TRANSCRIPTION OF BROADCAST NEWS

Jean-Luc Gauvain, Lori Lamel, Gilles Adda, Martine Adda-Decker

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{gauvain,lamel,gadda,madda}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this paper we report on our recent work in transcribing broadcast news shows. Radio and television broadcasts contain signal segments of various linguistic and acoustic natures. The shows contain both prepared and spontaneous speech. The signal may be studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), or may contain speech over music.

Transcription of this type of data poses challenges in dealing with the continuous stream of data under varying conditions. Our approach to this problem is to segment the data into a set of categories, which are then processed with category specific acoustic models. We describe our 65k speech recognizer and experiments using different sets of acoustic models for transcription of broadcast news data. The use of prior knowledge of the segment boundaries and types is shown to not crucially affect the performance.

1. INTRODUCTION

The goal of this research is to automatically transcribe radio and television news broadcasts. Transcription of such shows is a major step towards developing real-world applications to deal with the vast amounts of information generated on a daily basis.

Broadcast news shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic nature, with abrupt or gradual transitions between segments. The signal 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. 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 linguistic style ranges from prepared speech to spontaneous speech. Acoustic models trained on clean, read speech, such as the WSJ corpus, are clearly inadequate to process such inhomogeneous data.

Two principle types of problems are encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. In order to address variability observed in the linguistic properties, we analyzed differences in read and spontaneous speech, with regard to lexical items, word and word sequence pronunciations, and the frequencies and distribution of hesitations, filler words, and respiration noises. As a result of this analysis, these phenonema were explicitly modeled in both the acoustic and language models as described in [5].

Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training specific acoustic models for the different acoustic conditions. Our basic strategy for transcription is to use simple Gaussian mixture models to partition the data, and then to process each segment using category specific models. In this paper we address primarily the effect of data classification and accurate segmentations on the recognition performance.

2. WORD RECOGNIZER OVERVIEW

In this section we describe the task-specific training data and the word recognizer, which is described in more detail in [5, 4]. For task-specific acoustic training data we used about 50 hours of broadcast news data distributed by LDC, for which about 35 hours were transcribed. These data were obtained from 10 different sources: ABC Nightline, ABC World News Now, ABC World News Tonight, CNN Early Prime, CNN Headline News, CNN Prime News, CNN The World Today, CSPAN Washington Journal, NPR All Things Considered, and NPR Marketplace. The 1995 development data consisted of 10 half hour Marketplace shows. The 1996 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.

For language modeling data, we used 161 million words of newspaper texts from the 1995 Hub3 and Hub4 LM material, 132 million words of broadcast news transcriptions (years 92 to 96), and 430 K words corresponding to the transcriptions of the acoustic training data. The 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 to reprocess the text materials to be closer to the observed spoken form[5]. As we have done previously, the

training texts were processed to treat the 1000 most frequent acronyms as whole words instead of as sequences of independent letters[3]. We also added about 300 compound words such as "let me" and "going to", to allow reduced pronunciations for these common word sequences.

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 development test set and the resulting lexical coverage on the 1996 Hub4 dev test data is 99.34%. The pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises)[5]. 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 word recognizer for this task is based on a 65k-word system developed for large vocabulary, continuous speech recognition [1, 2, 3], which has been evaluated on read-speech data in the ARPA WSJ and NAB tasks annually since 1992. 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 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (0.3-3.5kHz for telephone speech models) every 10ms. For each 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 as done in PLP analysis[6]. The cepstral coefficients are normalized on a segment basis using cepstral mean removal and variance normalisation. Thus each cepstral coefficient for each segment has a zero mean and unity variance. Each phone model is a tied-state left-to-right, CDHMM with Gaussian mixture observation densities (about 32 components). The triphone contexts to be modeled 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 performed in three passes for each segment. In the first pass a word graph is generated using a bigram language model (with about 2M bigrams) and gender-specific sets of position-dependent triphones (with about 6000 tied states). The segment is then decoded using the word graph generated in the first step with a larger set of acoustic models (position-independent triphones with about 7000 tied states) and a trigram language model (including 8M bigrams and 16M trigrams). Unsupervised acoustic model adaptation is performed for each segment using the MLLR scheme, prior to the third decoding pass.

3. DATA PARTITIONING

The transcription of a broadcast show requires dividing the data into managable size segments. Since the shows contain segments of different acoustic and linguistic natures, we have tried to assess the utility of partitioning and classifying the segments prior to word recognition. In this section we describe the segmentation algorithm and the category specific acoustic models used to assess the importance of data partitioning. We have previously investigated segmentation using the 5 acoustically motivated categories proposed by BBN (background noise, pure music, speech on music, wideband speech, and telephone speech) [7]. Compared to the *focus conditions* used in the ARPA Nov96 evaluation (see Table 1) which distinguish both acoustic and linguistic categories[11], the acoustically motivated appear to be more easily obtained by an automatic algorithm than the focus conditions.

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

Table 1: Focus conditions for ARPA Nov'96 Hub4 evaluation.

A segment classifier was developed and evaluated using the 1995 Marketplace data. Nine of the shows were used to construct models for segmenting the test data, and the 10th show was kept aside for development. The segmenter uses a small left-to-right tied-mixture HMM with 64 Gaussians for each signal type. Viterbi decoding with the fully connected models is used to segment the data and assign each speech frame to one of the classes. High intermodel transition penalities are used to avoid cutting the signal into too many short segments.

Since long portions of signal are often of the same segment type, a chopping algorithm was developed to chop segments longer than 30 s into smaller pieces so as to limit memory required for the trigram decoding pass. To do this, 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 prior to word recognition.

The task-specific acoustic models (BN) are estimated using MAP adaptation of gender-dependent seed models trained on the WSJ0/1 corpus with the broadcast news training data (BN-WB). For the telephone speech models, the gender-dependent seed models were obtained by adapting bandlimited WSJ models with the bandlimited segments of the broadcast news training data (BN-TB).

Type-specific acoustic models were trained for the different categories of data defined for the Nov96 focus conditions. The WSJCAM0 corpus was also used to train models for condition F5, non-native speakers of American English. The LIMSI evaluation system[5] used type-specific models for all focus conditions with the exception of high quality prepared and spontaneous speech which were combined into one data-type. In the next section we compare the results obtained with the type-specific models to those obtained when only wideband and reduced

Show	*Duration*	*WordErr*
Development data		
ABC PrimeTime	25 min	27.1%
CNN WorldView	22 min	19.1%
NPR MorningEdition	10 min	24.2%
CSPAN WashJournal	30 min	30.4%
NPR TheWorld	16 min	35.0%
NPR Marketplace	24 min	15.7%
Overall	127 min	25.2%
Evaluation data		
CSPAN WashJournal	32 min	25.6%
NPR TheWorld	10 min	30.5%
NPR Marketplace	7 min	23.0%
CNN Morning News	31 min	29.7%
Overall	106 min	27.1%

Table 2: Average word error rates by show for the partitioned evaluation on the 1996 development data and official NIST results on the evaluation test data.

bandwidth models are used.

4. EXPERIMENTAL RESULTS

For the Nov96 ARPA evaluation, the test data came from 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 which did not appear in the training or development 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 defined *focus conditions*[11].

For the evaluation, LIMSI reported results only for the partitioned evaluation component. These results are summarized for the development and evaluation test sets in Table 2 for the partitioned evaluation condition. The evaluation test data were taken from 4 shows. The overall word error rate is 27.1%.

The PE condition assumes that both the segment boundaries and the data type (F0-Fx) are known, but automatically making some of the distinctions is not so evident. We therefore explored two alternatives to see the importance of such prior information: the first uses only the segment boundaries but not the data type classification; and the second uses no prior information (unpartitioned condition).

Results using type-specific model sets are compared with results obtained using only two model sets (for wideband and telephone band) in Table 3 for the development test data. These results are the output of the second decoding pass with a trigram language model, which made use of the word graphs generated with the first pass type-specific acoustic model sets. No segment-based adaptation was performed. The results are identical for focus conditions F0, F1 and F2 since the same acoustic models are used in both cases. Only small differences are

Label	*Duration*	*Type-Specific WordErr*	*BN-WB/TB WordErr*
F0	25 min	12.0%	
F1	28 min	27.6%	
F2	19 min	36.1%	
F3	11 min	23.8%	22.8%
F4	16 min	19.6%	20.3%
F5	9 min	21.9%	22.0%
Fx	19 min	46.3%	48.0%
Overall	127 min	26.8%	27.1%

Table 3: Word error rates for type-specific and wideband/telephone band models on the PE for the 1996 devdata. Results are from trigram decoding without segment-based adaptation. (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)

observed for the other data types, with the type-specific models performing slightly better except for the F3 data, where the type-specific models used in the evaluation are seen to perform slightly less well than the wideband models.[1] It is interesting to note that the word errors obtained for speech in background music are lower than the spontaneous broadcast speech, as the level of music is usually low enough for the speech to be clearly understood.

We explored the no prior knowledge (unpartitioned) condition using the Marketplace show from the 1996 development data. We have previously reported the segmentation and classification rate on a complete Marketplace show kept aside from the 1995 development data[3]. Compared to reference labels provided by BBN, the frame classification rate was 94%, with the majority of segmentation errors 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 light of these segmentation results, and the word recognition results in Table 3, we decided to use only a two way classification, dividing the data into wideband and bandlimited segments. The telephone segments in the show were correctly detected, with boundary locations close to those marked manually.[2] Each segment longer than 30 s was subsequently chopped into chunks, and each chunk was processed independently. A Gaussian mixture model is used to identify the sex of the speaker of each chunk. It should be noted that there can be multiple speakers of the same or different sex in a single chunk.

[1] At the time the evaluation was held, these models performed slightly better in second pass decoding with the development word graphs which had been generated using our 1995 NAB acoustic models[3]). After the evaluation new word graphs were generated for the development data using the final evaluation setup, and with these new word graphs the original small gain was no longer observed.

[2] One boundary was poorly placed, dividing a word, which given the small amount of telephone data accounts for the difference word error rates in Table 4.

	Partitioned		*Unpartitioned*
	Type-Specific	*BN-WB/TB*	*BN-WB/TB*
Label	*WordErr*	*WordErr*	*WordErr*
F0	11.3%		12.0%
F1	20.9%		20.8%
F2	17.1%		20.5%
F3	10.7%	10.0%	15.2%
F4	16.8%	17.1%	21.9%
F5	19.5%	24.0%	26.0%
Fx	50.6%	53.4%	64.4%
Overall	16.4%	16.7%	18.7%

Table 4: Word error rates for the dev96 Marketplace show. Type-specific models for PE and wideband/telephone band models for PE and UE. Results are from trigram decoding without segment-based adaptation.

The telephone speech segments are then decoded with the telephone speech models and all the other segments are decoded using the wideband models.

A surprising result is that even without explicitly trying to separate out pure music segments on the Marketplace data, the recognizer rejects most of this data by outputting essentially only non-speech or filler word models. However, decoding of pure music segments was found to be substantially slower than decoding speech, even under somewhat degraded conditions or in the presence of background music. Thus, it would be of interest to be able to reliably separate out pure music segments to reduce the computation needs.

The partitioned evaluation using type-specific acoustic models for the same Marketplace show was 16.4%, which can be compared to 16.7% using only 2 model sets (BN-WB/BN-TB). Using our simple segmentation scheme a word error rate of 18.7% was obtained for the unpartitioned condition. This 10% relative increase in error rate is surprisingly small since this approach should suffer from the following drawbacks: there is no explicit music modeling (for the PE there is evidently no error due to pure music segments); changes in acoustic conditions and speaker turns (i.e. same segment can contain data from speakers of different sex) are ignored, which should be a problem for the segment based cepstrum normalization. This shows that the accurate partitioning of the data is not crucial for transcription, at least at current state-of-the-art word error rates.

5. CONCLUSIONS

We have presented our work in developing a system to automatically transcribe television and radio broadcast news shows. Our initial findings are that quite reasonable word accuracies (greater than 70%) can be obtained on unrestricted broadcast news data, with acoustic models trained on about 50 hours of task-specific data, and with language models trained on broadcast news transcriptions and on newspaper texts adapted to better model the observed data by inserting breath noise and filler words.

Our experiments comparing type-specific models with more general task-specific models (BN-WB/TB) demonstrate that there is only a slight loss in performance, which does not justify the additional burden in training and decoding with specialized model sets. Similar results have been reported in [10]. This effect may be due to the relatively limited amounts of training data and test data for the different conditions, and may not be true if equal amounts of training data were available for all conditions.

Comparing the recognition error rates with known segmentations to those obtained without use of this information, only a moderate performance degradation was observed on the Marketplace data. Put in other terms, the prior knowledge of the segmentation does not allow us to significantly reduce the error rate. This is in line with the Nov96 results for the BBN and CMU systems for the partitioned and unpartitioned conditions[9].

This extreme approach of not attempting to partition the data allowed us to estimate the upperbound of performance loss. It was interesting that only a moderate degradation in performance was observed for speech in the presence of background music, and that pure music segments did not generate a lot of recognition errors. The long decoding time observed for pure music segments suggests that a music detector would be of interest for real applications. While we would like to reduce the difference in performance without prior knowledge of the partitions, it is certainly as important to improve the robustness of the underlying technology.

REFERENCES

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

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

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

[4] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News Shows ," *ICASSP-97*.

[5] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *ARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, **87**(4), 1990.

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

[8] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, pp. 171-185, 1995.

[9] D.S. Pallett, J.G. Fiscus, "1996 Preliminary Broadcast News Benchmark Tests," *ARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997.

[10] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, "Modeling Those F-Conditions – or Not," *ARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997.

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