# Advertisement Recognition Using Mode Voting Acoustic Fingerprint

Reza Fahmi[1], Hossein Abedi Firouzjaee[1], Ali Janalizadeh Choobbasti[1], and S.H.E Mortazavi Nafajabadi[1], Saeid Safavi[2]
[1]Miras Information Technologies, Tehran, IRAN,
{Reza, Hosein, Ali, Hani}@miras-tech.com,
WWW home page: http://miras-tech.html
[2]School of Engineering and Technology, University of Hertfordshire
College Lane Campus, Hatfield AL10 8PE, Hertfordshire, UK
s.safavi@herts.ac.uk

## ABSTRACT

Emergence of media outlets and public relations tools such as TV, radio and the Internet since the 20th century provided the companies with a good platform for advertising their goods and services. Advertisement recognition is an important task that can help companies measure the efficiency of their advertising campaigns in the market and make it possible to compare their performance with competitors in order to get better business insights. Advertisement recognition is usually performed manually with help of human labor or is done through automated methods that are mainly based on heuristics features, these methods usually lack abilities such as scalability, being able to be generalized and be used in different situations. In this paper, we present an automated method for advertisement recognition based on audio processing method that could make this process fairly simple and eliminate the human factor out of the equation. This method has ultimately been used in Miras information technology in order to monitor 56 TV channels to detect all ad video clips broadcast over some networks.

**Keywords:** Audio Fingerprinting, Mode Voting, Advertisement Recognition.

## 1. INTRODUCTION

Nowadays many companies are using media such as TV and the Internet as platforms for advertising their products and services. Among these platforms, TV is one of the oldest and despite passing a long period since its emergence, held its place as one of the most popular tools for advertisements [1]. One of the key challenges for the companies that advertise their products on TV is to measure ad effectiveness. In order to measure the ad effectiveness, marketing companies track specific ad information such as its duration, broadcasting time, over various TV channels for a specific period of time. The main challenge here is that the process of extracting various ad information over TV channels is done manually. This is a relatively costly process in terms of the number of individuals to be employed and time spent, it also makes the process highly unlikely to be scale-able.

For Automation purposes number of already deployed systems for automatic speaker characterization systems (from speech field) were evaluated [2,3], e.g. acoustic based methods like GMM-UBM [4,2], and factor analysis based approaches [5,6], but best results obtained when we have used a method called audio fingerprinting [7]. In this paper, a method called audio-fingerprinting is used for extracting the time-line information of broadcasting ad video clips in the stream of TV channels through a completely automated process using Miras Big Data platform [7] as a storage system.

Pinpointing the occurrences of a relatively short ad video clip in a longer video stream is not a trivial task. Most existing methods focus on heuristic features [8][9][10]. These methods mostly use heuristic features such as cut rates, soundtracks volume and blank frames to distinguish ads from TV programmes. However, these methods are not robust enough in ad recognition when we dealing with a large number of TV channels which have a diverse set of ad video clips that are similar in sound features. In [1] address this issue by detecting repetition across cost, which prevents this method to scale. In [11] different method for ad detection has been reviewed. Despite having same issues which mentioned before video processing based methods can only be used on video streaming while our method can be applied to radio streaming as well.

# 2. METHODOLOGY

In order to address the issues mentioned in the previous sections, especially the cost intensity of these processes for monitoring ads in the TV video streams, a new method is proposed here. Before describing the proposed method, we explain the methodology in three subcategories:

1. Generating a subsequence signal (SSS)[1] in a relatively larger signal.
2. Labelling the SSS based on the ad training set which includes all audio of ad video clips broadcasting on the TV channels.
3. Applying mode filter on the stream of an SSS in order to recognize the ads accurately.

The whole system procedure includes preprocessing, fingerprinting detection (also called audio fingerprinting) and mode voting modules. The methodology workflow is illustrated in figure 1. Bear in mind that the arrangement of ad broadcasting in the TV stream is independent of each other. This means specific ad does not depend on the next ad to be broadcast. This phenomenon gives us the ability to label signals on the fly.

## 2.1 Preprocessing Module

In the preprocessing stage, all video streams are converted to audio signals so that the input signals would be ready to be processed with audio fingerprinting. Please be noted this part is trivial for audio streaming.

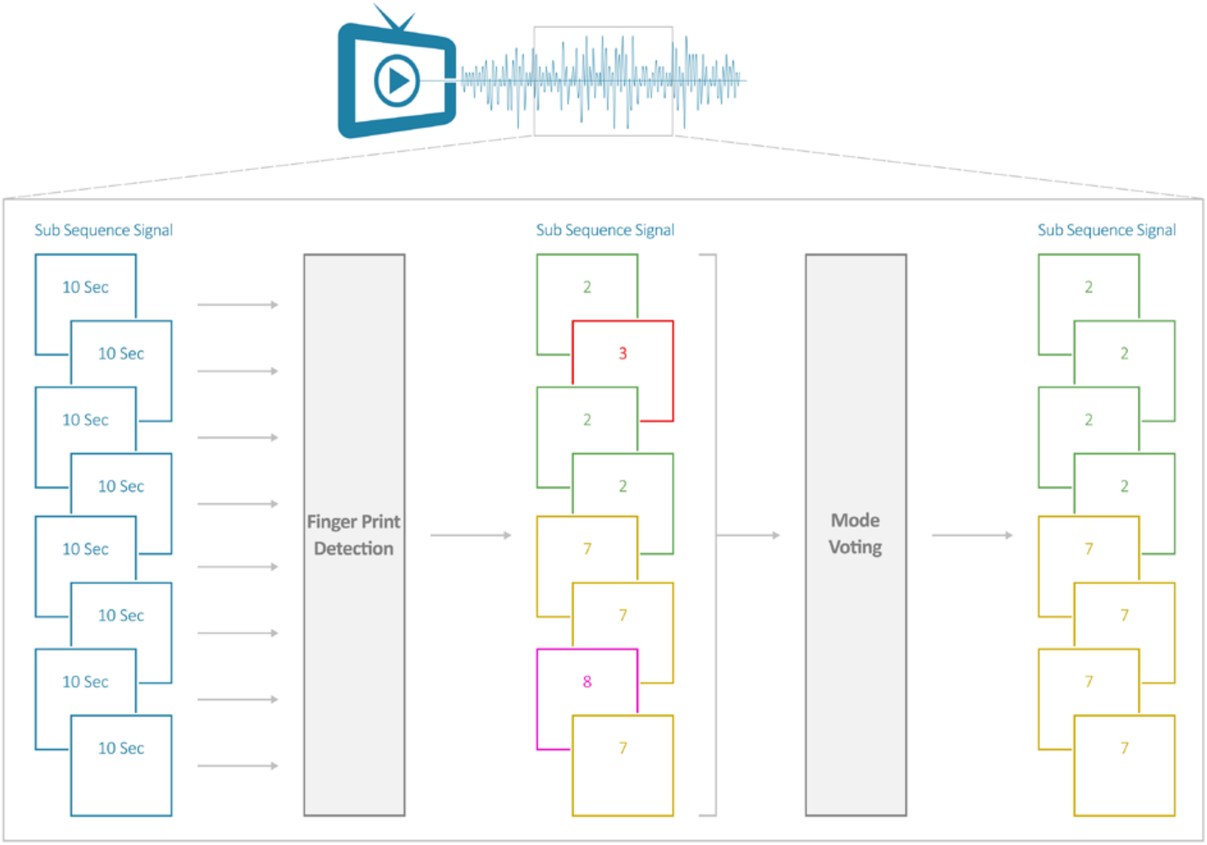


Figure 1. The methodology workflow.

Using this module, all signals are segmented into a set of small signals with length $W_L$ and distance $W_S$ from each other. This gives us the ability for batch processing on the signal. The permitted length of the streams could vary in (0, $W_L$]

whereas this parameter grows in value the loss of data in the processes of labeling SSSs would increase with choosing low values for $W_S$, and consequently increases the number of the small streams, computational cost increases.

## 2.2 Finger Printing Detection Module

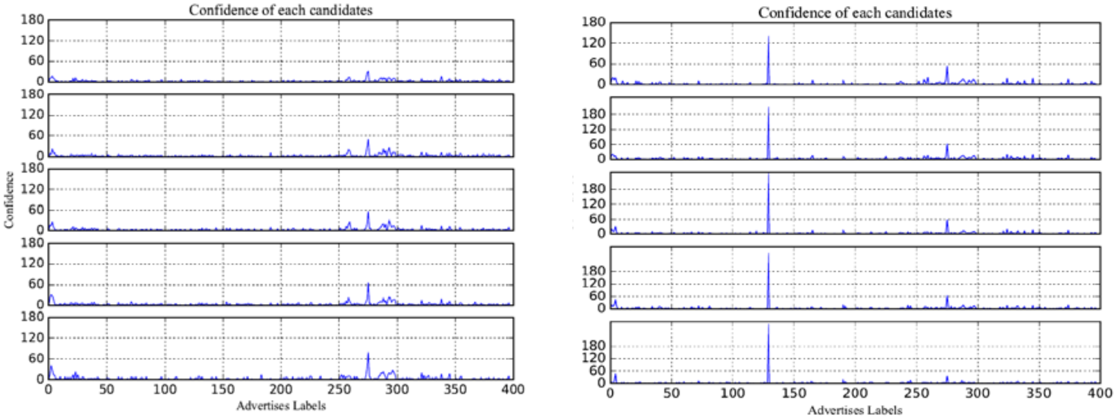
After preprocessing the inputs and converting them into SSSs, every SSS is labeled using audio fingerprint detection method. A well-known method for querying an audio signal in a database of reference signals [12], it is used to determine if an SSS is associated with a certain commercial video clip is present in our database or not. We have used hashing along with Fast Fourier Transform (FFT) on segmented signals [13], which results in producing a confidence vector with size $M$ where $M$ is the number of commercial video clips in the training set.

$$\begin{aligned} &ConfidenceVector_i^m = \sum_{w=0}^{W} Matching(SSS_i^w, m), m \in M . \\ &Matching(w, m) = \begin{cases} 1 & w \in m \\ 0 & w \notin m \end{cases} . \end{aligned} \tag{1}$$

Where *ConfidenceVector* expresses the number of fingerprints, which aligned [13]. The database for audio fingerprinting is created after the set of reference commercial video clips converted into audio signals.

## 2.3 Mode Voting Module

After performing fingerprint detection, each video stream will be represented as a sequence of labels where designate predictions for different SSSes (different parts of the stream, please see figure 1). Each label associates a certain SSS with a certain commercial video clip in the training dataset through the process of audio fingerprinting. There is a problem in which some certain parts of the stream could be very similar to more than one candidate signals, this may result in misclassification of these parts. To alleviate the problem of misclassification in a sequence of prediction a filtering process has been adopted. The filtering method used here is the mode filter that could be used to remove noise. During the filtering process a moving window with width $W_M$ slides on the sequence and as this, it proceeds, the predicted labels of each point is replaced with the mode of labels in the moving window. To differentiate between the broadcast time of commercial and other programs we use *ConfidenceVector* (please see Eq.1). It can be observed from Figure 2 while commercials are broadcasting (see Figure2 (b)) confidence value of correct candidate has much higher value compared to other candidates. However, during normal TV program broadcasting this value is negligible (see Figure2 (b)).



(a) The confidence of 5 consecutive SSSes during program commercial broadcasting time.

(b) The confidence of 5 consecutive SSSes during commercial broadcasting time

Figure 2. Presentation of ConfidenceVector

Therefore, the following equation represents the decision rule for labeling each SSS as TV program or commercial video clip.

$$\max(ConfidenceVector_i) > mean(ConfidenceVector_i) + std(ConfidenceVector_i) \quad (2)$$

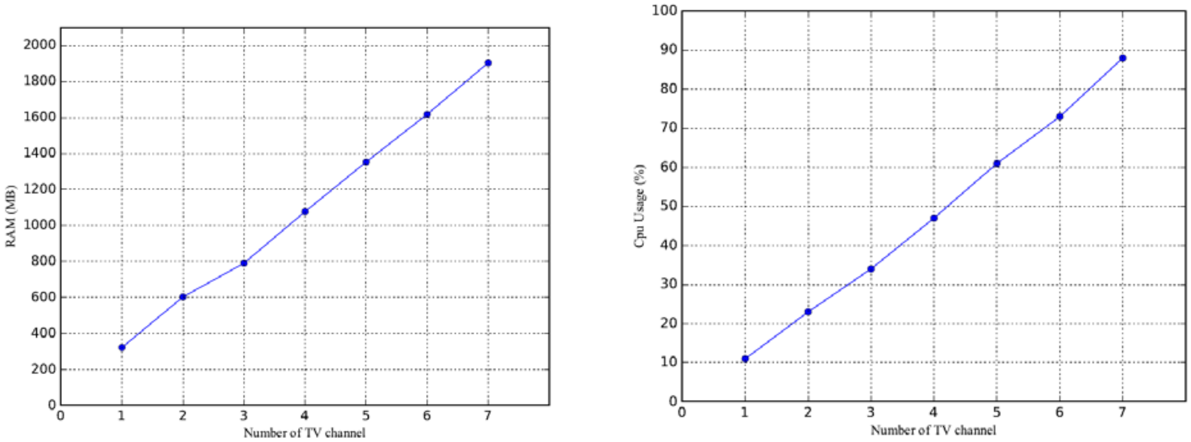
The whole procedure flowchart is presented in Figure 4.

## 3. EXPERIMENTAL RESULTS

For this study, we set up an experiment where the video stream of a single day from the IRIB 3 channel along with a set of 391 commercial video clips have been used as a benchmark for evaluating the proposed method. The parameters $W_L$ and $W_S$ are set to 10 and 2 seconds respectively. Furthermore, $W_M$ has been set to 8. The proposed method is evaluated based on the accuracy of detection and its computation cost. The video streams used for the experiment contain 218 minutes of commercials, which accounts for 106 commercial video clips in which during the experiment only 63 unique commercial video clips have been broadcast. The results of the experiment are presented in table 1. This experiment has been tested on the Intel(R) Core(TM) i7-6700HQ CPU 2.60 GHz, 2592 MHz. The memory consumption and CPU usage in this experiment have been illustrated in Figure 3.

**Table 1.**Experimental results.

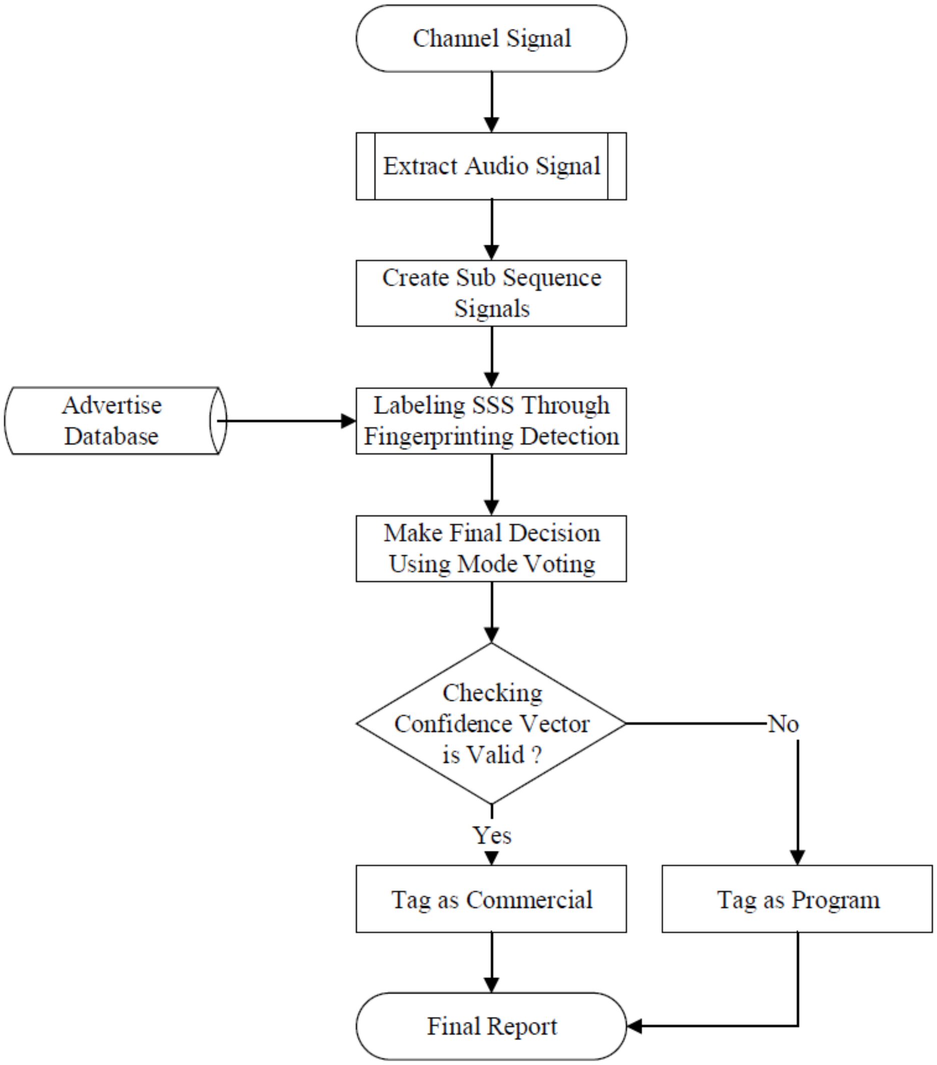
| Method | #SSS | Advertise #SSS | False Negative | False Positive | Ad Detection Accuracy | Program Accuracy |
|---|---|---|---|---|---|---|
| Without Mode Voting | 43196 | 6541 | 162 | - | 97.52% | - |
| With Mode Voting | 43196 | 6541 | 31 | 1736 | 99.53% | 99.62% |



(a) RAM consumption

(b) CPU usage

Figure 3. The impact of number of channels to be processed on computational resources



Channel Signal
Extract Audio Signal
Create Sub Sequence Signals
Advertise Database
Labeling SSS Through Fingerprinting Detection
Make Final Decision Using Mode Voting
Checking Confidence Vector is Valid ?
No
Yes
Tag as Commercial
Tag as Program
Final Report

Figure 4. The methodology flowchart.

As it can be observed from these figures, the rate of CPU usage and Memory consumption increases linearly which means this methodology can be scaled up easily.

## 4. CONCLUSION

In this paper, a simple method is presented for automating the process of pinpointing commercial video clips in a relatively large video stream and extracting the information about their occurrences. This method has fairly low computational cost because it uses the fingerprint algorithm as a model to recognize the video clips through sound features only. It has been shown that this methodology has the ability to scale up easily and can be distributed over many machines.

# REFERENCES

[1] Michele Covell, Shumeet Baluja, and Michael Fink. Advertisement detection and replacement using acoustic and visual repetition. In *Multimedia Signal Processing, 2006 IEEE 8th workshop on*, pages 461-466. IEEE, 2006.

[2] Mohamad Hasan Bahari. *Automatic Speaker Characterization*. Ph.D. thesis, Arenberg doctoral school, Faculty of Engineering Science, 2014.

[3] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, and M. Carey. Speaker recognition for children's speech. *Interspeech*, pages 1836-1839, 2012.

[4] S. Safavi, A. Hanani, M. Russell, P. Jančovič, and M. Carey. Contrasting the effects of different frequency bands on speaker and accent identification. *IEEE Signal Processing Letters*., 19(12):829-832, 2012.

[5] Saeid Safavi, Martin J. Russell, and Peter. Jančovič. Identification of age-group from children's speech by computers and humans. In *INTERSPEECH*, pages 2440-2444. ISCA, 2013.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788-798, 2011.

[7] Fred Korangy, Hamed Ghasemzadeh, Mohsen Arjmandi, and Reza Azmi. Actor system and method for analytics and processing of big data, May 10, 2016. US Patent 9,338,226.

[8] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Robust learning-based tv commercial detection. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 4-pp. IEEE, 2005.

[9] Pinar Duygulu, Ming-yu Chen, and Alexander Hauptmann. Comparison and combination of two novel commercial detection methods. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 1267-1270. IEEE, 2004.

[10] David A Sadlier, Sean Marlow, Noel O'Conner, and Noel Murphy. Automatic tv advertisement detection from MPEG bitstream. *Pattern Recognition*, 35(12):2719-2726, 2002.

[11] Brandon Satterwhite and Oge Marques. Automatic detection of tv commercials. *IEEE Potentials*, 23(2):9-12, 2004.

[12] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41(3):271-284, 2005.

[13] Audio Fingerprinting and recognition website. URL: https://github.com/worldveil/dejavu. Page accessed on February 04, 2017.

# AUTHORS' BACKGROUND

| Your Name | Title* | Research Field | Personal website |
|---|---|---|---|
| Reza Fahmi | Researcher | Artificial Intelligence | http://miras-tech.com/ |
| Hossein Abedi Firoujaee | Researcher | Artificial Intelligence | http://miras-tech.com/ |
| Ali Janalizadeh Choobbasti | Researcher | Artificial Intelligence | http://miras-tech.com/ |
| S.H.E Mortazavi Nafajabadi[1] | Researcher | Data Intensive Applications | http://miras-tech.com/ |
| Saeid Safavi | Associate professor | Speech Processing | https://www.octave-project.eu/ |

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor