Audio Fingerprinting System for Broadcast Streams

Cláudio Neves[†], Arlindo Veiga[†], Luís Sá[†,¥], Fernando Perdigão[†,¥]
[†] Instituto de Telecomunicações, Departamento de Eng.ª Electrotécnica e de Computadores,
Universidade de Coimbra – Pólo II, 3030-290 Coimbra, Portugal
[¥] Dep. de Eng.ª Electrotécnica e de Computadores, Universidade de Coimbra – Pólo II, 3030-290 Coimbra, Portugal
e-mails: {claudiorneves, aveiga, luis, fp}@co.it.pt

***Abstract* — This paper describes an audio fingerprinting system to search for jingles and advertisements in broadcast streams, using a fast bit-pattern matching algorithm. The system is described and evaluated using broadcast television streams. The results show that the system is extremely efficient and robust in situations where the audio items to locate do not overlap in time or suffer from time distortions such as compression or expansion.**

I. INTRODUCTION

Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A lot of examples and applications of fingerprinting can be found in the audio and even in video domain, e.g., broadcast monitoring [1], music identification (query-by-example) using a mobile phone [2], copyrighted multimedia control [3], television applications based on audio identification [4] and many others. These examples are the consequence of the growing availability of multimedia files in the last years. In the literature, the audio fingerprinting problem is referenced to by many different terms, such as "audio indexing", "audio identification", "jingle detection", or "detection of acoustic patterns".

Many media streams have the so called "jingles", a small identifier of the beginning or of the end of a broadcast program. Its detection can be used as a fast segmentation or classification of broadcast programs.

Any key sound that appears regularly in an audio stream with minor changes, such as jingles or advertisements, can be used to structure or monitor broadcast streams. Because these audio segments are almost always invariant in both the time- and frequency domain, and usually have only a few seconds of length, they could be easily identified by searching its "signature" in the fingerprinted stream. The searching should be many times faster than real-time and highly accurate.

In the literature there are several proposed solutions to this problem. Usually the fingerprint algorithm is very simple and consists in performing a spectral analysis in order to produce an efficient representation of the audio signal. The key-sound signature or the signature model is defined based on this representation. A general requirement is that the audio signature retains the perceptual cues which are invariant to signal degradations. Therefore the signal is almost always down sampled to a few kHz before the spectral analysis.

A simple search strategy is presented in [5]. It consists in the determination of Euclidean distances between signature and signal spectral vectors, accusing detection if the distance is below a given threshold for a small number of frames. In [6] a solution is presented based on sinusoidal modeling, where the signature is characterized by a small set of strong spectral components. A statistical model based on cepstral parameterization and a metric based on the covariance of the signature is given in [7]. A very fast solution based on energy peaks in the time-frequency plane using hash tables is presented in [8]. The method presented in [9] uses a very efficient signature representation, for indexing of musical excerpts, which enables it to run in low resource computational devices. The use of neural networks to detect acoustic patterns in broadcast news is reported in [10] and a comparison of algorithms for fingerprinting in audio is reported in [11].

Solutions based on hidden Markov models (HMM) were also proposed, e.g., [1] and [7]. We also tried such solution, using a Markov chain with no auto-loops, due to the expected time regularity of the observations. The main encountered problem was how to estimate robust parameters based on a single example. The method is less efficient and less robust than the one proposed here. However, it should provide superior results if the requirement is to find "similar" sounds, instead of "almost equal" sounds.

Among the proposed solutions the one presented in [9] has the great advantage of using a binary representation of spectral patterns, which permits not only an efficient representation of the audio content but also a very fast search. For that reason this representation was adopted in the present work. In [9] this fingerprinting scheme was tested with distorted signals, showing that the method is robust to noise, equalization and coding schemes, but less robust to time compression or expansion. In the present work, several experiments were conducted in order to find out the influence of the spectral analysis and bit pattern definition on the algorithm performance.

The purpose of this work was to identify the critical parameters of the algorithm and investigate which ones should be tuned in order to boost its performance. In this sense we evaluate the algorithm with different window length and several mask definitions. We also analyze if a lower frame rate could affect negatively the algorithm performance. The system was evaluated with a 25 hour corpus acquired from five Portuguese broadcast television stations.

The paper is organized as follows. In Section II the fingerprinting algorithm is presented. Section III describes a broadcast video database used to test the algorithm performance and presents some evaluation experiments. Finally, in section IV a discussion of the results is done and the conclusions are drawn.

II. ALGORITHM DESCRIPTION

A. Fingerprints

The first step is to down sample the audio signal to 8 kHz because we intend to preserve only a small audible bandwidth. Next, a spectral analysis of the audio signal is performed with a rectangular window[1] and a mel filterbank with 33 channels. This is done in the spectral domain, using filters with triangular responses. The window length and frame rate are important parameters that may affect the identification accuracy and search performance, as explained in section III. The last step corresponds to the binarization of the spectrogram. The main idea is to preserve, essentially, the spectral peaks or other information judged as robust or perceptually relevant against distortions. This can be done using a convolution of the spectrogram with a mask. The first mask in Figure 1 finds negative slopes on the spectrogram in two consecutive frames. The second mask corresponds to the proposal in [9]. "Mask3" identifies peaks in the sonogram, which tend to give evidence to tonal components. "Mask4" does the same thing but requires a longer tonal presence.

Mask1

–1	–1
+1	**+1**

Mask2

+1	–1
–1	**+1**

Mask3

–1	–1
+2	**+2**
–1	–1

Mask4

–1	–1	–1
+2	**+2**	+2
–1	–1	–1

Fig. 1. Convolution masks to binarize the spectrogram. The abscissa index represents time (in frames) and the ordinate index represents frequency (filterbank channels). The mask (0,0) point is indicated in bold.

The final bit pattern has ones if the convolution result is positive or zero otherwise. The final representation has 32 bits per frame which is conveniently represented by 32-bit words. Figure 2 shows a spectrogram and the corresponding binarization after the application of three masks. As can be seen, the bit pattern represents essentially the tonal components that are present in the jingle, in the case of "Mask1" and "Mask3".

B Searching

The searching method is very simple in the case of fingerprints composed by bit patterns. It corresponds to count the matching bits between the signature and audio binary patterns, in each frame, when the signature pattern slides over the audio pattern. Alternatively, we can obtain the error bits with X-OR operations (between 32-bit words) and use a lookup table as a fast method to count the error bits. Dividing the number of error bits by the total signature bits results in the bit error rate (BER). The method is equivalent to finding the correlation between the signature and the audio if the binarization values were "1" and "-1". The average BER is 0.5 because the masks have a zero DC term, which implies an equal number of "0" or "1" bits, in average. The BER variance depends on the mask and the length of the signature.

A signature is found if the BER is below a given threshold. Actually, we used a little hysteresis in order to increase the search accuracy. An example is shown in Figure 3, where five occurrences are below a given threshold.

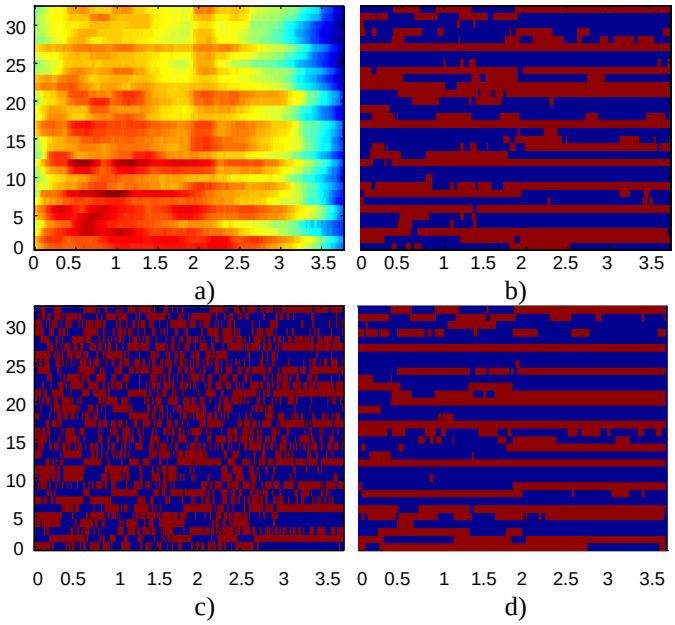

Fig. 2. a) Spectrogram of an audio signal ("jng1RTP1"). The abscissa index represents time (in seconds) and the ordinate index represents filterbank channels or bits. b)-d) The corresponding bit patterns for "Mask1", "Mask2" and "Mask3", respectively. Dark-red regions represent "1" bits and blue regions represent "0" bits.

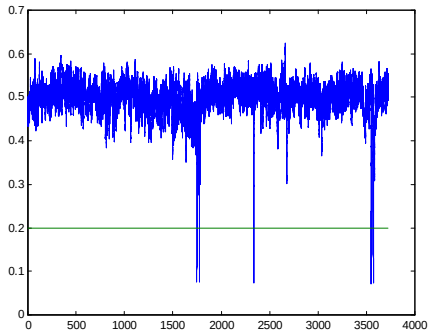

Fig. 3. Bit error rate when searching the jingle "jng1RTP1" in a one hour RTP1 stream. The abscissa index represents time (in seconds).

[1] As a first choice we decided to use a rectangular window to simplify the buffer management. Also, due to the low number of channels in the filterbank, the effect of the window choice on the final representation is reduced.

This method is indeed very fast. Using a common PC we can search a 10 second signature within a one hour signal in less than 0.46 s (with a frame rate of 50 frames per second).

III. EXPERIMENTS AND RESULTS

In order to calibrate the system we have carried out several tests using broadcast audiovisual streams collected by a media monitoring company.

A. Database description

The database is composed by 25 media streams, each one with one hour of duration. The files were acquired from five distinct Portuguese broadcast television stations, namely, "Rádio e Televisão de Portugal" (RTP1 and RTP2), SIC, "SIC Notícias" (SNOT) and TVI. From five different broadcast channel files, 16 signatures were selected with different lengths, as indicated in Table I. The prefix "jng" in the signature names represents jingles; "pubcut" refers to a small part of an advertisement and "pub" corresponds to a complete advertisement. As the names suggest, most signatures are composed of music, but some have speech parts in order to test the algorithm performance in these situations.

Table I
Signatures for testing

Signature Names	Station	Length	Number
jng1RTP1	RTP1	4s	23
jng2RTP1	RTP1	3s	11
pubcutRTP1	RTP1	6s	3
pubZonTVCaRTP1	RTP1	30s	4
pubModBombRTP1	RTP1	30s	4
jng1RTP2	RTP2	4s	4
jng2RTP2	RTP2	2s	2
jng3RTP2	RTP2	6s	2
jng1SIC	SIC	4s	14
pubcutSIC	SIC	3s	2
pubVodADSLSIC	SIC	25s	6
jng1SNOT	SNOT	4s	13
pubcutSNOT	SNOT	1s	2
jng1TVI	TVI	2s	6
jng2TVI	TVI	3s	2
pubcutTVI	TVI	3s	5
TOTAL:			103

The broadcast channel files were manually labeled with the occurrence of the signatures. The audio is taken from these files using the Windows Media Format SDK. The files were fingerprinted with the algorithm described above. The labels do not overlap in time, i.e. the "pubcut" labels were taken from advertisements different from the complete ones. There are 4 abnormal situations which deserve a mention. The first one is the occurrence of one incomplete jingle (about 1/3), followed immediately but another different jingle. We decided to keep this jingle as a reference label to test the algorithm in this situation. Another situation is the occurrence of two jingles embedded in speech, however with a low SNR (much less than 0 dB). The last situation is a distorted jingle embedded in a faded-out musical excerpt. In total, there are 103 occurrences of the 16 signatures in the corpus, 99 of them are regular and 4 are abnormal.

B. Experiments

Several experiments were made in order to investigate the effect of different window length, frame rate and mask definition in the fingerprint algorithm.

Usually in speech and audio analysis the window is about or shorter than 30 ms. However, longer windows seem to work better, especially for signals with tonal components. Window durations from 80 ms to 240 ms were tested.

The analysis of frame rate was also tested in order to evaluate the needed time resolution to accurately locate the signatures. The exact instant where the signature begins is usually not very important, but a low frame rate could lead to miss a signature due to phase differences.

Table II shows the obtained results. The results are quantified in terms of "hits", H (true detections). Deletions are D=103–H, because there are no insertions except for the three entries marked with "*" where there is just one.

A fixed threshold of 0.23 was used for all masks but "Mask2", where the threshold of 0.35 was used, as proposed in [9]. In fact, as can be seen in Fig.2, the "randomness" of bit pattern for "Mask2" is greater than for other masks, which implies that its BER has a lower variance and, therefore needs a higher threshold. We have verified that with a threshold of 0.23 for "Mask2", the performance degrades dramatically.

Table II
Performance results against window length, frame rate and mask type

Window Length (ms)	Frame Rate (fr./s)	Mask1	Mask2	Mask3	Mask4
240	25	98	98	99	99*
240	50	99	95	99	99*
240	100	99	91	99	99*
120	25	98	95	99	99
120	50	98	95	99	99
120	100	98	86	99	99
80	25	98	95	99	99
80	50	97	93	99	99
80	100	98	81	99	99

For better performance the detection threshold could be adapted to each mask and signature length, in order to give an equal false positive rate (insertion rate). In particular, the

insertions observed for “Mask4” could be avoided by lowering the threshold.

The 4 abnormal situations referred to above were never identified. All the configurations have a similar performance; however, “Mask1” and “Mask3” seem to have a consistent better performance than “Mask2”.

As can also be observed from Table II, the window length and frame rate have little influence on the results. This study suggests that a frame rate as low as 25 frames per second is enough for detection of signatures with lengths from 3 to 30 seconds. In our practical implementation we choose “Mask3” and a window length of 120 ms, which corresponds to 960 samples and allows performing FFTs of 1024 samples (instead of 2048, needed for a longer window).

IV. CONCLUSIONS

In this paper we have presented a fingerprinting system to detect jingles, advertisements or other audio short segments present in audio streams. The fingerprinting method generates a sequence of 32-bit patterns by applying a binarization mask to the spectrogram of the audio signals. Due to the binary representation of the signals, the search of signatures is very fast and can be implemented in general propose hardware. Several parameters of the algorithm were investigated in order to boost the performance. The system was evaluated with a 25 hour broadcast television corpus and all the normal occurrences of the present signatures were correctly identified.

REFERENCES

[1] E. Batlle, J. Masip and P. Cano, “System Analysis and Performance Tuning for Broadcast Audio Fingerprinting”, in *DAFx’03*, London, UK, Sep. 2003.

[2] A. Wang, “The Shazam music recognition service”, in *Communications of the Association for Computing Machinery*, 49(8):44–48, Aug. 2006.

[3] Koninklijke Philips Electronics, “Philips Content Identification Protecting Assets”, http://www.business-sites.philips.com/contentidentification/home/index.page.

[4] M. Fink, M. Covell and S. Baluja, “Social- and Interactive-Televison Applications Based on Real-Time Ambient-Audio Identification”, in *EuroITV’06*, Athens, Greece, May 2006.

[5] J. Pinquier and R. André-Obrecht, “Jingle Detection and Identification in Audio Documents”, in *IEEE ICASSP’04*, Montreal, Canada, May 2004.

[6] M. Betser, P. Collen and J.-B. Rault, “Audio Identification using Sinusoidal Modeling and Application to Jingle Detection”, in *ISMIR’07*, Vienna, Austria, Sept. 2007.

[7] S. Johnson, and P. Woodland, “A Method for Direct Audio Search with Applications to Indexing and Retrieval”, in *IEEE ICASSP’00*, Istanbul, Turkey, June 2000.

[8] J. Ogle and D. Ellis, “Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings”, in *IEEE ICASSP’08*, Las Vegas, USA, May 2008.

[9] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System”, in *ISMIR’02*, Paris, France, Oct. 2002.

[10] H. Meinedo and J. Neto, “Detection of Acoustic Patterns in Broadcast News using Neural Networks”, in *Acústica’04*, Guimarães, Portugal, Sept. 2004.

[11] P. Cano, E. Batlle, T. Kalke and J. Haitsma, “A Review of Algorithms for Audio Fingerprinting”, in *IEEE MMSP’02*, St. Thomas, USA, Dec. 2002.