Robust Radio Broadcast Monitoring Using a Multi-Band Spectral Entropy Signature

Antonio Camarena-Ibarrola[1] and Edgar Chávez[1] and Eric Sadit Tellez[1]

Universidad Michoacana
Mexico

Abstract. Monitoring media broadcast content has deserved a lot of attention lately from both academy and industry due to the technical challenge involved and its economic importance (e.g. in advertising). The problem pose a unique challenge from the pattern recognition point of view because a very high recognition rate is needed under non ideal conditions. The problem consist in comparing a small audio sequence with a large audio stream (the broadcast) searching for matches.
In this paper we present a solution with the *multi band spectral entropy audio fingerprint* (MBSES) which is very robust to degradations commonly found on amplitude modulated (AM) radio. Using the MBSES we obtained perfect recall (all audio ads occurrences were accurately found with no false positives) in 95 hours of audio from five different am radio broadcasts. Our system is able to scan one hour of audio in 40 seconds if the audio is already fingerprinted (with e.g. a separated slave computer), and it totaled five minutes per hour including the fingerprint extraction using a single core off the shelf desktop computer with no parallelization.

1 Introduction

Monitoring content in audio broadcast consists in tagging every segment of the audio stream with metadata establishing the identity of a particular song, advertising, or any other piece of audio corresponding to feature programming. This tagging is an important part of the broadcasting and advertising businesses, all the business partners may use a third party certification of the content for billing purposes. Practical examples of application of this tagging include remote monitoring of audio marketing campaigns, evaluating the hit parade, and recently (in Mexico at least) monitoring the media coverage of political campaigns among others.

There are several alternatives to audio stream tagging or media monitoring, current solutions are ranged from low tech, human listeners, to digital content tagging, watermarking and audio fingerprinting. In this paper we are interested in automatic techniques, where the audio stream can be analyzed and tagged without human intervention. There are several commercial turnkey solutions reporting about 97% precision with a very small number of false positives, the most renowned is *Audible Magic* `http://www.audiblemagic.com/` with massive

databases of ads, songs and feature content. The core of the automated techniques is the extraction of an audio fingerprint, which is a succinct and faithful representation of the audio stream, in both the audio stream and the content to be found in the broadcast. This change of domain serve two purposes, on the one hand it is faster to compare the succinct representation. On the other hand, since only significant features of the signal are retained very high accuracy can be obtained in the comparison. In this paper we present a tagging technique for automatic broadcast monitoring based on the MBSES. Our technique has perfect recall and is very fast, scoring from 12 to 40 times faster than real time broadcasting in a single-core standard computer with no parallelization. As described in the experimental part we were able to improve by some the recognition rate of trained human operators working on a broadcast monitoring firm.

2 Related Work

It is a fact that most audio sources can be tagged prior to the broadcasting, specially with the advent of digital radio. Even in the case of analog audio broadcasting it is possible to embed digital data in the audio without audible distortion and persistent to degradations in the transmission. This technique, called *audio watermarking*, is suitable for applications where the broadcast station agree to modify the analog content, and needs a receiver capable of decoding the embedded data on the end point. This type of solutions are described in [1] and [2]. Usually they are sold as a turnkey system with both the transmitter and the receiver included. Watermarking is not suitable for doing audio mining or searching in large audio logs because, in the past, audio was not transmitted with embedded data.

A more general solution consist in making a succinct and faithful representation of the audio, specific enough to distinguish between different audio sequences and general enough to allow the identification of degraded samples. Common degradations are white/colored noise adding, equalization and re-recording. This technique is called *audio fingerprinting* and has been studied in a large number of scientific papers and due to its flexibility it has been the first choice mechanism for audio tagging. When small excerpts of audio are used to identify larger pieces of the stream an additional artifact is introduced to the process, the time shifting effect. This is due to the discrete audio window being represented, and the failure to match the start of the audio window in both the excerpt and the stream. Audio fingerprinting must be resilient to all the above distortions without loosing specificity. Among the proposals for audio fingerprinting we can count the Mel-frequency Cepstral coefficients (MFCC) [3], [4]; the *Spectral Flatness Measure* (SFM) [5]; *tonality* [6] and *chroma values* [7], most of them are also analyzed in depth in [8]. Recently in [9, 10] the use of entropy as the sole feature for audio fingerprinting proved to be much more robust to severe degradations outperforming previous approaches. This technique is the *multiband spectral entropy signature* or MBSES described in some detail in the paper.

Once the fingerprint is obtained, it is not very difficult to build on this first piece a complete system for broadcast monitoring. Such a complete system is discussed in [11] using a fingerprint. In Oliveira's work [11] the relevant feature was the energy contained in both the time and the frequency of the signal. The authors reported a correct recognition rate of 95.4% with 1% of false positives. Other good example of a system for broadcast monitoring with excellent results is [12] where the relevant feature was the *spectral flatness* which is also the feature used in the MPEG-7 wrapper (see [13] for details) for describing audio content.

Due to the economic importance of media monitoring (up to 5% of the total advertising budget is devoted to monitoring services) several companies have proprietary, closed technology for broadcast monitoring. In this case we can only compare with the performance figures publicly reported in white papers.

We selected MBSES to build our system due to its anticipated robustness. Using this fingerprint we were able to have perfect recall and no false positives in very low quality audio recordings just by tuning the time resolution. This results outperform the reported precision of both academic and industrial systems. Audio tagging, particularly using a robust fingerprint like the one described in this paper, is a world class example of a successful pattern recognition technique. Several lessons can be extrapolated from this exercise.

The rest of this paper is organized as follows, first we explain how the MBSES of an audio signal is determined, then we describe the implemented system in detail, a description of the experiments performed to test our system follows, and finally some conclusions and future work directions are discussed in the last section.

3 Broadcast Monitoring with MBSES

The final product of a monitoring service is a tagged audio log of the broadcast. Assuming the role of the broadcast monitoring company, a particular client request counting a particular ad in a given number of radio stations. The search is for some common failures in the broadcasting of audio ads, namely the absence of the ad, airing it at a time different from the one paid (time slots have different prices depending of the time of the day, and the day itself) and airing only a fraction of the audio ad. Lack of synchronization between airing and marketing campaigns may lead to large loses, for example when a special offer last only for one day and the ads are aired the day after the special offer expires. The only legal bonding for auditing purposes is the audio log showing the lack of synchronization, hence recording is mandatory.

When designing a system for broadcast monitoring, the above discussion justifies having an offline design. Since recording is mandatory, the analysis of the audio can be done offline, we can assume the stream is a collection of audio files. Even low tech companies with human listeners can analyze audio three times faster than real time, playing the recordings at a higher speed and skipping feature programming when tagging the audio logs. The human listener memorize a set of audio ads and when playing the recording identifies one of them, she/he

makes an annotation in the broadcast station log, putting the time of occurrence, and the ad ID. In this case accuracy of annotations is within minutes. Human listeners can process 24 hours of audio in some 8 hours of work.

Our design replicates the above procedure in a digital way. We will compare the audio fingerprint of the stream with the corresponding audio fingerprint of the audio ads being monitored. We will have annotations accuracy in the order of milliseconds, and 12 to 40 times faster than real time.

3.1 The Multi Band Spectral Entropy Signature

We describe in some detail the MBSES to put the reader in the appropriate context. The interested reader can obtain more information in references [9, 10] and [14].

The reason behind the robustness of the MBSES is because it can be computed with just the histogram of the signal, and for overlapping windows the histogram changes very slowly because the two histograms share many samples. The net effect of the above observation is to have a representation of a signal as a weighted average of its value in a given window. A single value represents a segment of the signal. This gives the succinctness required by a fingerprint, and if we also observe that noise affect only a subset of the samples in a given period of time we fulfill other of the requirements. Obtaining the entropy of the signal directly in the time domain (more precisely the entropy of the energy of the signal) has proved to be very effective for audio fingerprinting in [10]. With this approach, called *time domain entropy* (TES) the recall was high; but some degradations, like equalization, it dropped quickly. To solve this problem in [9] the signal is divided in bands in a logarithmic scale (the Bark scale). The result was a very strong signature, with perfect recall even for strong degradations. Below we detail the extraction of the MBSES.

1. The signal is processed in frames of 256 ms, this frame size ensures an adequate time support for entropy computation. The frames are overlapped by 7/8 (87.5%), therefore, a feature vector will be determined every 32 ms
2. To each frame the Hann window is applied and then its DFT is determined.
3. Shannon's entropy is computed for the first 21 critical bands according to the Bark scale (frequencies between 20 Hz and 7700 Hz). To compute Shannon's entropy, equation 1 is used. σ_{xx} and σ_{yy} also known as σ_x^2 and σ_y^2 are the variances of the real and the imaginary part respectively and $\sigma_{xy} = \sigma_{yx}$ is the covariance between the real and the imaginary part of the spectrum.

$$H = ln(2\pi e) + \frac{1}{2} ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2) \tag{1}$$

4. For each band obtain the sign of the derivative of the entropy as in equation (2). The bit corresponding to band b and frame n of the AFP is determined using the entropy values of frames n and $n-1$ for band b. Only 3 bytes for each 32 ms of audio are needed to store this signature.

$$F(n,b) = \begin{cases} 1 \; if \;\; [h_b(n) - h_b(n-1)] > 0 \\ 0 \; Otherwise \end{cases} \tag{2}$$

A diagram of the process of determining the MBSES of an audio-signal is depicted in Fig. 1.

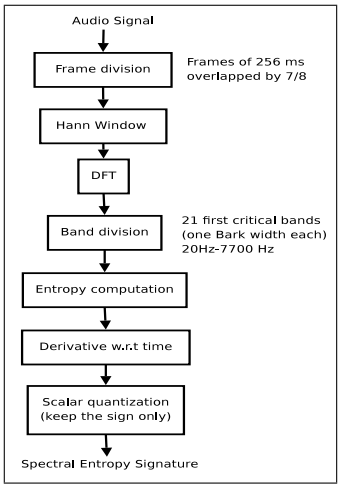

Fig. 1. Computing the Spectral Entropy Signature

The fingerprint of the signal is now a binary matrix, with one column representing each frame in the signal. The most interesting part is that now the Hamming distance (the number of non matching bits compared element by element) is enough to measure similarity between signals.

3.2 The Monitoring Procedure

Monitoring is quite simple when we have a robust way to measure similarity between the audio stream and the audio segment, such as now that we have extracted the MBSES for both parts.

Figure 2 exemplifies the procedure for searching for a particular ad occurrence in the stream. The smaller matrix (the audio ad) is slide one bit at a time to search for a match (a minimum in the distance).

We observed a peculiar phenomenon in searching for a minimum in the Hamming distance, there is a sudden increase just before there is a match, figure 3 illustrated this, an ad was found in minutes 3 and 41.. This is probably because the signature is not very repetitive, moreover, it is little compressible.

The Hamming distance can be efficiently computed with a lookup table counting the number of ones in a 21 bit string. This lookup table is addressed with the value of $x \oplus y$ with $\oplus$ the XOR operation between x and y the columns being compared.

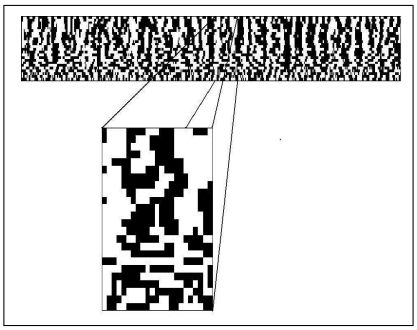

Fig. 2. The signature of the audio ad is the smaller matrix, the long grid is the signature of the monitored audio. When the Hamming distance falls below a threshold we count a match.

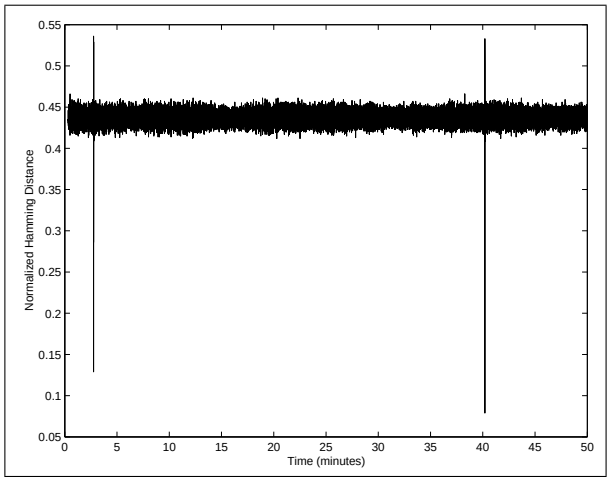

Fig. 3. This plot corresponds to the Hamming distance between the ad being searched and the corresponding segment in the audio stream. Notice a sudden increase followed by a decrease in the distance, both above and below a clear threshold.

4 Experiments

For our experiments we used all-day recordings from five different local AM (Amplitude Modulated) radio stations, this recordings were provided by *Contacto Media Research Mexico SA de CV* (CMR) in the lossy compression format mp3@64kbps spread in 95 files of approximately one hour each. Thirteen recordings of commercial spots were also provided to us as well as the results from manually monitoring these stations by their trained employees.

We determined the signatures of every one-hour mp3 file and stored them in separate binary files, generating 95 long signatures at this step. The process of checking spot's occurrences in one-hour files lasted 40 seconds approximately.

The whole process of checking 95 hours of audio generating the complete report lasted one hour approximately.

The report generated by our broadcast monitoring system was compared with the report provided by CMR. Wet found 272 occurrences while CMR reported only 231, the missed 41 ads were manually verified by us. It is noticeable that trained operators (human listeners) have failed to report those 41 spots, perhaps due to fatigue or distraction. On the other hand all of the ad occurrences detected by operators were detected by our system.

The recognition rate reported by Hellmuth *et al* in [12] for similar experiments since they also use off-line monitoring, degrading by lossy compression precisely in the format mp3@64kbps and excerpts of 20 seconds (e.g the size of most commercial ads) was 99.8%. In contrast, our experiments report a precision of 100% since no commercial ad occurrence was missed with our system. Table 1 compares this results including the results reported by Oliveira *et al* in, [11].

Table 1. Comparison with the reported results on similar research

System	True positives rate (recognition rate)	False positives Rate (recognition mistakes rate)
Proposed System	100%	0%
Hellmuth et al [12]	99.8%	-
Oliveira et al [11]	95.4%	1%

5 Conclusions and Future Work

We found our Multi-band spectral entropy signature (MBSES) to be adequate for robust automatic radio broadcast monitoring. The time resolution of the signature was adjusted to work with commercial spots with high speech content.

Instead of searching sequentially among the collection of spots for an occurrence of any of them, we will design a proximity index that would allow working with thousands of spots without affecting the speed of the monitoring process. On the other hand, preliminary results about using *graphic processing units* (GPU) for computing the fingerprint shows an important speedup with respect to single core computing. This also pose very interesting audio mining challenges in archived audio logs of several-year long recordings.

6 Acknowledgements

We want to thank the firm *Contacto Media Research Services SA de CV* in Guadalajara, México for providing us with the manually tagged recordings used in this paper.

References

1. Haitsma, J., van der Veen, M., Kalker, T., Bruekers, F.: Audio watermarking for monitoring and copy protection. In: MULTIMEDIA '00: Proceedings of the 2000 ACM workshops on Multimedia, New York, NY, USA, ACM (2000) 119–122
2. Nakamura, T., Tachibana, R., Kobayashi, S.: Automatic music monitoring and boundary detection for broadcast using audio watermarking. In: SPIE. (2002) 170–180
3. Sigurdsson, S., Petersen, K.B., Lehn-Schioler, T.: Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In: International Symposium on Music Information Retrieval (ISMIR). (2006)
4. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval (ISMIR). (oct 2000)
5. Herre, J., Allamanche, E., Hellmuth, O.: Robust matching of audio signals using spectral flatness features. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2001) 127–130
6. Hellman, R.P.: Asymmetry of masking between noise and tone. Perception and Psychophysics **11** (1972) 241–246
7. Pauws, S.: Musical key extraction from audio. In: International Symposium on Music Information Retrieval ISMIR. (October 2004) 96–99
8. Cano, P., Battle, E., Kalker, T., Haitsma, J.: A review of algorithms for audio fingerprinting. Multimedia Signal Processing,IEEE Workshop on (December 2002) 169–167
9. Camarena-Ibarrola, A., Chavez, E.: On musical performances identification, entropy and string matching. In: Fifth Mexican International Conference on Artificial Intelligence 2006 (MICAI2006). (November 2006) 952–962
10. Camarena-Ibarrola, A., Chávez, E.: A robust entropy-based audio-fingerprint. In: Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME, IEEE CS Press (2006) 1729–1732
11. Oliveira, B., Crivellaro, A., César, Jr, R.M.: Audio-based radio and tv broadcast monitoring. In: WebMedia '05: Proceedings of the 11th Brazilian Symposium on Multimedia and the web, New York, NY, USA, ACM (2005) 1–3
12. Hellmuth, O., Allamanche, E., Cremer, M., Kastner, T., Neubauer, C., Schmidt, S., Siebenhaar, F.: Content-based broadcast monitoring using mpeg-7 audio fingerprints. In: International Symposium on Music Information Retrieval ISMIR. (2001)
13. Group, M.A.: Text of ISO/IEC Final Draft International Standard 15938-4 Information Technology - Multimedia Content Description Interface - Part 4: Audio. (July 2001)
14. Camarena-Ibarrola, J.A.: Identificación Automática de Señales de Audio. PhD thesis, Universidad Michoacana de San Nicolás de Hidalgo (January 2008)