Automatic TV Advertisement Clustering Based on Audio Features from Broadcasted Signal

Arthur F. Sofiatti*, Natália W. Rovaris*, Igor G. Hoelscher*, Joel A. Luft*†, Tiago Balen* and Altamiro A. Susin*
*Electrical Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
{arthur.sofiatti, natalia.rovaris, igor.hoelscher, joel.luft, tiago.balen, altamiro.susin}@ufrgs.br

†Federal Institute of Education, Science and Technology of Rio Grande do Sul, Canoas, RS, Brazil
joel.luft@canoas.ifrs.edu.br

***Abstract*—New techniques are emerging to extract the relevant information from Big Data in order to guide decision-supporting systems. This paper presents the use of audio descriptors to characterize audiovisual advertising for automatic monitoring of television broadcast. Using cepstral features to describe audio tracks, the framework implements a context-based distance algorithm to measure the similarity between transmitted television and verified commercials. The algorithm uses a set of adaptive parameters based on the energy behavior of the audio signal of the advertisement to weight the relation between two coefficient arrays in each query. The prototype was developed in Matlab and after was then transcribed in C language. A test suite containing 96 hours of analog television broadcast was captured at several Brazilian TV networks from different regions and compared with more than 90,000 hours of TV programs and more than 20,000 verified commercials. Experimental results show 97.1% hit rate with test samples affected by noise at transmission, reception and acquisition, and containing short periods of silence.**

I. Introduction

Data Mining evolution is driving our civilization towards an Information Society [1], [2]. About 2.5 zettabytes of data were generated in 2012 and it will increase significantly each year, reaching nearly 45 zettabytes by 2020 [3], given mainly by the expansion of the Internet of Things.

The acquisition, processing, transmission, storage and retrieval of data – emerging from media, public transactions, web searches, browsing history and in particular social networks – reach levels unmanageable by the current methods and equipment. Different approaches to this problem have been addressed, under paradigms such as *cloud computing* and *data mining*, to deal with what has been called ”data deluge” [4].

Processing systems that intelligently track information are becoming more sophisticated, some to select only the relevant information on data provided by a sensor, others however, driven by the interest in discovering the users profile for the purpose of direct offering products, starting a revolution in the marketing techniques. This scenario dictates new standards in several areas like culture, education, fashion and entertainment, being used to give more realism and visibility to ads.

Therefore, advertising companies started to upgrade their production more frequently, resulting on a gradual increase in the volume of electronic advertisements. Thus, decision-supporting systems may dispense the hard and boring work of a human observer and assist companies with a more efficient audit of TV contracts.

The field of TV commercials analysis covers two main problems: detection and clustering/classification. The detection is responsible for finding new commercials or commercial breaks in different broadcast streams. To solve this problem [5] used video-only features, and more recently [6], [7] presented solutions based on audiovisual information.

Clustering or TV advertisement classification, covers the search of known commercials over different transmitted signals. Works have been dealing with this problem using matching algorithms and audio, video or audiovisual features as fingerprints to each commercial [8]–[12].

This article presents a TV advertisement (adv) classification framework, in order to monitor the proper transmission of commercials. The system uses the audio tracks of broadcasting media transmissions in order to assist the auditing of adv contracts.

The next section will explain the method, with theoretical background. In section III we discuss the implementation process of the tool, covering some details about this application. In section IV, we present the data set used for tests and the corresponding results. Finally, section V show our conclusions about the work.

II. Methodology

In the audio processing field, data mining has been used mainly for voice recognition and classification. However, before that, studies were directed to process non-speech audio signals allowing the generation of an extensive set of information on different events. Therefore, identification and event classification problems in non-speech audio signals occupy an area of active research in audio processing, targeted to a wide range of topics: surveillance, human-computer interaction and context sensitive applications [13].

Among the main challenges of this project is the fact that the audio signals in TV broadcast may present varied behavior, containing silent or speech periods, background music or inaudible signals, and have their quality affected by the acquisition and transmission conditions, sub-sampling and compression. Performance is another important requirement,

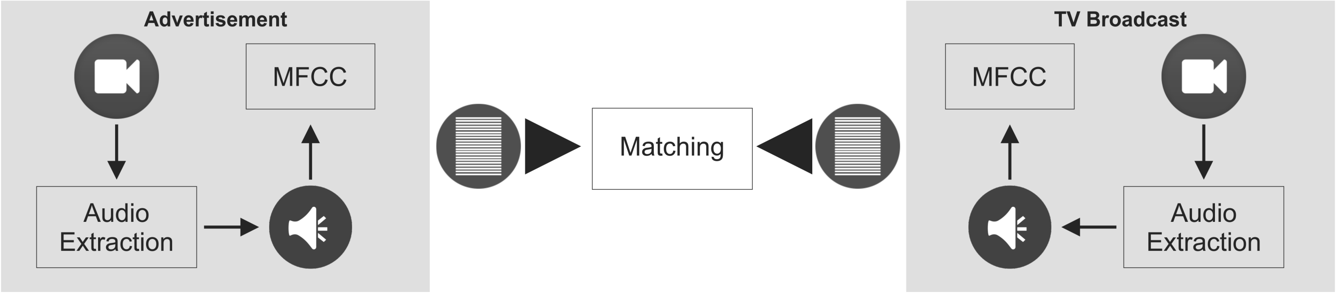


Fig. 1. Block diagram of the tool.

considering a scenario in which thousands of advertisements may be audited in approximately 12,000 hours of daily television programming, generated by more than 500 broadcasters from Brazil [14].

Aiming to reduce the comparison time and obtain greater reliability in the search, the system performs the extraction of the Mel-frequency Cepstral Coefficients (MFCC). The choice of MFCC to describe audio signals is due to the widespread use in the sound classification, justified by their orthogonality, linearity and multidimensionality [15].

Davis and Mermelstein published one of the first studies showing the efficacy of the method in speech recognition [16]. Recently, [17] allied used of MFCC parameters for a statistical analysis to classify common sounds at sporting events. Reference [13] also presented some advances in recognition of non-speech audio, classifying events like ”glass breaking”, ”dog barking”, ”scream”, ”gunshot”, ”engine noise” and ”rain”.

Finally, [9], [10] reported the use of MFCC features for classification of advertisements. Using only the position of the highest coefficient as a feature for each window of the audio track, [10] was able to recognize advertisements in Digital Television Broadcasts.The technique is still used in speech recognition [18], [19] and audio in general [20]–[22].

The samples used to validate the implementation were obtained from analog broadcast transmissions in different regions and conditions acquired from analog transmissions. Those samples can be affected by energy variation, noise, distortion, degradation, interference and human intervention. Although cepstral coefficients show good results for audio classification, the treatment of non-defined sounds (like silence or noise) still is a challenging task.

To deal with these problems, we present a similarity function between cepstral coefficients, with adaptive parameters based on the energy behavior of the audio. Matching the array with energy coefficients, this tool is able to avoid false-positives and confirm the occurrence when the relation between all others coefficients have a high similarity value.

III. Implementation

The framework was designed to follow all the steps of a data mining tool: selection, pre-processing, processing, mining and interpretation/validation. In the selection and pre-processing stage, the tool performs the extraction of audio signals from video files, converting to *WAV* format, 16-bit uncompressed, using the FFmpeg package. Then in the processing step, an audio fingerprint is generated using MFCC transformation. Mining is executed by the matching algorithm, and finally an interpretation step defines if the advertisement is present or not on the TV broadcast.

Figure 1 is the system block diagram, that performs comparison between advertisement cuts and television broadcasting periods. In prototype stage, the algorithms were written in Matlab$^{\text{TM}}$. Thereafter the entire project has been translated to C programming language and organized in a dynamic-link library (DLL).

In the selection and pre-processing stages, the tool performs the extraction of audio signals from video files, converting to *WAV* format, 16-bit uncompressed, using the FFmpeg package. Then in the processing step, an audio fingerprint is generated using MFCC transform. A module opens the audio file, scrolls through the header and creates an array with the samples from audio. Then, another module extracts the audio coefficients to future comparison. Mining is executed by the matching algorithm, and finally an interpretation step defines if the advertisement is present or not on the TV broadcast.

Additional functionalities were implemented: in the eventuality of an advertisement starts near of the end of a broadcast file, it could happen the broadcast file finishes before the advertisement. In this case the system accepts a second file were is looked for the remaining part of the advertisement just in the very beginning.

Another useful information provided by the application is the size of the adv found. When an occurrence is detected a smaller window formed just by the last coefficients of the adv is used as input and searched in the broadcast vector in a region around it should be. It is used to confirm the entire adv has been transmitted.

To allow the final user changes some sensitive parameters, it was implemented configuration files, one to each delimited stage (extraction and matching). This gives the possibility in execution time to choose (within certain limits) between time consumed and accuracy.

A. Feature Extraction

Our method represents an audio track as a matrix of MFCC. Each query is a comparison between the adv matrix and the

TV broadcast matrix. Each line of the feature matrix is a mel-frequency cepstral transform of a window with 1024 samples and half window overlap. That way, the number of lines N can be defined as:

$$N = RTZ((ns - 512)/512) \tag{1}$$

where ns is the number of audio samples and RTZ is "round toward zero".

The number of columns M is the number of MFC coefficients for each window. Based on [23], our configuration generates 20 coefficients, but only the first 13 are used in the comparison. Experimentally it was notice using more than the first 13 coefficients there was not significant variation in the hit rate, therefor it was used this number of coefficients in the comparison to save processing and memory.

To exemplify the output of our extraction method, an advertisement with 30 seconds of duration and sampled at 8000 samples per second, will be described by a $467x20$ matrix of coefficients, and reduced to a matrix of $467x13$ coefficients defined as the fingerprint of the audio. A ".mfc" file was defined to save the samples extracted and calculated. A settings file is set, where the user can choose the parameters configuration. This file contains information like extensions accepted by the program, number of windows, FFT size and number of coefficients.

B. Matching Algorithm

The matching step works as a mining process in this system, performing a modified distance algorithm, calculating the similarity between each fingerprint of the advertisement and the features of each broadcast file, defined as targets.

Let A be the adv matrix of coefficients, with dimension of NxM, and B the target matrix of coefficients, with dimension of LxM. M is the number of coefficients generated per window sample, N is the number of windows forming the adv file and L is the number of windows forming the target file. The matching algorithm output is the information about the existence of the adv into the target, with time accuracy and degree of certainty.

Knowing that the first MFC coefficient is related to the energy behavior of the window, lower values for this coefficient describe windows with poorly defined sound, like silence or noise. In these cases, the remaining coefficients become less significant.

Based on this information, we developed a two-stage matching method, which generate two sets of weighted inverse euclidean distances: one between energy (CDe), and other between the 12 remaining features (CDc). This algorithm was developed to be robust to noise and short periods of silence. Equations 2 and 3 show the calculation of CDe and CDc for A and B, respectively.

$$CDe(i) = \frac{1}{\sqrt{\sum\limits_{n=0}^{N} W * (A(n,1) - B(i+n,1))^2}} \tag{2}$$

$$CDc(i) = \frac{1}{\sqrt{\sum\limits_{n=0}^{N} W * \frac{\sum(A(n,2{:}13) - B(i+n,2{:}13))^2}{12}}} \tag{3}$$

with $i = 1, 2, ..., L - N$.

The weights $W = \{w_1, w_1, ..., w_N\}$ are adaptive parameters responsible for indicating the level of significance of each window distance. These parameters are calculated based on the first coefficient of the adv, over a trapezoidal-shape function.

Based on our experiments, when the first coefficient is bellow -20, the window has an indistinct sound, and w_n is set to zero. For values between -5 and -20, the sound is classified as poor and w_n receives a value between 1 and 0, according to the upper and lower limits, respectively. When the first coefficient is above -5, the respective weight w_n is set to 1.

The position i indicates the window in B when starts to compare A, and each i will receive one value for distance. After tests, we defined that if both A and B have more than 60% of windows with non-defined sounds, the algorithm won't calculate similarity, keeping the behavior around i.

Figure 2 shows the outputs of the matching step, expressing the behavior patterns of the arrays of distances when there is and there is not occurrence of advs in the TV broadcast file. Figure 2 a and b represent the CDe and CDc, respectively, from a query, where was found a match, while c and d are the CDe and CDc arrays from another, where there was no match.

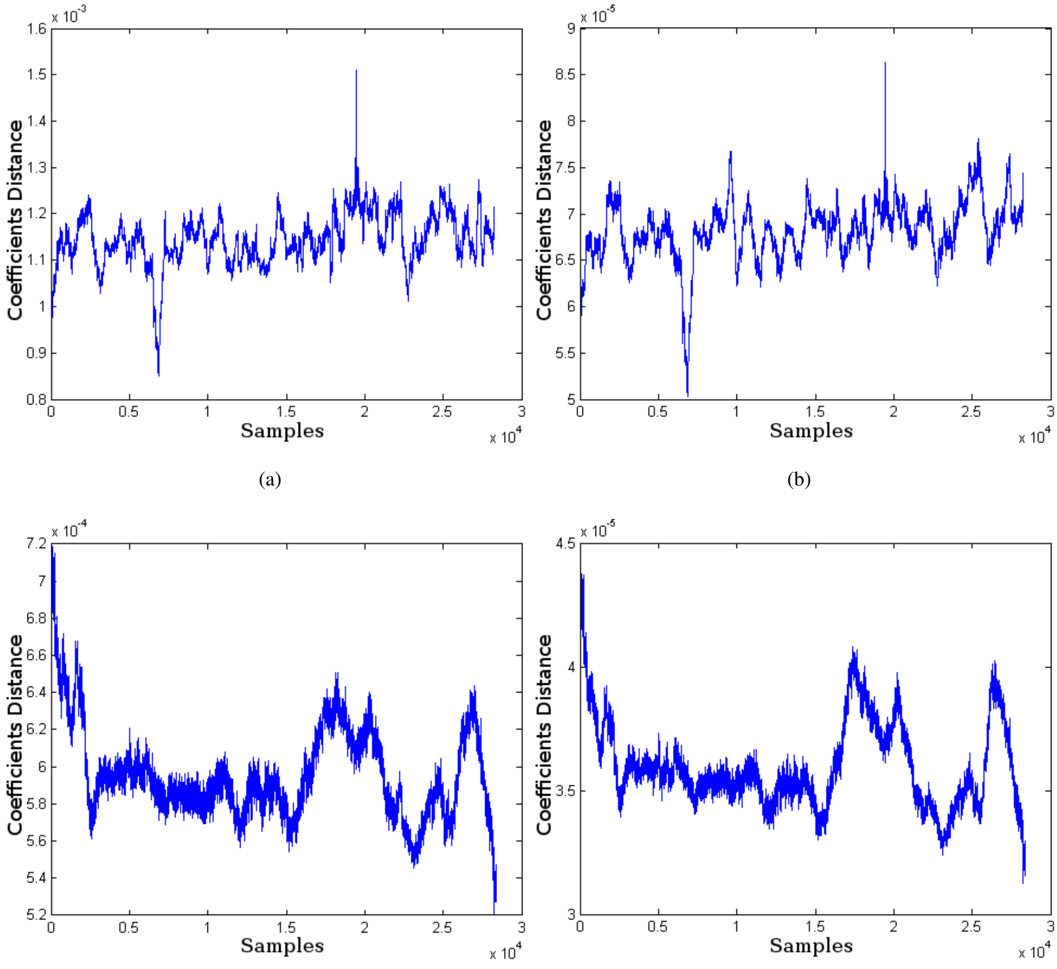
The visible narrow peak in Figures 2(a) and (b) indicates an occurrence. The occurrence point in arrays CDe and CDc indicates the time when the advertisement starts to be transmitted.

Also, Figures 2(c) and (d) display a comparison algorithm output (CDe and CDc) when there is no occurrences on the search target, being impossible to find a peak with the same behavior as the previous example.

But not always this peak stands out for its amplitude compared to the other distances in the matching stage output. Because of that the pattern that is to be searched is the sudden change in behavior of the distance array, which causes narrow peaks, high in relation to its neighborhood but not necessarily higher to the entirely signal.

Therefore, the system must interpret the comparison output to determine whether is there one or more occurrences of A in B and register the time at which each one was found. Is applied a high pass filter, which emphasizes abrupt transitions of a signal, highlighting the behavior shown in Figures 2(a) and (b). The filtered signals are shown in Figure 3.

This step works as an interpretation/validation stage in our data mining process. The peaks of the filtered signal are analyzed in order to determine with a degree of certainty the presence of the selected commercial in the TV broadcast. For this, we estimate an approximation to the common maximum of the signal, based on the RMS. This value is used as a threshold and is represented in Figure 3 by the red line.



(a)

(b)

(c)

(d)

Fig. 2. Output examples from matching stage. a and b: CDe and CDc of matching sequence. c and d: CDe and CDc of non matching sequence

All the peaks above the thresholds in both CDe and CDc are analyzed to determine if they correspond to a match. The degree of certainty (DoC) is calculated from the relation between the peak and the threshold amplitudes, given by:

$$DoC = 1 - \frac{threshold(CDc)}{peak(CDc)} \tag{4}$$

For example, for the visible peak above the threshold on both Figures 3(c) and 3(d) the DoC value is 0.8698. The signal CDe is used only to confirm that the energy behavior from advertisement and target are similar, while CDc gives robustness in calculating the DoC.

The DoC value also allow us to classify every possible occurrence in three different groups:

- Group A: $DoC \geq 0.8$ (confirmed occurrence).
- Group B: $0.5 \leq DoC < 0.8$ (possible occurrence).
- Group C: $DoC < 0.5$ (false occurrence).

Just matches at the second group asks for manual review, while occurrences from group one can be accepted and from group three can be rejected.

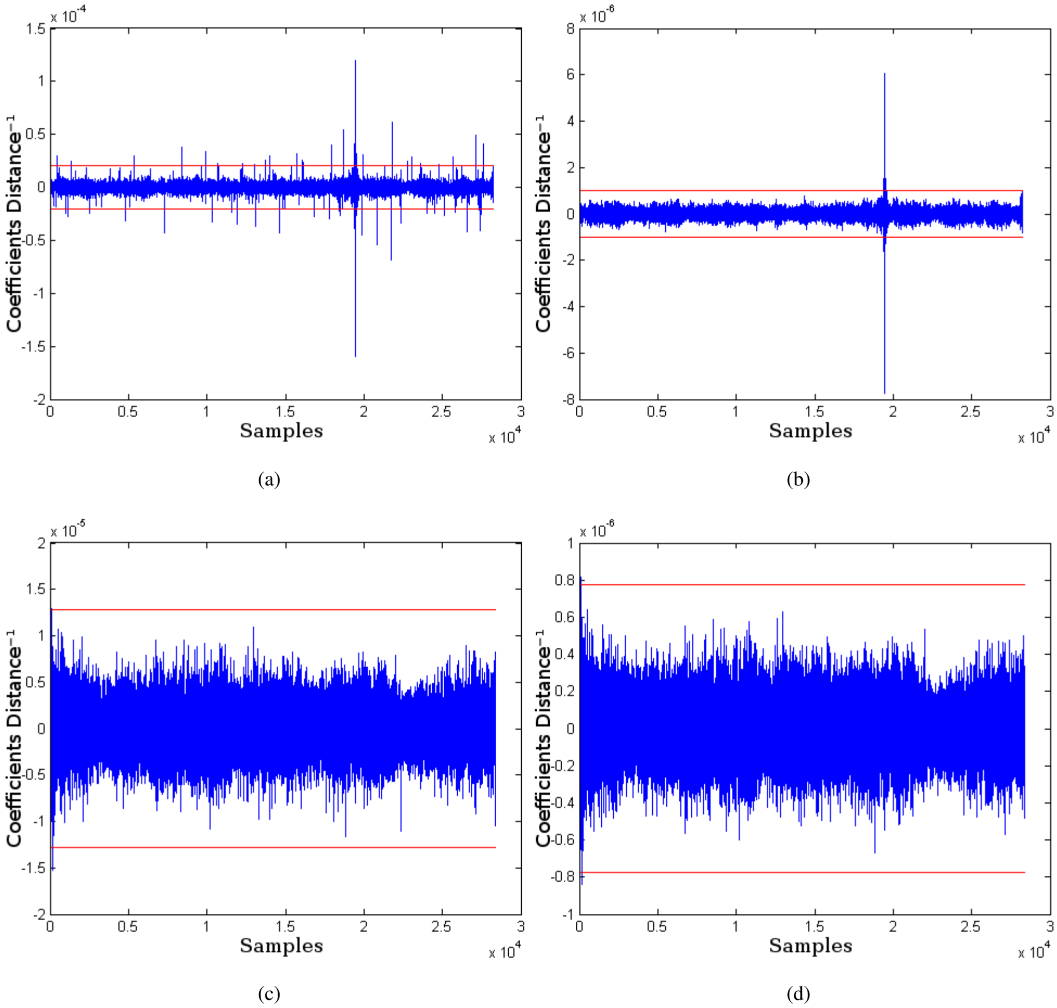
IV. Results

To validate the implementation, a random sample was generated from a database provided by a private company. This database contains more than 90,000 hours of local analog TV acquired in various regions of Brazil, compressed and relayed over the Internet, divided into 30 minutes files. Also, a set with more than 20,000 known advertisements were provided.

Experiments were conducted over a subset of this database, containing 86 advertising files, with an average duration of 30 seconds each, in 96 hours of television programming.

Table 1 presents the match results. Of all 450 occurrences manually registered in the selected population, the tool was capable of displaying 437 matches, indicating a hit ratio of approximately 97.33%. From these events, 410 were classified into Group A and 27 intro Group B.

In addition, the experiment showed 8 false-positive records, all with DoC below 0.8, which can help the operator on the



(a)

(b)

(c)

(d)

Fig. 3. Output examples from matching stage after high-pass filtering over the signals of Figure 2.

TABLE I
EXPERIMENT RESULTS.

Total occurrences	450
Successful matches	437
Matches in group A	410
Matches in group B	27
Missed detections	13
False Positives	8
False Positives in Group A	0
Hit Rate	97.1%

analysis of the report. Group C are not registered on the final report.

In addition to the tests with real broadcast described, that present more than noise, other intrinsic problems related to the analog transmission, it was performed tests adding white Gaussian noise to this already polluted signals. It was used Matlab function ”awgn”, with the option ’measured’. The results are showed in the Figure 4 for signal-to-noise ratios (SNR) of 10 dB and 0 dB.

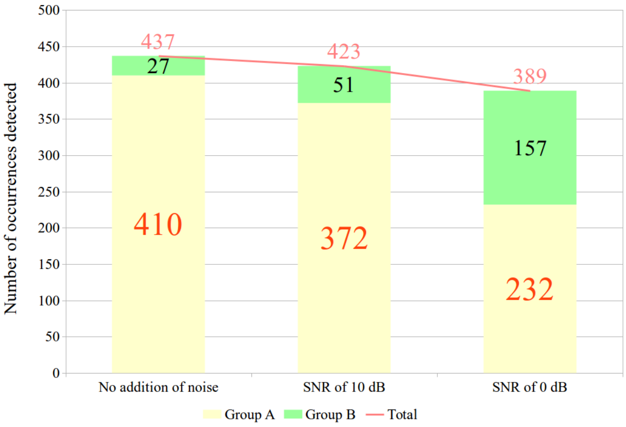


Fig. 4. Results of tests with noise addition.

The report allows us to observe some cases where the MFCC comparison technique has a peculiar behavior. One of the advertisements of the experiment has a background music, whose melody is repeated at a fixed period. This causes some windows to have similar descriptors that are repeated within each period with the melody, causing the recognition of the same adv cyclically, producing false positives.

The system is also able to indicates with precision of tenths of a second the instant at which the advertisement occurs, and to detect several occurrences in a single comparison. When it occurs, more peaks will appear in the same vector.

V. Conclusion

The MFCC showed effectiveness in characterizing small audio windows, generating a reduced set of coefficients that can describe the signal behavior on a scale that simulates the response of the human auditory system. With an explicit treatment of silent and quasi-silent segments, the framework achieved a hit rate of 97.1%, for the randomly selected test suite, matching 86 commercials with 96 hour of TV broadcast.

The algorithm also allows flexibility in choosing the window size, overlap size and the number of generated and crossed coefficients, which should be focus of a study to determine the optimal setting for a higher hit rate and efficiency, at the same time that reduces false-positive rate.

Nevertheless, the experiment showed good results on clustering advertisements from a challenging database, being robust to noise and short periods of silence.

The behavior of the system with the massive deployment of Digital TV, required by law in Brazil, should also be a case study in the future. Moreover, the low cost of an automatic auditing system based on the audio still allows to perform the analysis of commercials in radio broadcasting.

Acknowledgment

The authors would like to thank to the SPOT Company from Porto Alegre for the database used in this project.

References

[1] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011. [Online]. Available: http://science.sciencemag.org/content/332/6025/60

[2] S. I. Association and S. R. Corporation, "Rebooting the it revolution: A call to action," https://www.src.org/newsroom/rebooting-the-it-revolution.pdf, 2015, online, acesso em Jan/2016.

[3] Christian Hagen, Khalid Khan, Marco Ciobo, Jason Miller, Dan Wall, Hugo Evans, and Ajay Yadav. (2013) Big data and the creative destruction of today's business models. ATKearney. [Online]. Available: https://www.atkearney.com/documents/10192/698536/Big+Data+and+the+Creative+Destruction+of+Todays+Business+Models.pdf

[4] R. Kulkarni, A. Forster, and G. Venayagamoorthy, "Computational intelligence in wireless sensor networks: A survey," *Communications Surveys Tutorials, IEEE*, vol. 13, no. 1, pp. 68–96, First 2011.

[5] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," *Multimedia Computing and Systems' 97. Proceedings., IEEE International Conference on*, pp. 509—-516, 1997. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=609763

[6] B. Zhang, B. Feng, P. Ding, and B. Xu, "TV commercial detection using constrained viterbi algorithm based on time distribution," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, no. Fskd, pp. 2010–2014, 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6234003

[7] B. Zhang, T. Li, P. Ding, and B. Xu, "TV commercial segmentation using audiovisual features and support vector machine," *Proceedings - 2012 International Symposium on Instrumentation and Measurement, Sensor Network and Automation, IMSNA 2012*, vol. 1, pp. 326–329, 2012.

[8] Sung Hwan Lee, Won Young Yoo, and Young Suk Yoon, "A visual feature based video identifying system for the TV commercial's monitoring," in *2006 8th International Conference Advanced Communication Technology*. IEEE, 2006, pp. 4 pp.–883. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1625708

[9] H. Duxans, D. Conejero, and X. Anguera, "Audio-based automatic management of TV commercials," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1305–1308, 2009.

[10] J. E. Borras, J. Igual, C. Fernandez-Llatas, and V. Traver, "A TV Commercial Retrieval System based on Audio Features," *Icete2013*, 2013.

[11] H.-G. Kim, H.-S. Cho, and J. Y. Kim, "TV Advertisement Search Based on Audio Peak-Pair Hashing in Real Environments," in *2015 5th International Conference on IT Convergence and Security (ICITCS)*. IEEE, aug 2015, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7293031

[12] J. Y. Lee and H. G. Kim, "Audio fingerprinting to identify TV commercial advertisement in real-noisy environment," *14th International Symposium on Communications and Information Technologies, ISCIT 2014*, no. 1, pp. 527–530, 2015.

[13] B. D. Barkana, B. Uzkent, and I. Saricicek, "Normal and Abnormal Non-Speech Audio Event Detection Using MFCC and PR-Based Feature Sets," *Advanced Materials Research*, vol. 601, no. December 2015, pp. 200–208, Dec. 2012. [Online]. Available: http://www.scientific.net/AMR.601.200

[14] Anatel, "Relatório de radiodifusão completo - tv," http://sistemas.anatel.gov.br/SRD/TelaListagem.asp, 2016, online, acesso em Jan/2016.

[15] H. Terasawa, J. Berger, and S. Makino, "In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with mfcc-derived spectral envelopes," *J. Audio Eng. Soc*, vol. 60, no. 9, pp. 674–685, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16372

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1163420

[17] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, *HMM-Based Audio Keyword Generation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 566–574. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30543-9_71

[18] R. Sriranjani, B. M. Karthick, and S. Umesh, "Experiments on front-end techniques and segmentation model for robust Indian Language speech recognizer," *2014 20th National Conference on Communications, NCC 2014*, pp. 0–5, 2014.

[19] Christian Arcos Gordillo, Abraham Alcaim, and Marco Antonio Grivet Mattoso Maia, "Reconhecimento De Voz Contínua Com Atributos PNCC e Métodos De Robustez WD e MAP," in *SBrT 2013*. Sociedade Brasileira de Telecomunicações, 2011, pp. 2–5. [Online]. Available: http://www.sbrt.org.br/artigos?id=11

[20] R. Radhakrishnan, a. Divakaran, and T. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 5, pp. V–628–31, 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1200048

[21] L. Jiqing, D. Yuan, H. Jun, Z. Xianyu, and W. Haila, "Sports audio classification based on MFCC and GMM," *Proceedings of 2009 2nd IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT2009*, pp. 482–485, 2009.

[22] J. Vavrek, J. Juhar, and A. Cizmar, “Audio classification utilizing a rule-based approach and the support vector machine classifier,” in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, Jul. 2013, pp. 512–516. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6613985

[23] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.