RECOGNITION OF CONTINUOUS BROADCAST NEWS WITH MULTIPLE UNKNOWN SPEAKERS AND ENVIRONMENTS

Uday Jain, Matthew A. Siegler,
Sam-Joo Doh, Evandro Gouvea, Juan Huerta, Pedro J. Moreno, Bhiksha Raj, Richard M. Stern
{uj,msiegler,sjdoh,egouvea,juan,pjm,bhiksha,rms}@cs.cmu.edu
Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

Practical applications of continuous speech recognition in realistic environments place increasing demands for speaker and environment independence. Until recently, this robustness has been measured using evaluation procedures where speaker and environment boundaries are known, with utterances containing complete or nearly complete sentences. This paper describes recent efforts by the CMU speech group to improve the recognition of speech found in long sections of the broadcast news show Marketplace. Most of our effort was concentrated in two areas: the automatic segmentation and classification of environments, and the construction of a suitable lexicon and language model. We review the extensions to SPHINX-II that were necessary to enable it to process continuous broadcast news and we compare the recognition accuracy of the SPHINX-II system for different environmental and speaker conditions.

1. INTRODUCTION

Historically, speech recognition systems have tended to be evaluated under conditions where the following were assumed to be true:

1. The audio is presegmented, with each segment containing complete or nearly complete sentences or phrases.
2. There is a beginning and ending silence before and after the speech component.
3. The speaker, environment, and noise present in each utterance are constant throughout the utterance.
4. The text of each utterance is primarily read from written prompts.

The goal of the ARPA 1995 Hub 4 evaluation was to transcribe speech contained in audio from Marketplace broadcasts, with speech that is often inconsistent with all four of these assumptions. While this is a far more challenging domain than those used in previous continuous-speech evaluations, it compels the research community to confront a number of important problems including rapid adaptation to new speakers and acoustical environments, adaptation to non-native speakers, robust recognition to highly spontaneous and idiomatic speech, and robust recognition of speech in the presence of background music. Good solutions to all of these problems are needed in applications such as CMU's INFORMEDIA system which transcribes speech from television broadcasts and video archives.

Most of our effort in this work was directed at framing the task in a manner which is consistent with these assumptions. We first discuss some of the general issues involved with two aspects of continuous speech processing: the acoustic problem and the linguistic problem. We subsequently describe the implementation of the CMU's system used in the 1995 ARPA Hub 4 task.

2. THE ACOUSTIC PROBLEM

The audio in Marketplace broadcasts is an unbroken stream of up to 30 minutes of program material. As is common in broadcast news shows, there are overlapping segments of speech and music, with various speakers recorded in different environments.

We see changes in noise and channel as having a greater impact on recognition than changes in speaker identity, since our compensation schemes and acoustic models contain the principle assumption that the environment does not change within an utterance. Environmental classification schemes, to be discussed below, were geared towards discerning these changes rather than sentence boundaries.

The process of dividing a long stream of audio into smaller segments is referred to as segmentation. SPHINX-II in the configuration used for this evaluation could not tolerate segments shorter than 3 seconds or longer than 50 seconds without adverse effects on recognition performance. The 50-second limit was due to system memory constraints. The 3-second minimum duration limit was imposed because segments shorter than 3 seconds were found to be unreliable, especially in noisy regions of the broadcast. In addition, incomplete speech events at the very beginning or ending of each utterance can cause drastic recognition problems.

The goal of segmentation is therefore twofold: to provide audio within which the recording environment is the same throughout, and to begin and end each utterance during silence periods.

2.1. Environmental Classification

Preliminary studies using the training data for the 1995 Hub 4 evaluation showed that recording environments appearing in the Marketplace broadcasts can be grouped into four categories:

- Clean speech, 8 kHz bandwidth
- Degraded speech, 8 kHz bandwidth
- Speech with background music, 8 kHz bandwidth
- Telephone speech, 4 kHz bandwidth

Several Gaussian classifiers were trained to partition speech into category classes of male versus female speech, telephone versus non-telephone speech, and clean versus degraded speech.

2.2. Utterance Segmentation

The segmentation of a long stream of acoustic data (the news show) into manageable chunks was an important part of the Marketplace system. The segmentation was carried out at predicted silence points to ensure that segmentation did not occur in the middle of words. The process also incorporated classifier information so as to ensure that the final segments were acoustically homogenous.

2.3. Environmental Compensation

Results of pilot experiments showed that recognition error rate increased when the background environment was in the “music” or “degraded” categories. In these situations, we used the CDCN algorithm [1] to compensate for environmental effects.

2.4. Acoustic Modelling

Optimum recognition could be achieved if each of the environmental and speaker conditions would be recognized with fine-tuned models for the specific conditions. We used telephone-bandwidth speech models for the telephone speech and clean full-bandwidth models for all other speech.

3. THE LINGUISTIC PROBLEM

The Marketplace broadcast is a mix of prepared and extemporaneous speech. The nature of extemporaneous speech suggests that there will be sentence fragments, and a greater use of the personal pronouns *I* and *YOU* than would typically be found in written material. In addition, the classification-based segmentation process is not geared towards providing complete sentences, but constant environments. As a result, there is a good chance that sentences will be broken in the middle during speech pauses even during prepared speech. These considerations suggest that the best language model for the task would be a combination of models from several domains.

3.1. The Language Model

The language model (LM) is built from an interpolation [6] of a large “static” model with two smaller “adaptation” models. The static model is the publicly-distributed standard trigram model for the 1995 ARPA Hub 3 evaluation. The adaptation models contain out-of-domain text from the epoch of the test material (August 1995) and in-domain text occurring before the epoch of the test material. The out-of-domain adaptation LM is a trigram model created from the August 1995 financial and general news texts released by the LDC. The in-domain adaptation LM is a bigram model created from the 10 Marketplace shows distributed as a training set by the LDC.

Begin-of-sentence and end-of-sentence tokens were removed in the creation of the adaptation language models to facilitate the recognition of audio segments containing sentence fragments. The largest possible lexicon was used in constructing the language models: 64 k words. Tables 1 and 2 compare word error rates for the evaluation set obtained using the static Hub 3 model and the interpolated Hub 4 model.

3.2. The Lexicon

Although the LM is built with a particular lexicon in mind, the number of pronunciations available to the decoder is greater due to multiple pronunciations.

In addition, a large vocabulary task with more than 64k pronunciations has many confusable pronunciations. In this way, the benefit of out-of-vocabulary (OOV) reduction by increasing the vocabulary is offset by the increased complexity of the task. Figure 1 shows the OOV rate for the development test set as a function of lexicon size.

Six different lexicons were evaluated on two of the development test shows in an attempt to select an optimum size. We surmised that acoustically more difficult speech, such as telephone-bandwidth speech or speech in the presence of music, presents a greater mismatch to the recognition system than speech containing a few OOV occurrences. Table 3 summarizes the effect of dictionary size on recog-

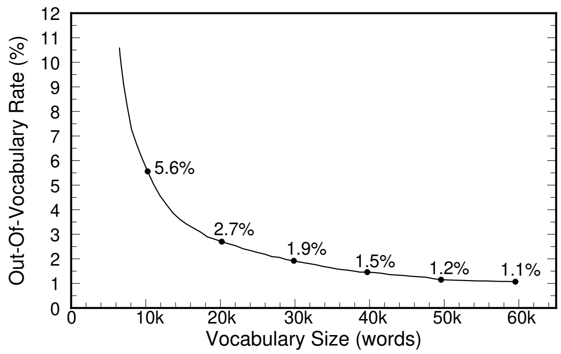

Figure 1. Out-of-Vocabulary (OOV) rates for the development test set for lexicons of different size. Each lexicon contains the top N words from the H3 lexicon mixed with the words found in the ten Marketplace training shows.

Speaker and Environment Type	Portion of Test	H3 LM WER (%)	H4 LM WER (%)
All Speakers/Envs	100 %	42.0	41.0
Anchor/Correspondent	51 %	30.5	28.7
Clean speech	39 %	29.4	27.7
Background music	7 %	44.8	41.3
Telephone speech	4 %	15.5	15.5
Other Speakers	32 %	52.8	52.3
Clean speech	17 %	46.4	46.8
Background music	0.5 %	–	–
Telephone speech	15 %	58.9	57.4
Foreign Accent	17%	55.9	55.9

Table 1: Comparison of word error rates (WER) for the "heads-and-tails" portion of the 1995 Hub 4 evaluation test set using two different language models. Size refers to the percent of total speech represented by a particular condition. Word error rates for conditions that represent less than 2% of the test set are not shown.

Speaker and Environment Type	Portion of Test	H3 LM WER (%)	H4 LM WER (%)
All Speakers/Envs	100 %	39.9	39.5
Anchor/Correspondent	68 %	29.5	28.6
Clean speech	41 %	24.3	23.7
Background music	18 %	36.0	34.6
Telephone speech	9 %	40.1	38.7
Other Speakers	26 %	65.5	66.3
Clean speech	12 %	52.6	54.7
Background music	1 %	–	–
Telephone speech	13 %	78.2	77.3
Foreign Accent	6 %	48.7	49.7

Table 2: Same as Table 1, but for the "whole-show" portion of the 1995 Hub 4 evaluation test set.

nition accuracy for American speakers of English, using data from the full shows 940401 and 940429 in the development test set.The lexicons were constructed by combining the N most frequent words from the H3 language model with all the words found in the ten Marketplace training shows.

Environment Type	Portion of Test	Size of Dictionary					
		10k	20k	30k	40k	50k	60k
All	93 %	38.8	37.0	36.6	36.2	36.2	36.3
Clean	59 %	33.4	31.4	30.8	30.5	30.7	30.8
Other	35 %	48	47	47	46	46	46
Noise	13 %	66	66	65	64	64	64
Music	14 %	36	36	36	36	36	36
Telephone	7 %	45	45	45	45	44	44

Table 3: Recognition accuracy for American speakers of English as a function of dictionary size and environment type.

Increasing the dictionary to its maximum size provided a significant improvement in recognition accuracy only for high-quality speech showed. As a result, two lexicons were constructed for the system, containing 60,000 words and 30,000 words. The 60,000-word lexicon is used to decode segments of speech classified as clean speech, and the 30,000-word lexicon is used for all other segments.

4. SYSTEM IMPLEMENTATION

The CMU H4 transcription system to process the Marketplace broadcasts is composed of the following stages:

1. Initial-pass classification and segmentation
2. Acoustic compensation
3. Initial-pass recognition
4. Decoder-guided segmentation
5. Final recognition

We discuss the processing of each stage in turn.

4.1. Initial-pass classification and segmentation

In early implementations of the system, segmentation was based only on silence detection. Segmentation points were created when a silence meeting a preset duration criterion was detected. While this procedure provided segments of suitable length, it tended to segment in the middle of words, especially in the presence of noise and music. This was a source of errors as the decoder assumes that there will be no incomplete speech events at the very beginning or ending of each utterance. Furthermore, there was no way of ensuring that the eventual segments would be acoustically homogeneous.

To ensure that segments were obtained from a homogenous recording environment, we developed a classification-based segmenter. This segmenter used the presence of silence at environment changes to provide segmentation points. It classified the acoustical content of the segments according to the categories of male versus female speech, telephone versus non-telephone speech, clean versus degraded speech and music versus non-music. The silence threshold was adaptive to provide reliable segmentation in the presence of background speech and music.

Because the durations between changes in acoustic source, environment, or background can vary widely, the system imposed hard limits on the minimum and maximum segment lengths. Some segments were still obtained from more than a single unique class because silence could not be detected at the class changes with confidence. This problem was addressed with decoder-guided segmentation discussed below.

4.1.1 Segmenter-Classifier features

The environment classifiers used multimodal Gaussian distributions that were trained from hand-segmented and labeled training data from six of the ten Marketplace shows in the training set. Gaussian mixtures with 16 components were used to characterize the probability densities used by the male/female, clean/noisy and music/non-music classifiers, but 16-component and 8-component Gaussian mixtures were needed for the telephone/non-telephone classifier.

To increase the accuracy and robustness of the classifiers the cepstral energy was averaged over a region of ten frames. This method improved the ability of the music/non-music classifier to distinguish speech with music from speech without music in the background.

4.1.2 Segmenter-Classifier performance

The performance of the classifiers for the initial-pass segmenter, based on hand classified utterances, is provided in Table 4 below. Inconsistencies between decisions based on manual classification and automatic classification were considered to be errors.

Classifier	Errors
Tel/Non-tel	4.7%
Male/Female	4.2%
Clean/Degraded	16.3%
Music/Non-music	7.8%

Table 4: Percentage of classification errors for the initial-pass segmenter.

In the actual Hub 4 evaluation, only the male/female, telephone/non-telephone and clean/degraded Gaussian classifiers were used to classify 1-second windows of incoming audio. Classification for the current window was determined based on a maximum likelihood decision using raw cepstral coefficients derived from the signal. When the output of any of these three classifiers changed for any of the three classes during the course of the audio, the segmenter searched for the presence of a silence within the 1-second window at the transition. Silence was detected by searching for minimum energy in the given window, and labelling as silence all contiguous frames with energy within a fixed threshold relative to this minimum. A segmentation point was defined when the silence was atleast 15 frames long. Consecutive segmentation points occurring less than 3.0 seconds apart were ignored. If a segment exceeded 50 seconds, the segmenter located another silence occurring anywhere within the segment in the manner just described. These were the limits on utterance length imposed by the decoder used in this evaluation. After all breakpoints were found, the segments were reclassified over each segment in its entirety rather than independently for each individual 1-second window.

4.2. Acoustic Compensation

Speech that is classified as either noisy or telephone-bandwidth is compensated using an improved version of the Codeword-Dependent Cepstral Normalization (CDCN) algorithm [1]. CDCN improves the recognition accuracy of speech when the recording environment is different from that of the speech used to train the acoustic models. CDCN distributions for the evaluation system were trained from SI-284 WSJ0 and WSJ1 Corpora for use with noisy speech. For telephone-bandwidth speech, the SI-284 WSJ0 and WSJ1 Corpora were passed through a filter representing an average telephone channel and then used to train the CDCN distributions. Table 5 shows how recognition in adverse environments improves with the addition of CDCN.

Environment	WER (%)	
	Baseline	CDCN
Music	58.4	40.9
Noise	56.2	47.0

Table 5: Changes in recognition performance for Show 940204 with the addition of CDCN environmental compensation.

4.3. Initial-pass recognition

A fast version of SPHINX-II [3], CMU's semi-continuous hidden Markov model recognition system, is used to decode the speech for each segment. The only modification to the SPHINX-II system as described in [3] is that reduced-bandwidth signal processing is used to process speech that the initial-pass segmenter determines to be of telephone bandwidth.

The baseline acoustic models used to recognize full-bandwidth speech are a gender-dependent set of full-bandwidth models trained from the SI-284 WSJ0 and WSJ1 Corpora .

In the Hub-2 component of the 1994 ARPA CSR evaluation we found that telephone-specific acoustic models were more effective than acoustic compensation schemes that manipulate the feature vectors [5]. For the Marketplace broadcasts we trained gender-independent telephone-bandwidth models with a subset of utterances from the Macrophone telephone speech corpus [2].

A duration-based rejection method is used to discard words falsely decoded during music-only passages. Phonetic duration models based on the SI-284 WSJ0 and WSJ1 Corpora were used to discard words where the probability of duration was less than 0.001.

4.4. Decoder-guided segmentation

In some cases it was not possible to find silences that were sufficiently long to ensure that segmentation did not occur in the middle of a word, even though the classifiers detected a change in acoustic conditions with a high degree of certainty. In these cases we ran the SPHINX-II decoder as a silence detector, and we looked for the closest silence to a change in detected conditions.

After the initial decoder pass, all regions of audio decoded as silence are collected and sorted in decreasing duration. A top-N search is used to determine new breakpoints which yield segment durations meeting preset criterion for minimum, maximum and average value. These criteria are 3 seconds, 30 seconds, and 10 seconds. These locations were used as break points in a second segmentation of the entire show.

Additional breakpoints are retained where transitions from telephone to non-telephone classifications occur using, decoder detected silence, in the manner described above. All the resultant segments are then reclassified as before.

4.5. Final recognition

Recognition in the final pass proceeds in the same fashion as in initial-pass recognition. Segments labeled as music are treated in the same manner as those labeled as degraded.

5. PERFORMANCE OF THE MARKETPLACE TRANSCRIPTION SYSTEM

During the course of our development, various improvements and innovations reduced the relative recognition error rate by 33%, as summarized in the figures cited in Table 6. The baseline system used in this Table was the implementation of SPHINX-II with which we began our development of the Marketplace transcription system. It included two gender-dependent full-bandwidth acoustic models, class-based segmentation, no environmental compensation, and the 1994 S2-P0 NAB-trained language model and dictionary.

Evaluation System Innovation	WER (%)	WER reduction
Baseline	60.6	–
Reduced-Bandwidth Models	54.5	10.1%
Long Word Rejection	53.3	2.2%
Resegmentation using Hypothesis	52.7	1.1%
CDCN Compensation	49.1	6.8%
H3 LM	41.8	14.9%
H4 LM	40.6	2.9%
Optimal Dictionary	40.0	1.5%

Table 6: Improvements in word error rate on the evaluation test set as improvements and new components were added to the baseline system.

Table 1 shows the overall performance of the system for the entire 1995 Hub 4 evaluation set, after adjudication procedures. The results are grouped according to speaker and environment type. As expected, speech from the "Other Speakers" category was recognized poorly compared to recognition error rates obtained for anchors and correspondents. We generally found that extemporaneous speech or speech from non-native speakers increased the word error rate by about 50 percent relative to the baseline of read speech in a studio environment, and the presence of background music appeared to increase the error rate by 35 to 50 percent.

Speaker and Environment Type	Portion of Test Set	WER (%)
All Speakers/Envs	100 %	40.0
Anchor/Correspondent	57 %	28.0
Clean speech	40 %	25.8
Background music	11 %	35.3
Telephone speech	6 %	28.7
Other Speakers	30 %	57.0
Clean speech	15 %	49.1
Background music	< 1 %	76.0
Telephone speech	14 %	64.3
Foreign Accent	13 %	54.6

Table 7: Recognition performance of different speakers and environments for the evaluation test set using the system described.

In a final post-evaluation analysis we compared the performance obtained using manual and automatic initial-pass segmentation and classification. These results are summarized in Table 8 below, which were obtained by running the evaluation system using the H3 language model on the training show 940204.

Segmentation	Classification	WER (%)
Manual	Manual	40.7
Manual	Auto	38.8
Auto	Auto	42.1

Table 8: Comparison of results obtained using automatic and manual initial-pass segmentation and classification.

As can be seen from Table 8, the use of manual segmentation reduces the relative word error rate by 4.7 percent, suggesting that further improvements could be obtained by better segmentation. The surprising result that automatic initial classification outperforms manual classification appears to reflect the fact that the automatic classifier provides a more helpful (although less "correct") classifications of speaker gender for this particular set of test material.

6. SUMMARY AND CONCLUSIONS

The transcription of continuous speech from radio broadcasts poses many new interesting challenges for developers of speech recognition system. Initial development of the CMU Marketplace Transcription System focussed of necessity on various aspects of the infrastructure needed to automatically segment and classify the different types of speech occurring the broadcasts. Improvements to the system reduced the relative error rate by 33 percent, with the greatest improvements provided by the addition of appropriate language models, acoustic models, and environmental compensation procedures. We expect that further substantial improvements to the system will be obtained by the incorporation of speaker adaptation, better compensation for the effects of background music, and a recognition system that makes use of continuous HMMs.

ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We also thank Ravishankar Mosur, Eric Thayer, Ronald Rosenfeld, Bob Weide the rest of the speech group for their contributions to this work.

REFERENCES

1. Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

2. Bernstein, J. and Taussig, K.,"Macrophone: An American English Telephone Speech Corpus for the Polyphone Project". *ICASSP-94*, May 1994.

3. Huang, X., Alleva, F. A., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., and Rosenfeld, R.: "The Sphinx-II Speech Recognition System: An Overview", Computer Speech and Language, Volume 2, pp. 137 - 148.

4. Hwang, M.-Y., *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University, 1993.

5. Moreno, P. J., Siegler, M. A., Jain, U., And Stern, R. M. "Continuous Recognition of Large-Vocabulary Telephone-Quality Speech", Proceedings of the ARPA Workshop on Spoken Language Technology, 1994, Austin, TX, Morgan Kaufmann, J,. Cohen, Ed.

6. Rudnicky, A., "Language Modelling with Limited Domain Data," Proceedings of the ARPA Workshop on Spoken Language Technology, 1994, Austin, TX, Morgan Kaufmann, J,. Cohen, Ed.