European Journal of Computer Science and Information Technology
Vol.7, No.6, pp.1-22, December 2019
Published by ECRTD- UK
Print ISSN: 2054-0957 (Print), Online ISSN: 2054-0965 (Online)

CLASSIFICATION OF PUBLIC RADIO BROADCAST CONTEXT FOR ONSET DETECTION

C. O. B. Weerathunga, K. L. Jayaratne, P. V. K. G. Gunawardena
University of Colombo, School of Computing

ABSTRACT: *This research focuses on the investigation of a unified methodology for the onset detection in Sri Lankan radio broadcast context with the approach of classification of the broadcast context. Various audio patterns in the broadcast context were observed and a supervised learning approach was employed in the classification process. Different audio features were examined with respect to the broadcast context. Identified audio semantics in the broadcast context were used in refining the output gained in supervised learning models. Onsets were predicted using the classification results. The evaluation method was carried out with ground truth data obtained from a Sri Lankan FM broadcast recording. The proposed approach provided the accuracies of 41% for news, 76% for radio commercials, 75% for songs and 59% for other voice related segment classification. The onset detection model was successful in predicting the onsets with an error rate of (+/-) 2.5s with approximately 82% of accuracy level.*

KEYWORDS: audio monitoring, radio broadcast, onset detection, audio feature analysis, supervised learning

INTRODUCTION

Radio broadcasting plays an important role in today’s communication by providing opportunities to people to keep up-to-date on the news and trends. It is a unidirectional wireless transmission over radio waves. Radio stations broadcast different content to receivers and here, the receiver is known as the listener. Large portion of the world population listen to AM/FM radio over the airwaves which is higher than TV viewership, PC use, smartphone and tablet usage [1].

Information which is given over the radio, broadcasts to a large number of listeners. Many countries consist of large number of radio broadcasting channels which transmit different content categories. Sri Lanka itself contains around 100 broadcasting channels which are generally being divided along linguistic lines with state and private media operators. Radio broadcasting is important for both developed and developing countries for information provision as well as for entertainment. The overall content which is broadcasted on a radio station can be categorized in to different sections. Some of them are as news, songs, commercials, radio dramas, discussions, interviews, sport commentaries, religious programs etc.

Monitoring radio broadcast content is an essential part in every country’s broadcasting act [2, 3, 4, 5, 6]. Authorized people in mass media and information corporations, singers, composers, lyricists, advertising agents, government defence organizations and law enforcement authorities etc. will need information in the broadcasting content for different purposes [7, 8]. Composers, singers and lyricists need monitoring of music/songs which are broadcasted in the radio to monitor the copyrights, royalty payments [9, 10, 11, 12, 13], and for security rights of their compositions. Advertising agents keep alerts on the broadcasting advertisements. Regulating authorities monitor broadcast content to see whether they accept the government enforced laws. Government defence organizations and law enforcement authorities keep an alert on the radio broadcast content, especially on news and radio discussions for different name referencing. With all these requirements of different stakeholders it is clear that an automated monitoring for radio broadcast context is a requirement in today's society.

This paper describes the attempts taken by the researcher in classification of public radio broadcast context into different pre-identified content categories and the onset detection of these content categories based on the classification result. The remainder of the research paper is organized as follows. Section II elaborates the related literature for onset detection and classification of broadcast context. In Section III, a detailed description of the proposed approach with the designed details is discussed. Section IV discusses the experimental setup, results and findings of the research. In Section V we conclude with future work of the research.

Commercially available solutions

As discussed above, it is clear that there is a massive requirement for the radio broadcast monitoring. For the monitoring purposes, there are several cheap and non-reliable audio analysis approaches are available in the current environment. Some of those approaches include reading the attached meta-data, asking for the broadcast report from broadcast stations, having a human observer to listen and monitor the content played in a particular broadcast channel etc. Since there are large number of radio channels exists in a country above mentioned methods are highly impractical and will lead to erroneous solutions. Today there exist some commercially available software solutions for the radio broadcast monitoring process.

Digital radio tracker (DRT) [14], ACRCloud broadcast monitoring service [15] and 'BeatGrid media' [16] can be identified as some of the commercially available software solutions for the radio broadcast content monitoring. Almost all of them are not for the Sri Lankan broadcast context and are not freely available. Most applications are not supporting for a unified approach for the broadcast monitoring. Many applications monitor separate content categories. Some applications require a content to be uploaded for the monitoring process and only selected items will be monitored. Therefore, it is clear that a unified methodology for the monitoring of Sri Lankan radio broadcast content has a knowledge gap.

As an initial approach for the broadcast monitoring system, it is essential to identify the different content categories in a FM radio broadcast. Since a radio stream is continuous it contains all these aforementioned categories (i.e. news, commercials, songs, radio dramas, phone conversations and other human voice content). Therefore, the classification of radio broadcast context and onset detection of it can be identified as the research problem which aids in facilitating a unified methodology to identify different content categories in the Sri Lankan FM broadcast context.

Onset detection

Onset detection is the technique of finding the starting point of the content in an audio context. The main goal of onset detection can be seen as the detection of events in an audio signal [17]. Different multimedia intelligent management systems use the techniques of onset detection for the classification of different content in audio files.

For realistic situations, it is hard to get an ideal audio signal. In the real world, the signal is polyphonic, which means there might be different sound sources appearing at the same time interval and might contain some noise from the outside environment. So, it is difficult to detect the onset locations directly with quantitative time varying in the transient region [18]. Because of that, there exist different approaches to find onsets in an audio signal. Approaches to onset detection can be of different domains as time domain, frequency domain, phase domain and complex domain [19]. In time/temporal domain the onset detection approach is based on the energy of the signal. Here the relative change in energy is considered in building the detection function [19, 20]. In the frequency domain the high-frequency content across a signal spectrum is observed and their magnitudes are used in building the detection function [19]. In phase domain, the phase change is observed across fast Fourier transformed frames [19, 21]. In complex domain functions, different detection domains are combined to build a single detection function [19, 22].

According to the research work; Tutorial on Onset Detection by Pablo Bello et al. [19] they have stated that if the signal is very percussive, time domain methods are usually adequate and if not, then the spectral methods like phase distributions and spectral differences are adequate. And also, they emphasized that a combination of different detection functions might work well with the application levels.

RELATED WORK

Auditing the content of audio transmitted by radio stations is of great interest for the government, publicists, musicians, and for the managers of radio broadcast stations among others. Monitoring radio broadcast serves several important purposes such as auditing audio marketing campaigns such as advertisements; ranking popular songs, ensure copyrights and royalty payments of music compositions, preventing banned content being broadcasted etc.

A smarter and more reliable approach to broadcast monitoring will need to classify the broadcast context into pre-identified categories. In the audio classification process, onset detection plays a major role. Many researchers have attempted in different ways in detecting the onset of audio files. Many of them include onset detection in music domain. Relatively fewer attempts have been made in onset detection and classification of radio broadcast context. In the analysis of radio broadcast context, the discrimination between speech/human voice and music/non-speech signals is an important problem. As a result, a lot of research has been conducted in a global context in this area.

The most common way of onset detection in an audio stream is using an energy based algorithm. Even though it is generally a fast algorithm, its effectiveness becomes low when the transients of the signals are not presented and when the energy of the signals get overlapped with polyphonic mixtures. An alternative to the standard onset detection was proposed by Juan Pablo Bello and Mark Sander [23]. They came up with a phase-based onset detection technique instead of the traditional energy-based detection. Since the phase carries all the timing information of an audio signal and as transients are well-localized in the time domain the proposed approach have made more meaningful results. According to the evaluated results, their proposed onset detection function only worked for instrumentals.

The neural network is another approach that has been attempted in onset detection. Jan Schlüter and Sebastian Böck have proposed a novel approach for musical onset detection using a convolutional neural network (CNN) [17] as an alternative to recurrent neural networks (RNN) and hand-designed method. They have used the spectral representation of the wave and onsets were characterized by a swift change of spectral content over time. With the selected musical domain the characterization of onsets through a spectral representation is possible. But the spectrogram representation of a radio broadcast wave would be quite different from the instrumental and musical composition. Hence the identification of onsets from the spectrogram representation would be inefficient.

In the process of extracting and classification of radio broadcast context, many researchers have tried to differentiate speech and music from radio streams by using different approaches. Omer Mohsin Mubarak, Eliathamby Ambikairajah, and Julien Epps have attempted in separating music with speech based on Mel frequency cepstral coefficient (MFCC). Gaussian mixture model (GMM) has been used as the classifier [18]. They discovered that music and speech gave minimum error rates in a different number of MFCC-based features.

Music note onsets can be classified as ‘soft’ onsets and ‘hard’ onsets. Hard onsets show a sudden change in energy while soft onsets show a very gradual change. Even though hard onsets are easily detectable using a time-frequency representation and energy based function, soft onset detection is somewhat hard as sound sources contain noise and oscillations associated with frequency and amplitude modulation [24]. As an approach to distinguish both hard and soft onsets from music pieces Ruohua Zhou and Joshua D. Reiss proposed a methodology for music onset detection by combining energy based and pitch based approaches [24]. Energy based functions are used for the signals identified to be hard onsets and pitch based algorithms are used for the signals with soft onsets. Their approach worked best for the classes of the solo drum, solo brass, and solo wind. The researchers have stated that by combining both energy and phase-based approaches together might work well for other onset detection classes.

The approaches presented in [17, 18, 23, 24] usually work in the offline mode as their peak picking algorithms rely on future information to determine the location of the onset. Only a few algorithms were designed to work for online scenarios by aiming to minimize the delay between the onset occurrence and reporting [25]. Sebestian Böck, Andreas Arzt, Florian Krebs, and Markus Schedl proposed a novel approach for online real-time onset detection with recurrent neural networks [25]. Their proposed system comprised of three main processing steps as signal pre-processing, neural network onset prediction and peak post-processing and the proposed approach was successful in achieving performance close to current state-of-the-art offline onset detection algorithms having a zero delay between the onset detection and reporting it.

Many researchers have focused on onset detection and segmentation of content in music compositions. Only a few have paid their concentration of that in radio broadcast context. John Saunders has proposed a novel approach in discriminating speech from music on broadcast FM radio [26]. Researchers were successful in identifying some of the features which can differentiate speech and music. Some of them are tonality, energy sequences, excitation patterns etc. In their proposed approach they have used the average zero-crossing rate (ZCR) of the time domain waveform and a multivariate Gaussian classifier to decide the class of the test token. Even though the dataset gathered contained different broadcast content categories like talk, commercials and many types of music the paper does not mention any evaluation results.

When considering the FM radio broadcast content radio commercials can be seen as a content type which has a very dynamic and unpredictable behavior. In modern times a larger portion of broadcast content contains advertisements. Shashidhar.G.K et al. have proposed a real-time identification of advertisement segments in radio broadcast [27]. The researchers have identified certain audio features like energy, pitch, and duration etc. that present in both advertisements and other audio streams. And also they have observed some other factors related to advertisements in the FM broadcast like the frequency of the occurrence of advertisements, appearance of the custom tune of the radio station before advertisements, pitch and speaking rate of the speaker in advertisements etc. unlike in previous research the researchers have used an ensemble approach with Hidden Markov Model (HMM) and Artificial Neural Network (ANN) for the detection process.

Audio classification and segmentation will provide useful information in audio content understanding. Lie Lu, Hong-Jiang Zhang and Hao Jiang proposed an approach in audio stream segmentation [28]. They have followed a hierarchical approach in the classification of speech and non-speech discrimination by employing a novel algorithm based on K-Nearest Neighbor (KNN) and linear spectral pairs-vector quantization. Further segmentation of non-speech content into music, environmental sound and silence were characterized by a rule based method. In order to improve the classification results, audio features like high zero crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), spectrum flux (SF), linear spectral pair divergence distance, band periodicity (BP) and noise frame ratio (NFR) were selected.

Use of Deep Neural Network (DNN) for the audio classification can be seen as an emerging research area. Zvi Kons and Olith Tolendo-Ronen have proposed an approach for the audio classification through DNN [29]. Audio event classification for outdoor events has been carried out using the designed deep neural network structure. Outdoor events like crowd noise, applause, noise of vehicles, and music have been used in the classification. According to the researchers' opinion, improvement of the audio features used for the classifier would lead to more precise results.

R.Kotsakis, G.Kalliris and C.Dimoulas have investigated on various audio pattern classifiers in broadcast-audio semantic analysis [30]. Researchers have employed a supervised learning approach with feature extraction and feature selection procedures. They have implemented different multi class classification and hierarchical schemes in the discrimination of radio broadcast content categories and were successful in efficient speaker recognition and speech-music discrimination with 90% positive recognition rate.

By examining most of the existing literature we can identify that there are many onset detection approaches which were basically tested and evaluated under the music domain. And also, comparatively there are less number of researches has been conducted in segmenting the radio broadcast context. Even though there are many approaches to find music over speech, in almost all the approaches they have addressed the instrumental work and not the songs with a human voice. Most of the approaches were tested on specific datasets which were free from noise. The proposed approach will target on identifying the relevant features for the classification of broadcast context into pre-identified categories and from the obtained classification results the onsets were detected for the content extraction.

PROPOSED APPROACH

The major component of this research is finding an onset detection technique which facilitates the onset detection in radio broadcast context for the separation of content types. The proposed onset detection technique follows a classification approach in predicting the onsets in a radio broadcast context. As the initial step, the audio features in a broadcast context were examined and a classification model was designed to get promising accuracies for the content labelling. Semantic rules which were observed in a radio broadcast context are applied in refining the classification results. Finally, the onsets were predicted according to the content changes in the refined classification results.

The evaluation model was based on the ground truth data which is being annotated by a human user according to his/her insight on the broadcast content types. Figure 1 illustrates the high-level diagram of the proposed approach.

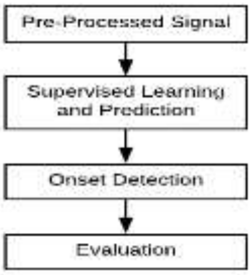


Figure 1: Proposed Methodology

Design Considerations

As an initial approach to analyze the radio broadcast context; Sri Lanka's FM broadcast content is analyzed prior to this research and the model will be trained according to the Sri Lankan broadcast context. For the preliminary stage of model construction and evaluation, one broadcast channel is being used. The used channel is known as 'SLBC-Commercial Service'. Since the FM radio broadcast context is very dynamic the scope is limited up to classification of four different content categories as News (i.e. pure news bulleting without any background sound effects), Songs, Radio Commercials and Other pure human voice content (i.e. human voice prominent parts which includes all the phone conversations, radio discussions, radio dramas with voice only parts etc.). A limited length (maximum 1 hour) of the audio streams will be used in the testing process of the model. And also, the model will be trained with the aforementioned content categories. All the recordings for the model will be taken in the format of '.wav' with the sample rate of 22050. Jingles (i.e. the small sound effects played in between different broadcast content) are avoided in the initial experimental process and they also considered as radio commercials. When annotating the dataset the assumption that offset of one content category will be the onset of the other content has been taken into consideration. No silence removals will be done for the audio streams in the pre-processing stage.

Design Overview

According to the proposed methodology, there are few main state-of-art approaches. They are as audio pre-processing, audio windowing, feature extraction and selection, supervised learning model building and training according to the pre-identified categories, evaluation of predicted content categories and finally the onset detection and post-processing. Figure 2 illustrates the design overview of the proposed approach.

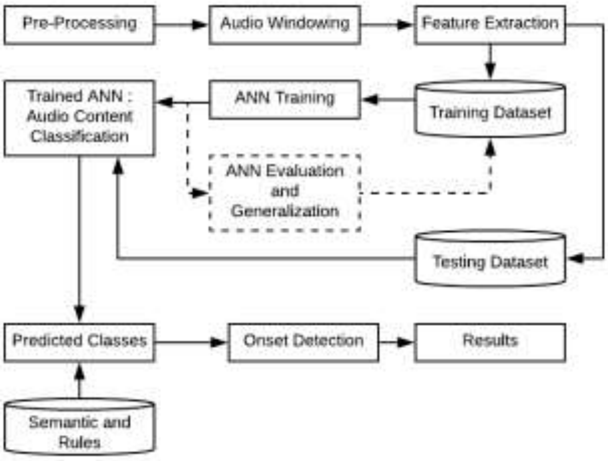


Figure 2: Design Overview

Preprocessing

In the preprocessing stage, the acquired signal is converted to the .wav format if it is not in the.wav format. Since most of the recordings are in stereo format they are being summed to a single channel (i.e. to a mono channel). The sample rate of each and every acquired audio is converted to 22050 Hz to make the consistency for the next stages.

Auto Windowing

Acquired audio signals are then subjected to time-windowing to serve short-term, non-stationary, signal processing. Use of non- overlapping orthogonal windows was selected as the simplest and the less computationally demanding approach. Various window sizes were tested to serve a better audio detection. Smoothing windows (i.e. Hamming, Hanning etc.) were avoided to accelerate the sharp changes and abrupt event detection. By the considering the fact that minimum duration for a radio broadcast event is not less than 1-3s [30], the window size of 2.5s [26, 28] was empirically selected as a good promise for the fine audio detection and classification.

Feature Extraction

In order to differentiate the pre-identified categories from a given radio broadcast stream, it is necessary to identify the features which can distinguish them from one another. For that feature extraction is very much essential. The selected features were successful in the classification of voice over music/song. Python ‘Librosa’ library is used in feature extraction process.

Table 1 illustrates the composition of the extracted features.

TABLE 1: COMPOSITION OF EXTRACTED FEATURES

Feature	Number of values per frame
Chroma Features	12
MFCC Features	13
RMS Feature	1
Spectral Centroid	1
Spectral Contrast	7
Spectral Roll-off	1
Zero Crossing Rate	1
Onset Strength	1
Tempo	1
Total number of features	38

Ground Truth Data Construction

In order to make the ground truth dataset manual annotation of data was required. Sample recordings from the aforementioned FM radio broadcast were acquired. Acquired audio clips were subjected to the previously mentioned pre-processing and feature extraction processes. Time which the content change happens was marked by listening to the clips carefully. Software named ‘Audacity’ is used to listen and get the time value of the content change points in the acquired samples. The feature vectors of the ground truth dataset were annotated accordingly with the content change time values. Annotated class labels were as follows.

News’ - News frames
‘Song’ - Song frames
‘Advert’ - Radio commercials / advertisements
‘Voice’ - Voice prominent content in radio discussions, radio interviews, phone calls, radio programs etc.

Figure 3 illustrates the feature vector annotation methodology. For an instance suppose there is a content change at the time value 29.757s. Figure 3 shows the onset location with respect to the frames. According to Figure 3, it is between 27.50s and 30.00s. That means it belongs to the F3 frame. Therefore, new content C2 will start from frame F3.

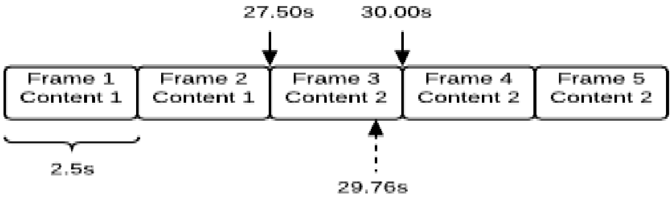


Figure 3: Feature vector annotation methodology

Supervised Learning Model

Since the proposed approach is based on the supervised learning model it is necessary to train a neural network which is capable of predicting the audio content classes with some high accuracy rates. The Neural network model in ‘Keras’ library is used in constructing the multi-layer perceptron (supervised learning) model. The model comprises with the input layer with 38 input nodes, 6 hidden layers with 200 neurons each and one output layer with 4 output neurons. The network structure is designed by the trial and error method. Rectified linear activation function (ReLU) was used in every layer except the output layer. Softmax activation function was used in the output layer. The neural network was fine-tuned with the ground truth data to get higher results for the test cases.

Semantics and Rules

Some set of rules have been proposed in-order to refine the prediction results of the ANN. Rules were generated according to the behavior of the selected content types of the FM broadcast stream. Following are the derived rules.

a) A single song frame cannot appear in between two news frames (i.e. a scenario like ‘News’ ‘Song’ ‘News’ cannot appear). They should be transformed into ‘News’ frames.

b) A single advertisement frame cannot appear in between two news frames (i.e. a scenario like ‘News’ ‘Advert’ ‘News’ cannot appear). They should be transformed into ‘News’ frames.

c) If there’s a series of continuous song frames appearing at the middle of the prediction results, then the time duration of that block should be at least 40s. That means there should be at least 16 blocks of continuous song frames if that to be considered as a song.

d) Song frames can only be refined to advertisement frames.

e) If there is a series of song frames and with some non-song frames at the middle and if the whole series can add up to at least to 40s then the misclassified frames can be replaced with song frames.

The prediction results are being refined using these rules to enhance the accuracies of the prediction. In the first phase of refinement the ‘News’ frames are being refined. In the second phase, ‘Song’ frames are refined. If the difference between two detected song frames is less than 5 then the intermediate content is transformed into song frames.

In the third phase, the ‘Song’ frames are refined according to the observation that a song should last at least 40s (16 frames) of duration. If not, the song frames are refined to the advertisement/radio commercials frames (‘Advert’ frames). Refined results are carried forward to onset detection.

Onset Detection

Onsets were identified by analyzing the refined results of the classification stage. The results were considered in a sequential manner and the places which the content change experiences were considered as onsets. Onsets were detected with an error rate of 2.5s (frame length). Figure 4 illustrates the onset detection methodology followed by the proposed approach.

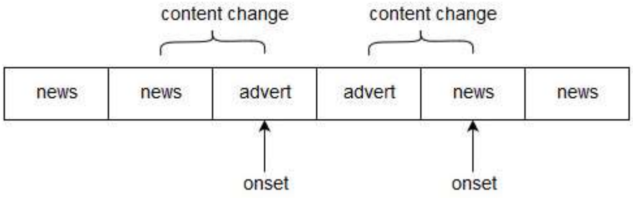


Figure 4: Onset detection methodology

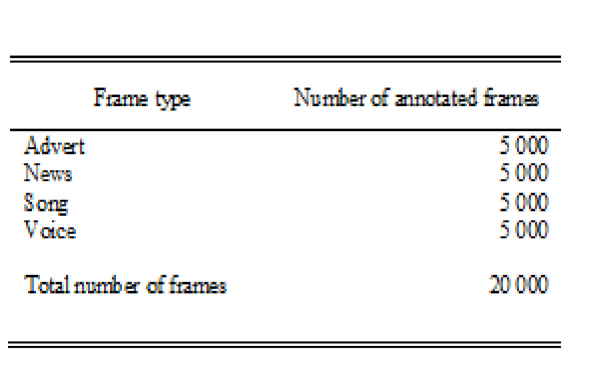
Evaluation Model

The evaluation process of the model is based on ground truth data. Radio broadcast recordings of the selected Sri Lankan broadcast channel (“SLBC- Commercial Service”) were taken and they were manually annotated. Accuracies for the test results were obtained with respect to confusion matrices. The statistical values obtained from the classification results before refinement and after refinement were used in the evaluation of the classifier. Similarly, a confusion matrix for the onsets was obtained with respect to the ground truth annotations.

Training Dataset

The training dataset comprises of the total of 20,000 manually annotated feature vectors generated from the FM broadcast recordings of the Sri Lankan broadcast channel ‘SLBC Welanda Sewaya’. An equal number of frames for each content category were chosen to reduce the biases of the classifier for a particular content type. Table 2 illustrates the composition of the training dataset.

TABLE 2: COMPOSITION OF THE TRAINING DATASET



Frame type	Number of annotated frames
Advert	5 000
News	5 000
Song	5 000
Voice	5 000
Total number of frames	20 000

Testing Dataset

Testing set contains two recordings of the length of 1 hour from the broadcast channel “SLBC Welanda Sewaya”. Both recording in the testing set contains all the defined content categories. Table 3 and Table 4 contain the compositions of the testing dataset I testing dataset II respectively.

TABLE 3: COMPOSITION OF THE TESTING DATASET I

Frame type	Number of annotated frames
Advert	460
News	295
Song	544
Voice	140
Total number of frames	1 439

TABLE 4: COMPOSITION OF THE TESTING DATASET II

Frame type	Number of annotated frames
Advert	135
News	10
Song	440
Voice	855
Total number of frames	1 440

EXPERIMENTS AND RESULTS

Experiments for the classifier accuracies and the onset detection accuracies were conducted using the test datasets mentioned in section 3.E. Each experiment value was obtained by repeating the experiment 3 times for the selected set of values. Statistical measures of the average values of the testing results are taken to obtain the correct accuracies for the classification results and onset detection results. Feature selection and neural network validation experiments were carried out using the training dataset illustrated section 3.D.

Feature Selection and Neural Network Validation

The proposed classification model is totally based on the audio features. Feature selection methods were employed in order to select the best set of features for the classifier. All the aforementioned features (i.e. 38 features mentioned in the section 3.B.3) were ranked by using the feature selection algorithms named InfoGainAttributeEval and OneRAttributeEval [30]. A software tool named 'Weka' has been used in feature ranking process. Table 5 illustrates the feature ranking results from the aforementioned algorithms.

TABLE 5: FEATURE RANKING RESULTS

Rank	InfoGainAttributeEval	OneRAttributeEval
1	Spectral Contrast - 4	Spectral Contrast -3
2	Spectral Contrast - 3	Spectral Contrast – 4
3	Spectral Contrast - 2	Spectral Centroid
4	MFCC - 9	Spectral Contrast - 2
5	Spectral Roll-off	MFCC – 2
6	Spectral Centroid	MFCC – 9
7	MFCC – 2	Spectral Contrast – 5
8	Spectral Contrast – 5	Spectral Roll-off
9	MFCC – 7	Spectral Contrast -1
10	MFCC – 8	MFCC - 8
11	MFCC – 13	MFCC – 4
12	MFCC – 4	MFCC - 7
13	MFCC – 3	Zero Crossing Rate
14	Spectral Contrast – 1	MFCC – 13
15	Zero Crossing Rate	Spectral Contrast – 7
16	Spectral Contrast – 7	MFCC -3
17	MFCC – 10	MFCC -10
18	Onset Strength	Chroma Feature – 6
19	Chroma Feature – 6	Onset Strength
20	Chroma Feature – 5	Chroma Feature – 9
21	Chroma Feature – 7	Spectral Contrast – 6
22	MFCC – 6	Chroma Feature – 5
23	Chroma Feature – 9	Chroma Feature – 4
24	Chroma Feature – 10	Chroma Feature – 2
25	Chroma Feature - 2	Chroma Feature – 7
26	Spectral Contrast – 6	MFCC – 1
27	Chroma Feature - 4	MFCC – 6
28	Root Mean Square	Chroma Feature – 10
29	Tempo	Chroma Feature – 8
30	Chroma Feature – 8	Root Mean Square
31	Chroma Feature – 12	Chroma Feature - 12
32	Chroma Feature – 1	MFCC – 5
33	MFCC – 1	Chroma Feature – 1
34	Chroma Feature – 11	MFCC -12
35	MFCC - 12	Chroma Feature -3

TABLE 6: NEURAL NETWORK PREDICTION RESULTS

		Number of hidden layers			
38 Features		3	4	5	6
200 neurons in each hidden layer	News	75.690	78.002	66.835	55.107
	Advert	60.408	60.844	67.631	68.996
	Song	56.924	58.946	50.689	77.145
	Voice	79.048	77.857	80.179	80.952
300 neurons in each hidden layer	News	64.310	73.625	63.524	52.525
	Advert	65.884	67.467	72.926	65.211
	Song	61.857	58.885	61.826	69.914
	Voice	80.714	80.000	66.667	74.762
30 Features		3	4	5	6
200 neurons in each hidden layer	News	78.227	67.677	57.351	44.108
	Advert	57.205	63.464	73.872	70.524
	Song	32.843	59.681	30.637	60.417
	Voice	74.762	79.524	77.143	75.714
300 neurons in each hidden layer	News	74.523	55.219	45.230	43.098
	Advert	58.661	75.619	80.349	70.160
	Song	58.333	65.686	29.044	81.985
	Voice	76.190	77.857	80.714	69.048
20 Features		3	4	5	6
200 neurons in each hidden layer	News	45.5667	51.403	31.987	26.936
	Advert	71.397	70.670	79.913	75.400
	Song	72.794	52.757	63.419	43.137
	Voice	46.429	65	66.429	78.810
300 neurons in each hidden layer	News	55.668	48.373	28.620	15.937
	Advert	68.122	74.381	79.767	73.071
	Song	31.679	55.821	57.169	74.877
	Voice	76.190	70.714	74.762	70.952

From both the aforementioned algorithms the order of the ranked features is almost the same except for few cases. Therefore, the first 20, first 30 and all 38 features from the InfoGainAttributeEval ranking are chosen to decide the best suitable neural network.

Neural network was trained according to the trial and error basis. A different number of hidden layers with the neuron count of 200 [29] and 300 each with the activation function 'ReLu' was selected for the experiments. The network is trained for 300 epochs. Each network model was tested with three trials and the averaged values were taken as the accuracies. Table 6 discusses the average prediction results for the used test dataset. Highest average accuracies obtained at each network model is highlighted.

As in table 6, the most promising results were obtained in the network model with 38 features, 6 hidden layers with 200 neurons in each hidden layer. Therefore the final network can be considered as a neural network with 6 hidden layers with 200 neurons in each layer with the activation function 'ReLu'.

10-fold cross validation for the training dataset

Once the network structure was decided, 10-fold cross validation was carried out for the training set to get an averaged prediction accuracy of the designed neural network. Table 7 illustrates the prediction results obtained after 10-fold cross validation of the training dataset. According to the averaged accuracy value we can say that the neural network model is capable in predicting the selected content categories with an accuracy of 58%.

TABLE 7: 10-FOLD CROSS VALIDATION ACCURACIES

Iteration	Accuracy value (%)
1	48.35
2	63.60
3	61.10
4	57.20
5	58.10
6	54.20
7	61.10
8	58.80
9	62.50
10	57.35
Average accuracy	58.23 (+/- 4.24)

Classification results with refinement rules

Test datasets mentioned in section 3.E were used in the evaluation of the model. For each test dataset, the model is executed for four times and the average results had been taken in calculating the accuracy measures. At first the average initial prediction accuracies were taken. Table 8 illustrates the average accuracies of the initial prediction results of test dataset I and test dataset II.

Model produced promising results for the song and advertisement/radio commercial classification and lower results for the news classification. The lower accuracies obtained for the test dataset II impacted a lot for the overall accuracy of news frame classification. Low composition of news frames in test dataset II paved the way for it to get a very low accuracy rate for the news classification.

Next the prediction results were transferred to the refinement phase I where the refinement of news was carried out. Table 9 displays the average accuracies obtained after news refinement phase. By applying refinement rules for news the classification accuracies of test dataset I show some improvement while test dataset II show no improvement. The major reason for not showing such improvement is that it's less composition for the news. Finally, the results obtained from refinement phase I were transferred to the refinement phase II and III where the song refinement was carried out. Result set obtained at the end of song refinement was taken as the final classification results of the model. Table 10 shows the average classification results after song refinement phase.

After song refinement phase, in both test datasets the prediction accuracies for songs and advertisement/ radio commercials have improved. That is because, in Sri Lankan broadcast context there is a high tendency for a radio commercials get misclassified with the songs. With the final refinement phase those misclassifications have been corrected to a considerable extent. And also, there can be instances where the news and voice frames get misclassified with the radio commercial frames. The major reason for that is the dynamic nature of the radio commercials in the Sri Lankan broadcast context.

The annotation of these frames depends on the human knowledge and the human has the capabilities of identifying the broadcast commercials over songs/music, radio talks and other voice prominent segments. But the proposed classifier is solely based on the audio features and has no knowledge base in distinguishing one content category over the other. Therefore, there is a high possibility for the misclassifications of selected content categories with the radio commercials.

According to the classification results it is clear that a very minute percentage of news and voice have misclassified as songs. So we can say that the proposed classification model works well with identifying song content over voice and news content.

TABLE 8: INITIAL CLASSIFICATION ACCURACIES OF THE MODEL

	Test dataset 1	Test dataset 2	Average (%)
News	54.040	17.500	35.770
Advert	69.651	70.185	69.918
Song	82.537	46.989	64.763
Voice	70.000	47.043	58.522

TABLE 9: CLASSIFICATION ACCURACIES OF THE MODEL AFTER NEWS REFINEMENT

	Test dataset 1	Test dataset 2	Average (%)
News	63.552	17.500	40.526
Advert	68.559	70.000	69.280
Song	82.537	46.989	64.763
Voice	70.000	47.043	58.522

TABLE 10: CLASSIFICATION ACCURACIES OF THE MODEL AFTER SONG REFINEMENT

	Test dataset 1	Test dataset 2	Average (%)
News	63.552	17.500	40.526
Advert	68.996	82.963	75.979
Song	93.107	55.852	74.479
Voice	70.000	47.043	58.522

TABLE 11: ONE TO ONE ONSET DETECTION ACCURACY MEASURES

	Test dataset 1	Test dataset 2	Average (%)
Precision	9.379	1.952	5.666
Recall	39.375	55.769	47.572
Specificity	89.332	74.246	81.789
False Negative Rate	60.625	44.231	52.473
Accuracy	87.94302	74.079	81.011

TABLE 12: ONSET DETECTION WITH (+/-) 2.5S ERROR RATE ACCURACY MEASURES

	Test dataset 1	Test dataset 2	Average (%)
Precision	13.497	2.621	8.059
Recall	56.250	75.000	65.625
Specificity	89.814	74.422	82.118
False Negative Rate	43.750	25.000	34.375
Accuracy	88.881	74.427	81.654

Onset detection results

Once the final classification results are obtained, onset detection is carried out. Whenever there is a content change in the classification results the changed frame is considered as an onset. At first the one to one onset detection is carried out. Table 11 illustrates the accuracy measures for one to one onset detection.

According to the above shown numerical figures precision score shows comparatively low results. The major reason for that is our proposed model predicts onsets which aren't actually the true onsets. That means the false positive rate of the model is comparatively high. Other accuracy measures are at a considerable level.

If the predicted onsets were considered with an error rate of (+/-) 2.5s from either side of the frame then the accuracy levels of correct onset predictions were at higher levels. Table 12 illustrates the accuracy measures when onsets were identified with the aforementioned error rate.

When the onsets were detected with an error rate of (+/-) 2.5s the overall accuracy measures showed an improvement in their values when compared with the one to one onset measures.

DISCUSSION

According to the observed classification results, it is visible that the 'News' frames do not get misclassified as 'Song' frames in a majority and 'Song' frames do not get misclassified as 'News' frames in the majority. The constructed model is successful in differentiating the 'News' frames from 'Song' frames. There were so many situations where 'Song' frames got misclassified as 'Advert' and vice versa. That is basically because of the dynamic nature of the radio commercials (i.e. 'Advert' frames). A more precise technique which aids in identifying the structural similarities of these selected content categories [31, 32] would help in improving the classification results. And also building a knowledge base on top of the classification model would do better in the classification of the content categories.

When examining the onset detections it is visible that in many instances the content change from 'Voice' to 'Song' is identified with high accuracies. High accuracies can be obtained if the onset change is observed with (+/-) 2.5s of error rate. At many instances, one to one onset detection does not happen frequently. It is clearly visible that the predicted onsets have moved one frame up or down from the actual onset values. Therefore, in the detection of the onsets of the radio broadcast content, the onsets with the (+/-) 2.5s error rate should be considered.

To reduce the false positives of the predicted onsets the classification model should further tuned to get good classification results for the identified content categories.

CONCLUSION AND FUTURE WORK

The main aim of this research is to assist a deep automated analysis for the application levels of the radio broadcast context monitoring process. In aid of this task, an initial approach was taken in identifying the content categories in a broadcast context and identifying the onsets of those identified content categories in a broadcast stream. The proposed approach focused on classifying the broadcast content into news, advertisements/radio commercials, songs and other human voice content which includes radio discussions, radio dramas, phone interviews etc. In the classification model, the main target was to identify a set of features and design a good network structure which helps in classifying the above-identified content categories with promising accuracies. The proposed classification model was successful in identifying news, advertisements/radio commercials, songs and other voice content with accuracies of 41%, 76% 75% and 59% respectively. According to the classification results which were discussed in Section 5.3.4, the proposed classifier is successful in identifying the song frames over news frames. The dynamic nature of the Sri Lankan radio broadcast context affected for the misclassification rates of songs with advertisements/radio commercials and other voice content.

The proposed onset detection methodology is solely based on the above classification results. Onsets were characterized by the positions where the classification results experience a content change. The proposed model was successful in detecting the onsets of (+/-) 2.5s of error rate with an accuracy of 82%. More improvements to the classification model will help in reducing the false positive rate for onset detection of the model.

As discussed, the classifier output has a huge weight in the accuracies of the detected content categories as well as onsets. More improvements to the classifier should be done in future works. For an instance, a classifier like a CNN can be used for the classification process. More experiments should be done in the identification of more features which can give promising results for the classification of broadcast content categories. A more comprehensive training set which can detect the content categories in many broadcast channels should be developed in the generalization of this approach. Well-mannered annotation procedure should be followed in making the training dataset. A knowledge base can be built on top of the classifier in advancing the accuracy results. And also, the experiments to identify the structural similarities of the content categories can be performed in improving the classification results. Some metadata like time can be used in the refinement process of the classification results and some more semantic rules should be identified when generalizing this approach to other broadcast channels.

REFERENCES

[1] “Celebrating Radio: Statistics,” Celebrating Radio: Statistics | World Radio Day 2015. [Online]. Available:http://www.diamundialradio.org/2015/en/content/celebrating-radio-statistics.html. [Accessed: 26-Dec-2017].

[2] E. Nishan, W. Senevirathna, and K. L Jayaratne, “A highly robust audio monitoring system for radio broadcasting, Proceedings of sixth Annual International Conference on Computer Games, Multimedia and Allied Technology” GSTF Journal on Computing (JoC), vol. 3, no. 2, pp. 87-98, 2013.

[3] N. Senevirathna and K. L Jayaratne, “Automated content based audio monitoring approach for radio broadcasting,” Proceedings of sixth Annual International Conference on Computer Games, Multimedia and Allied Technology (CGAT 2013), Singapore, pp. 110–118, CGAT, 2013.

[4] E. N. W. Senevirathna and K. L. Jayaratne, “Audio music monitoring: Analyzing current techniques for song recognition and identification,” GSTF Journal on Computing (JoC), vol. 4, no. 3, pp. 23-34, 2015.

[5] E. D. N.W. Senevirathna and K. L Jayaratne, “Automated Audio Monitoring Approach for Radio Broadcasting in Sri Lanka,” Proceedings of International Conference on Advances in ICT for Emerging Regions (ICTer 2017), Sri Lanka, pp. 92–98, 2017.

[6] E.D.N.W. Senevirathna and Lakshman Jayaratne, “Radio Broadcast Monitoring to Ensure Copyright Ownership”. International Journal on Advances in ICT for Emerging Regions (ICTer), 11(1), 2018.

[7] Dhanith Chaturanga and Lakshman Jayaratne, “Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches”. International Journal of Computing (JOC) by Global Science and Technology Forum (GSTF), 3(2):137-148, 2013.

[8] Dhanith Chaturanga and Lakshman Jayaratne, “ “Musical Genre Classification Using Ensemble of Classifiers”, Proceedings of fourth International Conference on Computational Intelligence, Modeling and Simulation (CIMSim 2012), Kuantan, Malaysia, 2012.

[9] Rajitha Amarasinghe and Lakshman Jayaratne, “Supervised Learning Approach for Singer Identification in Sri Lankan Music”, European Journal of Computer Science and Information Technology (EJCSIT) by European Centre for Research Training and Development UK, 4(6):1-14, 2016.

[10] Rajitha Peiris and Lakshman Jayaratne, “Musical Genre Classification of Recorded Songs Based on Music Structure Similarity”, European Journal of Computer Science and Information Technology (EJCSIT) by European Centre for Research Training and Development UK, 4(5):70-88, 2016.

[11] Tharika Madurapperuma, Gothami Abayawickrama, Nesara Dissanayake, Viraj B. Wijesuriya and K. L. Jayaratne, “Highly Efficient and Robust Audio Identification and Analytics System to Secure Royalty Payments for Song Artists”, Proceedings of

IEEE International Conference on Advances in ICT for Emerging Regions (ICTer 2017), Sri Lanka, 149-157, 2017.

[12] Rajitha Peiris and Lakshman Jayaratne, "Supervised Learning Approach for Classification of Sri Lankan Music based on Music Structure Similarity", Proceedings of ninth Annual International Conference on Computer Games, Multimedia and Allied Technology (CGAT 2016), Singapore, 84-90, 2016.

[13] M. G. Viraj Lakshitha and K. L. Jayaratne, "Melody Analysis for Prediction of the Emotion Conveyed by Sinhala Songs", Proceedings of IEEE International Conference on Information and Automation for Sustainability (ICIAfS 2016), Sri Lanka, 2016.

[14] "DigitalRadioTracker.com - Your Global Radio Airplay Monitoring Solution....," DigitalRadioTracker.com - Your Global Radio Airplay Monitoring Solution.... [Online]. Available: https://www.digitalradiotracker.com/. [Accessed: 26-Dec-2017].

[15] "Broadcast Monitoring - Music, Ads on Radio, TV | Media Monitoring," ACRCloud. [Online]. Available: https://www.acrcloud.com/broadcast-monitoring. [Accessed: 26-Dec-2017].

[16] "World's most efficient mobile Automatic Content Recognition," Beatgrid Media, 23-Nov-2017. [Online]. Available: http://www.beatgridmedia.com/. [Accessed: 26-Dec-2017].

[17] J. Schluter and S. Bock, "Improved musical onset detection with Convolutional Neural Networks," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[18] O. Mubarak, E. Ambikairajah, and J. Epps, "Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources," Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005

[19] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035–1047, 2005.

[20] X. Rodet and F. Jaillet, "Detection and modeling of fast attack tran- sients," in Proc. Int. Computer Music Conf., Havana, Cuba, 2001, pp. 30–33

[21] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "A Combined Phase And Amplitude Based Approach To Onset Detection For Audio Segmentation," Digital Media Processing for Multimedia Interactive Services, 2003.

[22] J. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain," IEEE Signal Processing Letters, vol. 11, no. 6, pp. 553–556, 2004.

[23] J. Bello and M. Sandler, “Phase-based note onset detection for music signals,” 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 03).

[24] Zhou, R., & Reiss, J. (2007). Music onset detection combining energy-based and pitch-based approaches. In *Music information retrieval evaluation exchange*
[25] Böck, S., Arzt, A., Krebs, F., Schedl, M. (2012). Online real-time onset detection with recurrent neural networks. In *Proceedings of the 15th international conference on digital audio effects*.

[26] J. Saunders, “Real-time discrimination of broadcast speech/music,” 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings.

[27] S. G. Koolagudi, S. Sridhar, N. Elango, K. Kumar, and F. Afroz, “Advertisement detection in commercial radio channels,” 2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS), 2015.

[28] L. Lu, H.-J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, pp. 504–516, 2002.

[29] Z. Kons, O. Toledo-Ronen, "Audio event classification using deep neural networks", *Proceedings of INTERSPEECH*, pp. 1482-1486, 2013.

[30] R. Kotsakis, G. Kalliris, and C. Dimoulas, “Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification,” Speech Communication, vol. 54, no. 6, pp. 743–762, 2012.

[31] J. Foote, M. Cooper, U. Nam, "Audio retrieval by rhythmic similarity", *Proc. Int. Conf. Music Information Retrieval*, 2002.

[32] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532).