# Transcribing Radio News

*Francis Kubala,*
*Tasos Anastasakos, Hubert Jin, Long Nguyen, Richard Schwartz*

BBN Systems and Technologies
70 Fawcett Street
Cambridge MA 02138

## ABSTRACT

We have recently extended the capabilities of BBN's large vocabulary discrete-utterance speech recognition system (BYBLOS) to operate on raw audio recordings of radio news programming. The recordings are given to the system as large monolithic waveforms without any additional side-information. Our goal is to transcribe all speech in the input with the highest accuracy possible. The problem is very challenging because radio news programming has frequent changes in speaker, speaking style, dialect, accent, topic, channel, and environmental conditions. Furthermore, the monolithic input presents new problems for recognition algorithms and language models since all useful boundaries (such as speaker turns or sentence ends) are unknown.

## 1. Introduction

The 1995 DARPA Hub-4 test introduced an entirely new test paradigm for speech recognition research, involving the transcription of studio recordings from radio news programs. In this test, long-duration monolithic recordings are given to the system without any side-information.

The input varies in almost every conceivable way. Speaking styles range from carefully read monologues to free conversation and dramatizations. Some speakers have regional dialects or non-native accents. Topics change unpredictably even though the broad domain is centered around business news. There are frequent bandwidth changes between studio and telephone channels. Background music and noise are common. Furthermore, the large monolithic input presents new problems for recognition algorithms and language models since all useful boundaries (such as speaker turns or sentence ends) are unknown.

Since our existing large vocabulary system, BYBLOS, had been developed primarily in the context of a discrete-utterance dictation task, with known speaker session boundaries and constant channel conditions, there were many new logistical and technical problems to address before this type of data could be handled at all.

In the next section, we describe the preliminary system used in the 1995 test. Results and conclusions from diagnostic experiments are presented in section 3.

## 2. 1995 Dry-run System

The Hub-4 test embodied a show-dependent or show-adaptive paradigm since test and training data came from multiple broadcasts of the same radio show, namely, Marketplace (MP) We chose to utilize some, but not all of the show-dependent knowledge at our disposal. Specifically, we did not deliberately add words from the MP training data to our recognition lexicon. We did, however, add the MP transcriptions (50K words) to our large language training corpus (213M words). Finally, we chose to study supervised speaker adaptation for one single speaker in the training data, the primary anchor speaker *David Brancaccio*, who had the most training data available.

Processing stages of the dry-run system were organized as follows.

- Training Phase
  1. Segment training data by hand.
  2. Classify each segment into male, female, or anchor speaker.
  3. Adapt a seed model to each class (supervised).
- Testing Phase
  1. Automatically segment and classify the test data.
  2. Recognize each segment with a coarse model.
  3. Chop long segments at silence locations.
  4. Recognize again with a detailed model.

In the dry-run system, we made several simplifying assumptions for expediency's sake which should not be construed as desirable for a broadcast news transcription system. For example, we made no attempt to deal with telephone bandwidth data nor did we try unsupervised adaptation on the test, even though our past experience indicates these will enhance performance significantly [4, 5].

### 2.1. Acoustic Modeling

We used gender-dependent models, estimated from the WSJ SI-284 training corpus, as the seed models for adaptation. The HMM densities were organized as Phoneme-Tied Mixtures (PTM) as described in [3]. In this system, each of the 46 phonemes shares a mixture of 256 Gaussian kernel densities. Thus, each gender model contains about 12K Gaussians in total. The center phone of each triphone ties it to the appropriate set of Gaussians. The triphones also share a clustered set of about 25K mixture weights. We chose the PTM system because it is very stable on the highly variable MP data and it is easy to adapt.

Six of the ten MP training shows were hand-marked with speaker-change boundaries and partitioned into four groups: male, female, anchor speaker and pure music/noise.

We adapted the WSJ gender models to each gender and the anchor class, using Maximum-Likelihood Linear Regression (MLLR) similar to the approach described in [2].

The number of regression classes are selected automatically by clustering the HMM Gaussian mixture components and arranging the clusters in a hierarchical tree structure. The final number of regression classes is determined by choosing all tree nodes whose Gaussian components are estimated from sufficient adaptation data. A matrix is constructed to transform the components tied to each selected tree node. Components that are derived from the allophones of silence are withheld from the clustering procedure altogether and transformed separately from all other models of the system.

We then construct 3 *channel-adapted* HMMs for males, females, and the anchor by transforming the gender seed models. Since the anchor transformation is estimated solely from anchor speech, the resulting model is also *speaker-adapted*.

We also created Gaussian mixture classifiers from the four groups of training data in order to identify segments of test data during recognition. This permits each segment to be dispatched to the appropriately adapted HMM or to be rejected as a music or noise segment if it contained no speech.

The classifiers were simple Gaussian mixtures whose components were estimated from 25 second, disjoint samples of data from the given class. On the training data, the classifier error was less than 3%, mostly due to confusions between segments of pure music and those containing both music and speech.

## 2.2. Acoustic Segmentation

We devised a simple method to chunk the monolithic input waveforms into segments having homogeneous acoustic characteristics. The method relies on a robust generalized likelihood ratio test employed in speaker-change detection work [1] to measure a distance between two speech segments.

Two Gaussian classifiers are repeatedly estimated from adjacent windows of analysed speech as the windows are shifted along the input, frame by frame. At each position in the input, we can determine the likelihood that the two Gaussians are generated from the same distribution. We then hypothesize a speaker/channel boundary at any position for which the segment distance measure peaks.

Figure 1 illustrates the behavior of the segmenting operation. The noisy trace is the value of the distance measure computed at each frame position in a 214 second segment containing 9 speaker turns. The smooth trace shows the value of a time varying threshold used to identify peaks that signify segment boundaries. It is computed as a function of the local mean and variance of the distance trace. The vertical lines show the locations of the putative hits, all of which are correct for this segment. There is one peak that is missed around frame number 15000, which separates a pure noise segment from a non-native speaker segment.

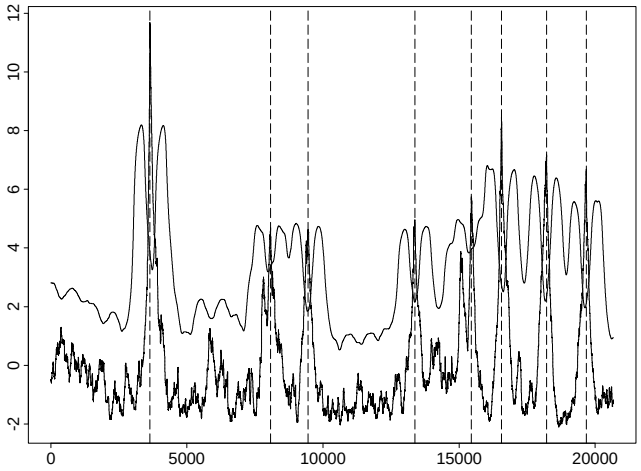


Figure 1: Putative segment boundaries with a 4 second data window.

We combined the putative hits from two passes of the segmenter using data window durations of 3 and 4 seconds. This enabled us to resolve some segments as short as two seconds long. Overall, this segmentation strategy accurately located nearly 80% of the true speaker-change boundaries when measured on hand-marked development test data. To achieve this hit rate, however, the segmenter hypothesized nearly 60% false alarms, primarily within the segments containing both speech and music or noise.

After locating the segment boundaries in the test, each segment is classified and passed to the matching class- or speaker-adapted HMM.

## 2.3. Linguistic Segmentation

After acoustically segmenting the input, many long duration segments remained, and these caused the detailed lattice recognition pass to grow very large. We chopped these long segments at hypothesized silences located by a preliminary recognition pass.

We studied the effect of chopping on perplexity (PP) as shown in table 1. We observe that chopping segments at the longer pauses reduces PP somewhat compared to segments cut at the true speaker turn boundaries. But more importantly, table 1 shows that chopping at true sentence boundaries is much better than either, reducing 3-gram PP by more than 20%. These observations led us to try to make linguistically-motivated segment cuts by using the language model (LM) to guide their placement. We modified the LM to allow sentence boundaries after every word. Then we chopped the segments at those silence locations with the highest likelihood of a sentence boundary within a duration window that varied as a function of segment length and boundary likelihood.

We also tried the simpler expedient of using the output of the preliminary recognition pass with the baseline (unmodified)

| Segment Boundaries | 2-gram PP | 3-gram PP |
|---|---|---|
| Speaker Turns | 302 | 199 |
| Long Pauses | 275 | 193 |
| Sentences | 249 | 157 |

Table 1: Effect of segment boundary locations on perplexity (PP).

LM. Segments were chopped at the longest duration silences in output that occurred within a fixed maximum separation limit of 10 seconds, typically.

In recognition experiments, both methods performed equivalently, so we used the simpler one. Moreover, neither method degraded performance relative to the unchopped segments.

After chopping, no unduly long segments remained, and we proceeded to recognize each segment with a more detailed lattice recognition pass.

## 2.4. Language Model

Transcriptions from 10 Marketplace shows, containing 50K words, were available for LM training. We studied the effect of combining this tiny show-dependent sample with a background corpus of over 200M words from other broadcast news and newspaper sources, and concluded that the show-dependent and background data can simply be combined with equal weight.

The background LM corpus was composed of 45M words from 1992-94 editions of the Wall Street Journal, 46M words from 1994-95 editions of several major North American newspapers, and 111M words from 1992-95 editions of numerous news and public affairs programs from television and radio broadcasts.

We defined the lexicon as the 45K most frequent words in the background corpus. Lexical coverage on the Marketplace test was 98.6%.

The perplexity of the LM, estimated from the background corpus only, was 260. This was reduced to 198 with the addition of the show-dependent Marketplace data, which ultimately resulted in a 5% reduction in word error rate.

# 3. Experimental Results

We conducted numerous diagnostic experiments in the course of preparing for the Hub-4 test and continued to do so afterwards. In this section, we present results from the most informative of these experiments.

## 3.1. Adaptation to Channel

It is well known that subtracting the cepstrum mean from the input is a simple and effective procedure for removing linear differences between channels [4]. We had unintentionally omitted this step in our dry-run system in which we adapted the seed model, estimated from WSJ acoustics, directly to the Marketplace data for each of the three prior classes. Since our MLLR adaptation procedure aligns the data to the model to estimate its transformations, it is likely that cepstrum normalization could improve the adaptation by improving the quality of the alignment.

| Model | norm | male | female | anchor |
|---|---|---|---|---|
| 1) WSJ Seed | no | 41.5 | 27.8 | 24.4 |
| 2) WSJ Seed | yes | 37.3 | 24.8 | 23.4 |
| 3) Adapted | no | 35.6 | 21.1 | 17.2 |
| 4) Adapted | yes | 31.4 | 18.8 | 16.6 |

Table 2: Effect of cepstrum mean normalization and MLLR adaptation on word error rate.

In table 2, we show results on development test data (2 complete shows) that was hand-segmented at true speaker turns with each segment correctly classified into male, female or anchor speaker. Segments containing only noise or music were omitted. This represents the ideal case of perfect segmentation, classification, and noise rejection.

Comparing conditions 1 and 2, the results show that cepstrum normalization yields a 10% relative improvement for the gender models. There is much less improvement for the anchor, but this is reasonable since there is little channel variation in samples of his speech.

Conditions 1 and 3 show that supervised MLLR channel-adaptation alone, as was done in the dry-run, was giving a 14–24% improvement for the gender classes. Here, the gain for the anchor is larger (29%), and this is also reasonable since the adaptation is to a single speaker.

Using normalization and adaptation together, the relative improvement between conditions 1 and 4 increased to 24–32% for males and females respectively, and 32% for the anchor. MLLR channel-adaptation appears very capable of removing channel differences and it is itself benefitted significantly by normalizing the cepstrum first. The MLLR procedure also appears to produce effective speaker-adapted HMMs.

## 3.2. Adaptation to Test

Given the great variety of conditions present in the MP data, we assumed that unsupervised adaptation on the test data would yield significant additional gains to the already large ones achieved by channel adaptation on the training.

For the three conditions shown in table 3, true segment boundaries were known, but classification was automatic. Supervised adaptation to the three classes of gender and anchor speaker was done for all three conditions. Condition 1 is the baseline which was adapted only on the training data.

Condition 2 reveals that there was virtually no gain for adapting to the 3 broad classes in the test. It appears that adapting to an ensemble of speakers is not effective once the channel differences have been removed as much as possible. Furthermore, the result for the anchor speaker suggests that

| Adaptation to Test | male | female | anchor |
|---|---|---|---|
| 1) None | 30.4 | 21.4 | 14.9 |
| 2) On 3 Classes in Test | 30.1 | 20.7 | 14.7 |
| 3) On Each Speaker in Test | 27.8 | 17.5 | 14.8 |

Table 3: Effect of unsupervised adaptation on the test data.

the supervised speaker-specific adaptation had already saturated on the large amount of training data available for this speaker.

Condition 3 represents the ideal case in which every speaker in the test is perfectly segregated from the rest for unsupervised adaptation to the test data. It shows that today we can expect something approaching an additional 8–15% improvement for adaptation to unknown speakers in the test if we can accurately segment and cluster them automatically. This may be a promising area for further investigation in the future.

## 3.3. Segmenter Performance

We have also evaluated the performance of our acoustic segmenter. In table 4, we compare two results that differ only in the use of the automatic segmenter. In both conditions, the segment classification is automatic, as is the music and noise rejection. Here we see that the degradation due to our

| Segments | male | female | anchor |
|---|---|---|---|
| 1) Known | 31.5 | 21.4 | 16.2 |
| 2) Automatic | 37.5 | 27.2 | 19.6 |

Table 4: Performance of the automatic segmenter.

automatic segmenter is 21%.

In order to understand the cause of such a large degradation, we performed another experiment on the same test set, but starting with true speaker-turn boundaries, correct classification, and perfect music and noise rejection. We then compared two methods of automatic segmenting *within* the known speaker-turns. The first condition in table 5 is the

| Intra-Turn Segments | male | female | anchor |
|---|---|---|---|
| 1) None | 30.3 | 19.7 | 14.8 |
| 2) Acoustic | 36.4 | 24.8 | 22.2 |
| 3) Linguistic | 30.5 | 19.2 | 15.2 |

Table 5: Comparison of two segmenting methods within true speaker-turns.

baseline performance for known speaker-turn boundaries and no further intra-turn segmentation. This condition can be compared to the known segment condition in table 4 to illustrate the small degradation incurred from automatic classification and noise rejection mentioned above.

Condition 2 in table 5, shows the performance after applying the acoustically-based segmenter to the known speaker-turn segments. This operation preserves the true turn boundaries but reproduces all the spurious false boundaries that would result from a fair segmenting of the entire monolithic input without known turns. Comparing conditions 1 and 2 in table 5, we discover that nearly all of the 21% degradation introduced by our segmenter is due to the *intra-turn* segments. This is an important piece of information because we have already discovered a method of segmenting within a turn that does not degrade performance.

As noted in section 2.3, we found that it was good enough to simply chop segments at the longest duration silences in output occurring within a fixed maximum separation limit. Typically, this limit was set at 10 seconds.

In condition 3, we show that using this approach, we can segment within the true speaker turns almost at will without degrading the performance relative to the baseline. Note that segmenting linguistically into 10 second chunks resulted in more intra-turn segments than when using our acoustic approach. This implies that we should be able to find a non-degrading, fully automatic method of segmenting the raw input as well. Clearly, this is an area in which our immediate future work will concentrate.

## Acknowledgements

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. Gish, H., M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, Oct. 1994, pp. 18-32.
2. Leggetter, C. J., P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of the Spoken Language Systems Technology Workshop*, Jan. 1995, pp. 110-115.
3. 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*, Jan. 1995, pp. 77-81.
4. Anastasakos, T., F. Kubala, J. Makhoul, R. Schwartz, "Adaptation to New Microphones Using Tied-Mixture Normalization", *Proceedings of ICASSP-94*, Apr. 1994, vol. 1, pp. 433-436.
5. Zavaliagkos, G., R. Schwartz, J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition", *Proceedings of ICASSP-95*, May 1995, vol. 1, pp. 676-679.