TRANSCRIBING BROADCAST NEWS SHOWS

J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,gadda,lamel,madda}@limsi.fr

ABSTRACT

While significant improvements have been made over the last 5 years in large vocabulary continuous speech recognition of large read-speech corpora such as the ARPA Wall Street Journal-based CSR corpus (WSJ) for American English and the BREF corpus for French, these tasks remain relatively artificial. In this paper we report on our development work in moving from laboratory read speech data to real-world speech data in order to build a system for the new ARPA broadcast news transcription task.

The LIMSI Nov96 speech recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on newspaper texts. The acoustic models are trained on the WSJ0/WSJ1, and adapted using MAP estimation with task-specific training data. The overall word error on the Nov96 partitioned evaluation test was 27.1%.

INTRODUCTION

Over the last 5 years significant advances have been made in large vocabulary, continuous speech recognition, which has been a focal area of research, serving as a test bed to evaluate models and algorithms. However, these tasks remain relatively artificial as they mainly make use of laboratory read speech data. In this paper we report on moving toward real-world speech data in order to build a system for the new ARPA broadcast news transcription task.

The goal of this task is to transcribe broadcast news shows which contain signal segments of various natures such as prepared speech and spontaneous speech, which may be of studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), as well as speech over music and pure music segments. Acoustic models trained on clean speech are clearly inadequate to process such inhomogeneous data. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc.

The work presented in the paper addresses two main aspects of the problem. The first concerns segmenting the shows into homogeneous regions, and the second, decoding the segments with multiple sets of acoustic models tailored to each type of data.

DEVELOPMENT WITH MARKETPLACE

For this work we make use of the materials used for the Nov95 Hub4 “dry-run” evaluation to recognize MarketPlace radio broadcasts[1]. For our development work on this task we started with the 65k word recognizer developed for the ARPA NAB November 1995 evaluation [4]. This recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on newspaper texts. Acoustic modeling uses cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms. Each phone model is a tied-state left-to-right, CDHMM with Gaussian mixture observation densities (about 32 components). The modeled triphone contexts were selected based on their frequencies in the training data, with backoff to right-context, left-context, and context-independent phone models. Word recognition is carried out in two passes for each speech segment. In the first pass a word graph is generated using a bigram language model and in the second pass decoding uses the word graph generated by the 1st pass and a trigram language model.

The wideband acoustic models were trained on the WSJ0/1-si355 training data containing a total of 46k sentences[4]. Only the data from the close-talking, Sennheiser HMD-410 microphone was used. For telephone speech models, we used telephone channel models developed for the Hub2 test in 1994[3]. These models were trained on a bandlimited version of the WSJ si284 corpus, and adapted using MAP estimation[5] with 7k WSJ sentences of telephone speech data taken primarily from the Macrophone corpus. No task-specific acoustic training data was used. For language modeling data, we used newspaper texts and read speech transcriptions predating July 30, 1995. This data includes the August’94 release of the CSR standard LM training texts distributed by LDC (years 88-94), the 1994 NAB development data (excluding the devtest data), the WSJ0/WSJ1 read speech transcriptions (85,343 sentences), and the 1994 and 1995 financial domain material (Hub3 LM material).

Nine half-hour MarketPlace shows were used as task-specific training data to construct models for segmenting the test data (1 show was kept aside). A small left-to-right

	Identified class			
Test data	S	T	MS	M
Wide-band speech (S)	99.9	0.0	0.0	0.0
Telephone speech (T)	1.2	98.8	0.0	0.0
Speech+music (MS)	32.0	0.0	66.4	1.6
Music (M)	7.5	0.0	1.7	90.8

Table 1: Segmentation results in terms of the percentage of frames correctly and incorrectly classified for each class of data.

tied mixture HMM with 64 Gaussians was built for each of the following signal types: background noise, pure music, speech on music, wide-band speech, and telephone speech. The models were trained using the segmentations and labels provided by BBN[6]. Viterbi decoding on the 5 models (fully connected) is used to segment the data and assign each speech frame to one of the 5 classes.

A show is transcribed as follows: First the show is segmented using the tied mixture models. Segments identified as background noise and pure music are discarded. The telephone speech segments are then decoded with the telephone speech models and all the other segments are decoded using the wideband models. Unsupervised MLLR adaptation [7] is performed using all the data of a given type in the current show. Since sentence boundaries are not known, each segment is decoded as a single unit.

The segmentation error at the 10 ms frame level on the complete Marketplace show kept aside for development was 6%. As can be seen in Table 1 most of the segmentation errors are due to the misclassification of the `speech+music` frames (32.0% are classified as `speech`) and the `music` frames (7.2% are classified as `speech`). `Speech+music` frames are often classified as `speech` when the music is fading out because the signal is not very different from a speech signal with slight backgound noise. In this show there were no segments labeled as noise (N) by the transcribers, and no noise segments were detected by the segmenter.

The overall word error rate of the transcription for the same Marketplace is 24.6%. The error rate is seen to be much lower on wideband speech (16.2%), and much higher on telephone speech (42.6%) and `speech+music` (37.1%). The higher error rate observed for the telephone speech is not only due to the channel (reduced bandwidth and possible distortions), but also to the fact that most of this speech is spontaneous in nature, whereas much of the wideband speech is prepared. Also contributing to the overall error rate are insertions due to words recognized in a few music segments which are erroneously labeled as `music+speech`.

TRANSCRIBING BROADCAST NEWS

For the Nov96 evaluation, the scope of the task was enlarged to include multiple sources of broadcast news (radio, TV) and different types of shows (such as CNN Headline News, NPR All things Considered, ABC Prime Time news). The test data included episodes of shows not appearing in the training material. The 1996 evaluation consisted of two components, "partitioned evaluation" component (PE) and the "unpartitioned evaluation" component (UE). All sites were required to evaluate on the PE, which contains the same material as in the UE, but has been manually segmented into homogeneous regions, so as to control for the following *focus conditions*[8]:

F0- Baseline broadcast speech
F1- Spontaneous broadcast speech
F2- Speech over telephone channels
F3- Speech in the presence of background music
F4- Speech under degraded acoustical conditions
F5- Speech from non-native speakers
Fx- All other combinations

There was about 35 hours of task specific training data for which transcriptions were available. These data were obtained from 10 shows, such as ABC Nightline, CNN Headline News, CSPAN Washington Journal, and NPR Marketplace.

The development data were taken from 6 shows: ABC Prime Time, CNN World View, CSPAN Washington Journal, NPR Marketplace, NPR Morning Edition, and NPR The World. Our first experiments with the development data had a word error rate around 38%, using models similar to those used to decode the MarketPlace data. After incorporating the Hub4 acoustic and language model training data, modifying the lexicon and phone set, and using segment specific acoustic models, the word error on the development data was reduced to about 25%. In the remainder of this section, we describe our 1996 Hub4 system.

Acoustic models

Various approaches were investigated to build acoustic models from the available WSJ-si355 and Hub4 training data. The most effective solution for our system was the following:

1. Train large sets of gender-dependent tied-state models on the secondary channel of the WSJ0/1-si355 data. The resulting acoustic model set contained 7000 mixture distributions.
2. Use MAP estimation techniques to adapt the si355 seed models to the Hub4 training data, providing the baseline Hub4 models sets M1 and M2 (bandlimited analysis). For F5 (non-native speakers), the si355 models were adapted with British English data (WSJ0CAM), prior to adaptation with the Hub4 training data to create model set M5.
3. For the F3 and F4 conditions, the M1 models were adapted using supervised MLLR and the F3 and F4 parts of the training data, resulting in models M3 and M4.
4. Unsupervised MLLR adaptation is carried for each test segment in the final decoding pass.

The M1 models were used to process the F0 and F1 segments. The M2 models were used to process the F2 segments, as well as all other segments labeled as telephone speech by the Gaussian classifier. The M3 and M4 models were used to process the F3 and F4 data respectively, and the M1 and M5 models were used in parallel for the F5 data and all Fx segments labeled as "non-native."

A 64-component Gaussian mixture was built for telephone and wideband speech. For gender selection, condition-specific Gaussian mixtures (64 components) were estimated.

In order to model filler words and breath noises, 2 new phones were added to the existing phone set. These new phones are only trained with the Hub4 acoustic data since they are infrequent in read-speech data.

Language models

The language models were trained on newspaper texts (the 1995 Hub3 and Hub4 LM material – 161M words), and on the broadcast news (BN) transcriptions (years 92 to 96 – 132 M words). All trigrams occurring in the BN training transcriptions and in last year's MarketPlace transcriptions were included in the LM. The addition of other newspaper texts from any date led to a degradation both in terms of perplexity on the Hub4 devtest texts and recognition accuracy.

The 65k recognition vocabulary included all words occurring in the transcriptions (17883 from the BN transcripts and 6332 from 1995 MarketPlace). The LMs and vocabulary selection were optimized on the 1996 Hub4 developement test set. The resulting lexical coverage on the 1996 Hub4 dev test data is 99.34%.

The BN training texts were cleaned in order to be homogeneous with the previous text materials. Since in the BN texts word fragments are represented with a "hyphen", compound words were not split. We retreated all the transcriptions in order to split hyphenated words, as the occurrence of word fragments was marginal compared to other situations where the hyphen needed be treated.

The 1996 training transcripts were processed to map filler words (such as UH, UM, UHM) to a unique form, and the frequencies of filler words and breath noises were estimated for the different types of segments. These estimates were used in reprocessing the text materials. For breath noises, the observed proportion is different for the different segments (about 4.5% for the F0 and F1 segments, but only about 3% in the F3 and F4 segments). We hypothesized that the lower proportion in the F3 and F4 segments was an artifact due to the background music and noise which may have masked the breath noises. We also observed that while most breath noises appear at phrase boundaries, they also occur at other locations. We thus decided to process all of the training texts (1995 Hub3 and Hub4 and BN training texts) adding a fixed proportion of breath (4%), mostly near punctuation markers, respecting a minimum and maximum distance between two breath markers. A larger difference across segment types was observed for filler words, from 0.25% in prepared speech to about 3% in spontaneous speech. However, even though the global proportions were different, the filler words tend to occur in similar contexts for the different segment types. After systematic examination of their relative proportions in the training transcriptions, we constructed a "degrading" filter which adds filler words in the text with a parametrizable global proportion, so that the relative proportion of the fillers near specific common words was similar to that observed in the training transcription.

The resulting language models were tested using perplexity and recognition word error. Construction of different LMs for prepared and spontaneous speech according to the proportion of fillers found in the transcriptions, led to a gain in terms of perplexity, but did not reduce the recognition word error. We found that adding a small proportion of filler words (0.5%) improved the recognition accuracy, but adding a large proportion (3-5%) reduced performance.

As was done last year, we processed the 1000 most frequent acronyms in the training texts in order to treat them as whole words instead of as sequences of independent letters, as well as adding compound words for common word sequences, such as "let me" and "going to".

We split the different segments into 2 homogeneous groups from the LM point of view: one group corresponding to prepared speech with F0, F3, F4, F5 segments, and the other to spontaneous speech with F1, F2 segments. For the 1st bigram decoding pass, different LMs were used for prepared speech (cut-off 8, 2M bigrams) and spontaneous speech (cutoff 3, 1.9M bigrams). In the latter case the newspaper training texts were not used. For the 2nd pass, while the use of different trigrams for prepared and spontaneous speech LMs led to a gain in terms of perplexity, the word accuracy was worse on the development data. We therefore used a single 65k trigram LM trained on all the texts mentioned above (cut-off 1-2, 7.6M bigrams and 13.4M trigrams).

Lexicon

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 effects, which are relatively frequent in the broadcast emissions, and are not used in transcribing other lexical entries. The training and test lexicons were created at LIMSI and include some input and/or derivations from the TIMIT, Pocket and Moby lexicons. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 65k vocabulary contains 64,968 words and 72,488 phone transcriptions. Frequently occuring inflected forms were verified to provide more systematic pronunciations. The use of compound words for frequent word sequences enabled us to provide reduced pronunciations such as /lɛmi/ for "let me" and /gʌnə/ for "going to", in addition to the representation of frequent acronyms as words already used in our 1995 Hub3 system.

Decoding

Prior to decoding, segments longer than 30ms are chopped into smaller pieces so as to limit memory required for the

	Development data		*Evaluation data*	
Label	*Duration*	*WordErr*	*Duration*	*WordErr*
F0	25 min	11.5%	31 min	20.8%
F1	28 min	25.6%	32 min	26.0%
F2	19 min	34.3%	10 min	27.1%
F3	11 min	22.0%	7 min	20.3%
F4	16 min	19.0%	9 min	33.3%
F5	9 min	19.5%	2 min	27.8%
Fx	19 min	43.7%	14 min	46.1%
Overall	127 min	25.2%	106 min	27.1%

Table 2: Word error rates for the PE on the 1996 devdata and official NIST results on the evaltest data. (F0: baseline broadcast speech, F1: spontaneous broadcast, F2: speech over telephone channels, F3: speech in background music, F4: speech under degraded acoustic conditions, F5: non-native speakers, FX: other)

trigram decoding pass. 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 locate probable pauses where the segment can be cut. A Gaussian classifier is then used to estimate the gender for each segment and to label the Fx as wideband or telephone band.

Word recognition is performed in three steps: 1) word graph generation, 2) trigram pass, 3) segment-based acoustic model adaptation. A word graph is generated using a bigram backoff language model. This step uses a gender-specific sets of position-dependent triphones with about 6000 tied states and a small bigram language model (about 2M bigrams). Differents acoustic models are used for the different segment types. The model set is chosen based on the segment label. The sentence is then decoded using the word graph generated in the first step with a large set of acoustic models (position-independent triphones with about 7000 tied states) and a trigram language model (including 8M bigrams and 16M trigrams). Finally, unsupervised acoustic model adaptation is performed for each segment using the MLLR scheme, prior to the last decoding pass.

Experimental results

The evaluation test data were taken from 4 shows. The overall word error rate is 27.1% and the per show word errors are the following: CNN Morning News (29.7%), CSPAN Washington Journal (25.6%), NPR The World (30.5%) and NPR MarketPlace (23.0%).[1] The word error by segment type is given in Table 2, along with the results on the development data. While there are substantial differences across the conditions, the overall error rates are comparable.

The word error on the F0 devdata is about half that of other conditions. The same is not true for the eval data, due primarily to a long weather report spoken very quickly and with a high OOV rate. Speech over background music (F3) appears to be easier to handle than speech in noisy conditions (F4). This may be because speech over music usually occurs at the beginning and end of broadcasts, and is meant to be intelligible.

SUMMARY

The problem of segmenting broadcast news shows has investigated using the 10 MarketPlace shows distributed by NIST as Hub4 training data prior to the Nov95 evaluation. Compared to reference labels provided by BBN, the frame classification rate was 94%. Transcription of this data using WSJ0/1 models had an overall word error of about 25%.

The Nov96 ARPA evaluation investigated transcription of broadcast news from a wider variety of sources. About 35 hours of task specific training data with transcriptions were used to improve the acoustic and language models of the system. The development test data were used to optimize the recognition vocabulary and LM. Over 8000 new words were added to the lexicon, as well as compound words to allow modeling of reduced forms observed in spontaneous speech. On the partitioned evaluation using data from 4 shows, an overall word error of 27.1% was obtained (official NIST score).

[1] Transcribing and scoring this type of data is difficult. The reference transcriptions contain non-speech events, word-fragments, alternate spellings (particularly of proper names), contracted forms, that all can influence the word error rates. In addition, the segments are extracted from continuous broadcasts to ensure that overlapping speech segments are eliminated. This can result in many short turns of only a few words, and potential misalignments in reference and hypothesized transcriptions. As a result the word error rates may be overestimated.

REFERENCES

[1] *Proc. DARPA Speech Recognition Workshop*, Arden House, Feb. 1996.

[2] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), Oct. 1994.

[3] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *ICASSP-95*.

[4] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[5] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, **2**(2), April 1994.

[6] F. Kubala et al., "Toward Automatic Recognition of Broadcast News," *Proc. DARPA Speech Recognition Workshop*, Arden House, Feb. 1996.

[7] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, **9**(2), 1995.

[8] R. Stern et al., "Specification for the ARPA November 1996 Hub 4 Evaluation," Nov. 1996.