BROADCAST NEWS TRANSCRIPTION

Francis Kubala, Hubert Jin, †Spyros Matsoukas, Long Nguyen, Richard Schwartz

BBN Systems and Technologies
70 Fawcett Street, Cambridge, MA 02138
†Northeastern University, Boston, MA
fkubala@bbn.com

ABSTRACT

In this paper we describe our recent work on automatic transcription of radio and television news broadcasts. This problem is very challenging for large vocabulary speech recognition because of the frequent and unpredictable changes that occur in speaker, speaking style, topic, channel, and background conditions. Faced with such a problem, there is a strong tendency to try to carve the input into separable classes and deal with each one independently. In our early work on this problem, however, we are finding that the rewards for condition-specific techniques are disappointingly small. This is forcing us to look for general, robust, and adaptive algorithms for dealing with extremely variable data. Herein, we describe the BBN BYBLOS recognition system configured to handle off-line transcription and we characterize the speech contained in the 1996 DARPA Hub-4 testbed. On the partitioned development test set, we achieved a 29.4% overall word error rate.

1. INTRODUCTION

The DARPA Hub-4 Broadcast News testbed consists of a large collection of audio recordings of a variety of news and issues-oriented programs from radio and television broadcasts. It is intended to spur development of robust recognition techniques to deal with highly variable, real-world speech.

This data varies in almost every conceivable way. Speaking styles range from carefully read monologues to fluent conversation and even dramatizations. Some speakers have regional dialects or non-native accents. Topics change unpredictably. There are frequent changes between studio and remote-location environments, often involving a telephone channel in half of a conversation. A variety of signal processing techniques are used by modern broadcasters to control gain. Background music, speech, and noise are commonly added to the signal to enrich its auditory appeal to the listener. The large monolithic input also presents new problems for recognition algorithms. All potentially useful boundaries between the changing conditions are unknown to the system.

This is a major departure from the great majority of past work in large vocabulary speech recognition, which generally assumes some knowledge of the specific conditions of the input and also that those conditions will remain fixed throughout recognition. The broadcast news problem forces system builders to choose between trying to separate and deal with each set of conditions independently or finding robust adaptive techniques that can operate reasonably over all of the various conditions found in the data. In our recent work, we are finding diminishing returns for the divide and conquer approach. The data is so variable that separation and classification of the input is not 100% accurate and the separate models adds complexity and size to the system. Therefore we have turned our attention toward more general approaches which do not require prior classification of the data.

In the next section, we describe the configuration of the BBN BYBLOS system used for transcribing broadcast news material.

2. BYBLOS TRANSCRIPTION SYSTEM

We have reconfigured the BBN Byblos system to handle the problem of off-line transcription of large digital audio files. A key component in the transcription system is a new 2-pass N-best decoder. This decoder uses a fast-match algorithm, Phoneme-Tied-Mixture (PTM) HMMs [7] and a bigram Language Model (LM) in the forward direction. The PTM model has about 12K Gaussian parameters. A backward pass uses State-Clustered Tied-Mixture (SCTM) HMMs [7], with 64K Gaussians, and a trigram LM. The backward pass is very fast because it has the trellis of forward word-ending scores available, permitting it to prune very aggressively and selectively. The backward pass also supports an integrated word-dependent N-best traceback [8], making the generation of multiple hypotheses an efficient process. The N-best output is then reordered with a SCTM model that includes between-word triphones.

The transcription procedure is organized into 3 logical stages:

- Gender classification, segmentation, channel normalization, and speaker clustering to break the monolithic waveforms into usable chunks
- Speaker-Independent (SI) recognition to create transcriptions for unsupervised adaptation
- Speaker-Adapted (SA) recognition to produce the final answer

2.1. Segmentation

A preprocessing stage is required to accomplish 3 things. First, the segments need to be cut at gender-change boundaries and classified as male or female, since our acoustic models are gender-dependent (GD). Next, we need to break the long-duration waveforms into short, fairly uniformly sized segments for computational efficiency in the N-best stage. Finally, we need to identify all segments belonging to the same speaker, channel, and background conditions to make maximum use of the test data in unsupervised adaptation.

Gender segmentation is accomplished with context-independent GD phoneme HMMs in a combined-gender model that has two separate sets of phonetic models – one trained on male speech and the other on female speech. The recognizer emits a sequence of gender-tagged and time-stamped phones which are used to identify locations of gender changes in the input. Each segment is then decoded with GD context-dependent phonetic HMMs and a 20K word LM to accurately locate pauses in the input. The GD segments are further reduced in duration by cutting them at pauses into fairly uniform sized chunks averaging about 20 words long. We have previously reported that no degradation results by chopping at pause locations without regard to the linguistic context of the words on either side of the cut [5]. We constrain the cuts to occur at pauses of 150 msec or longer unless the length of the segment becomes too long.

We compute and subtract a SNR-dependent cepstral mean for each frame in a chopped segment in the spirit of [1]. We observed a small gain on all of the clean speech conditions for this Cepstral Mean Subtraction (CMS), but small degradations on the music and noise conditions in our initial experiments.

We then cluster the segments using a weighted likelihood ratio criterion [4] on the cepstral parameters of the segments. Since consecutive segments are more likely to be from the same speaker, we apply a penalty on the temporal separation between segments. Evaluated after adaptation to each of the speaker clusters, we found that automatic clustering performed just as well as adapting to the true speaker clusters.

2.2. Adaptation

We have previously reported on a Speaker-Adapted Training (SAT) algorithm [3] designed to reduce the inter-speaker variability inherent in the commonly used, pooled-speaker SI model. This procedure iteratively reestimates a transformation of the Gaussian parameters of each training speaker to the pooled-speaker model and produces a pooled adapted-speaker model that has markedly reduced variances in each dimension, compared to the SI seed model. For the Broadcast News problem, we have recently implemented an efficient SAT procedure that permitted estimation of transformations for one thousand training speakers.

For the transcription problem, we perform the adaptation in two nearly identical recognition stages. A SI seed model is used to generate the hypotheses for unsupervised adaptation via Maximum Likelihood Linear Regression (MLLR) [6]. Both the PTM and SCTM models are adapted to each speaker cluster and used in a second recognition run over the data. The SA model produces a new N-best list which is used to adapt the SAT SCTM model. The list is then reordered to produce the final top1 result.

3. BROADCAST NEWS TESTBED

The 1996 DARPA Broadcast News testbed consists of three collections of recordings that are disjoint in date of broadcast. Acoustic training data consists of 87 episodes of about a dozen different news programs broadcast from May and early June of 1996. The amount of speech data in these recordings is about 38 hours. 2.5 hours of additional material, from 6 episodes broadcast in early July, are set aside as development test data. A similar amount of data, from the late summer timeframe, was used as evaluation test data in the November 1996 DARPA Hub-4 benchmark tests.

All data was annotated in detail by the Linguistic Data Consortium (LDC) and the National Institute of Standards and Technology (NIST). Conditions of the data, such as speaking style, presence of competing background sources, and subjective fidelity of the signal were labeled. All boundaries between these conditions were marked, as were locations of each change in speaker and topic. These annotations are useful for analyzing results. They were also used as explicit side-information that was given to the systems evaluated in the 1996 DARPA Hub-4 Partitioned Evaluation (PE) test. At BBN, we made no use of these annotations in the recognizer, preferring instead to focus our research upon techniques that would extend to the companion, Unpartitioned Evaluation (UE) test, for which no side-information was available. The UE test is specifically designed to mimic the real-world problem of large vocabulary broadcast news transcription.

3.1. Characteristics of the Test Data

The Hub-4 test data are composed of digital recordings of whole episodes of selected programs. Each 30–120 minute episode is contained in a single large waveform. The annotations provided with the data permit us to characterize some interesting dimensions of the problem. Speaking mode was classified as either spontaneous or planned speech, which is typical of straight news reportage from the anchor newsdesk. Dialect/accent was identified as native American English or non-native. Channel fidelity was subjectively characterized as high, low, or intermediate. The presence and subjective level was indicated for three background sources – music, noise, and speech from secondary speakers.

We looked at several variables as a function of these annotated categories. After decoding the data while constraining it to the correct answer, we had word and phoneme durations available for analysis. We also measured word-level perplexity for each condition using a 20 K-word language model.

Condition	1. % of total	2. phone length (csec)	3. % silence	4. pause length (csec)	5. % short phones	6. word length (phones)	7. 3-gram PP	8. effective PP
1. prepared	18	7.4	11	26	11	4.1	306	306
2. spontaneous	21	7.3	15	26	20	3.4	232	553
3. low fidelity	15	8.5	21	31	13	3.6	363	577
4. music	8	7.7	15	21	13	3.9	344	341
5. noise	14	7.6	14	24	11	4.0	495	460
6. non-native	8	7.6	10	19	11	4.0	265	223
7. mixed	16	8.0	23	29	14	3.4	305	639

Table 1: Characteristics of Broadcast News data.

In table 1, we show these measures as a function of the labeled features of the data. The seven conditions shown are mutually exclusive. The prepared and spontaneous speech includes only native speakers, on high fidelity channels with no competing background. The low fidelity condition is composed primarily of telephone and other reduced bandwidth or degraded channels. It contains only native speakers, and either prepared or spontaneous speech, but no background sources. The music, noise, and non-native conditions contain either prepared or spontaneous speech over a high-fidelity channel. Music and noise segments have only native speakers while non-native segments have no background. About 16% of the data falls outside of any of these seven conditions. This data, indicated as *mixed* in table 1, roughly divides in half along the dimensions of native/non-native, planned/spontaneous, high/low fidelity, and clean/background conditions, in various combinations.

The phoneme durations shown in column (2) expose the low fidelity (telephone) condition as containing significantly slower speech than any other condition. The mixed condition, which is half low fidelity data, shows a proportionately lowered speaking rate. The other conditions all have about the same average phone length.

The percent of silence in the data stream and the durations of those pauses, shown in the 3rd and 4th columns of the table, reflect the degree of hesitation in the speech. The percentage of silence is surprisingly high in the low fidelity and mixed conditions – more than 20% of the data in these conditions. The average durations of pauses in these two conditions are also elevated. Once again, speech over the telephone stands out from the rest. Non-native speakers are also exceptional in this feature, exhibiting the lowest average duration and percentage of silences.

The 5th column shows the percentage of phonemes that occurred at the minimum duration permitted by our recognizer (30 msec). The spontaneous condition is singled out here as the most highly coarticulated with a striking 20% of the phonemes realized at the minimum duration. In fact, all categories show higher short-phone counts than we expected. Moreover, we found that most of these short phones occur in groups of adjacent phones. This may well be an indication of a structural problem in our model and needs additional study.

Average word length roughly indicates the acoustic confusability of task since shorter words are harder to recognize in general. The average word lengths (in phones) shown in column (6) indicate that the spontaneous, low fidelity, and mixed conditions are the most difficult. All three of these conditions are distinguished by a high proportion of spontaneous speech. This appears to confirm the intuitive assumption that, as speech becomes more fluent, shorter words are used with greater frequency. This effect shows up again in the word-level perplexities (PP) listed in column (7) where spontaneous speech exhibits dramatically lower perplexity than any other condition.

We have observed, in several cases, that perplexity combined with average word length yields a reasonably good prediction of performance for a given system evaluated across domains in which the average word length differs. We exponentiate the PP of one domain by the ratio of word lengths, giving the *effective* perplexity of the other. Typically, we've found that recognition performance is proportional to $PP^{1/2}$ for a given domain and acoustic model. Using the prepared speech PP as the baseline, we show the effective PP in column (8) for each condition. Now the PP distribution looks very different. By this measure, recognition accuracy should be lowest for the spontaneous, low fidelity, and mixed conditions. This prediction was born out in an experiment described below.

4. EXPERIMENTAL RESULTS

Initially, we attempted to construct condition-specific acoustic models under the assumption that specific solutions would be more powerful than general ones. In particular, we were interested in working on the telephone data with a model trained on reduced-bandwidth data. As we had done on the WSJ corpus in [2], we band-limited the training data and retrained the SI HMMs. In our WSJ work, we had stereo recordings available from a high-quality wide-band microphone and a telephone handset. Using a

wide-band model to recognize the narrow-band data degraded by 340% compared to matched wide-band training and test. Simply bandlimiting the training data reduced that degradation by half. An additional small gain was achieved by adapting the narrow-band model to the test.

In contrast, for the telephone data in the Broadcast News corpus, we achieved only a 5% improvement for bandlimiting and adaptation compared to the wide-band on narrow-band baseline. Furthermore, this experiment assumed perfect knowledge of the telephone data segmentation and classification, so the gain would be even smaller on the real problem. Such a small gain calls into question the strategy of constructing condition-specific models since they introduce considerable additional complexity and size into the recognition system. In response, we have refocused our attention on more general methods that have the potential to improve all conditions. SNR-dependent CMS and SAT are examples of such globally applicable techniques. In a similar vein, we believe that GD models are not worth the effort any more and we intend to begin working with gender-independent adapted models in the near future.

In table 2, we show our final development PE test result before the November 1996 Hub-4 evaluation, broken out by condition. Recall that the PE test removes the necessity to do preliminary segmentation and classification by providing the system with segment boundaries and the condition identities for each segment (each segment belongs to only one condition class). We made no use of the segment class labels, however.

Condition	SI WER	SAT WER	relative gain
1. prepared	14.9	13.5	9.4
2. spontaneous	33.9	32.7	3.5
3. low fidelity	44.3	39.7	10.4
4. music	27.8	25.1	9.7
5. noise	23.2	21.1	9.1
6. non-native	26.4	23.2	12.1
7. mixed	53.0	48.3	8.9
OVERALL	31.9	29.4	7.8

Table 2: Word Error Rate by condition.

For this experiment, we used a 45 K-word lexicon that covered 99.1% of the test. The LM was estimated from 430 M-words of training from broadcast and newspaper sources. The results in table 2 show gains in each condition for unsupervised adaptation to the test, with SAT HMMs. This 8% overall gain for adaptation is considerably smaller that the typical 12-15% improvement that we've observed on other corpora (WSJ and Switchboard) and on the 1995 Hub-4 test. The gain for the low fidelity / telephone condition, however, is comparable to the improvement that we observed with the, condition-specific, narrow-band model described above. This result amplifies the lesson that general techniques should be the primary focus of our work.

At the time of this writing, our UE (Unpartitioned Evaluation) result was relatively 7.5% worse than the PE result above. This degradation can be due to our segmentation and classification components and/or due to the additional non-speech material occurring between the PE segments, which are included in the UE test.

We calibrated our progress from 1995 on the broadcast news problem by testing on the same *Marketplace* episode with our 1995 system and our current system. With last year's adapted-SI PTM-only system, we got a WER of 31.3% in the PE testing paradigm. Using the system described here, we achieved 18.9% PE WER on the same episode.

Acknowledgement

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. Acero, A., X. Huang, "Augmented Cepstral Normalization for Robust Speech Recognition", *Proceedings of IEEE Automatic Speech Recognition Workshop*, Snowbird UT, Dec. 1995, pp. 146-147.
2. Anastasakos, T., F. Kubala, J. Makhoul, R. Schwartz, "Adaptation to New Microphones Using Tied-Mixture Normalization", *Proceedings of ICASSP-94*, Adelaide, South Australia, Apr. 1994, vol. 1, pp. 433-436.
3. Anastasakos, T., J. McDonough, R. Schwartz, "A Compact Model for Speaker-Adaptive Training", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
4. Gish, H., M. Siu, R. Rolicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings of ICASSP-91*, Toronto, Canada, May 1991, vol. 1, pp. 701-704.
5. Kubala, F., T. Anastasakos, H. Jin, L. Nguyen, R. Schwartz, "Transcribing Radio News", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
6. Leggetter, C. J., P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 110-115.
7. Nguyen, L., T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 77-81.
8. Schwartz, R., S. Austin, "A Comparison of Several Approximate Algorithms For Finding Multiple (N-Best) Sentence Hypothesis", *Proceedings of ICASSP-91*, Toronto, Canada, May 1991, vol. 1, pp. 701-704.