An adaptive BIC approach for robust audio stream segmentation

Janez Žibert¹, Andrej Brodnik¹, France Mihelič²

¹ Primorska Institute of Natural Sciences and Technology, University of Primorska, Koper, Slovenia
²Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
janez.zibert@upr.si, andrej.brodnik@upr.si, france.mihelic@fe.uni-lj.si

Abstract

In this paper we focus on an audio segmentation. We present a novel method for robust estimation of decision-thresholds for accurate detection of acoustic change points in continuous audio streams. In standard segmentation procedures the decision-thresholds are usually set in advance and need to be tuned from development data. In the presented approach we tried to remove a need for using pre-determined decision-thresholds and propose a method for estimation of thresholds directly from the currently processed audio data. It employs change-detection methods from two well-established audio segmentation approaches based on the Bayesian Information Criterion. Following from that, we develop two audio segmentation procedures, which enable us to adaptively tune boundary-detection thresholds and to combine different audio representations in the segmentation process. The proposed segmentation procedures are tested on broadcast news audio data.

Index Terms: audio segmentation, Bayesian Information Criterion, speaker diarization

1. Introduction

In many speech and audio applications, that operate on continuous audio streams, it is necessary to partition and classify different audio sources and acoustic events prior to the application’s main processing tasks. For example, audio streams in broadcast news (BN) transcription systems should be firstly chopped into smaller acoustic-homogeneous segments, which are then classified as speech and non-speech, since the main task - speech recognition - is performed only on speech segments [1]. The same pre-processing steps are necessary in audio-diarization and audio-indexing systems, where the obtained speech segments are additionally clustered together by similar acoustic or speaker properties [2], or in speaker detection and tracking systems, that operate in multiple speaker environments, and therefore audio data should be organized in a way to enable speaker recognition on data, that belong to one speaker only [3].

A main development of audio segmentation approaches was achieved along with the development of systems for processing of broadcast news audio data. The task of partitioning the audio streams for speech recognition was first introduced during the NIST Broadcast News Transcription evaluation in years 1997-98 [1]. Audio segmentation and clustering procedures were then fully evaluated in the “NIST Rich Transcription speaker-diarization evaluation” in years 2003-04 [2]. In 2005, the ESTER evaluation was conducted on similar task using French radio BN data [4]. Another evaluation campaign on multilingual BN data was carried on in years 2004-05 during the COST278 action [5]. All these studies have contributed to the development of audio-segmentation approaches as well as they have demonstrated the importance of other speaker-diarization tasks in the pre-processing steps of various BN speech processing applications.

In this paper we present a novel audio-segmentation method for robust and accurate detection of acoustic-change points in continuous audio streams. It employs change detection methods from the two audio segmentation approaches based on the Bayesian Information Criterion: a standard approach, introduced in [6], and a distBIC approach [7]. By combining methods from both approaches we developed a technique, which enabled us to adaptively tune decision-thresholds from the currently processed audio data. This technique also provides an ability for a decision-score normalization, that was utilized in our second audio segmentation procedure, where a segmentation was performed by a score-based fusion of different acoustic representations of audio signals. These approaches are described in Section 2. All the segmentation procedures were tested and compared on BN audio data. The experiments and the results are discussed in Section 3.

2. Audio-Segmentation Approaches

2.1. The BIC Segmentation Approaches

Due to its several useful properties the Bayesian Information Criterion (BIC) has become widely used for the audio segmentation and clustering purposes, where it was first introduced by by Chen and Gopalakrishnan [6].

The BIC is a penalized maximum likelihood model selection criterion [6, 7]. When it is used as a model selection criterion in the audio segmentation, a problem of change points detection is reformulated as a model selection task between two competing models. If $X$ defines an analysis window in an audio stream, where a change point needs to be found, the $\Delta_{BIC}$ measure can be defined as:

$$\Delta_{BIC}(b) = BIC(M_2) - BIC(M_1)$$
$$= L(X; M_2) - L(X; M_1) - \lambda(C(M_2) - C(M_1)).$$

(1)

where $M_1$ represents a model, when the data $X$, that is presented as a sequence of frame-based acoustic feature vectors, i.e., $X = x_1, x_2, \ldots, x_N$, is modeled by a single Gaussian distribution. The $M_2$ stands for a model, which is composed by two Gaussian distributions: one is estimated from data $x_1, \ldots, x_b$ and the other from data $x_{b+1}, \ldots, x_N$. The first term $L(X; M_2) - L(X; M_1)$ in (1) computes the difference in log-likelihoods of the data $X$ according to the models $M_2$ and $M_1$.
and $M_1$, while the second term $C(M_2) - C(M_1)$ computes the difference in the complexities of the models $M_2$ and $M_1$. The penalty factor $\lambda$ allows tuning of the balance between both terms. In the case, when the $\Delta BIC(b) < 0$, the data in the analysis window $X$ is better modeled by a single Gaussian distribution, while the $\Delta BIC(b) > 0$, favors representation of the data by two distributions, which supports the hypothesis that a change in acoustics occurs at the point $b$.

To detect multiple segmentation boundaries in an audio stream the BIC measure needs to be applied across the whole range of data points in an audio stream. Since the $\Delta BIC$ score is computed for every point within the analysis window $X$, there exist two general approaches, how to construct and move these analysis windows throughout the data stream [2].

The first approach was proposed in [6] and was introduced to the segmentation task together with the $\Delta BIC$ measure. This technique works in a sequential manner and tries to find segmentation boundaries within analyzing segments of varying lengths. Segmentation is performed in the following way. At start, initial window $X$ of pre-determined length is set at the beginning of a stream and the $\Delta BIC$ from (1) is computed for every data point within this window. Candidates for segmentation boundaries are points, where the $\Delta BIC > 0$, and among them a point with the highest $\Delta BIC$ score is selected as a change point. In this case, the window $X$ is moved to position of the change point, and the computation of the $\Delta BIC$ continues within the new window with the same length. If there are no change points in the initial window $X$ (the $\Delta BIC < 0$ for all the points within the window $X$), the window $X$ is increased by additional length, and a computation of the $\Delta BIC$ is redone on the extended window. These steps are repeated until there are no more data for processing. This approach was proposed together with the $\Delta BIC$ measure and it is used as a standard procedure in the most of segmentation systems. We refer it as a standard BIC approach in the paper.

Another approach of computing segment boundaries from audio data streams is to apply the $\Delta BIC$ measure on the fixed-length analysis windows. The Gaussians were estimated on the fixed-length analysis window and on the each half of the window, which is centered at the location of the current computation point. The $\Delta BIC$ scores are then computed for every point in an audio stream by sliding the analysis window along the data stream. This produced a distance function and the peaks in the distance function were candidates for change points. Only those points, where their absolute values exceed a pre-determined threshold, set on a development data, are chosen as segmentation boundaries. Additional smoothing of the distance function or elimination of the smaller neighboring peaks within a certain minimum duration is needed to prevent over-generation of change points at true boundaries. This approach was applied in [8, 9], where the KL2 distance was used for the detection measure. In [7] the KL2 distance was exchanged by a generalized log-likelihood ratio, that is actually a simplified version of the $\Delta BIC$, when $\lambda = 0$.

While there have been several solutions proposed to improve the performance of the BIC approaches to be computationally more effective and to increase the accuracy in a detection of short segments, not so many efforts have been made to set proper decision-thresholds for optimal change-points detection. Setting an optimal threshold always presents a trade-off between producing long, pure segments of acoustic homogeneous data, and minimizing the rate of missed change points. A general principle of almost all segmentation procedures is to set thresholds empirically from development data each time, when either the overall audio conditions change and are not known in advance (e.g. different recordings, different BN shows, etc.) or when new (alternative) acoustic features are applied into the segmentation process. In such cases the fixed pre-determined thresholds, estimated from the development data, could not adequately fit to the acoustics in the processing data due to mismatched conditions between training and real data, which could severely degrade the segmentation performance. In next section we propose a method that aims to overcome these problems by introducing adaptively-tuned decision-thresholds into the BIC-based segmentation process.

### 2.2. Adaptively Tuned Decision Thresholds

By combining both BIC segmentation procedures with fixed- and varying-length analysis windows we developed a method for estimating thresholds directly from the processed audio data. Accordingly, a need for using development data for tuning the segmentation parameters was removed.

We joined the standard BIC approach with the time-varying analysis windows and the distBIC approach, where the $\Delta BIC$ measure was computed on the fixed-length analysis windows into two-pass segmentation procedure. In the first pass, the decision thresholds were estimated directly from the processed audio data by using the distBIC segmentation procedure, and in the second pass the standard BIC approach was applied by using the estimated thresholds from the first pass.

Estimation of a decision threshold in the first pass was performed by using the distBIC segmentation procedure, where we computed the distance function for each point in an audio stream by sliding a fixed-length analysis window. Parameters for the window length, a sliding step and the $\lambda$ in the $\Delta BIC$ were set in advance and remained fixed during the computation. All the $\Delta BIC$ scores were then collected and among them a global maximum and minimum value was found. The threshold $\theta$ for the next segmentation pass was evaluated as:

$$\theta = \max \Delta_{BIC} - \alpha \cdot (\max \Delta_{BIC} - \min \Delta_{BIC}).$$

where $\max \Delta_{BIC}$ presents the maximum and $\min \Delta_{BIC}$ the minimum value among all the $\Delta BIC$ scores produced by the sliding fixed-length analysis window on the given audio stream within the distBIC approach. The threshold $\theta$ was thus determined as a relative shift from the maximum $\Delta BIC$ score by a portion of the difference between the maximum and the minimum score, which was controlled by the factor $\alpha$. The values for $\alpha$ were typically chosen between 0.0 and 0.2.

This threshold was then used in the second pass of the segmentation procedure within the standard BIC approach, where the penalty factor $\lambda$ remained the same as in the first pass. A detection of segmentation boundary was then performed by selecting the time point in an audio stream, where the highest overall $\Delta_{BIC}$ score was achieved among all those scores, which were above the overall decision threshold $\theta$ in the analyzing window.

In such a way, we removed a need for tuning the penalty factor $\lambda$ in the $\Delta BIC$ measure by introducing the $\theta$ threshold. The $\theta$ was evaluated from the $\Delta BIC$ scores in the first pass within the distBIC approach. The scores distribution was estimated from the current processing audio data and the threshold $\theta$ was evaluated as the relative shift from the maximum $\Delta_{BIC}$ score controlled by the parameter $\alpha$. As such, the absolute thresholds, defined by the penalty factor $\lambda$ in the $\Delta_{BIC}$ measure, were replaced by the relative thresholds, controlled by the $\alpha$, that is not depend on any other open parameters of the segmentation procedure nor it was influenced by any changes in
acoustic environments due to mismatched acoustic conditions between development data and the processed audio streams. Consequently, no additional development data were needed for setting the decision thresholds for the BIC segmentation.

2.3. Score-level fusion for BIC segmentation

This phenomena can be also exploited even further. By assuming that nearly the same $\Delta_{\text{BIC}}$ score-ranges are obtained with both approaches one can predict the score range of the $\Delta_{\text{BIC}}$ measure used in one approach by the estimated scores from the other approach. We took advantage of this for a normalization of the $\Delta_{\text{BIC}}$ scores within the standard BIC approach. The normalization of the standard BIC scores was performed by using the \textit{min-max score normalization} method, [10], where the minimum and the maximum score were estimated from the $\text{distBIC}$ segmentation:

$$\overline{\Delta_{\text{BIC}}} = \frac{\Delta_{\text{BIC}} - \min \Delta_{\text{distBIC}}}{\max \Delta_{\text{BIC}} - \min \Delta_{\text{distBIC}}}.$$  \hspace{1cm} (3)

The $\overline{\Delta_{\text{BIC}}}$ presents a normalized version of the $\Delta_{\text{BIC}}$ score computed within the standard BIC approach. The $\min \Delta_{\text{BIC}}$ and the $\max \Delta_{\text{BIC}}$ are the minimum and the maximum score obtained within the $\text{distBIC}$ approach on the same audio data, respectively. In the same manner the threshold $\theta$, defined in (2), can be normalized: the threshold $\theta_{sBIC}$, computed in (2), is inserted instead of the $\Delta_{sBIC}$ score in equation (3) and the result is stored as a normalized version $\overline{\theta_{sBIC}}$.

This kind of normalization of the BIC scores and the corresponding thresholds enabled us to extend the proposed segmentation procedure to detect segmentation boundaries based on the fusion of the BIC scores obtained from different acoustic representations of audio data. We achieved this by fusing the scores in the following way:

$$\overline{\Delta_{\text{fusBIC}}} = \sum_{i=1}^{g} w_{i} \cdot \overline{\Delta_{sBIC}}.$$  \hspace{1cm} (4)

Equation (4) presents a weighted average of the normalized $\Delta_{\text{BIC}}$ scores along different feature-representation streams $S_{i}$. The fusion weights $w_{i}$ can be additionally tuned for the optimum segmentation performance, but they were manually set to be evenly proportioned in our experiments. The decisions about segmentation boundaries were made in this case by comparing the $\Delta_{\text{BIC}}$ scores, summed over all streams, with the overall decision threshold, that was computed in the same way by using the formula in (4). A detection of segmentation boundaries was then performed by selecting the time point in an audio stream, where the highest overall $\overline{\Delta_{\text{fusBIC}}}$ score was achieved among all those scores, which were above the overall decision threshold in an analyzing window. This was done in the same manner as it is done in the standard BIC approach, just a single-score decision criteria was replaced by the criteria based on the fusion of scores, defined in (4).

In the segmentation procedure we were thus able to include different kind of acoustic representations. Various combinations of mel-frequency cepstral coefficients were tested in our experiments, but also other representations can be used.

3. Evaluation Experiments

We tested four segmentation approaches: the standard BIC approach (abbreviated as the $\text{dBIC}$ approach), the $\text{distBIC}$ approach (abbreviated as the $\text{distBIC}$ approach), and two proposed versions of the standard BIC segmentation, one with adaptively-tuned decision-thresholds (abbreviated as the $\text{atBIC}$ approach) and the other with a fusion of different acoustic representations (abbreviated as the $\text{fusBIC}$ approach).

The experiments were performed on two BN speech databases. The first database served for setting all open parameters of the segmentation approaches and for a comparison, how good can all the segmentation procedures perform in controlled acoustic environments. The database included 7 hours of BN shows in Slovenian language and is considered as a development database [11]. When all the parameters of the procedures were tuned, the segmentations were then performed on completely different data - on the COST278 BN database, that is composed of 30 hours of BN shows in 9 European languages [12].

3.1. Experimental Setup

In all the segmentation procedures the mel-frequency cepstral coefficients (MFCCs) were used as a basic representation of the audio data. The MFCC features were composed from the first 12 cepstral coefficients (without 0th coefficient) and a short-term energy, estimated directly from the signal. Also the first order derivatives of the MFCCs were applied. Such features are noted as the MFCC+$\Delta$MFCC features in all the segmentation approaches. Note, that in the fusion case, the MFCC+$\Delta$MFCC representation corresponds to two stream of features, where in one stream the MFCC features and in other the $\Delta$MFCC features were used.

The probability distributions in the $\Delta_{\text{BIC}}$ measure, which were used in all the segmentation approaches, were set to be single full-covariance Gaussians. To speed up the segmentation processes the estimations of the Gaussians and the computation of the BIC measure were performed by implementing several improvements of the base approach [6] as were suggested in [13, 2].

Segmentation procedures were assessed by using the standard evaluation measures, the recall, the precision and the F-measure, defined in [8]. The correctly found boundary is defined at the segmentation point, which is spotted in a time tolerance region around a true boundary in the reference segmentation (in our experiment the tolerance region was set to 1.0 s).

3.2. Development Data Experiments

The development data experiments were conducted during a process of tuning the open parameters of the evaluated segmentation approaches. All the open parameters of the segmentation procedures were tuned in a way to achieve the highest F-measure scores on the development set.

The segmentation results are shown in Figure 1. The results of the $\text{sBIC}$ and the $\text{atBIC}$ approaches in are plotted against the penalty factor $\lambda$, that was tuned in the $\Delta_{\text{BIC}}$ measure. The $\lambda$ values were selected from the range between 0.50 to 3.50 by using a fixed increment step of 0.25. The $\text{dBIC}$ approach was tuned according to the threshold for selecting the local-maximum points. The threshold values were chosen from the interval $[-300, 0]$. The $\text{fusBIC}$ approach was optimized by setting the fusion weight $w_{i}$ in equation (4) to values between 0.1 and 0.9. For the evaluation purposes all the plots were produced in one graph, even though the plots could not be completely aligned to each other since the tunings were performed on different ranges of open parameters.

By examining a behavior of such plots we tried to estimate, how sensitive was each segmentation procedure to its open pa-
parameters, and consequently how good segmentation procedures would perform in situations, where open parameters would be sub-optimally tuned. Figure 1 shows that the results of the proposed segmentation procedures, the atBIC and the fusBIC, produced relative flat curves in a comparison to the sBIC and the dBIC approaches, which implies that there is no need to use any development data for threshold tuning in both of the proposed approaches.

3.3. Test Data Experiments

An assessment of the presented segmentation approaches was additionally performed on the COST278 test data. The test data was composed of 30 hours of BN audio data, where had to be additionally performed on the COST278 test data. The test data was composed of 30 hours of BN audio data, where had to be found 8012 segment boundaries. The results are summarized in Table 1.

Table 1: Segmentation results on the COST278 database.

<table>
<thead>
<tr>
<th>Segmentation approach</th>
<th>RCL (%)</th>
<th>FPR (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sBIC: MFCC+ΔMFCC</td>
<td>68.7</td>
<td>78.0</td>
<td>73.1</td>
</tr>
<tr>
<td>atBIC: MFCC+ΔMFCC</td>
<td>70.6</td>
<td>70.3</td>
<td>70.5</td>
</tr>
<tr>
<td>fusBIC: MFCC+ΔMFCC</td>
<td>84.1</td>
<td>69.8</td>
<td>76.3</td>
</tr>
</tbody>
</table>

The results in Table 1 are comparable to the segmentation results, that was presented in the study in [5], where the evaluation of several speaker diarization systems on the COST278 BN data was conducted. Even more, we achieved better segmentation scores with our proposed segmentation approaches in comparison to those segmentation systems.

Naturally, the performances of all the segmentation approaches were dropped in comparison to the development data experiments. This was expected since the audio data of both databases differ greatly. As such, the segmentation results in Table 1 expose big differences between the recall and the precision scores in all the approaches, but the F-measure scores speak in favor of the proposed approaches. Overall the highest scores were achieved with the fusBIC approach. This suggests that the performances of the standard segmentation procedures were more sensitive to how well their open parameters were tuned on the development data or how good the development data fit to the acoustic conditions in the test data.

On the other hand, the atBIC and the fusBIC approaches performed quite well on the test set in comparison to the development data experiments. This means, that the proposed approaches are less sensitive to optimal tuning and they can operate well even in the case of non-optimal settings of their open parameters, which suggests a relative independence of the approaches to the mismatched acoustic conditions between train and test data.

4. Conclusion

Our research aimed to explore audio segmentation procedures for speaker change detection in continuous BN audio-data streams. We implemented two standard segmentation approaches based on the Bayesian Information Criterion and developed a new method, that aimed to join two stages of both standard approaches in order to adaptively estimate the decision thresholds from the processing audio data. In such a way we were able to develop two segmentation procedures, that did not need any additional tuning of the decision thresholds on the development data. The experiments on the BN audio data showed, that the proposed segmentation approaches performed more reliable and stable across different range of acoustic conditions in comparison to the standard BIC and the DISTBIC approaches.

5. References