ADVANCES IN TRANSCRIPTION OF BROADCAST NEWS

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

BBN Systems and Technologies, Cambridge MA 02138
†Northeastern University, Boston MA 02115

ABSTRACT

In this paper, we describe our recent work in automatic transcription of broadcast news programming from radio and television. This is a very challenging recognition problem 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. We have chosen instead to rely on condition-independent models and adaptive algorithms to deal with this highly variable data. In addition, we have developed effective techniques to automatically segment the input waveform and cluster the segments into data sets containing similar speakers and conditions to support unsupervised adaptation on the test. Using this general approach, we achieved the best overall word error rate of 31.8% on the 1996 DARPA Hub-4 Unpartitioned Evaluation.

1. INTRODUCTION

Automatic transcription of broadcast news programming from radio and television is a very challenging recognition problem because of the frequent and unpredictable changes that occur in speaker, speaking style, topic, channel, and background conditions. For example, wide and narrow band signals are often interleaved within a conversation between a studio anchor and a caller. Also, music and ambient noise are frequently present in the audio track. 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. But doing so requires that separate condition-specific models be created and dispatched to appropriate segments of test data. This implies additional computation and system complexity and requires an accurate procedure for jointly segmenting and classifying the data into the known condition types. None of this is very appealing for practical reasons.

We would much prefer to employ a single seed model that can be quickly adapted to any condition found in the test data. If we could do this, a general transcription system could be organized into five logical stages:
1. Segment the data
2. Cluster the segments
3. Decode with a speaker-independent model to get transcriptions for adaptation
4. Adapt the model to each cluster
5. Decode with adapted models to produce the final answer

Note that this system needs no prior knowledge about the specific conditions contained in the test. However, it does require that we be able to automatically segment the data into small enough chunks to allow them to be automatically clustered into acoustically similar sets.

We have recently built such a general transcription system and found it to be competitive with more complicated approaches that attempt to model specific conditions expected in the test. In the following section, we describe an automatic acoustic segmentation algorithm that is effective on broadcast news. In section 3, we introduce a fully automatic blind speaker-clustering algorithm. In section 4, we demonstrate our successful *condition-independent* approach to the transcription of broadcast news.

2. ACOUSTIC SEGMENTATION

Since our acoustic models are gender-dependent, we need to cut the large monolithic input waveforms at gender-change boundaries and then classify the resulting segments as male or female. We also need to break the long segments into shorter ones for computational efficiency in our decoder. We accomplished both of these with a dual-gender, context-independent phoneme decoder. That is, male and female HMMs were decoded in parallel in a single pass over the data. The model incorporated a bigram on the phone transitions and permitted a gender transition anywhere. The pause model was common to both genders. The output of this decode was a sequence of pauses and gender-tagged phones with accompanying time-stamps.

The desired segments were then produced by cutting the input at pause locations and gender changes indicated in the phone transcription. Boundary decisions were guided by several heuristics. No speech segment was permitted to be shorter than 2 seconds. And boundaries were not located within pauses shorter

than 150 milliseconds, unless the hypothesized segment grew beyond about 10 seconds. These constraints resulted in an average segment length of 8 seconds over the test data.

This simple model proved to work very well. In the 1996 DARPA Hub-4 Unpartitioned Evaluation, this segmentation strategy was only slightly worse (5% relative) than the idealized test using known speaker/condition change boundaries. Moreover, the segmenter effectively rejected segments of pure music or noise by labeling them as pauses (non-speech intervals were included in the pause training). The gender classification was very stable - less than 1% of the data was misclassified for gender.

3. SPEAKER CLUSTERING

The goal of speaker clustering is to group segments from the same speaker and condition together to improve the effectiveness of unsupervised adaptation. We have developed a fully automatic blind clustering algorithm [4] to accomplish this. In practice, we regard a speaker as a generic concept which really means speaker with channel and background condition. Thus, speech from the same physical speaker with significantly different channel and/or background conditions should be treated as speech from two different speakers in speaker clustering. Conversely, we may want to classify speech from two speakers in the same cluster if their acoustic characteristics are not significantly different. In any case, the ultimate effectiveness of speaker clustering in this transcription task is measured by how well the clusters behave under unsupervised adaptation.

3.1. Clustering Algorithm

Consider that we have a collection of segments $S = \{s_1, s_2, ..., s_n\}$, and each s_i represents a sequence of spectral feature vectors, i.e. the Cepstral vectors in our implementation. Speaker clustering attempts to find a partition $P = \{p_1, p_2, ..., p_k\}$ of S such that each p_j contains only segments from the same speaker/condition and also speech segments from this speaker are classified into p_j only. Assume that the vectors in each of these sequences can be modeled as coming from a multivariate Gaussian distribution and that the vectors are statistically independent. A good clustering solution should have relatively small dispersion within clusters. The within-cluster dispersion [6] is defined as

$$W = \sum_{j=1}^{k} N_j * \Sigma_j$$

where Σ_j is the covariance matrix and N_j is the total number of feature vectors in cluster p_j. There are several good clustering criteria [2]. We prefer to use the determinant of W to measure the goodness of speaker clustering. Some penalty against too many clusters must be used to avoid the degenerate solution of one segment per cluster.

There are three stages in the clustering algorithm.
1. A distance matrix is created by computing the distance between every pair of segments. We use an acoustic segment distance measure from earlier work in speaker-identification [3]. We scale the distance by a parameter that favors segments occurring close in time, since nearby segments are more likely to have come from the same speaker.
2. A cluster dendrogram is constructed from the distance matrix using the routine, *hclust*, from the Splus statistical software package. For any given number k, this cluster tree can be pruned subsequently with only k leaves left, which corresponds to the k tightest clusters in the solution.
3. We then choose the best solution for k by minimizing the penalized determinant of the within-cluster dispersion matrix.

Figure 1 illustrates how the penalty interacts with dispersion to select the desired solution.

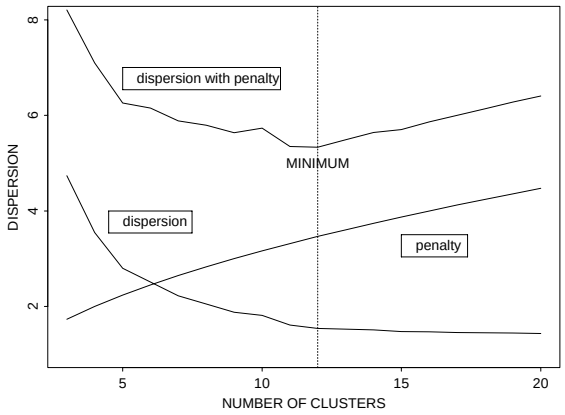

Figure 1: Model selection based on penalty against the number of clusters.

3.2. Clustering Experiments

We evaluated the effectiveness of the clustering algorithm on the 1995 Hub-4 development data consisting of half-hour segments of the program, *Marketplace*. In table 1, we compare a Speaker-Independent (SI) baseline against four adapted results which use different partitions of the test data for unsupervised adaptation. In experiment (2), the unsupervised adaptation is performed on each speaker turn individually and, as expected, this performs better than the SI baseline. In (3), the adaptation is done on ideal clusters composed of whole speaker turns and this performs marginally better than adapting to each turn individually. Experiments (2) and (3) are idealized cases from hand marked data for comparison to the

Experiment	WER
1. SI baseline - no adaptation	28.2
2. Individual speaker turns (ideal)	25.6
3. Clustered speaker turns (ideal)	25.0
4. Clustered segments	25.8
5. Clustered segments with adjacency	24.8

Table 1: Clustering results on Marketplace episode, 940523, of Hub-4 1995.

next two experiments in the table. In (4), we show the performance when the adaptation is performed on automatically segmented and clustered data, but without the adjacency bias on the distance measure. The final experiment (5) includes the adjacency bias and shows that our automatic speaker clustering algorithm improves unsupervised adaptation as much as the hand labeled ideal case. This experiment was repeated on the six episodes of the 1996 Hub-4 development test data with the same conclusion.

4. CONDITION-INDEPENDENT MODELS

There are many different speaking and recording conditions present in broadcast news programming. For the 1996 Hub-4 test, all conditions were collapsed into six categories or focus (F) conditions determined by five binary attributes: spontanteous speech, degraded channel, music, noise, and non-native accent. Many of the 32 possible combinations actually occurred in the data. It is tempting to develop specific solutions for each of these F-conditions but the cost for doing so is very high in terms of system complexity and dilution of research effort. It is much more desirable to have a general adaptive system that makes no assumptions about the specific conditions of the data. To determine whether we could achieve such a system with today's speech recognition technology, we needed to compare condition-independent models against two competing strategies: condition-specific training and supervised adaptation to condition.

4.1. Condition-Specific Approaches

One approach to dealing with the many conditions is to train a separate model for each one. In particular, one might suspect that including noisy degraded data in a model for the clean data could be detrimental. To test this supposition, we made a condition-specific model for "clean" speech from all of the data marked as F0 (clean, wideband, read, native accent) and F1 (+spontaneous). Broadcast news data divides roughly in half between clean and degraded data. In Table 2 we measure the effect of training the model with only F0 and F1 data (i.e., discarding the other 50% of the data). In both conditions, we did not use adaptation on the test data.

	Training Data	
Condition	All	F0,F1
F0. prepared	19.1	20.0
F1. spontaneous	42.8	42.7

Table 2: Error rate on clean wideband speech (F0 and F1) when training on all speech vs. only on F0 and F1 speech.

These results show that it is better to include the data from other conditions than it is to discard it in training. In a separate series of experiments, in which we made a narrow-band model specifically for telephone speech, we found that we could improve the performance on the narrow-band data by only a small amount (less than 10% relative). Since the telephone data is only 10% of the test data, the resulting tiny overall improvement did not justify the added system complexity required. So we concluded that condition-specific training offered little for broadcast news.

Another strategy would be to adapt a model trained on all of the speech to each of the marked conditions using supervised adaptation. This only requires training the system once, and quickly produces many condition-specific models. Since the training is supervised and there is substantial training for each condition, the adaptation can be quite detailed. In the following section, we compare this condition-adapted strategy to our desired condition-independent system.

4.2. Comparative Experiments

We propose to use a single model as the basis for adapting to any condition observed in the test data. Commonly, a pooled Speaker-Independent (SI) seed model is used. We have improved on the SI model by removing the characteristics of each training speaker (and condition) in an iterative procedure that we call Speaker-Adaptive Training (SAT) as described in [1]. SAT finds the "compact" model that results in the highest likelihood for all the speakers' data, given their corresponding transformations to the previous model. We are continuing to investigate SAT algorithms and have recently developed more practical methods that can handle thousands of training speakers [5].

In Table 3, we show the results obtained for systems configured with various combinations of adapting the model to the condition, unsupervised speaker-adaptation on the test, and Speaker Adaptive Training (SAT). Results are broken out for each test condition and averaged over all conditions.

Column	1	2	3	4	5
Adapt on Train Adapt on Test	NO NO	YES NO	NO YES	YES YES	SAT YES
models/gender	1	7	1	7	1
F0. prepared	16.6	16.1	15.3	14.9	14.8
F1. spontaneous	39.4	37.8	36.7	35.2	35.1
F2. low fidelity	45.4	44.5	40.1	40.4	40.2
F3. music	32.0	30.8	30.2	29.6	30.2
F4. noise	25.6	24.8	24.1	23.4	25.0
F5. non-native	30.8	31.0	25.9	25.6	23.4
FX. mixed	58.4	57.4	54.8	54.0	53.7
OVERALL	35.2	34.3	32.3	31.7	31.6

Table 3: Word error rate by test condition, for SI, supervised condition adaption, unsupervised adaptation on test, supervised condition adaptation plus unsupervised adaptation on test, and SAT adapted training with adaptation on test.

The SI baseline result is shown in column 1. The baseline uses a single model for all conditions without any adaptation. In the second column, the SI model is adapted to each condition in the training with supervision. This approach shows improvement in each condition but creates 7 different models which must be properly matched to the test segments by some unknown means. For this experiment and those in columns 3 and 4, we chose the correct model-to-condition assignments by hand. In column 3, we observe that unsupervised adaptation to the test is more powerful than supervised adaptation on the training even though it uses only a single seed model. Furthermore, the result in column 4 shows that the two adaptation strategies combine additively, but once again at the cost of creating a model for each condition in the training. In column 5, we use a single SAT model as the seed for unsupervised adaptation and the result is as good as the condition-adapted models of column 4. Moreover, for this system there is no need to determine the specific class of test segment – automatic clustering into blind classes is sufficient.

So at this time, we believe that condition-independent models can achieve state-of-the-art performance on highly variable data such as is found in broadcast news. We do not dispute that further progress could be made on condition-specific modeling, but we believe that the overall gains that can be achieved by these methods cannot justify the costs in terms of system complexity, computation during recogniton, and fragmented research effort.

5. SUMMARY

We have successfully employed a dual-gender context-independent phone model to automatically segment the large monolithic broadcast news input waveforms. Compared to hand segmented data, the performance on automatically segmented data was only 5% worse in relative terms.

We have demonstrated a completely automatic speaker clustering algorithm that is used to group acoustically similar segments together for unsupervised adaptation. We compared the performance of unsupervised adaptation using our clustering algorithm versus using the ideal clustering from known segment conditions and found no significant difference in performance.

Most importantly, we have shown that general adaptive techniques can effectively deal with the extraordinary variability found in broadcast news programming. In training, we remove as much speaker- and condition-specific variability as we can with Speaker Adaptive Training. This results in a single model for all conditions that is better suited for unsupervised adaptation to the test data than the common pooled Speaker-Independent model. Using this *condition-independent* approach, we achieved the best overall word error rate of 31.8% on the 1996 DARPA Hub-4 Unpartitioned Evaluation (UE test).

Acknowledgements

This work was funded under NRad contract N66001-97-D-8501. 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. Anastasakos, T., J. McDonough, R. Schwartz, "A Compact Model for Speaker-Adaptive Training", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
2. Everitt, B., *Cluster Analysis*, Halsted Press, New York, 1980, pp. 24-35.
3. 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.
4. Jin, H., F. Kubala, R. Schwartz, "Automatic Speaker Clustering", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
5. Matsoukas, S., R. Schwartz, H. Jin, L. Nguyen, "Practical Implementations of Speaker-Adaptive Training", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
6. Wilks, S., *Mathematical Statistics,*, Wiley and Sons, New York, 1962.