AUDIO-BASED AUTOMATIC MANAGEMENT OF TV COMMERCIALS

Helenca Duxans, David Conejero and Xavier Anguera

{hdb,dco,xanguera}@tid.es
Telefonica Research,Via Augusta 177, 08021 Barcelona, Spain

ABSTRACT

Although TV commercial identification and clustering are suitable applications for automatic multimedia indexing technology, they remain as problems still unsolved. Most current systems either require a big computational load and therefore can not be executed online, or just perform a detection, without clustering nor identification. In this paper two advertisement indexing approaches are presented: an off-line detection and clustering system and an on-line identification system, both based only on audio features for computational reasons. For the off-line clustering two metrics are evaluated, and an initial commercial boundary detection algorithm, based on identifying drop energy points which are also acoustic change boundaries, is presented. For the on-line system we analyze the response-time/identification scores constraints. Experiments performed on real data validate both off-line and on-line implementations as well as that audio only features are enough discriminant to detect and classify TV commercials.

Index Terms— Multimedia systems, Identification, Clustering methods, Real time systems

1. INTRODUCTION

Automatic multimedia indexing is an emerging technology with multiple potential applications. One of the application fields is advertisement management, where revision and manipulation of huge amount of audio and video data is currently performed daily by publicity companies and the broadcasting industry by hand. The main needs of such companies include monitoring how many times and when target commercials are aired (*off-line advertisement identification*), how many different commercials (and their characteristics) are aired during a target period of time (*off-line advertisement detection and clustering*) and to do on-line notification of target commercials for augmented publicity applications (*real-time advertisement identification*).

Several previous studies have been carried out to detect TV commercials using either video only or audio+video cues. In [4, 6, 9] a combination of rules identifying the dynamics of commercial insertions and image features are used, although video only systems are usually computationally expensive and cannot achieve the performance of systems that also use audio features ([5, 7, 2, 3]). To the authors knowledge, no previous studies about advertisement clustering and real time identification have been published.

The goal of this study is to evaluate the viability of audio-based TV commercial detection, clustering and identification systems. On the one hand, an off-line algorithm is introduced to detect and cluster advertisement repetitions present in prerecorded material. On the other hand, advertisement identification will be performed on-line based on the previous introduced algorithms optimized according the real-time constraints. The proposed on-line identification system does not require the analysis of the whole advertisement, because the identification will be performed using as less discriminant data as possible, allowing for augmented publicity applications to be carried out.

The outline of the papers is as follows. In section 2 the proposed off-line advertisement detection and clustering system is explained, introducing the boundary detection algorithm and two different metrics for the clustering. Section 3 describes the on-line advertisement identification system. Experiments on advertisements detection, identification and clustering, and on the relationship between computational load and efficiency of the real-time implementation are presented in section 4. Conclusions are found in section 5.

2. OFF-LINE AUTOMATIC DETECTION AND CLUSTERING OF ADVERTISEMENTS

An automatic off-line TV commercial detection and clustering system should determine the start/end points of every advertisement present in a broadcasted media recording and group the repetitions of the same advertisement together. In this section we reason the approach taken for audio-based advertisement boundary detection and the different metrics used for the clustering.

2.1. Advertisement boundary detection

In order to detect advertisement boundaries a prior analysis of broadcasted media in Spain concluded that advertisements are always isolated by a decrease of the audio signal energy (about $10ms$ to $30ms$) occurring just before and after them. Moreover, advertisements usually had standard defined lengths (10s, 20s, 30s) although there were some exceptions such as TV channels self-promotions, very long TVShop-like commercials, etc.

Based on these observations, we designed the three step detection system presented here. During the first step, the drop energy points within the audio signal are found using a narrow average energy window (about tens of milliseconds) and are considered candidates to commercial start/end points. The optimum window length is a tradeoff between detection efficiency and false alarms.

The second step of the system only validates the candidates that separate different acoustic environments, to filter out false alarm drop energy points corresponding to digitally-edited material. In order to find out if a drop energy point is also an acoustic change point an algorithm based on the Bayesian Information Criterion (BIC [8]) is used. BIC dissimilarity (ΔBIC) compares two probability mod-

els in order to select the model that better represents the data via a likelihood criterion penalized by the model complexity, i.e. the number of parameters in each model. The expression of ΔBIC is shown in Eq. 1, where Θ_0 and Θ_1 are the two compared models, $\mathcal{L}_\Theta$ is the log likelihood function, ΔK the difference of the number of the model parameters, N the number of data points we are modeling and λ the penalty weight (typically $\lambda = 1$).

$$\Delta BIC(\Theta_0, \Theta_1) = \mathcal{L}_{\Theta_0} - \mathcal{L}_{\Theta_1} - \frac{\lambda}{2}\Delta K \log N \tag{1}$$

When $\Delta BIC > 0$ the model Θ_0 is selected as the model that better represents the data, otherwise Θ_1 is selected.

For the advertisement boundary detection application, we have formulated two hypothesis H_0 and H_1, modeled by Θ_0 and Θ_1 respectively. H_0 considers that both sides of the change point, $\mathcal{X}_a$ and $\mathcal{X}_b$, share the same acoustic environment and H_1 considers that both sides belong to different acoustic environments. Each hypothesis is modeled by a Gaussian Mixture Model (GMM), following the BIC-like algorithm proposed by [1] where H_1 is modeled by a GMM per side (Θ_{1a} with $Q_{\Theta_{1a}}$ components and Θ_{1b} with $Q_{\Theta_{1b}}$ components), and H_0 is modeled by only one GMM (Θ_0, with $Q_{\Theta_0} = Q_{\Theta_{1a}} + Q_{\Theta_{1b}}$ and therefore $\Delta K = 0$). Under these considerations, the BIC distance (ΔBIC) is computed as shown in Eq. 2:

$$\begin{array}{l}\Delta BIC(H_0, H_1) = BIC(H_0) - BIC(H_1) = \\ \mathcal{L}(\mathcal{X}_a, \mathcal{X}_b|\Theta_0) - \mathcal{L}(\mathcal{X}_a|\Theta_1 a) - \mathcal{L}(\mathcal{X}_b|\Theta_1 b)\end{array} \tag{2}$$

where the GMM's are trained on MFCC feature vectors. Only those drop energy points whose $\Delta BIC < 0$ will be considered acoustic change points and be analyzed further.

The third step of the advertisement boundary detection system validates the candidates whose time distance D to another change point fulfills

$$\left| D - c_l^{'} \right| < \epsilon \tag{3}$$

where $c_l^{'}$ belongs to the considered advertisements lengths $C_l^{'} = \{10s, 20s, 30s\}$ and ϵ is the accepted time error, which corresponds to the imprecision at detection of the beginning and the end of an advertisement. In the case that more than one distance $c_l^{'}$ may be fitted, only the smallest is considered.

The resulting segments after step three define the set of detected advertisements. These are stored in a database together with the time-stamp and emission channel. Further information is extracted via analysis by the clustering system described below.

2.2. Clustering of advertisement repetitions

Once the advertisement boundaries have been detected the clustering is carried out to group all the instances of the same advertisement across all stored media channels. For every new advertisement the similarity to all the advertisements of the clusters with the same length is computed. If the mean similarity to one cluster is greater than a preset threshold, the new advertisement is added to that cluster. If not, a new cluster is created.

Detected instances of the same advertisement will not be usually identical due to two main factors. First, not all instances contains the same background noise or have the same TV channel characteristics (the audio volume for example). Second, the detection process includes or deletes random amounts of audio frames at the beginning and ending of the advertisements. Two different metrics have been evaluated as similarity measures: the inverse of the total cost of a Dynamic Time Warping (DTW) MFCC alignment and the maximum value of spectral cross-correlation.

DTW finds the optimal alignment path between two sequences, and also the total distance between them, according to some distance measure an under some alignment restrictions defined for each application. The DTW used in this application has been designed taking into account the restriction that when comparing two instances of the same advertisement the alignment selected for the central part will be always diagonal. According to this fact, the cost of all diagonal alignment paths with initial points $\{y, 0\}$ for $y = y_{max}, y_{max}-1...0$ and initial points $\{0, x\}$ for $x = 0, 1...x_{max}$ are computed and normalized by the corresponding frame length.

The similarity measure S_{DTW} computed by the DTW algorithm corresponds to the maximum value of the inverse cost of the diagonal paths (see Eq. 4).

$$S_{DTW} = 1/min\{DTW_i(x, y), DTW_j(x, y)\} \tag{4}$$

with

$$DTW_i(x, y) = \begin{cases} 0 & x = x_i; y = 0 \\ DTW(x-1, y-1) + & x_i < x \leq X_{max} \\ + \frac{D(x,y)}{X_{max} - x_i} & 0 < y \leq Y_{max} - x_i \end{cases}$$

and

$$DTW_j(x, y) = \begin{cases} 0 & x = 0; y = y_j \\ DTW(x-1, y-1) + & 0 < x \leq X_{max} - y_j \\ + \frac{D(x,y)}{Y_{max} - y_i} & y_j < y \leq Y_{max} \end{cases}$$

where $D(x, y)$ are the distance between x^{th} and y^{th} MFCC components.

As a design criterion, Y_{max} and X_{max} were fixed to the frame length of the advertisements minus the sampling rate, in order to allow not to take into account the complete first and last seconds of one or both advertisement instances.

The second metric evaluated was the standard spectral cross-correlation implementation over the acoustic signal. Both signals to be compared were first multiplied by a Hamming window in order to decrease the influence of the initial and ending regions. The similarity measure corresponds to the maximum of the spectral cross-correlation normalized by the signal powers.

3. REAL-TIME ADVERTISEMENT IDENTIFICATION

Although the off-line advertisement detection and clustering system is a good solution for media indexing and archival, it is not applicable to on-line systems where real-time constraints apply. For this reason we extended the previous technology to build a real-time advertisement identification system which is designed to work continuously inspecting a TV broadcast channel in order to detect, as fast as possible, when one of a set of target advertisements is on the air. Such target advertisements can be manually defined by the content provider or automatically by an off-line clustering system.

Given that a real-time system has a very constrained requirements about computational load and time latency, the identification block is only activated when it is probable that a beginning of an advertisement occurs and performs the identification using the minimum required data with the fastest metric to assure optimum reaction time/identification score balance.

3.1. Activation of the identification

In order to activate the identification process, the audio stream of the TV broadcast channel is inspected every second looking for a drop in the mean energy. To determine the drop, each second (*activation gap*) is divided into shorter non-overlapping windows and the ratio between every window mean energy and the mean energy of the complete second is calculated. Only when the minimum ratio is lower than an activation threshold the system performs the identification.

The amount of time that the identification is working will be called *active time*. The window length used to calculate the mean energy and the activation threshold should be optimized to assure the maximum reduction of the active time.

3.2. Advertisement identification

Once the identification system is activated, the N seconds of the audio stream following that point are compared with the first N seconds of the target advertisements, which have been already stored in the system database. The identification is considered positive when the distance between the audio stream and the target advertisement is bellow a threshold.

From the two different metrics proposed in section 3.2 for advertisement comparison, we selected the correlation between the Fourier coefficients of the audio stream and the target advertisements given its direct applicability to the raw signal, and therefore smaller processing time in the current implementation. We allow ± 0.5 seconds shift between the signals (half of the activation gap) to assure maximum overlap between the audio stream and the target advertisement. The identification is positive when the maximum correlation value is over a predefined threshold.

4. EXPERIMENTS

Experiments with the presented algorithms were performed on data collected from six TV broadcast channels in Spain. The recorded data was split into a development and test sets. The development set corresponds to eight video files, with a total length of 8h55min and the test set corresponds to three video files, with a total length of 3h50min, containing 212 and 135 advertisements with various lengths within C_l', defined earlier.

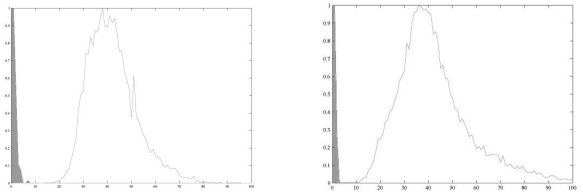
4.1. Advertisement boundary detection performance

The advertisement boundary detection system is composed by three interdependent functional blocks, each one with several free parameters: the window length and threshold for the drop energy points detection, the number of GMM components for the acoustic change detection and the accepted time error for the advertisement length filtering. As it does not exist a global optimization process to determine all the parameters together, our experiments have been focused on optimizing each block separately using the development database.

Table 1 summarizes the performance of our system for the best working point evaluated. The detection performance is measured in terms of the precision (PRC), defined as the percentage of well detected boundaries (#well detected boudaries/#boundaires) and the recall (RCL), defined as the percentage of detected points corresponding to true advertisement boundaries (#well detected boudaries/#detected points).

	#Ads	#Detected	PRC	RCL
dev	212	181	85.38%	85.38%
test	135	112	81.16%	82.96%

Table 1. Advertisement boundary detection results. *Ads* column indicates advertisement appearance count and the *Detected* column the number of correctly detected advertisements.



a) DTW distances b) SCC distances

Fig. 1. Histograms for equal (shadowed plot of each figure) and non-equal (no-filled) advertisement distance.

The system correctly detects 82% of the advertisements on the test database. These results improve when using the detection and clustering together, by adjusting the detection block to reach 100% precision and using the clustering block to increase the recall by eliminating the clusters with too few advertisements.

4.2. Advertisement clustering performance

To study the performance of the similarity metrics proposed for the clustering, the distances (inverse of the similarity) between all the detected advertisements have been computed. On Figure 1 are represented the distances between equal and non-equal advertisements for the DTW cost and the spectral cross-correlation (SCC). It can be observed that there is a considerable distance between the equal and non-equal distances for both metrics, although it is greater for the DTW cost. If the threshold value to determine whether a new advertisement belongs to one cluster is selected to not to group together any different commercial, 99.12% all the material for DTW and $97,37\%$ for SCC are well classified.

For these experiments both development and test databases have been used together since there are not enough repetitions in the test database to validate the conclusions.

4.3. Identification activation performance

The goal of the activation block in a real-time advertisement identification system is to reduce the computational load by activating the identification only when it is most probable that a beginning of an advertisement occurs. In this section we study the relationship between the length of the window used to compute the energy drops, the energy threshold used and the active time of the identification system.

To determine the energy threshold for each one of the evaluated window lengths we have collected the drop energy values of the intervals with advertisement start points of the development database. The maximum energy value, incremented by a safety interval of 5%,

Length	Threshold	%reduction
50ms	0.0315	56.02%
100ms	0.2139	28.04%
200ms	0.5623	18.36%

Table 2. Energy threshold values and percentage of active time reduction for the test database.

has been used as a energy threshold in the test database. The value of the threshold has been taken so conservative in order no to miss any advertisement due to not activating the system.

Results for the test database are shown in Table 2. It can be seen that with a window length of $50ms$ the number of seconds that the identification block is activated, and therefore consuming CPU, is reduced by 56.02%, without missing any start point.

4.4. Advertisement identification performance

In order to evaluate the performance of the advertisement identification block and the minimum audio data required for the identification we have run experiments with different comparison interval lengths on the test database, computing the number of well identified advertisements. The target advertisement used for the experiments were the repeated ones in the development database, eliminating the advertisements that, although are different, share the initial seconds (usually modified commercials of the same publicity campaign). According to the results shown in Table 3 the minimum length to assure 100% identification score is $2s$.

Length	M_{ne}	m_e	Threshold	%Ident
5s	0.3253	0.9092	0.7632	100%
4s	0.4764	0.8988	0.7932	100%
3s	0.5668	0.7880	0.7327	100%
2s	0.6200	0.6599	0.6499	100%
1s	0.5392	0.2301	—	—

Table 3. M_{ne} and m_e for the development database, threshold values and percentages of well identified advertisements of the test database.

To determine the threshold to decide when the audio stream corresponds to a target advertisement we have collected all the distance values obtained when the identification system is fed with the development database and the target advertisements correspond to the repeated ads present in the recordings. The selected threshold (*Th*) is as follows:

$$Th = m_e - 0.25(m_e - M_{ne}) \quad (5)$$

where m_e is the minimum distance between equal segments and M_{ne} is the maximum distance value for non-equal segments. This bias to m_e is due to the design criterion to prefer not to identify an advertisement than to miss-identified an audio segment.

It must be remarked that for an interval of $1s$ length it is not possible to determine a threshold. This is due to the fact that the activation of the identification is determined every second.

5. CONCLUSIONS

In this paper we introduce several algorithms for TV commercial detection, identification and clustering, that all converge into two systems, for on-line and off-line data, using only acoustic features. Our proposal differs from currently proposed solutions in that other solutions fall short in the analysis they perform of the detected advertisements, usually not providing any clustering of the found advertisements. Moreover, extensive computations (for example using video information) are performed not allowing for the systems to be executed in real-time with online data.

In this paper we have proposed an advertisement detection and clustering system with a three step off-line refinement process, based on energy-drop and acoustic change detection, which achieves a precision and recall over 81% (with room for improvement by combining it with the clustering result analysis). In the clustering step, we have compared two possible metrics. Both DTW cost and spectral cross-correlation allow clustering with almost 100% precision, with greater discriminatingly between clusters using DTW. Moreover, a real-time advertisement identification has been presented, where we have introduced a system activation algorithm that reduces computation up to 56.02% and with a response-time below three seconds.

6. REFERENCES

[1] J. Ajmera, I. McCowan, H. Bourlard. Robust speaker change detection. Tech. report, IDIAP, 2003.

[2] M. Covell, S. Baluja, M. Fink. Advertisement detection and replacement using acoustic and visual repetition. In *Proc. IEEE 8th Workshop on Multimedia Signal Processing*, pp. 461-466, Oct. 2006.

[3] P. Duygulu, M. yu Chen, A. Hauptmann. Comparison and combination of two novel commercial detection methods. In *Proc. ICME*, Taiwan, 2004.

[4] A. G. Hauptmann, M. J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *Proceedings ADL'98*, Santa Barbara, USA, 1998.

[5] X.-S. Hua, L. Lu, H.-J. Zha. Robust learning-based tv commercial detection. In *Proc. ICME*, 2005.

[6] R. Lienhart, C. Kuhmnch, W. Effelsberg. On the detection and recognition of television commercials. In *Proc of IEEE Conference on Multimedia Computing and Systems*, pages 509–516, Otawa, Canada, 1997.

[7] D. A. Sadlier, S. Marlow, N. O'Connor, N. Murphy. Automatic tv advertisement detection from mpeg bitstream. *Journal of the Pattern Recognition Society*, 35(12):2–15, 2002.

[8] S. Shaobing Chen, P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[9] J. Sanchez, X. Binefa. Audicom: a video analysis system for auditing commercial broadcasts. In *Proc. of ICMCS'99*, Firenze, Italy, 1999.