A BIC-POISSON-SVM Segmentation System for the Albayzin'10 Audio Segmentation Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo
plopez@gts.uvigo.es, ldocio@gts.uvigo.es, carmen@gts.uvigo.es

Abstract

In this paper, the system submitted by the UVigo-GTM for the Albayzin 2010 Audio Segmentation Evaluation is described. The task is performed in two different stages, a segmentation stage and a classification stage. The segmentation stage employs an approach called BIC-POISSON, which consists in a modification of the BIC algorithm oriented to reduce the false alarm rate in the acoustic change detection stage. The classification is performed with Support Vector Machines, using the segments obtained in the segmentation stage.

Index Terms: audio segmentation, BIC, poisson, GMM-UBM, SVM

1. Introduction

The increasing amount of multimedia information that is available nowadays makes it necessary to develop techniques to structure and classify this kind of information. These multimedia contents include music, video and spoken documents, so speech technologies become involved. In monitoring spoken documents it might be necessary to know the background acoustic condition in order to implement an online adaptation of the monitoring/recognition system to the acoustic conditions in order to improve the overall system performance. The work presented in this paper focuses in the segmentation of audio documents into homogeneous regions according to the background conditions.

Common audio and speech segmentation techniques are based on the Bayesian information criterion (BIC) algorithm [2] [4], which seems to be the fastest and the best performing algorithm when it is compared to techniques such as the adaptation and comparison of Gaussian mixture models (GMM) [3]. One of the problems encountered when segmenting with BIC algorithm is the high number of false alarms (i.e. change-points that are detected, although they do not exist), because it degrades the quality of the results.

The Albayzin 2010 Audio Segmentation Evaluation consists in the segmentation of audio broadcast news programs into homogeneous regions according to the backgroud conditions: clean speech, music, speech with background noise, speech with background music, and other. The segmentation strategy chosen by the UVigo-GTM group for this task was the BIC-Poisson [1] one. The basic idea in this approach is to discard change-points found by the BIC algorithm when they are likely to be false alarms by assuming that changes occurring in the audio stream constitute a Poisson process, so changes that have a low ΔBIC value will be discarded with a probability that follows a Poisson cumulative density function (cdf).

The classification stage is performed by using Support Vector Machines (SVMs), representing the audio segments as supervectors composed by the stacked means of Gaussian Mixture Models (GMMs). The one-against-one technique for multiclass classification with SVMs was employed in this work.

The structure of this paper is as follows: section 2 makes an overview of the whole audio segmentation strategy; in section 3 the segmentation algorithm and the BIC-Poisson technique are described; section 4 explains how the classification of the different segments was performed; section 5 describes the experimental framework used for the experiments and its purpose; section 6 shows the results of the whole system; and section 7 concludes the work and gives prospects on future work.

2. System overview

The developed system performs the segmentation of the data and the classification of the segments independently: first the data is segmented, and after that the segments are classified into the different classes. It is possible that, after classifying the segments, adjacent segments assigned to the same class appear, so these adjacent segments are merged, deleting the acoustic change-point between them.

The segmentation of the data is performed using an approach called BIC-Poisson [1], specially designed to reduce the number of false alarms in audio and speaker segmentation tasks. The classification is done by using Support Vector Machines (SVM). The whole procedure is summarized in figure 1.

3. Segmentation

3.1. Segmentation strategy

The segmentation strategy employed in these experiments is similar to the one described in [2]. It has a coarse segmentation stage where a window of data is segmented with a low resolution. If an acoustic change-point is found, a fine segmentation stage is performed, using a higher resolution. In the first stage, if there is no acoustic change-point, the window grows until it gets to a fixed maximum size, and then it slides. When a change-point is found, the window returns to its initial size and slides. In the second stage, a window of a fixed size is centered in the change-point found in the first stage and BIC is applied again using a higher resolution. If the hypothesis test indicates that there is a change-point in the window, this change-point is confirmed.

Subsequently, in the proposed segmentation technique a change-point rejection strategy is applied. The ΔBIC value is assessed in order to evaluate the significance of the change-point detected. if this is located above a given threshold no action is taken; otherwise the change-point is discarded with a certain probability.

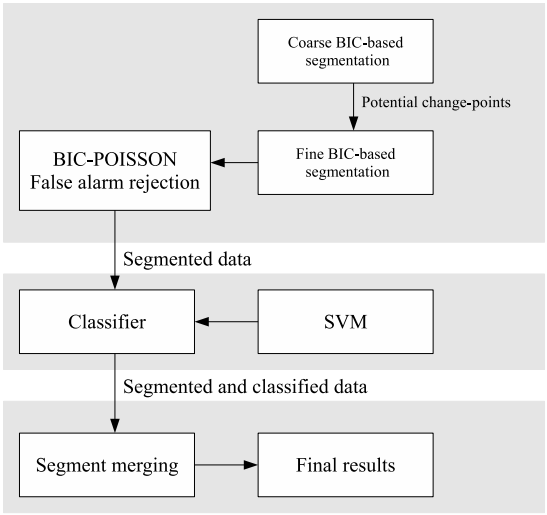

Figure 1: *Summary of the audio segmentation and classification system.*

3.2. BIC-Poisson: Poisson distribution-based change rejection

In this false alarm suppression strategy, it is assumed that the change-point instants follow a Poisson process. A Poisson process is an independent occurrence process where the number of occurrences in two disjoint time intervals is independent, the probability of an occurrence is proportional to the observed interval and occurrences are not simultaneous [5].

The speaker segmentation process fulfills those properties, since it is a process where arrivals (change-points) occur independently and in random instants. Poisson processes have a probability density function (pdf):

$$f(\mu, x) = \frac{e^{-\mu}\mu^{x}}{x!} \tag{1}$$

Its cumulative density function (cdf) is the sum of the probability density function in all points below a given value:

$$F(\mu, x) = \sum_{i=0}^{x} f(\mu, i) = \sum_{i=0}^{x} \frac{e^{-\mu}\mu^{i}}{i!} \tag{2}$$

The parameter μ, which represents the mean of the distribution, will in this case represent the expected number of changes.

The properties of the Poisson distribution will be used as follows: μ occurrences are expected over a given period of time, so initially a change is accepted with a very high probability. However, as the number of accepted changes increases and approaches or exceeds the expected number of changes, they are more likely to be rejected. This process is easily modeled using the cumulative density function $F(\mu, x)$ as a discard probability. The discard probability will be very low at first, but as the mean is approached or exceeded, it will steadily approach 1 (meaning that all the occurrences will be rejected).

Figure 2 shows graphically the increase of the discard probability with the number of accepted changes. As the amount of accepted changes reaches the value of 2μ approximately, the discard probability becomes close to one, i.e. almost every change-point will be discarded.

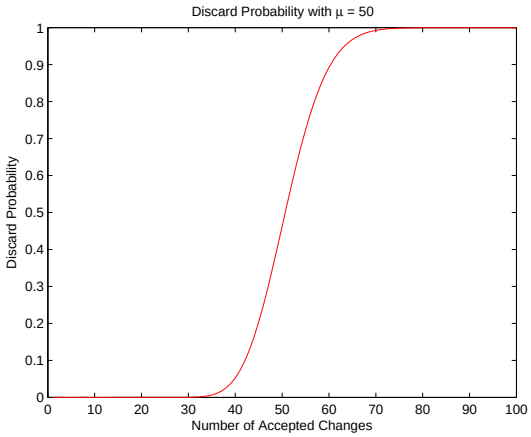

Figure 2: *Discard probability for the Poisson-based rejection algorithm.*

4. Classification

After segmenting the audio data, each audio segment has to be classified into one of the five pre-defined audio classes:

- Speech [sp]. Clean speech in studio from a close microphone.
- Music [mu]. Music is understood in a general sense.
- Speech with noise in background [sn]. Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
- Speech with music in background [sm]. Overlapping of speech and music classes or speech with noise in background and music classes.
- Other [ot]. This class refers to any type of audio signal (including noises) that doesnt correspond to the other four classes.

The classification scheme chosen in this study is based on Support Vector Machines (SVM). The SVM is a supervised learning method which, given a set of labeled samples, apply a nonlinear transform through a kernel function on the samples into a higher dimension space where the two classes are linearly separated under the margin maximization constraint.

The classification that has to be performed implies splitting the data into six different classes and SVMs, basically, can only solve binary classification problems. Nevertheless, it is possible to reduce a multiclass problem to a binary one. There are two main different approaches to solve this multiclass problem with SVMs: the one-against-one technique and the one-against-the-rest technique. To allow for multiclass classification the library for Support Vector Machines called libSVM [6], which was used in these experiments, uses the one-against-one technique by fitting all binary subclassifiers and finding the correct class by a voting mechanism.

Choosing accurate model parameters is very important to SVM training. Radial basis function (RBF) was used as kernel

function in the SVMs.

$$K(x_i, x_j) = exp(-\sigma||x_i - x_j||^2) \qquad \sigma > 0$$

Model selection in this class of SVM involves two hyperparameters: the penalty parameter C and the kernel width σ. The σ in the RBF kernel controls the shape of the kernel and C controls the tradeoff between margin maximization and error minimization. A grid-search on C and σ using 5 fold cross-validation is performed.

The SVM considers the sample data as points in a space of a given dimensionality, so a way to represent audio segments as sample data for the SVM training has to be chosen. In this case, first a Universal Background Model based on a Gaussian Mixture Model (GMM-UBM) of M mixture components is trained using data from all classes. Then, given an audio segment represented by N feature vectors of dimension D, the GMM-UBM is adapted using MAP adaptation to that audio segment [3]. By stacking the resultant means, a supervector of dimension $M \cdot D$ is obtained, being this supervector the representation of the audio segment as a point in a space of dimension $M \cdot D$.

5. Experimental framework

5.1. Database

The training and evaluation database consists of Catalan broadcast news data from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporaci Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), was splitted into 2 parts: one part for training/development (2/3 of the total amount of data), and the other part for evaluation (the remaining 1/3).

The distribution of classes within the database is the following: Clean speech: 37

The 16 available files to perform the training/development of the segmentation system were splitted as follows:

- Sessions 1 to 8 and 10 to 15: training of the GMM-UBM and the multiclass SVM.
- Sessions 9 and 16: selection of the parameters that give the best performance. The parameters to select were μ, λ, M, σ and C.

After testing on the development data the selected parameters were: $\mu = 20.0$, $\lambda = 3.5$, $M = 64$, $\sigma = 0.0078$ and $C = 2$.

5.2. Metric

The metric used to evaluate the system performance is defined as a relative error averaged over all acoustic classes:

$$Error = average_i \frac{dur(miss_i)dur(fa_i)}{dur(ref_i)}$$

where

- $dur(miss_i)$ is the total duration of all deletion errors (misses) for the i^{th} acoustic class.
- $dur(fa_i)$ is the total duration of all insertion errors (false alarms) for the i^{th} acoustic class.
- $dur(ref_i)$ is the total duration of all the i^{th} acoustic class instances according to the reference file.

It is worth noting that an incorrectly classified audio segment (a substitution) is computed both as a deletion error for one acoustic class and an insertion error for another. In the case when the system output is Other (non-Other) and the corresponding reference label is non-Other (Other), the audio segment is computed as a deletion (insertion) error only for the non-Other acoustic class.

A forgiveness collar of 1 sec (both + and -) will not be scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an acoustic class begins/ends.

5.3. Acoustic features

The data to perform the evaluation is given as a set of waveforms, and these are going to be represented by 12 Mel-frequency Cepstral coefficients (MFCC), extracted using a 25 ms Hamming window at a rate of 10 ms per frame. In the segmentation stage, these features are augmented by 0-th order cepstral coefficient. In the classification stage, this 13-dimensional feature vector is augmented with first and second order dynamic coefficients resulting in a 39-dimensional feature vector. Cepstral mean and variance normalization is also applied.

6. Experimental results

Table 1 provides the results obtained by the system submitted by the UVigo-GTM research group. It can be observed that the error in pure music is low. Indeed the main cause of error is, by far, the confusion between speech with background music, speech with background noise and clean speech.

Table 1: *Audio segmentation results on the development (session9 and session16) and evaluation corpus.*

Corpus	Error				
	Tot	music	speech	speech-music	speech-noise
Dev	35.95	18.73	50.48	34.40	40.20
Eval	33.15	22.41	41.80	27.47	40.93

7. Discussion and future directions

The audio segmentation and classification system submitted to Albayzin 2010 Evaluation was described in this paper. The audio segmentation task focuses in the context of broadcast news. According to the results obtained by the proposed system on the evaluation data, the validity of the combination of the BIC-poisson based audio segmentation approach with the SVM-based multiclass classification is confirmed.

Future work will focus on combining the traditional short-term MFCC features with prosodic and other acoustic features in order to discriminate better between speech with background music, speech with background noise and clean speech. Related to the classification stage, future work will focus on the analysis of other strategies for multi-class SVM based classification.

8. Acknowledgements

The UVigo-GTM group would like to thank the organizers and coordinators of the Albayzin 2010 Audio Segmentation Evaluation for their help and for kindly providing the data corpus. This

work was partially sponsored by the Spanish Ministry of Science and Innovation under the project TEC2009-14094-C04-04.

9. References

[1] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, "Reducing the False Alarm Rate in a Speaker Segmentation System", In preparation.

[2] M. Cettolo and M. Vescovi and R. Rizzi, "Evaluation of BIC-based algorithms for audio segmentation", Computer Speech and Language, vol. 19, 147–170, April 2005

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, January 2000.

[4] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, September 2000.

[5] A. O. Allen, "Probability, Statistics, and Queueing Theory with Computer Science Applications", 2nd ed. Academic Press, 1990.

[6] C.-C. Chang, C.-J. Lin, "LIBSVM: a Library for Support Vector Machines", Online: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.