Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 19 (2005) 147–170
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Evaluation of BIC-based algorithms for audio segmentation

Mauro Cettolo [a,*], Michele Vescovi [b], Romeo Rizzi [b]

[a] *ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Via Sommarive, 18 I-38050 Povo, Trento, Italy*
[b] *Università degli Studi di Trento, Facoltà di Scienze MM.FF.NN. I-38050 Povo, Trento, Italy*

Received 24 June 2003; received in revised form 24 May 2004; accepted 26 May 2004
Available online 1 July 2004

Abstract

The Bayesian Information Criterion (BIC) is a widely adopted method for audio segmentation, and has inspired a number of dominant algorithms for this application. At present, however, literature lacks in analytical and experimental studies on these algorithms. This paper tries to partially cover this gap.

Typically, BIC is applied within a sliding variable-size analysis window where single changes in the nature of the audio are locally searched. Three different implementations of the algorithm are described and compared: (i) the first keeps updated a pair of sums, that of input vectors and that of square input vectors, in order to save computations in estimating covariance matrices on partially shared data; (ii) the second implementation, recently proposed in literature, is based on the encoding of the input signal with cumulative statistics for an efficient estimation of covariance matrices; (iii) the third implementation consists of a novel approach, and is characterized by the encoding of the input stream with the cumulative pair of sums of the first approach.

Furthermore, a dynamic programming algorithm is presented that, within the BIC model, finds a globally optimal segmentation of the input audio stream.

All algorithms are analyzed in detail from the viewpoint of the computational cost, experimentally evaluated on proper tasks, and compared.

* Corresponding author. Tel.: +39-0461-314-551; fax: +39-0461-314-591.
E-mail addresses: cettolo@itc.it (M. Cettolo), vescovi@kirk.science.unitn.it (M. Vescovi), romeo@science.unitn.it (R. Rizzi).

0885-2308/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2004.05.008

1. Introduction

In the last years, efforts have been devoted amongst the research community to the problem of audio segmentation. The number of application of this procedure is considerable: from the extraction of information from audio data (e.g., broadcast news, recording of meetings) to the automatic indexing of multimedia data, or the improvement of accuracy for recognition systems. Typically, these tasks are performed by complex systems, consisting of a number of modules, some of them computationally expensive. Although in these systems the audio segmentation does not represent a computational bottleneck, the response time of the segmentation module can become an important issue under specific requirements. For example, ITC-irst delivered to RAI (the national Italian broadcasting company) an off-line system for the automatic transcription of broadcast news programs. Part of the requirement was also to supply a real-time version of the transcription station, able to guarantee adequate performance. Given these constraints, the speed of each component needs to be maximized without affecting its accuracy.

The segmentation problem has been handled in different ways that can be roughly grouped in three classes:

energy based methods: each silence occurring in the input audio stream is detected either by using an explicit model for the silence or by thresholding the signal energy. Segment boundaries are then located in correspondence of detected silences.

metrics based methods: the segmentation of the input stream is achieved by evaluating its "distance" from different segmentation models. Distances can be measured by the Hotelling's T^2 test (Wegmann et al., 1999), the Kullback Leibler distance (Kemp et al., 2000; Siegler et al., 1997), the generalized likelihood ratio (Gish et al., 1991), the entropy loss (Kemp et al., 2000), and the Bayesian Information Criterion (BIC) (Schwarz, 1978).

explicit models based methods: models are built for a given set of pre-determined acoustic classes – e.g., female and male speakers, music, noise, etc. Typically, the input data stream is classified from the maximum likelihood principle, through a dynamic programming decoding. The time indexes where the classification changes from one class to another are assumed to be the segment boundaries (Hain et al., 1998). Alternatively, in (Lu et al., 2001) Support Vector Machines are employed to learn class boundaries; (Scheirer and Slaney, 1997) compares the Gaussian Maximum A-Posteriori estimator, the Gaussian Mixture Model, the Nearest-Neighbor classifier and a spatial partitioning scheme based on K-d trees.

The main limitations of the first and the third approach are evident. Regarding the energy based methods, there is only a partial correlation between changes in the nature of the audio and silences. On the other hand, these methods are simple to implement and can perform their (limited) task in linear time and with sufficient precision – provided the absence of significant variations in the background acoustic conditions. With explicit models of acoustic classes, changes occurring within the same class are undetectable; for example, if only one model for female voices is employed, no change can be detected within a dialogue between female speakers. Moreover, it is required both to know in advance the classes of interest and the availability of suitable data for their training. On the other hand, these methods can reach very high accuracy rates with linear time cost complexity.

This paper concentrates on a specific class of metric-based methods. Metric-based methods do not require any prior knowledge, nor a training stage. They are efficient (linear in time if some