A TV Commercial Retrieval System based on Audio Features

Jose E. Borras[1], Jorge Igual[2], Carlos Fernandez-Llatas[1] and Vicente Traver[1]
[1] *ITACA-TSB, Universidad Politecnica de Valencia, Valencia, Spain*
[2] *ITEAM, Universidad Politecnica de Valencia, Valencia, Spain*

Keywords: Pattern Recognition, TV Commercial, Audio Features, Detection.

Abstract: In spite of new digital platforms, television (TV) continues to be the most influential advertising medium. The advertisers need to verify that their commercials are broadcasted on TV in the number and time they pay for them. Nowadays, this job is done manually by visual inspection of recordings of the broadcasted signal every day, consuming a lot of human resources. We present a system that automatize the process of identification of TV commercials. It is based on the detection of target commercials using their audio features. With the purpose of reducing the time of detection and the storage requirements, it uses audio features in a compact transformed domain. The algorithm is based on the similarities in the cepstral domain of the commercial to be detected and the audio recording of the TV signal. The results show that the system is able to obtain a satisfactory detection rate in a short time (detection rate above 90% with no false alarms), allowing the analysis of long recordings in a fast way.

1 INTRODUCTION

This paper focuses on the development of a system able to localize known TV commercials in large databases in a fast and effective way.

The advertising paradigm is changing in the digital era. Traditional mass media advertising is revealing less efficient than digital interactive marketing. It does not mean the end of mass media and general purpose advertising in a short time. It just means that marketing is becoming more complicate and that tools that can help in the automatizing of processes are becoming very valuable. Traditional advertising includes direct mail, television (TV), magazines, outdoor, newspapers, and radio. Among all of them, TV and radio are the ones that probably will never disappear, although their percentage will decrease, increasing the investment in the new digital media, basically social networks (internet), online video, and mobile marketing.

The transformation of marketing to the digital era has the advantage that pattern recognition techniques (Duda et al., 2000) can be applied to solve classical detection, identification or classification problems in advertising in a more efficient way. Machine learning can provide a series of very helpful automated monitoring tools to the companies on charge of auditing advertising campaigns.

Traditional TV is also moving to a new paradigm where interactivity and, as a consequence, personal TV is substituting typical broadcasting model. The fragmentation of audience in different technologies and specific content based channels complicates basic tasks in advertising management.

The measurement of the effectiveness of an advertising campaign is very controversial (Lavidge and Steiner, 2000). The first and most obvious task in advertising management is to verify that the commercial campaign has been broadcasted correctly, satisfying the signed advertising contract about number of commercials, time, duration, etc. This work is usually carried out manually by experts that visualize hours and hours of TV and extract the information of interest, i.e., in channel X the commercial Y was broadcasted from Monday to Friday at 20:00 PM. This procedure requires the use of a huge amount of resources, not only human.

Another problem is that real time supervision is almost impossible since it requires even more resources. The typical procedure is that at the end of the day (depending on the contract), reviewers supervise the recorded video looking for the commercial breaks and identifying the corresponding commercials of interest. Considering the growing number of channels, the increasing number of technologies involved in the broadcasting of multimedia con-

E. Borras J., Igual J., Fernandez-Llatas C. and Traver V..
A TV Commercial Retrieval System based on Audio Features.
DOI: 10.5220/0004511800650070
In *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications and 10th International Conference on Wireless Information Networks and Systems* (SIGMAP-2013), pages 65-70
ISBN: 978-989-8565-74-7

tent and the corresponding logistic problems in the recording of these contents and human supervision, advertising agencies are requiring the automating of many of these processes or at least the development of tools than can alleviate the load of the expert. Although the final decision is taken by the human expert, any application that can reduce the time dedicated by the expert to the supervision is very useful.

In the field of TV commercials, there are two main different problems. One is the identification of the commercial breaks. This is done by the detection of some audio and video features such as blank screens, change in the volume and some other discriminating characteristics. In the literature we can find many algorithms to detect the commercial breaks; see, e.g., (Lienhart et al., 1997), (Satterwhite and Marques, 2004) or (Zhang et al., 2012).

The second problem is the analysis of the commercial. In this case, the pattern recognition problem can be stated in different ways. The most basic one is the detection (identification) of a known commercial to assure that the campaign is broadcasted according to the terms of the agreement with the publicity agency (Duan et al., 2006). Another one is an unsupervised classification approach, i.e., there is not a target advertisement to look for but some clusters that can represent different classes of commercials, e.g., attending to their content (Hua et al., 2009).

In this paper, we address the detection problem or commercial retrieval, i.e., multimedia search in long recordings of broadcasted TV signals by a given commercial query. We will assume that the commercial breaks are correctly localized in the time domain and the goal is the identification of some known commercials in the recorded segments in a fast and effective way.

The key point, as in most pattern recognition problems, is to find a domain where the classes are separable or, in a more realistic approach, to find out the most discriminating features according to the problem under consideration. In the case of TV commercials, we have two different kinds of signals: the audio and video information. Since we are interested on a system that works rapidly and not demanding too many resources in terms of memory and computational load, we will explore in this paper a solution based on the audio features.

2 DETECTION OF TV COMMERCIALS BASED ON AUDIO FEATURES

The system is composed of three stages: in the first one, the recorded broadcasted signal is preprocessed in order to reduce its length and to obtain an input signal that is composed only of commercial breaks; in the second one, the descriptor of the query commercial is calculated and, in the third one, the detector is applied.

2.1 Preprocessing

The data come from real recordings of broadcasted TV signal. The first task is the extraction of the commercials. As we mentioned, there are many algorithms to carry out this work. We will assume that it is done in advance. In our case, we use Comskip (Comskip, 2012), a free MPEG commercial detector. It is a windows console application that reads a MPEG file and using information related to logo, black frames, silences, changes in aspect ratio and so on is able to indicate the time where a commercial break starts and ends.

Comskip is a configurable software, so some parameters must be set previously by trial and error depending on the broadcaster, i.e., TV channel. We use a set of parameters that pursue the goal that no commercial fragment is missing. In order to reassure this, we add a minute of broadcasting before and after to every commercial break detected by Comskip. This implies that the input data can contain some content that does not correspond to advertising. This is a trade off between the rapidness of the detection procedure and the overall system type II error or false negatives, i.e., commercials that are not detected in the broadcasting.

Finally, the different time periods of advertising breaks are concatenated obtaining a signal that is composed mostly of commercials $x[n]$. We will use only the audio part of the commercial, reducing the storage requirements and the time of computation. Some systems to detect and recognize commercials using the video content can be found in the literature, such as (Putpuek et al., 2010) or (Wu et al., 2010).

Note that the duration of commercials on TV range from a few seconds to a minute or even more in exceptional circumstances, depending on each country, broadcaster, time of the day or TV program. Considering the audio content, the diversity we can find is enormous. Some are based on people talking; some others are based on music; even there may be silent in very few occasions. This diversity means that, in order to obtain a general procedure, we can not exploit specific properties of speech or music waveforms. The goal of our proposal is to obtain a general purpose algorithm with an effective detection rate and a low computational cost.

The last step in the preprocessing is the reduction of the amount of data. Since one of the restrictions of the detection algorithm is its low computational cost, it is useful to reduce the dimension of the data. In addition, considering that the input data $x[n]$, i.e., the merging of commercial breaks, is extremely huge in nowadays TV with a lot of different channels bradcasting 24 hours a day, the dimensional reduction of the input is a must.

We use wavelets for this purpose. A wavelet is a waveform of limited duration giving location in time with some vanishing moments and finite energy (Mallat, 2009). The waveform depends on which wavelet family is chosen. Its localization features in time and frequency domain make them a very attractive tool in analyzing non stationary signals such as speech signals.

The discrete wavelet transform of $x[n]$ is obtained passing it through a set of filters. At the first level, the input signal is decomposed as the combination of some approximation coefficients (obtained with a low pass filter) and the detail coefficients (obtained with a high pass filter). In order to keep the same the length of the input signal and the decomposition, the low and high pass filtered signals are downsampled by 2. Therefore, the first level approximation of $x[n]$ can be expressed such as:

$$v[n] = \sum x[k]g[2n-k] \qquad (1)$$

where $g[n]$ is the impulse response of the low pass filter. In our case, $g[n]$ corresponds to the Daubechies wavelet of order 9.

The signal $v[n]$ looks like an approximation of the original one with half the length, i.e., we have reduced the time resolution by a factor of 2. In our case, we work with audio signals sampled at 11025 Hz, so this reduction is irrelevant in terms of accuracy in the time location of the commercial.

Next step is the representation of the commercials in a very compact way; it requires the transformation to a domain where the discriminative characteristics of the audio part of a commercial can be represented by a short feature vector or descriptor. At the same time, this vector must be as easy to obtain as possible since we will have to apply the same transformation to the time series $v[n]$.

2.2 Descriptor of the Commercial

In the case of audio recognition, the feature vector is usually composed of spectral descriptors and dynamic time characteristics. In our case, we are not limited to a kind of particular audio signals, as it is the case in speech recognition systems or music information retrieval. It implies that the information is not concentrated in a particular type of acoustical signal parameters, such as for voiced or unvoiced sounds. Any kind of sound content, including its absence, i.e., a silent segment, can be helpful in the identification of the commercial.

We use the cepstrum as the starting point to obtain the descriptor; the cepstrum coefficients are being used as the feature vector in many pattern recognition applications working on audio input signals, such as in (Furui, 1986). A variation that takes into account the way the human ear filters the sound in the frequency domain is the mel-frequency cepstrum, that uses a bank of filters with triangular shape and different bandwidth.

The real cepstrum is defined for a real signal $z[n]$ such as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of $z[n]$:

$$c_z[n] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \log\left|Z(e^{j\omega})\right| e^{j\omega n} d\omega \qquad (2)$$

where $Z(e^{j\omega})$ is the Fourier transform of $z[n]$.

In order to compress the signal, it is common to keep only the first 10 or 20 coefficients. But this is not possible in our problem since the application of the same transformation to $v[n]$ would produce a time series still too long. So we have to reduce the number of coefficients. To minimize the length of the descriptor, we finally order the coefficients by magnitude and keep the index position of the largest coefficient. This index becomes one element of the descriptor.

Repeating the procedure to every 40 milliseconds of the signal, we obtain the definitive vector $w[n]$ that describes the commercial. The 40 milliseconds fragments are obtained using a Hamming window to avoid transients in the border but keeping the signal quasi stationary in the interval. The window is moving overlapping one third between blocks.

As an example, in figure 1 we show two commercials and their corresponding representation in the feature domain. The top figures correspond to a case where the audio content is mostly music. We show the time waveform (left) and the feature vector (right). The bottom figures are an example of a waveform where the audio content is basically a person talking. Since we are working with commercials of different

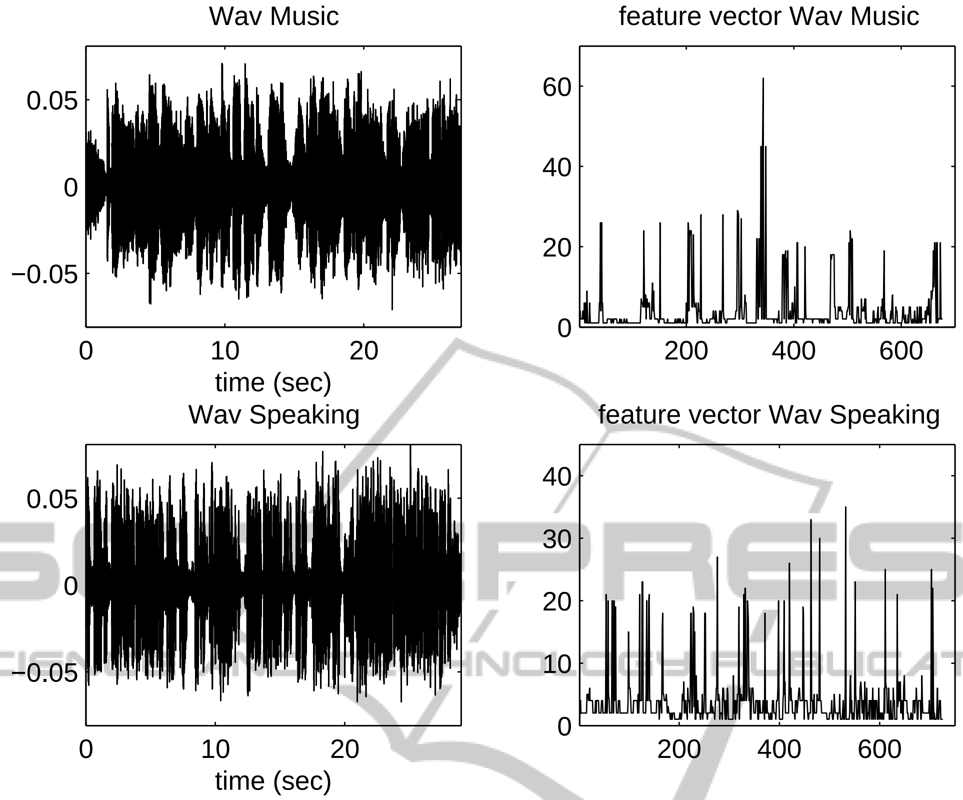

Figure 1: Top-left: commercial with music. Top-right: feature vector of it. Bottom-left: commercial with speaking. Bottom-right: feature vector of it.

duration, the length of the descriptor depends on the duration of the commercial.

The same transformation is applied to the preprocessed input signal $v[n]$, obtaining the vector $y[n]$.

2.3 Detector

The detection is carried out using classical detector based on the cross-correlation between the feature vector of the commercial $w[n]$ and the preprocessed audio signal $y[n]$. It is defined such as:

$$r(m) = \frac{\sum_{i=1}^{N}(w(i)-\hat{w})(y(m+i)-\hat{y})}{\sqrt{\sum_{i=1}^{N}(w(i)-\hat{w})^2 \sum_{i=1}^{N}(y(m+i)-\hat{y})^2}} \qquad (3)$$

where $\hat{w}$ is the mean value of the vector $\mathbf{w} = (w[1],\ldots,w[N])$ and $\hat{y}$ is the mean value of the corresponding fragment of $y[n]$, i.e., $\mathbf{y} = (y[m],\ldots,y[m+N-1])$.

The detection rule reads as: a commercial is detected at time m when correlation function $r(m)$ is greater than certain threshold λ. The value of λ is established empirically since we do not have a probabilistic model for the underlying hypothesis.

Note that the correlation can be greater than the threshold for some consecutive time indexes. The algorithm keeps as the detection instant the sample m_0 where $r(m)$ is maximum. The actual time location of that index in the original broadcasted signal is easily obtained since we keep track of all the transformation of the preprocessed signal. The algorithm can be personalized if we include the duration of the commercial as a parameter that helps to select the instants where the detector is applied.

The only parameter of the system is the threshold value λ. Its value can be updated in a dynamic way using some supervised detection periodically, e.g., when the broadcaster has changed. In any case, the system is tested periodically to check that the false alarm and detection rates are adequate according to the requirements of the system.

3 RESULTS

The input data are recordings from different digital television channels. Comskip is applied to the MPEG signal to extract the commercial blocks and the audio part is digitized using a sampling frequency of 11025 Hz and 16 bits. Since we are adding one minute before and after Comskip detects a commercial break, our detection algorithm can give duplicate positives when Comskip fails and our searched commercial is in this extra time and the time between commercial blocks is one minute or less (a very rare situation).

As an example of the detection procedure, in figure 2 we show the correlation between the searched commercial and 30 seconds of broadcasted signal after preprocessing. It is clear that the peak in the function indicates the presence of the commercial at that time. As we can appreciate, the signal to noise ratio is very large, so the threshold in the decision making can be set in a large confidence interval.

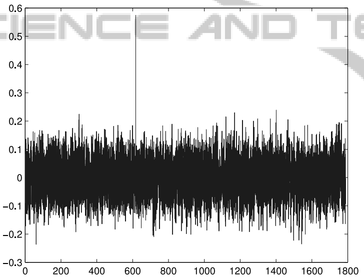

Figure 2: Cross-correlation function. The peak around sample $n = 800$ corresponds to a detection.

The algorithm has been tested in different conditions. In table 1 we show the results for the following experiment. We obtained 22 recordings from a general purpose TV channel at different times of the day. The duration of every recording is between thirty minutes and one hour. The database includes 350 commercials, including some of them very similar since they belong to the same advertising campaign. The duration of each of them is between 6 and 70 seconds. The threshold value is set to 0.4 to minimize false alarms. As we can see, there are recordings without commercials and some others with a lot of them (prime time). The average detection rate is 92% and the false alarm is zero, i.e., there are not false positives.

The most important advantage of the presented method with respect to the application of correlation between the commercial time series (or corresponding Fourier transform) and the input signal is its rapidness. Thus, an important factor to be considered is how to divide the input audio stream in order to reduce the time of computation.

Table 1: Number of commercials that are (are not) detected in corresponding recording. There are not false alarms.

Record#	Detected	Non Detected
Record 1	12	0
Record 2	14	2
Record 3	9	0
Record 4	0	0
Record 5	0	0
Record 6	0	0
Record 7	0	0
Record 8	0	0
Record 9	20	3
Record 10	0	0
Record 11	26	1
Record 12	19	2
Record 13	9	0
Record 14	42	2
Record 15	5	1
Record 16	10	3
Record 17	24	4
Record 18	24	1
Record 19	39	4
Record 20	42	3
Record 21	25	1
Record 22	24	1
Total	335	28

The time of computation of the commercial descriptor is fixed; but we can compute the time required by the overall system attending to two variables. First, when we transform the input audio signal to the feature domain, how do we divide the signal?, i.e., the duration of the fragments where we apply our transformation in order to obtain $y[n]$. Second, the number of commercials to be detected.

In order to evaluate these times, we applied the algorithm to a recording of one hour of commercials. We divided the one hour audio stream in non overlapping fragments of duration 5, 20, 30 and 60 minutes (the complete signal). We applied the algorithm for a different number of commercials, from a single one to fifty.

The total computational time is shown in figure 3. This time includes the calculation of the commercial signatures, the transformation of the audio signal to the feature domain and the correlation. As we can see, the best results are obtained when the input audio stream is divided in blocks of 20 minutes each one.

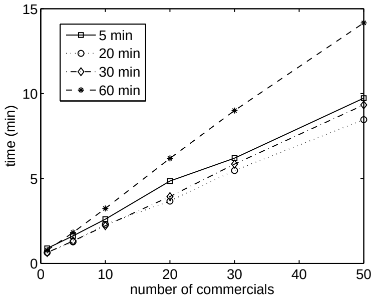

Figure 3: Time of computation in minutes vs. number of commercials to be detected for one hour of commercials. The input signal (60 minutes) is divided in blocks of 5, 20, 30 or 60 minutes.

4 CONCLUSIONS

We have presented a system for the commercial recognition in TV signals. We focused on the feature extraction procedure, where each entry of the feature vector corresponds to the position of the largest cepstral coefficient of a windowed segment of the commercial. This descriptor allows the representation of the commercial in a very compact and discriminative way, so the detection algorithm based on classical similarity measure obtained through the correlation between the input data and the commercial in the transformed domain achieves a very good performance in a short time.

In future work, the algorithm can be optimized considering specific additional information about the commercial, such as duration or content characteristics.

REFERENCES

Comskip (2012). http://www.kaashoek.com/comskip.

Duan, L.-Y., Wang, J., Zheng, Y., Jin, J. S., Lu, H., and Xu, C. (2006). Segmentation, categorization, and identification of commercial clips from tv streams using multimodal analysis. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 201–210. ACM.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(1):52–59.

Hua, X.-S., Lu, L., Li, M., and Zhang, H.-J. (2009). Learning-based automatic commercial content detection. US Patent 7,565,016.

Lavidge, R. J. and Steiner, G. A. (2000). A model for predictive measurements of advertising effectiveness. In *Advertising & Society Review 1, pp. 59–62*.

Lienhart, R., Kuhmunch, C., and Effelsberg, W. (1997). On the detection and recognition of television commercials. In *Multimedia Computing and Systems' 97. Proceedings., IEEE International Conference on*, pages 509–516. IEEE.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing*. Elsevier.

Putpuek, N., Cooharojananone, N., Lursinsap, C., and Satoh, S. (2010). Unified approach to detection and identification of commercial films by temporal occurrence pattern. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3288 –3291.

Satterwhite, B. and Marques, O. (2004). Automatic detection of tv commercials. *Potentials, IEEE*, 23(2):9–12.

Wu, X., Putpuek, N., and Satoh, S. (2010). Commercial film detection and identification based on a dual-stage temporal recurrence hashing algorithm. In *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval*, VLS-MCMR '10, pages 25–30, New York, NY, USA. ACM.

Zhang, B., Li, T., Ding, P., and Xu, B. (2012). Tv commercial detection using audiovisual features and support vector machine. In *Instrumentation & Measurement, Sensor Network and Automation (IMSNA), 2012 International Symposium on*, volume 1, pages 322–325. IEEE.