23,234 words). The lexical coverage is 99.14% and 99.53% on the eval96 and eval97 test sets respectively, and 99.73% on the eval98 test data.

Pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). The filler and breath phones were added to model these events, which are relatively frequent in the broadcast data and are not used in transcribing other lexical entries. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. As done in previous years, the lexicon contains compound words for about 300 frequent word sequences, as well as word entries for common acronyms. This provides an easy way to allow for reduced pronunciations[4].

Word Decoding

The word decoding procedure is shown in Figure 1. Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required for the trigram and 4-gram decoding passes[4]. To do so a bimodal distribution is estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames of the segment. This distribution is used to determine locations which are likely to correspond to pauses, thus being reasonable places to cut the segment. Cuts are made at the most probable pause 15s to 30s from the previous cut. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation, each with two passes.

Initial hypothesis generation (fast decoding), is carried out in two passes. The first pass of this step generates a word graph using a small bigram backoff language model and gender-specific sets of 5416 position-dependent triphones. This is followed by a second decoding pass with a larger set of acoustic models (27506 triphones) and a trigram language model (8M trigrams and 15M bigrams) to generate initial hypotheses which are used for cluster-based acoustic model adaptation. Band-limited acoustic models are used for the telephone speech segments.

	Test set (Word Error)		
System Step	*Eval96*	*Eval97*	*Eval98*
Step1 3gram	24.7/25.3	18.2/18.4	18.0/18.3
Step2 3gram	20.2/21.0	14.2/14.6	13.5/14.2

Table 1: Word error with manual/automatic segmentations using the Nov98 system.

The second step generates accurate word graphs. Unsupervised acoustic model adaptation (both means and variances) is performed for each segment cluster using the MLLR technique[9]. The mean vectors are adaptated using a single block-diagonal regression matrix, and a diagonal matrix is used to adapt the variances. Each segment is decoded first with a bigram language model and an adapted version of small set of acoustic models, and then with a trigram language model (8M bigrams and 17M trigrams) and adapted versions of the larger acoustic model set.

The final hypothesis is generated using a 4-gram interpolated with a category trigram model with 270 automatically generated word classes[8]. The first pass of this step uses the large set of acoustic models adapted with the hypothesis from Step 2, and a 4-gram language model. This hypothesis is used to adapt the acoustic models prior to the final decoding step with the interpolated category trigram model.

EXPERIMENTAL RESULTS

This section provides experimental results on the segmentation and partitioning process, compares recognition performance with manual and automatic partitioning, and summarizes the recognizer performance after each step on three test sets. All of our system development was carried out using the eval96 data.

The frame level segmentation error (as used in [7]) was evaluated on the eval96 test data using the manual segmentation provided in the reference transcriptions [6]. The average speech/non-speech segmentation frame error rate on the 4 half-hour shows was 3.7%, and the gender label frame error was 1%. The first show had a significantly higher frame error rate of 7.9% due to deletion of a long, very noisy speech segment.

In general more clusters are found than true speakers in the show, as a cluster can represent a speaker in a given acoustic environment. We looked at two measures of cluster homogeneity: the cluster purity, defined as the percentage of frames in the given cluster associated with the most represented speaker in the cluster (A similar measure was proposed in [2], but at the segment level.); and the "best cluster" coverage which is a measure of the dispersion of a given speaker's data across clusters. The average cluster purity for eval96 test data was 96%. Impure clusters tend to merge data with similar acoustic conditions. The best cluster coverage was obtained by averaging the percentage of data for each speaker in the cluster which has most of his/her data. On average 80% of the speaker's data goes to the same cluster. In fact, this average value is a bit misleading as there is a large variance in the best cluster coverage across speakers. For most speakers the cluster coverage is close to 100%, i.e., a single cluster covers essentially

	Test set (Word Error)		
System Step	*Eval96*	*Eval97*	*Eval98*
Step1 3-gram	25.3	18.4	18.3
Step2 3-gram	21.0	14.6	14.2
Step3 4-gram	20.2	14.3	13.7
4-gram class	19.8	13.9	13.6

Table 2: Word error rates after each decoding step with the Nov98 system.

	Test set (Word Error)		
System	*Eval96*	*Eval97*	*Eval98*
Nov96 system	**27.1***		
Nov97 system	25.3	**18.3**	
Nov98 system	19.8	13.9	**13.6**

Table 3: Summary of BN transcription word error rates. Official results shown in bold. *Nov96 system used a manual partition.

all frames of their data. However, for a few speakers (for whom there is a lot of data), the data is split into two or more clusters containing comparable amounts of data.

Table 1 compares the word recognition performance with automatic and manual (NIST) partitions on three evaluation data sets. The performance loss is about 1.5% relative after the first decoding step (ie. no adaptation). It is higher (2.4%) on the eval96 data due to the same deleted segment in show 1. After adaptation (step 2) the relative performance loss is about 4%, indicating that the clustering process is inappropriately merging or splitting some of the speakers' data. It appears that clustering errors are more detrimental to performance than segmentation ones.

Word error rates for the Nov98 system after each decoding step are given in Table 2. The first decoding step (used to generate the initial hypothesis) has a word error of about 25% on the eval96 data, and 18% on the eval97 and eval98 sets. A word error reduction of about 20% is obtained in the second decoding step which uses the adapted acoustic models. Relatively small gains are obtained in the 4-gram decoding passes, even though these also include an extra acoustic model adaptation pass.

Transcription results on the eval test sets from the last three years are reported in Table 3. The results shown in bold are the official NIST scores obtained by the different LIMSI systems. Only the Nov96 system used a manual partition. In Nov97 our main development effort was devoted to moving from a partitioned evaluation to the unpartitioned one. The Nov97 system did not use focus-condition specific acoustic models as had been used in the Nov96 system[4]. This system nevertheless achieved a relative performance improvement of 6% on the eval96 test data. The Nov98 system has more accurate acoustic and language models, and achieves a relative word error reduction of over 20% compared to the Nov97 system.

SUMMARY & DISCUSSION

This paper has reported on recent advances in transcribing radio and television news broadcasts. Most of the work was carried out in preparation for the Nov98 DARPA evaluation. A main contribution to the improved recognition performance is the generation of more accurate word graphs with adapted acoustic models (based on an initial hypothesis obtained in a fast decoding pass). This step is essential for obtaining graphs with low word error rates. Unsupervised HMM adaptation is performed prior to each decoding pass using the hypothesized transcription of the previous pass. This strategy leads to a significant reduction in word error rate. More accurate language models are obtained by interpolation of LMs trained on different data sets rather than training a single model on weighted texts. More training data has been used for both acoustic and language modeling. Concerning the acoustic models, state-tying uses divisive decision tree clustering instead of agglomerative clustering. This is particularly interesting when there are a very large number of states to cluster. All these improvements have led to a performance gain of over 20% compared to our Nov97 system. The overall word transcription error on the DARPA Nov98 unpartitioned evaluation test data (3 hours) was 13.6%. Although substantial performance improvements have been obtained, there is still plenty of room for improvement of the underlying speech recognition technology. On unrestricted broadcast news shows, such as the 1996 dev and eval data, the word error rate is still about 20%.

REFERENCES

[1] G. Adda et al., "Text Normalization and Speech Recognition in French," *Eurospeech'97*, Rhodes, pp. 56-59, Sept. 1997.

[2] S.S. Chen, P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 127-132, Feb. 1998.

[3] 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.

[4] J.L. Gauvain et al., "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *ARPA Speech Recognition Workshop*, Chantilly, pp. 56-63, Feb. 1997.

[5] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 75-79, Landsdowne, Feb. 1998.

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

[7] T. Hain et al., "Segment Generation and Clustering in the HTK Broadcast News Transcription System," *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 133-137, Landsdowne, Feb. 1998.

[8] M. Jardino "Multilingual stochastic n-gram class language models," *ICASSP-96*, Atlanta, May 1996.

[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] M. Siegler et al., Segmentation, Classification and Clustering of Broadcast News Audio," *DARPA Speech Recognition Workshop*, pp. 97-99, Chantilly, Feb. 1997.

[11] P.C. Woodland, T. Neieler, E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR", presented at the 1998 Hub5E Workshop, Sept. 1998.