EFFICIENT SPECTROGRAM-BASED BINARY IMAGE FEATURE FOR AUDIO COPY DETECTION

Chahid Ouali[1,2], Pierre Dumouchel[1] and Vishwa Gupta[2]

[1] ÉTS (École de Technologie Supérieure, Montreal), Canada
[2] CRIM (Computer Research Institute of Montreal), Montreal, Canada
{Chahid.Ouali, Vishwa.Gupta}@crim.ca, Pierre.Dumouchel@etsmtl.ca

ABSTRACT

This paper presents the latest improvements on our Spectro system that detects transformed duplicate audio content. We propose a new binary image feature derived from a spectrogram matrix by using a threshold based on the average of the spectral values. We quantize this binary image by applying a tile of fixed size and computing the sum of each small square in the tile. Fingerprints of each binary image encode the positions of the selected tiles. Evaluation on TRECVID 2010 CBCD data shows that this new feature improves significantly the Spectro system for transformations that add irrelevant speech to the audio. Compared to a state-of-the-art audio fingerprinting system, the proposed method reduces the minimal Normalized Detection Cost Rate (min NDCR) by 33%, improves localization accuracy by 28% and results in 40% fewer missed queries.

Index Terms— Content-based copy detection, audio fingerprints, spectrogram, TRECVID

1. INTRODUCTION

Content-based copy detection (CBCD) is a task with growing interest in academic and industrial fields. It provides the ability to identify duplicate audio content without the need to insert external information to the audio like watermarking techniques [1]. The idea behind CBCD is that the content itself contains enough unique information to detect copies. Audio copy detection has shown great value in a wide variety of applications including music identification, copyright control, broadcast monitoring and music library organization. Although some progress has been made in the last decade, it is still a challenging task. Audio signals subjected to a variety of distortions make audio copy detection difficult.

Many different audio features for CBCD have been used in the past. In a well known method [2], a binary fingerprint of 32 bits encodes the energy differences along the frequency and the time axes. The search uses a lookup table, allowing a very fast search. These fingerprints have been used in several other papers [3, 4, 5]. In [6], 12 Mel-Frequency Cepstral Coefficients (MFCCs) plus energy and its delta coefficients are used as audio features. A nearest neighbor search between the reference frames and query frames achieved good results in very difficult evaluation conditions. A copy detection system proposed in [7] achieved excellent performance in TRECVID 2010 and 2011 audio+video copy detection tasks. For the audio part, they used Weighted Audio Spectrum Flatness (WASF) features introduced in [8]. In [9], audio features are computed from 64 filter bank values, and made more discriminative by concatenating features for 3 successive frames.

In another related work, music identification is transformed into computer vision problem [10]. Local regions of the spectrogram image are transformed into a set of 32 bit vectors, and a classical hash table is used to perform the search. In [11], Haar wavelets are computed from spectrogram image and only wavelets with the largest magnitude are selected. Scale Invariant Feature Transform (SIFT) image descriptors are used in [12] to treat time scale modification and pitch shifting problem.

In more recent work, local spectral energies around salient points chosen from the maxima in the Mel-filtered spectra are selected [13]. Regions around each selected point are encoded to generate binary fingerprints. Compared to [2], this approach improved significantly the detection accuracy. The idea of constructing fingerprints based on spectrogram peaks has been used before in the Shazam system [14], where several time-frequency points are chosen from the spectrogram. A point is selected if it has higher energy than all its neighbors in a region centered on the point. Compact signatures representing peak pairs are then generated to form fingerprint hashes.

In this paper, we have improved the Spectro system introduced in [15, 16]. We propose a new feature extraction scheme of binary images. These images are obtained by converting the audio signal into binary spectrogram matrix based on a spectral energy threshold. We quantize this binary image by applying a tile, and transform each quantized image into an n-dimensional vector containing the positions of salient regions (or tiles) of the quantized image. This approach differs from methods like MASK [13] or Shazam [14] where salient points are selected directly from the spectrogram. In our work, features are extracted from the binary image that describes the shape of the signal after noise suppression and elimination of signal amplitude. Resulting images are robust to audio distortions. We reduce search time by quantizing the binary image and selecting the most relevant regions. We believe that salient regions are more robust than salient points, especially for transformations that add irrelevant speech to the signal. Different regions of the binary image are less likely to be distorted than different points in the spectrogram. We evaluate this method on TRECVID 2010 CBCD data, and we show that this new feature reduces min NDCR. The proposed method outperforms a state-of-the-art audio fingerprinting system [6] in terms of min NDCR, number of missed queries, F-measure and run time.

2. THE SPECTRO SYSTEM

In this Section, we describe our Spectro system in some detail, and then describe the improvements related to this paper in Sec. 3. The Spectro system transforms the audio signal into a time-frequency

representation. We convert the resulting spectrogram into a sequence of 2-D binary images and represent each image by an n-dimensional vector. We generate different versions of fingerprints by using different thresholds based on the average of the spectral values (in this work we need only one version). We create query fingerprints in the same way as reference fingerprints. However, we produce query fingerprints that have been speeded up or slowed down by 9% to reduce speed difference between queries and references. Finally, we perform nearest neighbor search between the query and reference fingerprints. In this section the Spectro system is briefly presented. More details can be found in [15,16].

Spectrogram generation: We transform the audio signal into a spectrogram matrix containing the intensity of the signal at any given time and frequency: We first down sample input audio to 8 KHz, apply Hamming window of length 96 ms, and then generate a spectrogram by computing the short time Fourier transform in this 96 ms window. We reduce this spectrogram to 257 frequency bins in the range of 500 Hz to 3000 Hz. We compute these 257 frequency bins every 3 ms.

Binary image generation: The resulting spectrogram matrix is divided into overlapping windows of size 257 x 333 (i.e. 1-sec window length) every 24 ms. We compute the mean intensity value of this 1-sec spectrogram matrix. Then, we replace the intensity values of this matrix by either 0 or 1 using this strategy: if the intensity is greater than the mean then we replace it by 1, otherwise we replace it by 0. We generate different versions of this binary image from the same spectrogram matrix by using different thresholds (e.g. 0.4 x mean, 0.6 x mean).

Fingerprint representation: We convert each binary image into an n-dimensional vector as discussed in Sec. 3. This vector is the compact fingerprint of the 257x333 binary spectrogram matrix.

Fingerprint matching: During retrieval, we first label each reference frame by the number of its closest query frame. To find the closest query frame, we use nearest neighbor algorithm with Manhattan distance as a measure of similarity. After the closest query frame has been found for each reference frame, the total number of fingerprints that match the query frame-synchronously is computed: We move the query over the reference. For each alignment, we count the number of reference fingerprints that match exactly query frame number (see Figure 1).

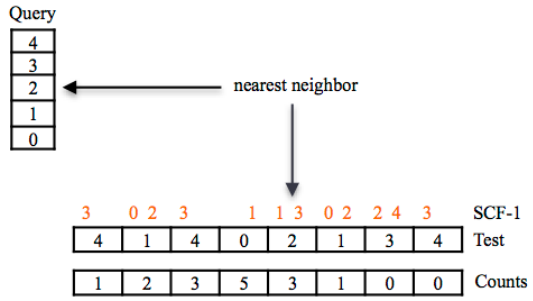

Fig. 1. Fingerprint matching.

Our algorithm differs from [6] as follows: with each reference frame, we take N frames before and after the closest frame. For example, in Figure 1 (SCF-1: successive closest frames with N=1), the closest query frame to the fifth test frame is frame 2. We update the count for not only query frame 2, but also for frames 1 and 3. This reduces the matching error due to a large overlap between frames that generates similar fingerprints for successive frames. For example, if the correct closest query frame to the reference frame is frame n, then the nearest neighbor matching algorithm may wrongfully label frame n-1 as its closest frame since query frames n and n-1 are similar due to the large overlap (24 ms frame advance and 1 sec window). We discuss in detail the influence of the parameter SCF in section 4. To accelerate this search; we use a parallel implementation of the nearest-neighbor algorithm on a Graphics Processing Unit (GPU).

3. FEATURE AND SEARCH ENHANCEMENTS

3.1. Improvements to binary image feature extraction

In the baseline version of Spectro system described above, a fingerprint is a 48-dimensional vector obtained by dividing the binary image into 24 horizontal slices and 24 vertical slices. We then take the sum of each slice to obtain a vector of 48 dimensions [15, 16]. In this section we describe a new fingerprint representation scheme that is more robust to audio degradations.

In [15, 16] we showed that increasing the dimensionality of the n-dimensional vector (where n is 48 above) reduces min NDCR. As we increase the number of dimensions we add more information about the localization of salient points in the binary image. To improve this representation, instead of using horizontal and vertical slices, we divide the binary image into small square tiles of a fixed size, so each element of the n-dimensional vector is the sum of a small square. This strategy results in a more accurate representation of the image. In the previous image representation [16], we processed each region of the image twice (one to compute the vertical slice and one for the horizontal slice). In this scenario, if a noise was introduced in the image, then two elements of the n-dimensional vector were affected. Here, only one element of the tile-based representation is affected. Secondly, when using horizontal and vertical slices, an image can be divided into a maximum of 590 slices (257 horizontal slices + 333 vertical slices) compared to 85581 (257 x 333) regions with tile-based representation (when using 1 x 1 tile).

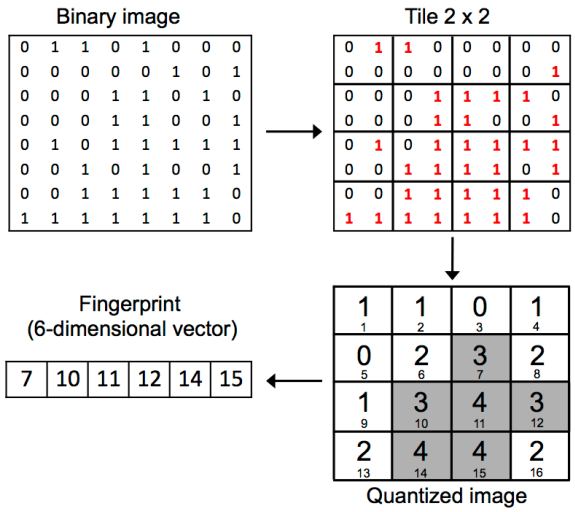

Fig. 2. Improved binary image feature extraction.

In this work we use a tile of size 10 x 10, which results in 745 tiles. However, a 745 dimensional vector is very large, increasing considerably the search processing time. Therefore, we select only a few salient tiles and discard the rest of the image. We select N tiles that represent the highest values. To encode the fingerprint, we keep only the position of the selected tile and eliminate its value. In other words, we divide the image into 745 tiles and we compute the sum in each tile. We number each tile of the image from 1 to 745, and then we look for the N tiles that have the highest values. The fingerprint represents the positions of these N tiles. Fig. 2 shows an illustration of this binary image feature extraction step.

3.2. Search algorithm

To search for a test segment that matches the query we perform a nearest neighbor search as described in section 2. Since the new fingerprint encodes positions of salient tiles in the binary image, the similarity between two fingerprints is equal to the number of positions that coincide instead of a Manhattan distance. This is equivalent to determining the intersection between query and test fingerprint elements. Finding this intersection set has a time complexity of $O(n^2)$. Pairwise comparison has a linear complexity, but in this case we must use all the 745 positions to perform pairwise operations. Since we perform the search algorithm on a GPU, transferring 745-dimensional fingerprints from CPU to GPU is time consuming. To reduce this transfer time, we transfer query and test fingerprints that encode only the positions of salient regions (N values with N << 745) from CPU to GPU, and then reorganize the fingerprints into a 745 dimensional vector once on GPU as illustrated in Fig. 3. In this figure, test fingerprint elements (i.e. positions of salient tiles) have a value of 1 when represented on the GPU by 16-dimensional vector as an example. Such a representation is appropriate for fast parallel calculation on the GPU in order to run in linear time.

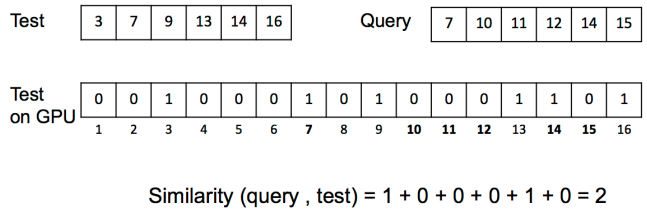

Fig. 3. Fingerprint similarity computation on the GPU.

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed method using the TRECVID 2010 audio copy detection dataset. We first present results of Spectro system using this new binary image feature that we will call Salient-Regions (SR). We study the impact of the number of successive closest frames (SCF) and the number of dimensions on min NDCR. We also compare results given by SR feature to the best results achieved by Spectro system using both Global and Local Means [15, 16]. Finally, we compare our system to a state-of-the-art NN-based system proposed in [6, 17]. This system achieved the best audio copy detection results compared to all participated in TRECVID 2009 (see TRECVID 2009 proceedings).

4.1. Datasets

The TRECVID 2010 CBCD data provided by NIST [18] consists of a reference collection of more than 11000 videos for a total of 400 hours of videos. There are 201 original audio queries, each query altered with 7 different transformations for a total of 1407 transformed queries. These transformations are: (T1) nothing, (T2) mp3 compression, (T3) mp3 compression and multiband companding, (T4) bandwidth limit and single band companding, (T5) mix with speech, (T6) mix with speech, then multiband compress, (T7) bandpass filter, mix with speech and compress.

When we examined this dataset we found other audio transformations not mentioned by NIST. For example, many queries are distorted by replacing some part of the signal by small silent segments at different places (e.g. query 3353 and 3771). In addition, some queries have undergone speed modification; either by speeding up or slowing down the query (e.g. query 3030, 4245 and 3857). Many reference audio files in this dataset have duplicates that skew the results. Therefore, we have removed these duplicate files.

4.2. Evaluation metrics

To evaluate the accuracy of locating a copied fragment within a video, we use F-measure that is defined as the harmonic mean of precision and recall. We use min NDCR to evaluate the copy detection effectiveness. NDCR is a weighted cost combination of the probability of missing a true copy and the false alarm rate. In the TRECVID evaluation, different parameters are defined for the "balanced" and "no false alarm" (NOFA) profiles. In NOFA profile, which is the more difficult, the cost of an individual false alarm is 1000 times the cost of an individual missed query, while in the balanced profile they both have a cost of 1. We report results here using the NOFA profile.

4.3. Experimental results

4.3.1. Results with the new feature

In order to reduce run time, we generated only one version of SR fingerprints with a threshold equal to the spectral *mean*, compared to four versions generated in our previous work with Global Mean and Local Mean. Table 1 shows min NDCR for Salient-Regions feature averaged over all 7 audio transformations using varying SCF values. As expected, min NDCR decreases with increasing dimensions. However, unlike our previous experiments with Global Mean and Local Mean, best results are achieved with SCF-0. The new feature seems to better represent the binary image and generates different fingerprints for successive frame, reducing the impact of similar successive frame problem.

Table 1. Min NDCR for SR features averaged over all transformations with varying SCF values.

Dimensions	SCF-0	SCF-1	SCF-3
12	0.149	0.145	0.149
24	0.135	0.136	0.143
44	0.129	0.129	0.132

Table 2 compares the best results given by Salient-Regions (SR), Global Mean (GM) and Local Mean (LM) features. From this table we can see that Salient-Regions with 44 dimensions (SR-44) outperforms both feature parameters and decreases min NDCR averaged over all transformations by 29% compared to Global

Mean and 42% compared to Local Mean. Although the lower min NDCR averaged over all transformations is given by SR-44, SR-12 and SR-24 also give good results and lower min NDCR compared to Local Mean and Global Mean. However, the lowest min NDCR for T1 and T2 transformations is achieved with the Global Mean features. Global Mean gives good results for transformations that do not add irrelevant speech to the query. However, its performance degrades for transformations that add irrelevant speech (T5, T6 and T7). Salient-Regions feature reduces the impact of adding irrelevant speech to the queries and achieves results comparable to those transformations that do not add irrelevant speech.

Table 2. Min NDCR for SR feature with varying dimensions compared to Global Mean and Local Mean features.

Feature	T1	T2	T3	T4	T5	T6	T7	Average
SR-12	0.104	0.112	0.149	0.112	0.179	0.187	0.201	0.149
SR-24	0.09	0.104	0.134	**0.097**	**0.172**	0.172	**0.179**	0.135
SR-44	0.09	0.09	**0.112**	**0.097**	**0.172**	**0.157**	0.187	**0.129**
GM	**0.075**	**0.075**	0.179	0.127	0.201	0.343	0.269	0.181
LM	0.149	0.179	0.209	0.157	0.284	0.313	0.276	0.224

4.3.2. Spectro versus NN-based

Table 3 compares min NDCR of our Spectro system (SR-44 and SCF-0), with the NN-based system introduced in [6]. The comparison shows that our system significantly outperforms NN-based system for all transformations. In fact, the Spectro system lowers min NDCR averaged over all transformations from 0.193 to 0.129, which is a relative improvement of 33%.

Table 3. Comparison of min NDCR for Spectro (SR-44) and NN-Based systems.

System	T1	T2	T3	T4	T5	T6	T7	Average
Spectro	**0.09**	**0.09**	**0.112**	**0.097**	**0.172**	**0.157**	**0.187**	**0.129**
NN-Based	0.179	0.187	0.194	0.187	0.201	0.194	0.209	0.193

Similarly, Spectro reduces the number of missed queries for all transformations and results in 69 fewer missed queries, a reduction of 40% (see Table 4).

Table 4. Comparison of the number of missed queries for Spectro and NN-Based systems

System	T1	T2	T3	T4	T5	T6	T7	Total
Spectro	**10**	**10**	**11**	**12**	**21**	**20**	**20**	**104**
NN-Based	22	25	25	24	25	24	28	173

Finally, the F1-measure for locating a query within a reference for Spectro and NN-based systems are compared in Table 5. We notice that Spectro outperforms NN-based system for all transformations. Our system improved F-measure by 28% relative to NN-based system.

Table 5. Comparison of F-measure for Spectro versus NN-Based systems.

System	T1	T2	T3	T4	T5	T6	T7	Average
Spectro	**0.869**	**0.885**	**0.889**	**0.9**	**0.896**	**0.891**	**0.867**	**0.885**
NN-Based	0.685	0.695	0.701	0.691	0.685	0.691	0.703	0.693

4.3.1. Run time

Running times in secs/query for each audio transformation and for SR-12, SR-24, SR-44 and NN-based fingerprints are shows in Figure 4. This figure shows that SR-12 is the fastest followed by NN-based features with a small difference. We notice that the running time increases as the numbers of dimensions increase but not proportionately. For example, running time is multiplied by a factor of 16 when using SR-44 instead of SR-12, even though the number of dimensions is only 4 times greater. This anomaly is related to the software implementation of the search algorithm on the GPU, and is primarily due to memory limitations that lead to this drastic increase in running time. However, even with SR-12, Spectro reduces min NDCR averaged over all transformations by 23% compared to the NN-based system.

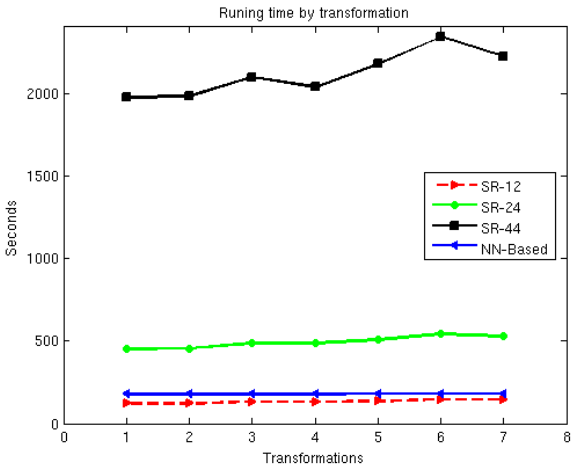

Fig. 4. Runtime in seconds/query for different systems.

5. Conclusion

We have presented latest enhancements of our spectrogram-based audio fingerprinting system. Specifically, we describe a new binary image feature extraction scheme that is highly robust to audio distortions. These images are obtained from a spectrogram matrix of the audio signal using a threshold based on the average of the spectral values. Fingerprints encode the positions of salient regions of the quantized binary image. We evaluate this method on TRECVID 2010 CBCD data, and we present results with varying dimension (12, 24 and 44 salient regions) and with varying SCF (successive closest frame) values. We show that this new feature is a better representation of the binary image and it solves the successive closest frame problem. The proposed method reduced min NDCR achieved by Spectro system using Global Mean and Local Mean features by 29% and 42%, respectively. Compared to a state-of-the-art nearest neighbor audio fingerprint system, Spectro reduced min NDCR (averaged over all audio transformations) by 33%, reduced the total number of missed queries by 40% and improved localization accuracy by 28% (when using 44 dimensions). Finally, we showed that SR-12 (with 12 dimensions) is a good trade-off between detection performance and run time. SR-12 gives 23% lower min NDCR compared to the NN-Based system, and is slightly faster.

6. REFERENCES

[1] Hartung, Frank and Martin Kutter. "Multimedia Watermarking Techniques." Proceedings of the IEEE 87, no. 7 (1999): 1079-1107.

[2] Haitsma, Jaap and Ton Kalker. "A Highly Robust Audio Fingerprinting System." In *Ismir*, 2002.

[3] Saracoglu, A., E. Esen, T. K. Ates, B. O. Acar, U. Zubari, E. C. Ozan, E. Ozalp, A. A. Alatan and T. Ciloglu. "Content Based Copy Detection with Coarse Audio-Visual Fingerprints." In *2009 Seventh International Workshop on Content-Based Multimedia Indexing (CBMI), 3-5 June 2009*, 213-18. Piscataway, NJ, USA: IEEE, 2009.

[4] Lebosse, J., L. Brun and J. C. Pailles. "A Robust Audio Fingerprint Extraction Algorithm." In *Proceedings of the Fourth IASTED International Conference on Signal Processing, Pattern Recognition and Applications, 14-16 Feb. 2007*, 269-74. Anaheim, CA, USA: ACTA Press, 2007.

[5] Lezi, Wang, Dong Yuan, Bai Hongliang, Zhang Jiwei, Huang Chong and Liu Wei. "Contented-Based Large Scale Web Audio Copy Detection." In *2012 IEEE International Conference on Multimedia and Expo (ICME), 9-13 July 2012*, 961-6. Los Alamitos, CA, USA: IEEE Computer Society, 2012.

[6] Gupta, V. N., G. Boulianne and P. Cardinal. "Crim's Content-Based Audio Copy Detection System for Trecvid 2009." *Multimedia Tools and Applications* 60, no. 2 (2012): 371-87.

[7] Jiang, Menglin, Shu Fang, YoungHong Tian, Tiejun Huang and Wen Gao. "Pku-Idm@ Trecvid 2011 Cbcd: Content-Based Copy Detection with Cascade of Multimodal Features and Temporal Pyramid Matching." In *TRECVID workshop*, 2011.

[8] Chen, Jianping and Tiejun Huang. "A Robust Feature Extraction Algorithm for Audio Fingerprinting." In 9th Pacific Rim Conference on Multimedia, PCM 2008, December 9, 2008 - December 13, 2008, 5353 LNCS, 887-890. Tainan, Taiwan: Springer Verlag, 2008.

[9] Jegou, H., J. Delhumeau, Yuan Jiangbo, G. Gravier and P. Gros. "Babaz: A Large Scale Audio Search System for Video Copy Detection." In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), 25-30 March 2012*, 2369-72. Kyoto, Japan.

[10] Yan, Ke, D. Hoiem and R. Sukthankar. "Computer Vision for Music Identification." In *Proceedings. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 20-25 June 2005*, vol. 1, 597-604. Los Alamitos, CA, USA: IEEE Comput. Soc, 2005.

[11] Baluja, S. and M. Covell. "Audio Fingerprinting: Combining Computer Vision Data Stream Processing." In *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing, 15-20 April 2007*, 213-16. Piscataway, NJ, USA: IEEE, 2007.

[12] Zhu, Bilei, Wei Li, Zhurong Wang and Xiangyang Xue. "A Novel Audio Fingerprinting Method Robust to Time Scale Modification and Pitch Shifting." In *18th ACM International Conference on Multimedia ACM Multimedia 2010, MM'10, October 25, 2010 - October 29, 2010*, 987-990. Firenze, Italy: Association for Computing Machinery, 2010.

[13] Anguera, Xavier, Antonio Garzon and Tomasz Adamek. "Mask: Robust Local Features for Audio Fingerprinting." In *2012 13th IEEE International Conference on Multimedia and Expo, ICME 2012, July 9, 2012 - July 13, 2012*, 455-460. Melbourne, VIC, Australia: IEEE Computer Society, 2012.

[14] Wang, A. L. C. "An Industrial-Strength Audio Search Algorithm." *Proceedings of the SPIE - The International Society for Optical Engineering* 5307, no. 1 (2003): 582-8.

[15] Ouali, Chahid, Pierre Dumouchel and Vishwa Gupta. "A Robust Audio Fingerprinting Method for Content-Based Copy Detection." In *International Workshop on Content-Based Multimedia Indexing*. Austria, 2014.

[16] Ouali, Chahid, Pierre Dumouchel and Vishwa Gupta. "Robust Features for Content-Based Audio Copy Detection." In Fifteenth Annual Conference of the International Speech Communication Association. Singapore, 2014.

[17] Heritier, Maguelonne , Vishwa Gupta, Langis Gagnon, Gilles Boulianne, Samuel Foucher and Patrick Cardinal. "Crim's Content-Based Copy Detection System for Trecvid." In *Proc. TRECVID-2009*. Gaithersburg, MD., USA, 2009.

[18] Smeaton, Alan F., Paul Over and Wessel Kraaij. "Evaluation Campaigns and Trecvid." In 8th ACM Multimedia International Workshop on Multimedia Information Retrieval, MIR 2006, co-located with the 2006 ACM International Multimedia Conferenc, October 26, 2006 - October 27, 2006, 321-330. Santa Barbara, CA, United states: Association for Computing Machinery, 2006.