Automatic Segmentation, Classification and Clustering of Broadcast News Audio

Matthew A. Siegler, Uday Jain, Bhiksha Raj, Richard M. Stern
ECE Department - Speech Group
Carnegie Mellon University
Pittsburgh, PA 15213
msiegler@cs.cmu.edu uj@cs.cmu.edu bhiksha@cs.cmu.edu rms@cs.cmu.edu

ABSTRACT

Automatic recognition of broadcast feeds from radio and television sources has been gaining importance recently, especially with the success of systems such as the CMU Informedia system [1]. In this work we describe the problems faced in adapting a system built to recognize one utterance at a time to a task that requires recognition of an entire half hour show. We break the problem into three components: segmentation, classification, and clustering. We show that *a priori* knowledge of acoustic conditions and speakers in the broadcast data is not required for segmentation. The system is able to detect changes in acoustics, recognize previously observed conditions, and use this to pool adaptation data. We also describe a novel application of the Symmetric Kullback-Leibler distance metric that is used as a single solution to both the segmentation and clustering problems. The three components are evaluated through comparisons between the Partitioned and Unpartitioned components of the 1996 ARPA Hub 4 evaluation test set.

1. INTRODUCTION

Most speech recognition tasks to date have dealt with discrete utterances, from a single speaker over a single channel, with labeled or implied beginning and ending points. In such tasks, the problem of determining where the speech region lies is merely one of silence detection. In addition, there is assurance that the speaker and channel will remain fixed over the length of each utterance. For the purposes of compensation and adaptation, speech recognition systems could easily take advantage of these factors.

However, the task of automatic transcription of broadcast news as in the CMU Informedia system [1] is more challenging since there are no explicit cues for changes in speaker and channel. In fact, either may change independently of the other such as, for example, an increase in background noise during an anchor's monologue. In order to transcribe the speech content in audio streams of this nature several new technologies must be developed.

For the CMU evaluation system for the 1996 ARPA Hub 4 task, the problem was divided into three processes: segmentation, classification, and clustering. In each step, the goal was to tune performance to the needs of the evaluation system.

2. THE *KL2* DISTANCE METRIC

Relative Cross Entropy, or the Kullback Leibler (*KL*) distance between two Random Variables *A* and *B* is an information theoretic measure equal to the additional bit rate accrued by encoding random variable *B* with a code that was designed for optimal encoding of *A* [2]. The larger this value, the greater the distance between to PDFs of the two Random Variables. It is formulated in the following way:

$$KL(A;B) = E_A\langle \log(P_A) - \log(P_B)\rangle$$

where $E_A\langle\ \rangle$ is the expectation operation performed with respect to the PDF of A.

Since this expression is not symmetric, it is not strictly a distance metric. We therefore define the *KL2* metric as:

$$KL2(A;B) = KL(A;B) + KL(B;A)$$

When both A and B have Gaussian distributions we obtain:

$$KL2(A;B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (\mu_A - \mu_B)^2\left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right)$$

Once again, the greater this value, the greater the distance between the two PDFs. When *A* and *B* have the same PDF this distance is zero.

3. SEGMENTATION

The ultimate goal of segmentation is to produce a sequence of discrete utterances with particular characteristics remaining constant within each one. The characteristics of choice depend on the overall structure of the recognition system.

In last year's Hub 4 system, recognition performance was enhanced by constructing several different models and compensation schemes for a fixed set of acoustic *classes* [5]. The segmentation algorithm was directed at locating times in the audio stream where there was a change in the acoustic class. To accomplish this, Gaussian models of each class were constructed from training data, and Maximum Likelihood selection of the class for the incoming audio was performed over a sliding window of audio. Each segment was coupled with the models which were designed to best recognize the speech for that class.

A shortcoming of this method was that it required *a priori* selection of the acoustic classes. Since there was no difference in training and testing broadcast programs, this was not difficult.

The 1996 Hub 4 CMU evaluation system has an adaptation component which works best when utterances contain exactly one speaker over an unvarying channel. However, the testing domain includes many different shows, both from television and radio. It would not have been possible to construct a compensation scheme and set of recognition models for every speaker and channel combination in the training data.

To perform segmentation without acoustic classes we employed a similar technique to Gish and Schmidt [4], but used the KL2 distance instead of the Generalized Likelihood Ratio. To accomplish this, means and variances were estimated for a two second window placed at every point in the audio stream. When the KL2 distance between bordering windows reached a local maximum, a new segment boundary was generated.

To evaluate the metric, comparisons of the segment beginning and ending points (segment boundaries) were made between automatic and hand-generated segments on the Hub 4 training corpus. Only comparisons to hand-generated segments that were at least 2 seconds long were made. For these segments, 64% of the boundaries were detected. However, 60% of the automatic segment boundaries were placed within segments, mostly during silences.

4. CLASSIFICATION

In comparison to last year's system, pre-recognition compensation techniques proved to be ineffective in this year's pilot experiments with the development test set. It was decided there was no need to build separate compensation schemes and speech models for a variety of predefined classes. Even so, experiments with the development test set indicated that two sets of models, one trained with full-bandwidth speech and the other with half-bandwidth speech could reduce the word error rate of telephone speech from 72% to 61% on the development test set, for a 15% relative improvement.

In order to automatically classify segments of audio as full or half bandwidth, a pair of Gaussian mixture models was constructed. The Full Bandwidth model was a 16-mixture model trained in a maximum likelihood framework from the F0, F1, F3 and F4 conditions in the H4-**1996** corpus. The Gaussian mixture for the Half Bandwidth model had 8 components and was trained from the H4-**1995** training data. The mixtures trained on this data were found to provide better classification of telephone speech than mixtures trained from H4-1996 training data that had been labelled as F2. We believe this is because many of the segments labelled as F2 in the H4-1996 training corpus were not spoken over telephone lines.

Maximum likelihood selection of the class given a segment of data was used to classify incoming speech. Table 1 shows the performance of automatic classifier on the development test set. The overall classification error, given the amount of full and half bandwidth speech was 2.8%.

	Reference Classification	
Automatic Classification	**F0, F1, F3, F4, F5**	**F2**
Full	98.1%	10.4%
Half	1.9%	89.6%

Table 1. Performance of automatic classifier on the development test set. Show n960715p was excluded because the *F2* data were not telephone bandwidth.

5. CLUSTERING

In the adaptation module, parameters of the recognition system are adjusted to optimize performance for each particular speaker and channel. If the same speaker and channel occur several times in a broadcast, more adaptation material is available. Since the identity of the speaker and channel is unknown for the segments of the incoming broadcast, collecting these segments together requires an unsupervised clustering technique.

To accomplish this, a simple agglomerative clustering method was chosen [3]. The critical element of this clustering technique is the distance metric used to compare the elements, and clusters with each other. Many possibilities are available, including the Mahalanobis distance and the generalized likelihood ratio (*e.g.* [4]).

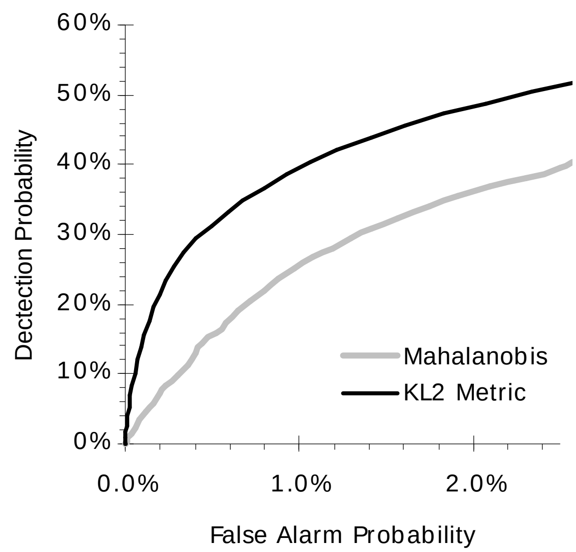

Figure 1. Comparison of performance of two different distance metrics for use in clustering F0 and F1 data in the training corpus.

For the distance metrics, the cepstral features used for speaker clustering were the same as those used for recognition. An utterance was clustered with an existing cluster if it was within a threshold distance, otherwise it was used as a seed for a new cluster. The threshold had to be small enough such that the clusters created were made up of utterances from only one speaker and yet large enough to boost the adaptation performance.

If the cluster submitted to the adaptation module contains more than one speaker or channel, there will be a degradation in performance. Because of this, false alarms are far more costly than missed detections in automatic clustering and therefore only low false alarm probabilities are tolerable.

We compared the clustering decisions of the Mahalanobis and the KL2 distance metrics on the Hub 4 training set. Figure 1 shows the performance of the two metrics on pairwise comparisons of segments in the F0 and F1 components of the training corpus with at least 2 seconds of audio. A *false alarm* occurs if segments from two different speakers are detected as the same. A *detection* occurs if segments from the same speaker are detected as the same. The detection probability is much higher for the KL2 distance metric than it is for the Mahalanobis distance at all false alarm probabilities.

The performance of the automatic clustering, illustrating the trade-off between expected cluster size and false alarm probability is shown in Table 2. As can be seen, the expected cluster size remains relatively fixed over a variety of false alarm probabilities.

KL2 Threshold	False Alarm Probability	Expected Cluster Size (seconds)
0.020	< 0.1%	32.6
0.038	0.2%	36.8
0.066	1.7%	36.6

Table 2. Trade-offs between false alarm probability and expected cluster size for automatic clustering on the entire training corpus.

6. EXPERIMENTS AND RESULTS

To examine the performance of automatic segmentation, classification and clustering, comparisons were drawn between the recognition error rates for the partitioned evaluation (PE) and unpartitioned evaluation (UE) components of the 1996 ARPA Hub 4 Broadcast News evaluation. In the PE, hand-labelled segment boundaries were used to provide discrete utterances. In addition, the focus condition of each segment was used to partition the data so that utterances could be clustered only with others from the same focus condition. For the UE system, automatic segmentation was performed, and no side information was used in the clustering process.

Table 3 shows the recognition error rates, both before and after adaptation. Degradation in performance from PE to UE before adaptation (2.4% relative) is due only to automatic segmentation errors. Degradation after adaptation (4.1% relative) comes from both automatic segmentation and clustering errors. This represents an improvement over last year's evaluation, where there was an 8.5% relative increase in word error rate due to automatic segmentation [5].

	Before Adaptation		After Adaptation	
Focus Condition	**PE**	**UE**	**PE**	**UE**
F0	25.9	26.0	24.5	24.7
F1	32.4	33.5	32.1	33.1
F2	43.2	44.7	38.6	39.1
F3	43.3	48.4	36.6	48.4
F4	45.7	45.0	43.7	42.1
F5	45.8	40.8	36.5	35.5
FX	61.8	62.9	55.8	58.3
All	**36.9**	**37.8**	**34.5**	**35.9**

Table 3. Word error rates for the evaluation test set, before and after unsupervised adaptation.

7. CONCLUSIONS

The Symmetric Kullback-Leibler Distance is an effective distance metric to facilitate the detection of long-term statistical differences in speech signals. When used for locating a change of speaker or channel in the 1996 Hub 4 evaluation, 64% of the hand-labelled segment boundaries in the training set were detected. In clustering segments belonging to the same speakers and channels the Symmetric Kullback-Leibler Distance provided segments with a very low error probability (< 0.1%), that were of reasonably large (33 seconds).

Although not insignificant, it is clear that performance losses due to automatic segmentation and clustering are small, and not a serious obstacle toward improving recognition of broadcast news audio.

REFERENCES

1. Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H., 1995. “Informedia, “Digital Video Library”, *Communications of the ACM*, 38(4):57-58.

2. Cover, T. M. and Thomas, J. A., 1991. *Elements of Information Theory*, New York: John Wiley and Sons.

3. Duda, R. O. and Hart, P. E., 1973.*Pattern Classification and Scene Analysis*, John Wiley and Sons.

4. Gish, H., and Schmidt, N., 1994. “Text-Independent Speaker Identification”, *IEEE Signal Processing Magazine*, Oct. 1994, pp. 18-32.

5. Jain, U., Siegler, M. A., Doh, S.-J., Gouvea, E., Huerta, J., Moreno, P. J., Raj, B., and Stern, R. M., 1996. “Recognition of Continuous Broadcast News With Multiple Unknown Speakers and Environments”, *Proceedings of the 1996 ARPA Speech Recognition Workshop*, Morgan Kaufmann Publishers.

6. Parikh, V., Raj, B., Stern, R. M., 1997. “Speaker Adaptation and Environmental Compensation for the 1996 ARPA Broadcast News Task”, in these proceedings.