

Acoustic Change Detection and Segment Clustering of Two-Way Telephone Conversations

Xin Zhong Mark Clements

Sung Lim

Georgia Institute of Technology
Atlanta, USA

Fast-Talk Communications
Atlanta, USA

{xinzh, clements}@ece.gatech.edu

slim@fast-talk.com

Abstract

We apply the Bayesian information criterion (BIC) to unsupervised segmentation of two-way telephone conversations according to speaker turns, and then proceed to produce homogeneous clusters consisting of the resulting segments. Such clustering allows more accurate feature normalization and model adaption for ASR-related tasks. In contrast to similar processing of broadcast data reported in previous work, we can safely assume there are two distinguishable acoustic environments in a call, but new challenges include a much faster changing rate, variation of speaking style by a talker, and presence of cross-talk and non-meaningful sounds. The algorithm is tested on two-speaker telephone conversations with different genders and via different telephony networks (land-line and cellular). Using the purities of segments and final clusters as the performance measure, the BIC-based algorithm approaches the optimal result without requiring an iterative procedure.

1. Introduction

Recently, a research effort has been devoted to automatic segmentation and clustering of audio databases. Such a task usually has two steps, as depicted in Figure 1. First, without prior acoustic information about the audio document, the system identifies and labels changes in speaker, channel, or other background environments. As in all detection problems, the errors include false alarms and misses, and a popular performance measure is the equal-error-rate (EER). This acoustic change detection (ACD) procedure produces a set of pure audio segments. Applications here include automatic indexing which facilitates information retrieval, as well as allowing an ASR model to be reset at each changing point. Depending on the ultimate goal, this is usually followed by a second step, termed as segment clustering (SC), where homogeneous segments at different locations of the time axis are grouped together. There are two contributing factors to the impurities of final clusters: a segment delivered by the ACD procedure might be impure to begin with, or sometimes it is simply assigned to the wrong cluster. Since smaller clusters generally tend to be more homogeneous, the performance measure must consider both cluster sizes as well as purities. With such a grouping, accurate feature normalization and model adaptation can be applied.

One of the earliest reported studies in this area [1] used the generalized log-likelihood ratio (GLR) as a similarity measure between adjacent acoustic segments, and recorded a changing point when the score fell below a threshold. Subsequent work includes using other measures, such as the Kullback-Leibler and the Bhattacharyya distances, and different frame works, such as GMM-based and model-selection-based approaches. In

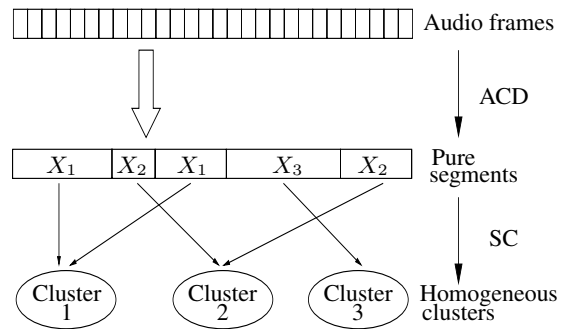


Figure 1: Acoustic change detection and segment clustering.

particular, the dominant method in recent years has been the Bayesian information criterion (BIC), a maximum likelihood model-selection problem with a penalty for model complexity. The BIC was first used in segment clustering of broadcast news (Hub-4 1996 and 1997) [2], where 20 of the final 32 clusters achieved purities close to 1. In [3], an iterative procedure was implemented, where the result of the SC was softly incorporated into the ACD stage to achieve a reduced EER of 18% for the Hub-4 data.

Our goal is to perform acoustic change detection and segment clustering of telephone conversations, which have received little attention thus far. The challenges here include a much faster changing rate, variation of speaking style even by the same speaker, and presence of cross-talk and non-meaningful sounds. The first factor is particularly crucial, as duration of a segment is vital to correct detection of its end points. Segmentation and clustering of AT&T service call recordings was conducted in [4]. A GMM was used to represent each of the two adjacent speech segments, and the GLR, which is similar to the BIC but without a penalty factor for model complexity, was used as the similarity measure. It reported that a minimum segment duration of two seconds was needed to achieve acceptable results as the GLR tends to be variable and inconsistent for small amounts of data. Unlike the broadcast data, however, we can safely assume *a priori* that there are two objectively distinguishable acoustic conditions in telephone conversations, categorized by the two speakers, even though the background noise and speaking styles might vary during a call as well. This provides the “stopping point” for the segment clustering stage which is not available when the input is broadcast news.

The rest of the paper is organized as follows: Section 2 summarizes the Bayesian information criterion; Section 3 de-

tails the ACD and SC procedures; Section 4 establishes a relevant performance measure and reports experimental results; Section 5 provides conclusions.

2. The Bayesian Information Criterion

Given a set of data points, $\mathbf{X} = X_1 X_2 \cdots X_N$, the Bayesian information criterion is formulated as a model selection problem. The BIC score for a model M is defined as:

$$BIC(M) = \log[L(\mathbf{X}, M)] - \frac{\lambda}{2} \cdot \#(M) \cdot \log(N) \quad (1)$$

where the two terms are the log-likelihood and model complexity, respectively ($\#(M)$ represents the number of free parameters in the model). The null hypothesis is that the whole audio segment can be modelled by a multivariate Gaussian process. Alternatively, if there is a changing point at X_B , then the segment is represented by two parts, $[X_1 \dots X_B]$ and $[X_{B+1} \dots X_N]$. Thus the two model candidates are:

$$\begin{aligned} H_0 : & \quad X_1 X_2 \cdots X_N \sim N(\mu, \Sigma) \\ H_1 : & \quad X_1 X_2 \cdots X_B \sim N_1(\mu_1, \Sigma_1); \\ & \quad X_{B+1} \cdots X_N \sim N_2(\mu_2, \Sigma_2) \end{aligned} \quad (2)$$

If the feature set used is of dimension d , each normal distribution requires $\#(M) = [d + \frac{1}{2}d(d+1)]$ free parameters. Therefore, taking only the variances of the distributions into consideration, the difference in BIC scores between the two models is:

$$\begin{aligned} \Delta BIC = N \cdot \log |\Sigma| - B \cdot \log |\Sigma_1| - (N - B) \cdot \log |\Sigma_2| \\ - \frac{\lambda}{2} \cdot [d + \frac{1}{2}d(d+1)] \cdot \log(N) \end{aligned} \quad (3)$$

First derived in the area of statistics, the BIC-based algorithm has recently been used in other speech processing applications such as mixture size-selection and decision tree state-tying. As the number of parameters in a model increases, the likelihood it represents the data also tends to go up. When this number is too large, however, it might result in overtraining. The criterion places a sensible balance between likelihood and complexity to determine the optimal model at any given time instant. For our purpose, the BIC-based algorithm is robust in detecting changes in acoustic conditions, particularly the speakers. In the ACD stage, a local maximum in the ΔBIC plot indicates possible presence of a changing point; inversely in the SC stage, a smaller increase in the score implies greater tendency for two segments to be merged into one cluster. Figure 2 illustrates the ΔBIC of 60 seconds of telephone conversation, where every local peak is a possible speaker turn.

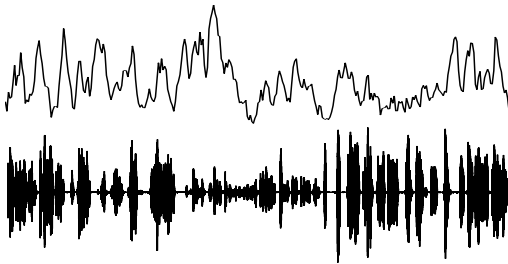


Figure 2: ΔBIC vs. waveform

3. Segmentation and Clustering

We now describe the procedures of acoustic change detection and segment clustering, depicted by Figure 3.

- *Feature Extraction*

The audio stream is broken into frames, each covered by a 32 ms Hamming window, and the frame-rate is 10 ms. The feature set extracted from each frame consist of the first 18 of 24 mel-frequency cepstral coefficients (MFCC), as well as the log-energy term, for a feature-dimension $d = 19$. Unlike traditional ASR, which usually captures the first 12 MFCCs, some high-time cepstral coefficients are also used because they contain speaker-dependent information which serves as a cue for speaker changes. When processing broadcast data, a voice activity detector (VAD) was employed to take out the silent frames. Due to the presence of background noise or other audio signals during a phone call, however, a VAD is not precise in labelling frames without speech. For this reason we send all frames into the ACD procedure.

- *The Criterion-score Calculation*

Similar to previous works, the ΔBIC score of a segment \mathbf{X} is calculated successively along the time axis. The issue here are placement of the window to extract a segment for testing, as well as setting the value for λ .

In [5] and others, the analysis window is of variable length, and adaptively extends depending on the location of the next change detection in an attempt to lower the EER. This is not the performance measure we use (see Section 4). In our work, a 4-second analysis window is used, and the assumed changing point is always at the middle. This provides 200 feature sets for each of the two sub-segments, and the window is shifted 200 ms forward for the next calculation.

The penalty factor λ is set to 1 in the formal definition of the BIC, although better results in ACD are reported when this value varies according to the local characteristics of the signal. Since a definite procedure for adaptively obtaining the optimal λ is not known, it is set to 1 in our work.

- *Peak-picking and Segmentation*

The local peaks of the ΔBIC values are candidates for speaker changing points. Some previously work placed a limit on the minimum time separating successive turns. For broadcast materials, such consideration is reasonable and reduces false alarms. In telephone conversations where interruptions are frequent, this procedure is less reliable. Also, since a significant portion of the false alarms are likely to be eliminated when neighboring segments are assigned to the same clusters, they are less harmful than missing an authentic changing point. For these reasons we consider all local peaks above a certain threshold.

This is the end of the ACD stage which delivers a set of time indexes. Every two adjacent values in this set defines an audio segment ready for the clustering stage.

- *Local/Global Clustering*

The segments are first clustered locally in a time-span of 30 to 40 seconds, depending on the number of segments it contains. It is then followed by global clustering where all portions of the conversation must be assigned until only two clusters remain. The reason for this two-step approach is as follows: the factors for correct clustering include the segment purities, which are very high (see Section 4), as well as the segment lengths, which are quite short (about one second on average). Since the speaking styles and background environment tend to

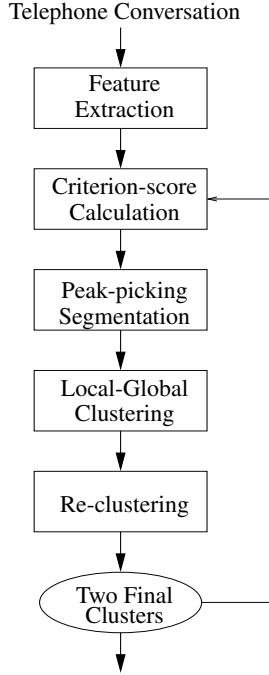


Figure 3: The ACD & SC procedures.

remain the same within a certain time span, such local clustering allows similar segments to be merged first. This produces intermediate clusters of larger sizes before merging them with segments/clusters further away along the time axis where, due to changes in other acoustic conditions, the similarity measure might have been weakened even among utterances by the same speaker. At any stage, clusters C_i and C_j are merged to produce a new cluster, where

$$[i, j] = \min_{i,j} [\omega_i \omega_j \cdot \Delta BIC(C_i, C_j)] \quad (4)$$

with ω_i being a weighting factor proportional to cluster size measured by number of frames. This local clustering step is stopped at a pre-determined number of clusters (greater than two), because the BIC-algorithm (or any measure) becomes less reliable as merging continues and impurities increase.

The intermediate clusters are then merged together on the global scale, also according to Equation 4. One of the challenges in ACD/SC of broadcast news is that the number of speakers presented is never known *a priori*. Here, we know the final result should have two clusters. The most straightforward method is to simply allow this procedure to continue until there are two clusters remaining. At the stage where there are three clusters left, however, there must be one which is of the highest impurity, and we can do better than simply augmenting it to another cluster. Our procedure first identifies this cluster (by a weighted BIC-score of high-energy frames), then breaks it up by assigning all of its member segments to the other two according to the ΔBIC scores. The improvement of this procedure is reported in the next section.

- *Re-clustering*

Correctly assigning a segment to a cluster also depends on purity of the latter, which in turn is inversely proportional to its size. After the system generates two clusters from the previous stage, we merge segments (again according to Equation 4)

within each until the largest sub-cluster exceeds a certain size, say 30 seconds. These two sub-clusters are still large enough, but each is now of higher purity, therefore re-assigning all other segments to them achieves better results.

- *Iteration*

We now have two final clusters, where each ideally contains utterances by a single speaker. This information can be incorporated back to the BIC-scoring stage, either by a hard or soft decision [3]. Instead of simply computing the ΔBIC score, the calculation for criterion-score is modified as:

$$(1 - \alpha) \cdot F + \alpha \cdot \frac{\sigma_F}{\sigma_{\Delta BIC}} (\Delta BIC) \quad (5)$$

where F is the distance measure between the two sub-segments within one analysis window and their respective clusters. Ideally, if both sub-segments are very close to the same cluster, then a breaking point should not be declared even if the ΔBIC score suggests otherwise, and vice versa. The ensuing processing stages remain unchanged.

4. Experimental Results

In this section we describe the database used for this research, define an appropriate performance measure, and report the results at various processing stages.

- *The Database*

“Switchboard” data recorded and compiled by the Linguistic Data Consortium were used in this work. Each call was six minutes in duration and contained English conversation between two speakers via telephone. Although a topic was suggested to start each call, the participants spoke in a normal and unscripted manner, with frequent pauses and non-meaningful sounds that are not common in broadcast news. Originally each call had two files, one for each channel/speaker. Before processing, data from these two channels were mixed into one, thus each input was a seamless two-speaker telephone conversation consisted of speech, pauses (silences), cross-talks, and non-speech sounds. The whole data pool was well-diversified in genders and ages of speakers, telephony networks, cellular handsets, and background noises. A total of ten calls were tested: two were of the same genders and same networks (land-line or cellular), two were of different types in both aspects, and the remaining three were the same in one aspect but different in the other.

- *Performance Measure*

Traditionally the EER is the figure-of-merit test for the ACD procedure. Due to its non-scripted nature, however, speaker turns within telephone conversations are much more frequent and thus less significant. Furthermore, a portion of the false alarms are likely be eliminated when two adjacent segments of the same speaker are merged at a later stage. For the clustering process, the performance is on cluster sizes (or number) and purities. Since we have pre-determined the final number of clusters to be two, purities should be the ultimate measure.

From the provided time stamps, four types of utterances are objectively distinguishable: speaker A, speaker B, cross-talk, and background noise. Their sizes within one single entity of audio stream (a segment or a cluster at any stage), in terms of number of frames, are denoted by X_A , X_B , X_c , and X_n , respectively. If there are N such entities remaining, the relevant purity is defined as:

$$P_r = \sum_{i=1}^N \omega_i \frac{\max[X_{A,i}, X_{B,i}]}{X_{A,i} + X_{B,i}} \quad (6)$$

where ω_i is weight of the entity in the whole six-minute conversation. This is the performance measure we use at any stage. Silence remained within each cluster can be taken out by a VAD if desired. Cross-talk is excluded because it is of small quantity, and there is no correct cluster assignment for such an utterance.

We want to point out that when the process is stopped with three remaining clusters, one of them usually contains most of the cross-talk and non-meaningful sounds (laughing, coughing, etc.) if the amount of such utterances is large enough to build up a cluster along the way. This is a great testament of the BIC's robustness. In general, however, stopping at 3 clusters is too risky without prior knowledge of the content or style of a two-way conversation. Also, for the two calls that give the worst performance, the reason is clearly due to the fact that one speaker noticeably altered the speaking style, and greatly resembled that of the other participant. If the final purpose is better feature normalization or model adaptation for robust ASR instead of speaker verification, such "incorrect" classification by the BIC might even be more appropriate. Since there is no objective similarity measure of speaking styles, however, we only calculate purity by precisely following speaker identities specified by the time stamps.

Since there are only two speakers, a random clustering would do no worse than $P_r = 0.5$, and the perfect result of two completely pure clusters would give $P_r = 1$. To estimate the effectiveness of our algorithm, we have also calculated the achievable P_r under optimal condition as follows. Given the BIC-based algorithm, the likelihood of a segment being assigned to the correct cluster depends on four factors, namely the sizes and purities of the segment and of the cluster. Using the time stamps, we assemble a cluster of pure speech by speaker A, and a similar one by speaker B, without any cross-talk or silence in either. We then successively take out a segment from a cluster, and use the ΔBIC score to determine how it is assigned. Under such arrangement, three of the four conditions needed for correct clustering are absolutely perfect. Since the average segment duration after the ACD procedure is about one second, this is the segment length we extract for testing each time. The P_r achieved in this "optimal" case provides the theoretical performance limit of the BIC-based algorithm.

- *Purity Results*

The P_r at various stages are summarized by Table 1. Stage 0 is the "optimal" result obtained by the procedure described above. Stage 1 is for the segments delivered by the ACD step of the first iteration. Stage 2 is when there are three clusters remaining at the global clustering procedure. For stage 3, the two final clusters are obtained by directly merging one more time from stage 2. For stage 4, the two clusters are produced by breaking up the one with the lowest purity at stage 2, as described in Section 3. The two final clusters at stage 5 are at the end of the first iteration, after the re-clustering step. Stage 6 are the ACD-produced segments from the second iteration, where the criterion score calculation is modified according to Equation 6.

Several observations can be drawn. First, the segments produced by the ACD procedure are of very high purity in terms of excluding utterances from the other speaker. This is due to the fact that we impose no limit in peak-picking of the ΔBIC scores. It produces a large amount of false alarms which is acceptable for our performance measure because of the merging that follows. The purity then takes a significant drop from stage 2 to stage 3 as some segments are incorrectly clustered. Second, by taking advantage of knowing the final cluster numbers

Table 1: Summary of the purity measures

Stages	Conditions	P_r
0	Optimal	97.50%
1	Segments (1st iteration)	97.43%
2	3 clusters by merging	90.54%
3	2 clusters (a)	87.81%
4	2 clusters (b)	90.17%
5	2 clusters (c)	92.16%
6	Segments (2nd iteration)	97.46%

beforehand, the two techniques we employ improve the purity from 87.81% at stage 3 to 92.16% at stage 5, which reaches 95% of the theoretical limit (stage-5 result over the "optimal" result). Finally, the ACD procedure of the second iteration, even with information of two very pure clusters obtained from the previous run, does not produce segments which are on average of higher purity (97.46% at stage 6 compared to 97.43% at stage 1) or of longer duration. Since the following procedures remain unchanged, the process is terminated after one iteration.

5. Conclusions

We apply the Bayesian information criterion to segmentation and clustering of two-way telephone conversations according to speakers. It proves to be a robust algorithm for different genders, background noises, and telephony networks. By knowing the final number of clusters, we devise two procedures to improve the overall cluster purity from 87.81% to 92.16%. The algorithm approaches its optimal limit and does not show improvement in purity after one iteration.

6. References

- [1] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 873–876.
- [2] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. Broadcast News Trans. and Under. Workshop*, 1998, pp. 127–132.
- [3] J. Lopez and D. Ellis, "Using acoustic condition clustering to improve acoustic change detection on broadcast news," in *Proc. Int. Conf. Spoken Language Processing*, 2000.
- [4] A. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, "Unsupervised speaker segmentation of telephone conversations," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 565–568.
- [5] P. Sivakumaran, A. M. Ariyaeinia, and J. Fortuna, "An effective unsupervised scheme for multiple-speaker-change detection," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 569–572.