Efficient Broadcast Monitoring using Audio Change Detection

Rohan Shah, Kedar Chandrayan, and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay,
Powai, Mumbai - 400076.
{rohanshah,prao}@ee.iitb.ac.in

Abstract. Digital audio broadcast monitoring is the task of detecting and locating occurrences of specific audio content in broadcast streams. It has important applications in the media industry such as automatic monitoring of the airing of advertisements or commercials. The monitoring is accomplished by searching the streaming audio to detect regions where similarity with the specified audio clips of interest is high. The process involves matching perceptually relevant features extracted by a frame-level analysis of the audio with stored reference templates. The searching of the entire audio stream with sufficiently small time step makes the process very computationally intensive. In this work, a method is proposed to reduce computation time by restricting the search to a limited set of candidate locations obtained using blind audio segmentation. The system is evaluated on real broadcast audio recordings to demonstrate the computational advantage achieved.

Key words: Broadcast Monitoring, Fingerprinting, Segmentation.

1 Introduction

Audio broadcast monitoring involves detecting and locating occurrences of specific audio content in audio streams. It has important applications in the media industry such as automatic monitoring of the revenue-generating airing of advertisements (commercials). It is typically accomplished by matching of reference audio templates of the commercials with the streamed broadcast audio based on similarity of acoustic features. The similarity matching process is computationally very intensive due to the fact that long audio streams have to be searched exhaustively for the occurrence of specific audio templates corresponding to one or more audio items of interest. Continuous monitoring would require real-time performance. Another important requirement of the broadcast monitoring system is robustness to typical degradations that broadcast audio is subject to such perceptual coding, time scaling, time shift, band-pass filtering and additive noise. This is an important consideration in the design of acoustic features that must provide a distinctive representation of the underlying audio and yet be insensitive to the degradations. Several systems have been proposed for broadcast monitoring in the recent literature. Acoustic features that provide a compact yet robust representation (or "fingerprint") of the audio are the subject of active research. Fingerprints extracted from the audio stream are compared via a distance measure to stored reference fingerprints in order to detect occurrences of the reference within the audio stream. Apart from robustness to common degradations of the audio signal, the fingerprint must be easily computable and provide for an easily computable distance measure. Since the objective is to identify every occurrence of the reference template (for example, the audio segment corresponding to a particular ad or commercial) in the broadcast audio, an exhaustive search over the full audio stream with small time step is required. In order to minimize any losses of accuracy due to frame boundary misalignment between the test and reference audio templates, it is essential to restrict the time step to values of about 10 ms. Search complexity is thus another important aspect to be considered in the practical deployment of a monitoring system. In the present work, we focus on search complexity reduction methods which can work, in principle, with any type of fingerprint or distance measure.

A fingerprint refers to the set of signal feature vectors corresponding to a fixed number of contiguous time intervals (frames) spanning the duration of the reference audio template. Acoustic signal features most suited to the task of similarity matching of audio segments, across signal modifications occurring in broadcast audio, are expected to be perceptually motivated. Several features have been proposed in the literature which can serve as the basis for a perceptual similarity measure including spectral envelope descriptors such as MFCC, spectral flatness and crest measures [1], perceptual attributes such as loudness, pitch, bandwidth, harmonicity, brightness [3], and spectral centroid [5]. To reduce the dimensionality and improve robustness, feature summarization by averages over time such as mean, variance, auto-correlation [3] and post-processing such as differencing and quantization [2], [7] have also been employed. Lower dimension features lead to more compact fingerprints and thus lower search complexity in terms of computation of the similarity measure.

For reliable detection of the reference audio fingerprint in the broadcast audio stream, it is necessary to perform similarity matching with fingerprints computed at small time steps throughout the streaming audio. Attempts have been made to reduce search complexity by performing fingerprint matching once per several frames [6], [4] and [5]. The algorithm in [6] uses the notion of landmarks to identify events in input which can be reproduced accurately even when audio is degraded. The events may correspond to, for example, local maxima the in spectral norm. Fingerprint computation and matching are then implemented only at the landmark frames. In [4] during the search process, fixed numbers of frames are skipped between similarity computations. Number of frames skipped depends on robustness of distance metric and should not affect the detection. The search algorithm of [5] is combination of skipping frames and candidate selection. N2 consecutive frames are selected from set of N1 frames of broadcast audio where N2 $\leq$ N1. Templates from the database whose features match with one of N2 feature vectors for at least one frame are selected as probable candidates. Detailed fingerprint matching is performed only for those candidates. This process of selection and matching is performed once every N1 frames. In [8], Lert Jr. et al. proposed the use of pre-determined events in video and audio to identify the possible locations of advertisement.

On the lines of the computation reduction approaches discussed above, the focus of the present work is to reduce the search complexity by selecting locations for template matching in a manner that yields a small set of highly probable locations of the reference templates in the audio stream. Restricting the similarity computations to the selected locations only can lead to a significant reduction in computation time. The specific method investigated for candidate frame selection is blind audio segmentation. Based on the assumption that the onset of a commercial would be marked by an abrupt change in the underlying signal statistics, audio segmentation methods are used to locate boundaries between acoustically homogenous regions of the audio stream. The proposed system is described in the next section followed by the presentation of an experimental evaluation of its performance on broadcast audio recordings in terms of computation and accuracy.

2 System Overview

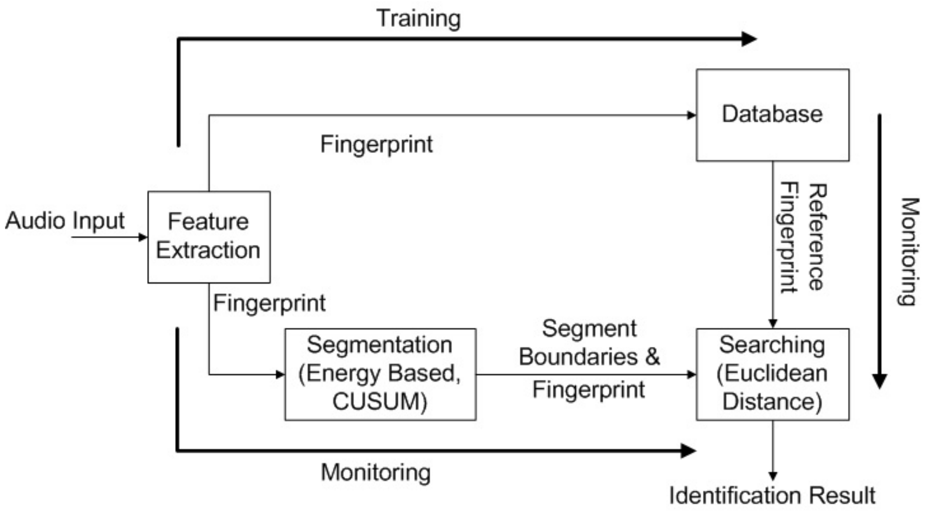

Fig. 1. Audio broadcast monitoring system based on segmentation.

The audio broadcast monitoring system depicted in Figure 1 comprises of feature extraction, segmentation and search blocks. The reference audio templates (corresponding to the advertisement clips) are processed off-line i.e. during the training of the system, to obtain reference fingerprints for the database. During on-line monitoring, the input broadcast audio stream is processed for feature extraction followed by segmentation. The segment boundaries are used as candidate locations in the search to detect the presence, if any, of a reference audio fingerprint. Although the segmentation can, in principle, be based on acoustic features that are different from those used for the fingerprint, it is preferable to use the same feature set for both modules in the interest of keeping the computational overhead from segmentation to a minimum. The implementation of the individual modules of the system is discussed next.

2.1 Feature Extraction

Our interest lies in finding features that provide a distinctive signature of the underlying audio content that can be used to estimate similarity between two audio fingerprints based on these features. Such features can also be used in the blind audio segmentation module which is based on detecting changes in the feature vector sequence to locate boundaries between relatively homogenous segments of the audio stream.

A variety of acoustic features have been proposed for use in audio similarity matching tasks. Since perceived similarity is what is sought, perceptually relevant features are used. Feature extraction is loosely based on the stages of human auditory processing, most prominently on the frequency filtering by the non-uniform cochlear filterbank. In this work we consider filterbank output powers as well as the cepstral transform of the filterbank powers in the form of the well-known Mel-frequency cepstral coefficients (MFCC).

Mel-frequency cepstral coefficients: Psychoacoustic studies have shown that the frequency resolution of human hearing is best at low frequencies and decreases with increasing frequency. This aspect has been incorporated in the derivation of perceptually relevant features in the form of output powers of a non-uniform filter bank. The Mel filter bank is widely used to simulate the cochlear time to frequency transformation. A cepstral transform is applied to the filter powers to achieve decorrelation and thus obtain a compact representation known as the Mel-frequency cepstral coefficients vector (MFCC).

Among various available implementations of MFCC, we use that of the auditory toolbox [10] wherein the initial 13 bands are linear in the range 133 Hz to 1 kHz followed by 33 log-spaced filters separated by a factor of 1.07 spanning a frequency range up to 10.3 kHz. The low-order 13 coefficients obtained after a Discrete Cosine Transform (DCT) is applied to the band energies comprise the MFCC feature vector. A data window of 46 ms is used for the spectral analysis with frame rate of 100 per second (10 ms step size). The feature vectors of contiguous frames over a selected duration are concatenated to obtain the audio fingerprint.

Binarized filterbank output powers: Time and frequency differencing of the spectral energy distribution can serve as a compact and robust signature of an audio segment [2]. The differences are further subjected to two-level quantization (binarization) to obtain a fingerprint that facilitates very simple similarity computation by way of the Hamming distance between binary sequences.

Based on the implementation of Haitsma and Kalker [2] for the power spectrum distribution, the spectrum is divided into non-overlapping logarithmic bands covering the frequency range of interest, i.e. 33 non-overlapping logarithmic bands in the range 300 Hz - 2 kHz are used to obtain the power spectrum distribution of each frame with a data window of duration 92 ms. An additional set of analysis parameters similar to that of MFCC was also investigated. The data window length was reduced to 46 ms and power spectrum distribution was computed over 33 non-overlapping logarithmic bands in range of 200 Hz to 10 kHz. In both cases, the step size used is 10 ms.

The binarized feature vector was computed for each of the above two parameter settings as described next. If energy of band 'm' in frame 'r' is denoted by E(r,m) then the actual bit in the binarized feature vector for band 'm' in frame 'r' F(r,m) is given by

$$F(r,m) = sign\left\{ \begin{array}{l} (E\,(r,m) - E\,(r,m+1)) - \\ (E\,(r-1,m) - E\,(r-1,m+1)) \end{array} \right\} \quad (1)$$

Thus if the frequency spectrum has M bands, the feature vector consists of M-1 bits. The bit vector obtained by this method has randomized pattern due to high pass filtering between adjacent frames. This randomized pattern requires precise time alignment between query and reference template for detection. This problem can be overcome by applying low pass filtering between adjacent frames [7].

$$F(r,m) = sign\left\{ \begin{array}{l} (E\,(r,m) - E\,(r,m+1)) + \\ (E\,(r-1,m) - E\,(r-1,m+1)) \end{array} \right\} \quad (2)$$

Resultant bit pattern after applying low pass filter has bit value of '1' in corresponding band enhancing the tonal component present in underlying audio. For further reference, binarization with high pass filter (Eq. 1) is termed as Binarized Band Energy Dif while other one is termed as Binarized Band Energy Avg (Eq. 2). Table 1 summarises the analysis parameters for the different feature vectors investigated in this work.

2.2 Segmentation

Blind audio segmentation is applied to locate boundaries between acoustically homogenous segments of the audio stream. Statistical distance measures have been used to detect instants of change in speaker and speech-music segmentation tasks. In the present case, we explore the utility of blind segmentation in the context of audio fingerprinting with a view to restrict the search to segment boundaries only. The underlying assumption is that the onset of an advertisement is related to an abrupt and significant change in the acoustic characteristics of the signal. Statistical models fitted to the observations on either side of each frame instant in the audio track are compared using a distance measure to select one of the two hypotheses for that instant: change or no change. In the interest of computation reduction in the subsequent search, it is desirable to have the reliable detection of advertisement audio boundaries with low false alarm rate. With this in view, we select the cumulative sum (CUSUM) algorithm for change detection, previously found to minimize miss probability for a given false alarm rate with respect to competing algorithms [9]. In addition, we incorporate an energy-based criterion to exploit the characteristic that a brief duration of silence often marks the change of audio track in a broadcast stream. The CUSUM algorithm involves fitting a model such as a Gaussian Mixture Model (GMM) to the statistics of observations on either side of the postulated change instant. An observation window duration n_0 is selected based on the assumption that it spans only one change point at most. The input feature vector stream is used to train two GMMs, one with the first α feature vectors within the observation window, and the next, with the final α feature vectors. The parameter α is chosen to be small enough to span a strictly homogenous segment of audio. A log-likelihood ratio (llr) is computed at each frame instant in the observation window given by

$$llr(x_i) = \log\left(\frac{likelihood\,(GMM_{end}|x_i)}{likelihood\,(GMM_{start}|x_i)}\right) \tag{3}$$

where x_i is the feature vector of the ith frame. The CUSUM at frame i is given by

$$CUSUM\,(x_i) = \sum_{k=i}^{l} llr(x_k) \tag{4}$$

where l is the last considered frame index in the observation window. If maximum value of CUSUM in the window exceeds a threshold then that location is marked as a change instant and the observation window is shifted to that location. The duration of the observation window is increased further by n_0. The process is repeated until the end of the broadcast audio.

In the overall segmentation procedure, first the frame-energy is thresholded to mark silence regions. Then the above CUSUM based change point detection is implemented over the continuous segments between silences to obtain the locations for the fingerprint search. When MFCC is used as base feature for fingerprint generation, the same MFCC vectors are used for CUSUM segmentation. While in case of binarized band energy, the vector of band energies computed across the 33 bands is used for segmentation.

2.3 Search

Each reference audio fingerprint in the database is matched with the same duration of audio located at a detected change instant via a normalized distance between feature vectors. For the MFCC, Euclidean distance between corresponding frame-level feature vectors serves as distance measure. In the case of the binarized fingerprints, the Hamming distance which gives the number of bit errors in a vector is used. If the minimum distance across reference templates is below a threshold, the corresponding reference template is declared as detected at the audio stream location. Once a particular template is identified, the searching algorithm skips the next several frames depending on the duration of the template (i.e. a minimum distance constraint). This method reduces the time taken by the algorithm but puts a limit on maximum hit rate achieved by the system. As the threshold is increased, the hit rate as well as the false alarm rate of the system increase. But after a point, false detections start masking the actual location due to the minimum distance constraint and reduce the hit rate achieved by the system.

3 Experimental Setup

In order to test the system, three different broadcast audio tracks were obtained as described in the following.

1. Audio track of an IPL (cricket) match of duration 3 hours obtained from the internet. It contained 33 commercial advertisements each of 30 sec duration. The number of distinct advertisements was 5.
2. News channel broadcast of duration 4 hours as recorded from a television set-top box. It contained 98 advertisements of which there were 15 distinct ones ranging from 5 to 30 sec.
3. News channel broadcast of duration 10 hours as recorded from a television set-top box. It contained 559 advertisements of which there were 60 distinct ones ranging from 5 to 30 sec.

With 5 advertisements common between tracks 2 and 3, the database contains a total 690 occurrences of 75 different advertisements. All recordings are sampled at 22.05 kHz with 16 bit resolution. For the reference fingerprint generation, one of the occurrences from the recording itself was taken. Either the entire advertisement or initial 10 sec of audio whichever is shorter is selected as the reference template. For testing, the 17 hours of audio was divided into 34 clips of half an hour duration. Though average duration of silence before advertisement was more than 100 ms, there were a few cases where the actual duration of silence was about 10 ms only, prompting us to set the silence duration threshold to 1 frame. The CUSUM parameters n_0 and α were 100 and 30 respectively. These parameters were selected based on the observed rapidity of the variation of audio content in the tracks. The threshold for segmentation was varied and finally 1000 for MFCC and 500 for band energy was selected empirically which marks maximum number of actual advertisement location marked with minimum average number of candidate locations.

Table 1. Summary of parameters for the different features investigated

Feature	Window length
MFCC	46 ms
Binarized Band Energy Dif (300 Hz - 2 kHz)	92 ms
Binarized Band Energy Avg (300 Hz - 2 kHz)	92 ms
Binarized Band Energy Dif (200 Hz - 10 kHz)	46 ms
Binarized Band Energy Avg (200 Hz - 10 kHz)	46 ms

In case of the MFCC fingerprints, the optimal distance threshold was found to be signal dependent. This is clear from the Figure 2 and Figure 3 where the minimum distance value in recording 1 is lower than distance value in recording 2. Thus threshold was selected as a fixed percentage of the average of the distance obtained across the particular test audio track. In the case of the binarized fingerprints as shown in Figure 4, the bit error rate (normalized Hamming distance) remained in range of 0.45 to 0.55 which enabled in selecting common threshold across the test audio streams. The threshold was systematically varied in order to study the trade-off between hit rate and false alarm rate in fingerprint detection.

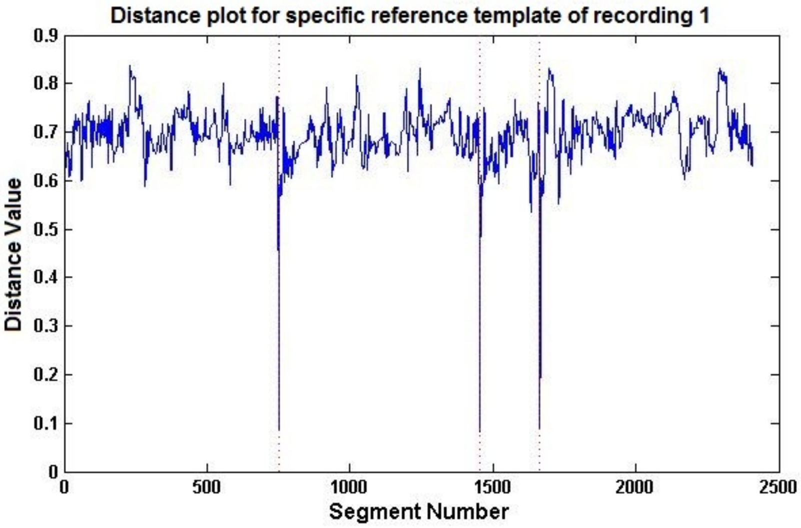

Fig. 2. Distance plot for specific reference template of recording 1 with MFCC feature

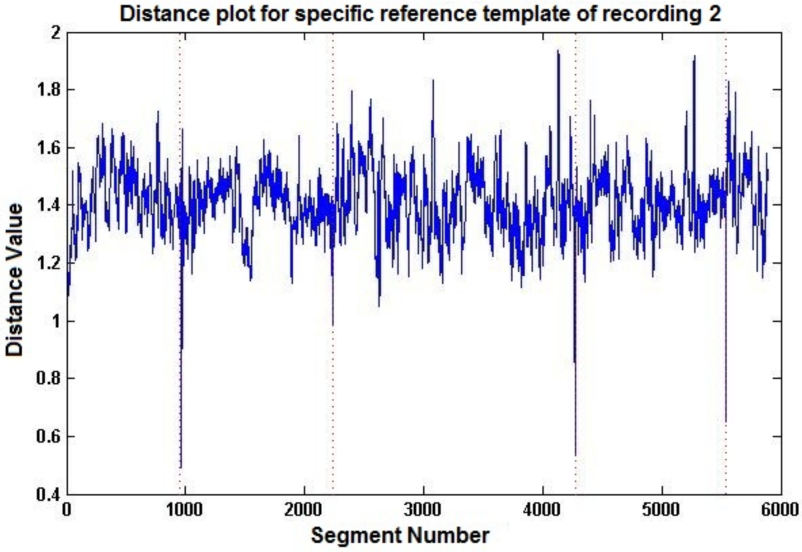

Fig. 3. Distance plot for specific reference template of recording 2 with MFCC feature

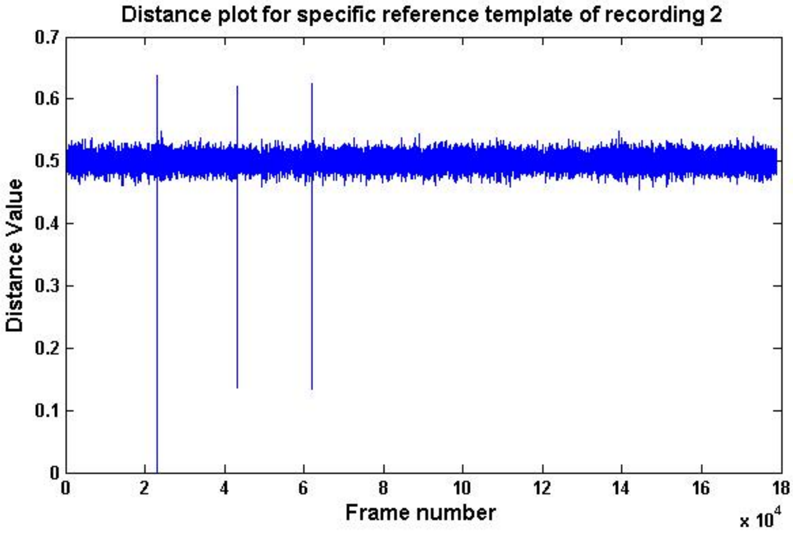

Fig. 4. Distance plot for specific reference template of recording 2 with Binarized Band Energy Dif (200 Hz 10 kHz) feature

4 Results

The performance of the system is measured by its precision and recall. Let c be the number of correctly identified reference templates (advertisements), d be the number of reference template occurrences that are not detected or identified wrongly and i be the number of false alarms (reference templates detected at locarions where there are none). We can then define precision and recall as follows.

$$Precision(P) = \frac{c}{c+i} \tag{5}$$

$$Recall(R) = \frac{c}{c+d} \tag{6}$$

The performance for 3 MFCC feature based systems with different segmentation methods (no segmentation, silence only and silence + CUSUM) appears in Table 2. These results are achieved by running the full audio monitoring system with MFCC features only on Recordings 1 and 2. The baseline system does not use change detection and performs fingerprint matching exhaustively at each frame. When silence detection is incorporated, the number of frames selected for matching computations decreases drastically as observed from the search time reduction. However there is an increase in the missed detections of the advertisements. With the further incorporation of CUSUM based change detection, we observe that the missed detections are reduced at the cost of an increase in false alarms and in search computation time.

Table 2. Comparative performance of 3 systems (Base feature: MFCC, Recording 1 and 2 only)

System	Max. recall	Precision	Computation time for 30 min duration
Baseline	1	0.938	2500 seconds
Silence	0.918	1	180 seconds
Silence + CUSUM	0.943	0.950	550 seconds

CUSUM alone cannot be used to mark probable advertisement location as it fails to mark exact start location near silence. This makes silence based segmentation necessary. In terms of computational time, the change detection algorithm takes about 180 seconds of computation. With around 5000 probable location per 30 min audio stream duration, the search algorithm takes about 550 seconds of computation. This gives around 5 times overall speed gain in the monitoring due to the incorporation of segmentation.

Table 3. Segmentation performance for different features

Feature	**Correct locations detected (Out of 690)**	**Average number of locations marked per 30 min duration**
MFCC	639	5000
Band Energy (300 Hz - 2 kHz)	630	7500
Band Energy (200 Hz -10 kHz)	629	10000

Table 3 provides the comparative performance of different features for segmentation algorithm. When band energy is used for CUSUM segmentation, average number of probable locations obtained was around 10000 with decrease in detected advertisement locations. The MFCC feature vector performs the best for segmentation probably due to its compactness and decorrelation. Although the segmentation performance of band energies is worse than that with MFCC, we use band energies in conjunction with the binarized audio fingerprints to minimize the overhead that would be incurred from the computation of different feature sets for different modules.

Table 4. Detection recall and precision (all 3 recordings)

Feature vector	**Segmentation**	**Comp. time (sec.) for 30 min audio stream**	**Recall with full precision**	**Max. recall**	**Precision at max. recall**
MFCC	No	2500	0.928	0.963	0.824
	Yes	550	0.907	0.940	0.887
Binarized Band Energy Dif (300 Hz - 2 kHz)	No	875	0.962	0.991	0.898
	Yes	320	0.885	0.904	0.958
Binarized Band Energy Avg (300 Hz - 2 kHz)	No	875	0.962	0.991	0.955
	Yes	320	0.915	0.973	0.793
Binarized Band Energy Dif (200 Hz - 10 kHz)	No	850	0.960	0.986	0.892
	Yes	250	0.701	0.871	0.953
Binarized Band Energy Avg (200 Hz - 10 kHz)	No	850	0.966	0.994	0.793
	Yes	250	0.914	0.957	0.904

A performance comparison of systems with segmentation and without segmentation for different feature vectors appears in Table 4. For all four feature vectors, the use of segmentation reduces the computation time with various amounts of decrease in the accuracy. More detailed performances in terms of precision and recall for the different systems are shown in Figure 5 and Figure 6. It is clearly visible from the figures that for a given precision, systems without segmentation have higher recall. The loss in recall has been introduced due to the failure of segmentation algorithm to mark all the locations effectively. For a few cases of the reference template, a sufficiently large change in acoustic property or silence was not observed which led to missed detection at that particular location. One more notable point was reduction in hit rate for Binarized Band Energy Dif. As pointed out earlier, when the detected segment location is one or two frames away from the actual boundary location, the feature vector obtained can be misaligned with the template leading to missed detection. This problem is largely overcome by the use of the Binarized Band Energy Avg.

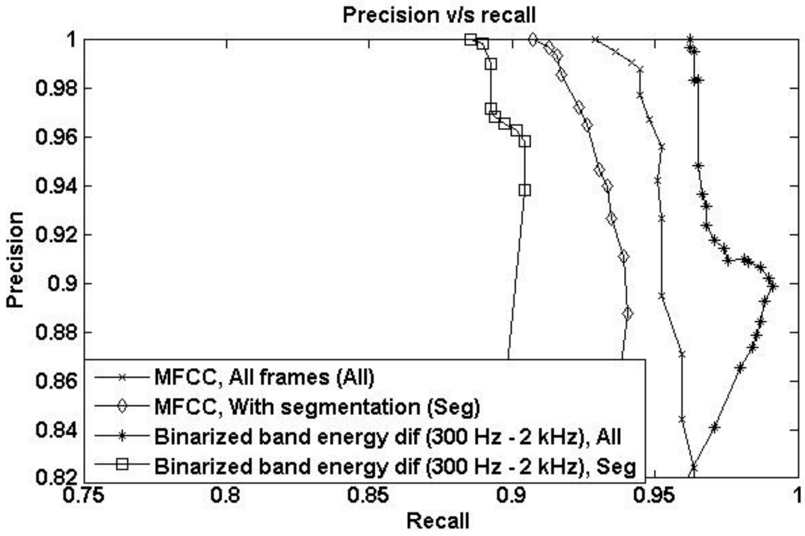

Fig. 5. Precision v/s recall for MFCC and Binarized Band Energy Dif (300 Hz - 2 kHz) with and without segmentation

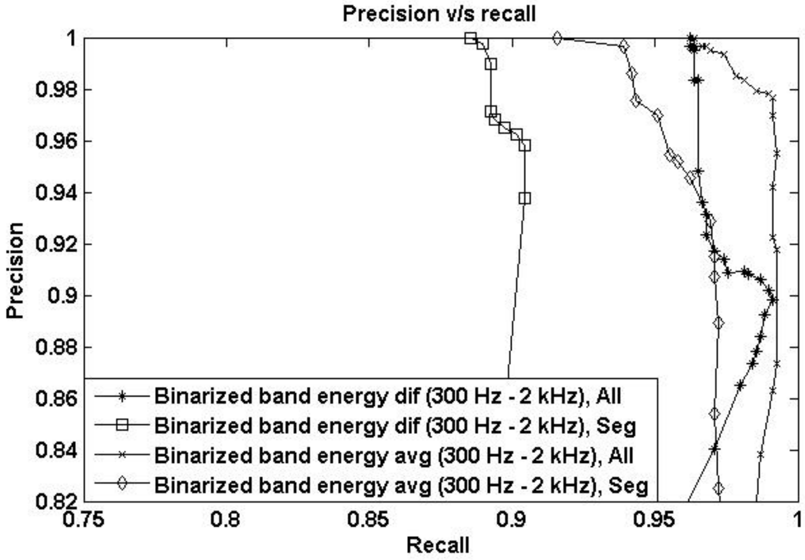

Fig. 6. Precision v/s recall for Binarized Band Energy Dif (300 Hz - 2 kHz) and Binarized Band Energy Avg (300 Hz - 2 kHz) with and without segmentation

5 Conclusion

Audio broadcast monitoring is implemented by matching fingerprints of reference audio templates with the fingerprints extracted from the streamed audio. In this work, a method to reduce the computational complexity of the search by restricting the similarity matching to candidate locations obtained by a prior stage of blind audio segmentation was investigated. Evaluation on a broadcast audio database with a total duration of 17 hours containing 690 occurrences of 75 distinct commercial clips demonstrated computational reduction by a factor of 3 or more with a loss in accuracy (missed detections) of less than 5%. Future work will address improvements to the segmentation module to reduce the missed detections as well as further improvements in overall computational efficiency.

References

1. Allamanche, E., Herre, J., Hellmuth, O., Froba, B., Kastner, T., Cremer, M.: Content-based Identification of Audio Material Using MPEG-7 Low Level Description. In: ISMIR proceedings of the Annual International Symposium on Music Information Retrieval, pp.197-204 (2001).
2. Haitsma, J., Kalker, T.: A Highly Robust Audio Fingerprinting System. In: ISMIR proceedings of the 3^{rd} International Conference on Music Information Retrieval, (2002).

3. Wold, E., Blum, T., Keislar, D., Wheaton, J.: Content - Based Classification, Search and Retrieval of Audio. IEEE Multimedia, pp. 27-36, (Fall 1996).
4. Pinquier J., Andre-Obrecht R.: Audio Indexing: Primary Components Retrieval - Robust Classification in Audio Documents. Multimedia Tools and Applications, Springer-Verlag Vol.30 N. 3, pp313-330 (2006)
5. Jang, D., Lee, S., Lee, S. J., Jin, M., Seo, J. S., Lee, Sunil, Yoo C. D.: Automatic Commercial Monitoring for TV Broadcasting Using Audio Fingerprinting. In Proceedings of The 29th Audio Engineering Society Conference, (2006).
6. Wang, A. L. C., Smith, J. O. III.: Systems and Methods for Recognizing Sound and Music Signals in High Noise and Distortion. U.S. Patent, Patent no. US2006012839A1, 8th June, 2006.
7. Neves, C., Veiga, A., Sa, L., Perdigao, F.: Audio Fingerprinting System for Broadcast Monitoring. In 7th Conference on Telecommunications. Santa Maria da Feira - Portugal, (May 2009).
8. Lert Jr., J. G., Lu Dunedin, D.: Broadcast Program Identification Method and Apparatus. U. S. Patent, patent no. , 30th June, 1987.
9. Omar, M., Chaudhari, U., Ramaswamy G.: Blind Change Detection for Audio Segmentation. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 501-504, (2005).
10. Slaney, M.: Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work Version 2. Technical Report of Interval Research Corporation, (1998).