GTM-UVigo System for Albayzin 2014 Audio Segmentation Evaluation

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

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the GTM-UVigo systems for Albayzin 2014 audio segmentation evaluation, which consist on segmentation followed by classification approaches with the same segmentation stage, but different classification approaches. Segmentation is performed by means of a Bayesian Information Criterion (BIC) strategy featuring a false alarm rejection strategy: the process of acoustic change-points is supposed to follow a Poisson process, and a change-point is discarded with a probability that varies in function of the expected number of occurrences in the time interval formed by the previous and candidate change-points. The classifier of the primary system represents the audio segments in a total variability space and then classifies them using logistic regression; contrastive system 1 represents the audio segments by means of Gaussian mean supervectors and classification is performed using a support vector machine; and contrastive system 2 models the different classes with Gaussian mixture models and performs maximum likelihood classification.

Keywords: Audio segmentation, iVectors, false alarm rejection strategy

1 Introduction

Audio segmentation is a task consisting on dividing an audio stream into homogeneous regions according to some criteria. Audio segmentation systems can be divided in two groups: those that perform segmentation followed by classification, and those that perform audio segmentation by classification.

Albayzin 2014 audio segmentation evaluation consisted on the detection of speech, music, noise or any combination of these three classes in a set of recordings. This paper describes the audio segmentation systems developed by the GTM-UVigo team for this evaluation, which consist on segmentation followed by classification approaches.

The segmentation stage was carried out by means of the Bayesian information criterion (BIC) approach for acoustic change detection [3]. A technique to reduce the false alarm rate of the BIC algorithm was added to this system [8]: the acoustic change-point process is supposed to follow a Poisson process and,

according to this, a candidate change-point that is suspicious of being a false alarm is accepted or rejected with a probability that depends on the length of the observed interval. This observed interval is, in this case, the length of the audio segment that would be formed if the candidate change-point were accepted.

Three systems with different classification approaches were submitted. The primary system represents the audio segments in a total variability subspace, also known as iVector representation [4], and classification is performed by means of logistic regression [2]. In contrastive system 1, classification is performed using a support vector machine (SVM); to do so, each audio segment is represented by means of a Gaussian mean supervector obtained by adapting a universal background model (UBM) to the feature vectors of the audio segment [10]. In contrastive system 2, a Gaussian mixture model (GMM) is trained for each class and the likelihood of a speech segment with each model is computed, selecting the one that obtains the maximum likelihood.

The rest of this paper is organized as follows: Section 2 presents an analysis of the database used in this work; Section 3 describes the audio segmentation approach in detail; Section 4 presents some preliminary results obtained on the training data; and Section 5 depicts some conclusions and future work.

2 Preliminary analysis of the database

Table 1 shows an overview of the two datasets of the Albayzin 2014 audio segmentation evaluation. As there is not a development partition, we decided to perform four different experiments in order to tune the parameters of the proposed system: four partitions of five recordings each were made and, on each experiment, three partitions were used for training while the remaining one was used for testing. Once the parameters of the system were tuned, the whole training dataset was used to train the system.

Table 1. Summary of the datasets of Albayzin 2014 audio segmentation evaluation.

Dataset	# recordings	Duration
Training	20	21 h 16 min 11 s
Test	15	15 h 37 min 50 s

Before developing the system, an analysis of the training data was performed in order to make some design decisions. Albayzin 2014 audio segmentation evaluation consisted on detecting when the classes speech, music and noise were present, which can appear individually or simultaneously. The first design decision consisted on, instead of detecting each class individually, defining a set of seven classes: speech (s00), music (0m0), noise (00n), speech with music (sm0), speech with noise (s0n), music with noise (0mn) and speech with music and noise

(smn). Figure 2 shows the percentage of time that each of this seven classes appears on the training data: it can be seen that there is almost no data of classes noise and music with noise, as they appear in less than 1% of the whole training data. Thus, as the amount of data for these classes was two little to properly train a classifier that detects them, we decided to ignore these classes and keep on with the remaining five.

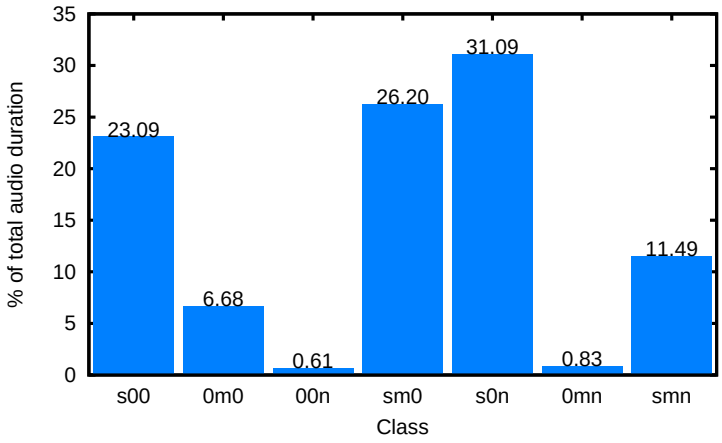

Fig. 1. *Distribution of the duration of the different classes on the training dataset of the database.*

3 System description

The audio segmentation systems presented in this paper have two main stages: segmentation and classification. First, the audio stream is segmented, and then these segments are classified using a classification approach. The specific techniques developed for this work are described in the rest of this Section.

3.1 Segmentation

Before performing segmentation, features were extracted from the waveform; specifically, 12 Mel-frequency cepstral coefficients (MFCCs) plus energy were obtained, leading to feature vectors of dimension $N = 13$. The features were computed using a 25 ms window and a time step of 10 ms, and cepstral mean substraction was applied, computing the mean using all the frames in each file.

The segmentation approach used in this system has three main steps:

– Coarse segmentation. A Bayesian information criterion (BIC) approach is applied in order to select candidate change-points. The BIC criterion is a

hypothesis test to decide whether there is a change-point in a window of data (H_1) or not (H_0) by observing a value ΔBIC: $\Delta BIC > 0$ means that hypothesis H_1 is stronger than hypothesis H_0, i.e. there is a change-point in the window; $\Delta BIC \leq 0$ means that there is no change-point in the window. A BIC segmentation system as described in [3] was implemented to perform audio segmentation: a window of data that slides and grows is analyzed in order to detect a candidate change-point in it by applying the BIC criterion [11]. The BIC algorithm has a tuning parameter λ, which was tuned on the training dataset.

- Change-point refinement. Anytime a candidate change-point is found, a fixed-size window is centered on this change-point and the BIC criterion is applied again in order to refine its position or to discard it. If the change-point is discarded, the system returns to the coarse segmentation stage.
- False alarm rejection strategy. A technique to reduce the number of false alarms was implemented on the BIC algorithm [8]. In this strategy, it is assumed that the change-point occurrences 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 [1].

  An homogeneous Poisson process is characterized by its rate $\gamma = \frac{1}{\tau}$, where τ is the mean time between occurrences (this rate is usually represented as λ, but in this paper we will refer to it as γ in order to avoid confusions with the penalty of the BIC strategy). In a Poisson process, the probability of having n occurrences in a time interval t is:

$$p_n(t) = \frac{\gamma^n e^{-\gamma}}{n!} \tag{1}$$

  Thus, the probability of not having an occurrence in a time interval t is

$$p_0(t) = e^{-\gamma t} \tag{2}$$

  and the probability of having one or more occurrences in a time interval t is

$$p_{\bar{0}}(t) = 1 - p_0(t) = 1 - e^{-\gamma t} \tag{3}$$

  The specific false-alarm rejection strategy is depicted in Figure 3.1: given the ΔBIC value obtained when refining the candidate change-point, we consider that the greater this value, the more likely the candidate change-point is a true change-point. Thus, if ΔBIC is lower than a threshold Θ_{BIC}, we consider this change-point as suspicious of being a false alarm. If this happens, we will discard this change-point with probability p_{discard} which, in this case, is equal to $p_0(t)$, where t is the time interval between the last confirmed change-point and the suspicious change-point. The value of τ was estimated as the median of the segment duration on the training dataset, and Θ_{BIC} was tuned on that data.

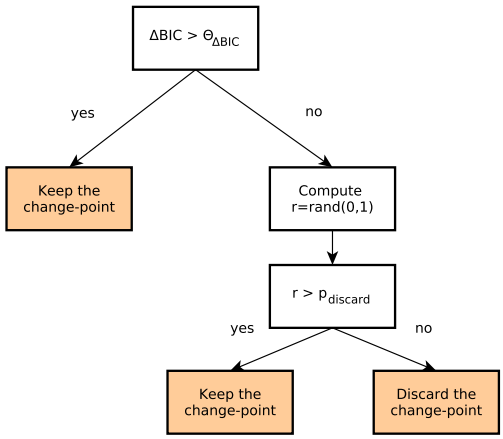

Fig. 2. *Block diagram of the proposed audio segmentation system.*

3.2 Classification

Different classifiers were developed using the segments obtained from the previous step, which are described below.

Primary system: iVector representation with logistic regression. Perceptual linear prediction (PLP) analysis was used to extract 13 cepstral coefficients, which were combined with two pitch features as described in [6], and augmented with their delta features. Hence, the dimension of the feature vectors was 13 PLP + 2 pitch + Δ = 30 features. After extracting the feature vectors, the segments were represented in a total variability subspace: given a Universal Background Model (UBM) with N mixtures, this UBM is adapted to the feature vectors of each segment using Maximum a Posteriori (MAP) adaptation, and the means of the resulting Gaussian Mixture Model (GMM) are concatenated in order to obtain a Gaussian mean supervector for each segment. The iVector technique is applied to the Gaussian mean supervectors, which defines a low-dimensional space, named total variability space, in which the audio segments are represented by a vector of total factors, namely iVector [4]. A Gaussian mean supervector $\mathbf{M}$ is decomposed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{4}$$

where $\mathbf{m}$ is the speaker and channel independent supervector, $\mathbf{T}$ is a low-rank total variability matrix, and $\mathbf{w}$ is the iVector corresponding to the Gaussian mean supervector. In this training stage, the matrix $\mathbf{T}$ is trained as described in [7].

Once the total variability matrix $\mathbf{T}$ is obtained, this matrix and the UBM can be used to extract iVectors from the acoustic features corresponding to

the different segments. In this system, the iVectors were classified using logistic regression with L-BFGS method [2]. Feature extraction, iVector representation and logistic regression were performed using the Kaldi toolkit [9]; the UBM had 512 mixtures and the dimension of the iVectors was set to 400.

Contrastive system 1: Gaussian mean supervector representation with support vector machine classification. The feature vectors used in this classifier were the 13 MFCCs described in Section 3.1, augmented with their delta and acceleration coefficients, leading to feature vectors of dimension 39.

Before classifying the segments obtained from the previous stage, they must be represented in a proper way. In this system, we chose to use a Gaussian mean supervector (SV) representation: a universal background model (UBM), which is a Gaussian mixture model (GMM) with M mixtures, is adapted to the feature vectors of the audio segment [10], and the obtained means are concatenated, forming a supervector of dimension $M \times N$. In this system, the number of mixtures of the UBM was 256, leading to supervectors of dimension 9984.

Classification was performed employing a support vector machine (SVM), which was trained using a set of supervectors and their groundtruth class labels. Specifically, an SVM with a linear kernel and L2-regularized logistic regression was trained for this task, and its cost parameter was tuned on the training data. Once the SVM is trained, it can be used to obtain the class labels of the test audio segments. SVM training and classification were performed using the library for large linear classification LIBLINEAR [5].

Contrastive system 2: GMM-maximum likelihood classification. The audio was represented by means of the 13 PLP cepstral features combined with pitch features and delta coefficients used in the primary system. Classification was performed doing maximum likelihood classification using Gaussian mixture models (GMMs). A GMM of 512 mixtures was trained for each of the five classes mentioned above and, for each segment to classify, the log-likelihood between the feature vectors of the segment and each of the GMMs was computed, selecting the class that achieved the highest log-likelihood [10]. Feature extraction, training and log-likelihood computation were performed using the Kaldi toolkit [9]; the training of the GMMs was performed by doing MAP adaptation of a universal background model (UBM) with full-covariance matrix.

4 Preliminary results

This Section describes different experiments that were performed to make design decisions about the segmentation and classification stages. The four experiments mentioned in Section 2, which are summarized in Table 2, were performed using different audio segmentation approaches. The experimental results are presented in function of the missed class time (MCT), false alarm class time (FACT), class error time (CET) and segmentation error rate (SER).

Table 2. Description of the four audio segmentation experiments. Recording XX stands for file "trackXX" of the training dataset.

Experiment	Training recordings	Test recordings
1	01-15	16-20
2	06-20	01-05
3	01-05, 11-20	06-10
4	01-10, 16-20	11-15

Tables 3 and 4 show the audio segmentation performance achieved when using the classic BIC segmentation approach and when applying the proposed false-alarm rejection strategy, respectively. The classifier used in these experiments was the Gaussian mean supervector representation with SVM classification. Comparing the two Tables, it can be seen that the false-alarm rejection strategy obtained a reduction of the SER by 1% or more in all the experimental cases, proving the validity of the proposed technique. Table 4 also shows that, on two experiments, a SER below 14% was obtained, while a SER by 16% was obtained on the remaining two, leading to a SER of around 15% on the whole training data.

Table 3. Audio segmentation results on the training data using the BIC segmentation stage and Gaussian mean supervector SVM classification.

Experiment	MCT	FACT	CET	SER
1	6.0%	7.3%	3.0%	16.30%
2	5.9%	6.2%	2.7%	14.75%
3	6.1%	8.1%	4.3%	18.54%
4	7.1%	5.8%	2.5%	15.39%
Total	6.3%	6.8%	3.2%	16.27%

Tables 5 and 6 show the results achieved when performing classification using the iVector representation with logistic regression and the GMM-maximum likelihood classification, respectively. Comparing these Tables with Table 4, we can observe that the best audio segmentation results were obtained when using the iVector representation with logistic regression, which improved the results obtained with Gaussian mean supervector SVM classification by 1.5%. A general improvement of all the types of errors was performed, but it can be noted that the lowest false alarm class time was achieved with the GMM-maximum likelihood classifier.

Table 4. Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and Gaussian mean supervector SVM classification.

Experiment	MCT	FACT	CET	SER
1	7.0%	6.8%	2.2%	15.89%
2	6.3%	5.5%	1.9%	13.65%
3	6.3%	6.6%	3.5%	16.39%
4	7.3%	4.8%	1.8%	13.85%
Total	6.4%	6.2%	2.3%	14.96%

Table 5. Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and iVector representation with logistic regression.

Experiment	MCT	FACT	CET	SER
1	6.8%	5.7%	1.4%	13.87%
2	6.2%	4.8%	1.7%	12.72%
3	6.1%	5.4%	2.2%	13.73%
4	5.0%	6.7%	1.7%	13.40%
Total	6.0%	5.6%	1.8%	13.43%

Table 6. Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and GMM-maximum likelihood classification.

Experiment	MCT	FACT	CET	SER
1	6.9%	4.8%	2.7%	14.46%
2	6.0%	4.8%	2.7%	13.45%
3	7.2%	4.8%	3.8%	15.77%
4	7.9%	4.1%	2.4%	14.29%
Total	7.8%	4.4%	3.0%	15.16%

5 Conclusions and future work

This paper described the GTM-UVigo system developed for Albayzin 2014 audio segmentation classification, which consisted on a segmentation followed by classification approach.

The segmentation stage introduced a false alarm rejection strategy which was based in the principle that the occurrence of acoustic change-points follow a Poisson process. Preliminary results on the training data showed that the use of the proposed false alarm rejection strategy led to a 1.5% reduction of the segmentation error rate with respect to the classic BIC approach.

Three different classification stages were submitted, which used different audio segment representation and classification approaches. The best results were obtained when using feature vectors with cepstral PLPs and pitch features represented in a total variability subspace, and performing classification by means of logistic regression. Further analysis must be performed in order to extract stronger conclusions about the segment representation and the classification techniques, as each classifier used a different feature representation, making it difficult to conclude whether the results depend on the representation, on the classification approach or on both of them.

The classification approach that obtained the highest segmentation error was the one that obtained the lowest false alarm class time, which leads to believe that a fusion of the different classification approach may result in a reduction of the segmentation error. Thus, we plan to perform fusion experiments using the different strategies presented in this work in order to improve the audio segmentation performance.

References

1. Allen, A.O.: Probability, Statistics, and Queueing Theory with Computer Science Applications. Academic Press, second edn. (1990)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA (1995)
3. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the BIC. In: Proceedings of ICASSP. vol. VI, pp. 537–540 (2003)
4. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing (2010)
5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification,. Journal of Machine Learning Research 9, 1871–1874 (2008)
6. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2494–2498 (2014)
7. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13(3), 345–354 (2005)

8. Lopez-Otero, P., Fernndez, L.D., Garca-Mateo, C.: Novel strategies for reducing the false alarm rate in a speaker segmentation system. In: ICASSP. pp. 4970–4973. IEEE (2010)
9. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
10. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
11. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)