Classification of Voice Content in the Context of Public Radio Broadcasting

G.A.G.S.Karunarathna[#1], K.L.Jayaratne [#2], P.V.K.G.Gunawardana[#3]

***Abstract*— With the rapid development of mass media technology, content classification of radio broadcasting has emerged as a major research area facilitating the automation of radio broadcasting monitoring process. This research focuses on the voice dominant content classification of radio broadcasting by employing a multi-class Support Vector Machine (SVM) in order to automate monitoring of radio broadcasting in Sri Lanka. This study investigates the performance of "One Vs. One" and "One Vs. All" methods known to be two conventional ways to build a multi-class SVM. These two multi-class SVM models are trained to classify five voice dominant classes as news, conversations, and advertisements without jingles, radio drama and religious programs.**

One of the substantial measures in creating such a classification is selection of the optimal feature sets. For that purpose, time domain features, frequency domain features, cepstral features, and chroma features are manually analyzed for each binary SVM classifier independently. Two multi-class SVM models are trained based on the selected features and the "One Vs. One" model was able to better classify the recordings with an 85% accuracy compared to 83% accuracy achieved by "One Vs. All" model. Further, the results revealed the importance of careful feature selection in order to achieve higher classification accuracies.

Keywords*— *Audio monitoring, Audio classification, Radio Broadcasting, Audio feature analysis, Support Vector Machines

I. INTRODUCTION

The radio is a dynamic and amiable communication device to people for many decades since its invention. Unlike other communications devices such as computers and smartphones, anyone can easily use the radio without an age bracket. According to the statistic in 2015, more than half of the population use the radio while they are driving, women tend to listen to the radio while they are cooking, and most of the people listen to the radio even at their workplaces [1]. Therefore, unlike the rest of the communication mediums radio plays an important role in sharing information.

In radio transmission, radio station and the listeners are the two endpoints. Radio stations broadcast unidirectional wireless signals over space to the multitudes of individual listeners with radio receivers. Radio stations broadcast a sequence of content categories such as songs, advertisements, news, interviews, conversations, and radio dramas. The number of listeners of a radio channel always relies on the programs that the radio stations are broadcasting. Thus, in order to grasp the audience to a program, the program should be performed well, and fit into the audience. Hence, for the purpose of measuring the performance of a broadcast program, radio stations need to monitor the broadcasting content regularly. Broadcast content monitoring helps to verify when and where the broadcast content placed, protect copyrights by knowing precisely how the content is being used, and measure performance across other broadcast channels [2, 3, 4, 5, 6]. Therefore, broadcast content monitoring is a necessary thing for radio stations.

Furthermore, the stakeholders of radio channels also need to monitor the broadcast content for different purposes such as business, political and legal needs [7, 8]. Authorized people in mass media and information corporations need to track the FM channels regularly to ensure whether the broadcasting contents adhere to the rules and regulations and has a diversity of available programs. Singers and composers need monitoring of songs to claim their rights [9, 10, 11, 12, 13]. Advertising agents are keen on the frequencies of the advertisement broadcasting that have a huge impact on their corporate income. Political parties also keep an alert on their name referencing in the radio broadcasting content, especially on the news and political discussions. Therefore, it is reasonable to state that many stakeholders are interested in monitoring the radio broadcasting content for a wide range of reasons.

In the monitoring of radio broadcasting, both manual monitoring and automated monitoring are used. In manual monitoring techniques such as having an observer to listen to the radio content, reading the attached meta-data, asking for the broadcast report from broadcast stations are used. These techniques become inefficient and resource intense when the amount of content that needed to be monitored is high. In automated radio monitoring processes, well-trained machine monitors the radio content effectively and efficiently than the manual monitoring process. Most of the time, developed countries use automated radio monitoring process [14]. As a developing country, Sri Lanka has not yet established such a technology to monitor radio broadcasts. Since there is a large number of radio channels in Sri Lanka, manual monitoring is not practical. Unfortunately, the mechanisms used in developed countries cannot be substituted for FM channels in Sri Lanka, due to the differences in languages and pronunciations. Hence, it is imperative for the Sri Lankan Broadcasting context to have an automatic radio monitoring program.

As the initial step to build an automated monitoring process, identifying different content classes (i.e. songs, advertisements, news, interviews, conversations, and radio dramas) in radio broadcasting content is essential. When analyzing the current situation of the above-mentioned problem, classifying broadcast context for onset detection is recognized as the closest research work [15]. Onset detection

Manuscript received on 22nd Feb. 2019. Recommended by Dr. D.N. Ranasinghe on 30th Dec. 2019.

G.A.G.S.Karunarathna, K.L.Jayaratne, P.V.K.G.Gunawardana are from the University of Colombo School of Computing, Sri Lanka.
(gothamikarunarathna@gmail.com, klj@ucsc.cmb.ac.lk, kgg@ucsc.cmb.ac.lk).

is the mechanism which is used to identify the places where the content changes are happening in a musical note or other sound streams. The researchers have proposed a unified methodology to automate radio broadcasting monitoring which detects onsets of radio broadcasting context with the assist of the classification of the broadcasting content. The proposed mechanism distinguishes songs, commercial advertisements with jingles, news, and other contents in a radio stream. However, the issue with this unified method is, it is unable to identify voice dominant content classes in the broadcasting context. Hence, the classification of different voice dominant contents in the radio broadcasting stream such as news, advertisements without jingles, conversations, radio dramas, and religious programs are identified as the knowledge gap in between the requirement and existing solutions. Therefore, in the context of radio broadcasting, this research proposes a methodology to classify voice dominant contents in radio broadcasting.

II. RELATED WORK

As audio classification has emerged as a demanding research area, a considerable amount of related works can be examined. Based on the approaches that the researchers followed in audio classification, these works can be divided into two parts as algorithmic approaches and machine learning approaches.

A. *Algorithmic Approaches*

Lie Lu, Stan Z. Li and Hong-Jiang Zhang [21] proposed an algorithm which is able to classify an audio stream into speech, music, environmental sounds and silence. Silence detection is performed based on short-time energy and zero-crossing rate (ZCR) in a one-second window. Linear Spectral Pairs (LSP) distance analysis is used to apply refinements over the proposed algorithm. The result of this research has some misclassifications between music and environment sound due to the overlaps in the distribution of the features.

An algorithm for discriminating speech from music on broadcast FM radio based on ZCR of the time domain waveform is proposed by John Saunders [16]. This technique emphasized the characteristics of speech such as limited bandwidth, alternate voiced and unvoiced sections, energy contour between high and low levels are well capable of separating speech from music.

Barzilay et al [17] proposed an algorithmic approach for the speaker's role identification in radio broadcasting context. This approach classified anchor, journalist and guest programmer by considering lexical features, features from the surrounding context and the duration features.

Though the algorithmic approaches show promising results, when the number of classes in the classification is increasing the problem becomes non-trivial. Identifying the threshold values to discriminate each class is also difficult. In order to avoid these negatives on the algorithmic approaches, researchers recently moved to machine learning approaches.

B. *Machine Learning Approaches*

The machine learning community has done numerous works the under both supervised learning and unsupervised learning. The learning approaches associated with supervised learning are Neural Networks (NN), Hidden Markov Model (HMM), and Support Vector Machine (SVM). Unsupervised learning approaches are K means and Gaussian Mixture Model (GMM). Since our problem domain refers to the supervised learning approaches, more attention goes to NN, HMM, and SVM.

The most recent and closest work of addressing the same problem is, "Classification of public radio broadcast context for onset detection" conducted by C. Weeratunga et al [15]. In this approach, the onset detection mechanism along with a classification model is proposed to predict four classes (i.e. songs, voice-related segments, news, and radio commercials). A supervised neural network model with 38 extracted features has included in the classification framework. Radio commercials, songs, news, and other voice contents are classified with accuracies of 76%, 75%, 41%, and 59% respectively. In this approach, the output of the onset detection largely depends on the accuracy of the classification. Currently, it has 82% accuracy for onset detection with respect to prior mentioned audio classes in radio broadcasting context. In order to automate the radio broadcasting monitoring process, the existing onset detection method should be improved. Therefore as a further step, the focus should be on the voice dominant content classification in radio broadcasting events.

Another supervised neural network approach has used by Khan et al [22] to classify speech and music. As the classification framework, multilayer perceptron neural network and back-propagation learning algorithms are used. The experimental results have shown an overall accuracy of 96.6%, with 100% accuracy in recognizing music from speech.

In the research done by R. Kotsakis, G. Kalliris, C. Dimoulas [14], various audio pattern classifiers in the broadcast-audio semantic analysis are investigated using radio program-adaptive classification strategies with supervised ANN system. In the evaluation, Kotsakis et al found ANN and KNN classifiers quite effective than tree complex and SMO methods.

Kons et al [23] suggested a Deep Neural Network (DNN) as a solution for classifying four classes as the crowd of people, cars/road noises, applause yelling/cheering, and various kinds of music recorded in outdoor. The overall performance of the DNN classifier achieved the best in most of the classes, except for the music class where the SVM performs better.

Same as Neural Networks approaches, the Hidden Markov Model is also shown high performance in radio broadcasting content classification problems. HMM is used in a radio commercial classification by G.Koolagudi et [19]. As they observed, in some situations where ANN failed (i.e. background music follows an advertisement), HMM performed well. Another work related to HMM has conducted by Yang Liu [18], identify the roles of speakers in radio broadcasting news contents. Well-structured news content is used in this research which highlights the speaker role sequences. Accuracy of 80% is obtained and they found the beginning and the end of the sentences in the voice of the speaker as a good heuristic for role identification.

SVM basically designs for binary classification problems. As extensions, multi-class SVM obtains by compromising set of binary SVM classifiers. There are 3 ways to design multi-class SVMs as One Vs. All, One Vs. One, and DAGSVM [24]. The main advantage of SVM when compared to other machine learning approaches is that SVM performs much better in many cases because it finds the best hyperplane/s that separates all data into different classes, no matter even the dataset is small [25]. Aurino et al [26] have proposed a One-class SVM based approach to detect anomaly events that are considered as abnormal sounds in the environment like a gunshot, screaming and broken glass. The proposed methodology consists of two stages. At the first stage, the researchers introduced a new mechanism called “Majority Voting and Rejection” to classify short time frames into predefined classes. At the second stage, aggregated the results of the first stage into longer time frames and reclassified.

In the work of Bouril, A. et al [27], 3000 phonocardiograms from 9 locations of the body of both adults and children were taken to identify normal and abnormal heart sounds using SVM. Here, 74 features of time and frequency domain were considered. The SVM model was utilized by a Gaussian Kernel where it allows three different classifications; -1 for the normal heartbeat, 0 for ambiguous sounds due to noise and 1 for abnormal heartbeat sound. In this research, a binary SVM is chosen to be effective in normal and abnormal classification. Audio-based event detection in live office environments using optimized MFCC features with SVM model has implemented by Kucukbay et al [28]. Sixteen classes such as alert beeping, clear throat, keyboard and switch on/off sounds were classified. One Vs. All multi-class SVM is used as the classifier. Martin Morato et al [29] conducted a case study on feature sensitivity for audio event classification using One Vs. All multi-class SVM. Same as the above [28], sixteen classes have been differentiated by MFCC features and MFE features using 2.5s frame length with 1s overlaps and 44.1 kHz sampling frequency. Wang, J.et al [30] have used a frame based multi-class SVM classifier to differentiate fifteen audio classes including both male, female voices. A frame-based classifier segmented one audio file into several frame sizes and trained the classifier for each. Even though this method improves accuracy from 13.9%, the pre-processing and training time was considerably high.

Lie Lu, Stan Z. Li and Hong-Jiang Zhang [31] have proposed a method called hierarchical binary support vector machine for employing an audio segmentation and classification. Here the researchers furthermore considered five pre-defined classes as silence, music, background sound, pure speech and non-pure speech including speech over music and speech over the noise. In the evaluation, it has shown the accuracy of the SVM based method is better than the method based on KNN and GMM. But the major disadvantage in this approach is misclassifications of upper levels can be propagated to the classifiers at the lower level.

Since the broadcasting FM channels demand the news content classification of broadcasting context, Vavrek, J. et al [20] also proposed a hierarchical tree to address the news content classification problem. This hierarchical classification strategy is used as a particular feature set for each SVM binary classifier. Therefore, the F-score feature selection algorithm is used to obtain optimal features for each SVM. The drawback of this work is the error of upper levels of the tree were propagated to the bottom levels of the tree. To prevent that, misclassifications of upper levels have not considered.

The work of Zhu, Y., Ming, Z. , Q. Huang [32] is classified six audio classes using clip based SVM method. Here, the researchers classified pure speech, music, silence, environmental sounds, speech with music, and speech with environmental sounds. The key finding of this work is, the researchers found that the performance of SVM shows good results in similar cases than Decision Trees, KNN, and Neural Networks.

The potentials of these approaches vary from problem domain to domain. Based on the research question and past studies in the domain, the following choices were made in order to carry out this study. Since the dataset consists of a set of pre-defined classes, a supervised learning approach is proposed for the classification. Therefore unsupervised classifiers were eliminated. As mentioned before, unique sets of features to discriminate each class from the rest has identified. Hence, if all the features are input to the classification model together, it will reduce the accuracy of the model because of some irrelevant features input to the classification of some classes. Therefore, another facet of this research is input different feature sets to discriminate each class. When considering the ANN approach, it is impossible to provide unique sets of features for each class separately. Moreover, according to the research conducted by C. Weeratunga et al [15], the ANN model is not the best approach to distinguish voice dominant categories such as news. In other hands, HMM was rejected in view of the fact that the sequence of the audio events appearing is not beneficial to our problem. Accordingly, the SVM classification model is selected after considering all aspects.

Since SVM's are originally designed for binary classification, the multi-class SVM builds as a compound of binary SVM classifiers. As we already identified specific features for each class, we can input only the relevant features separately in the case of using a multi-class SVM model because it holds multiple binary SVM models. Accordingly, multi-class SVM is chosen as the most suitable classifier which fits into our problem domain. As it is a composition of several binary SVMs, multi-class SVM can be designed as one of the following methods [24],

- One Vs. One
- One Vs. All
- Dynamic Acyclic Graph SVM (DAGSVM)

One Vs. All constructs N number of binary SVM models where it has N number of classes. Every single binary SVM is trained with all of the data in the one class with positive labels and rest with negative labels. The decision function which has the largest value is taken as the predicted class. One Vs. One constructs N(N-1)/2 number of binary SVM models where each one is trained only for two classes and a class is predicted using the "Max-winning" strategy. Same as One Vs. One, DAGSVM also constructs N(N-1)/2 number of binary SVM models where each one is trained for two classes. These binary SVMs are structured as a top to the bottom hierarchical tree where it has (N-1) number of leave nodes. It starts at the root node, then a binary decision function is evaluated, and it moves to either left or right depending on the output value of the previous node.

Since the DAGSVM is a hierarchical graph, the misclassifications of upper levels in the graph can propagate to lower levels in the graph [33]. This will lead to an erroneous situation. Hence DAGSVM was rejected at the very first step. One Vs. One and One Vs. All both have benefits as well as limitations [24]. It depends on the application domain. Hence this research attempts to obtain the most reliable method by modeling the multi-class SVM in both ways.

III. PROPOSED APPROACH

The proposed approach mainly focuses on the classification of different voice dominant classes in radio broadcasting of Sri Lanka. Since the dataset consists of a set of pre-defined classes (i.e. news, advertisements without jingles, radio dramas, conversations, and religious programs), a supervised learning approach is proposed initially for the classification. SLBC (Sri Lanka Broadcasting Cooperation) audio recordings are used as the dataset in order to represent all the Sri Lankan FM channels. The length of the dataset is 5 hours and 50 minutes and contained both male and female voices. As shown in Figure 1, initially the whole dataset is divided as 60% and 40% for training and evaluation purposes respectively. Again the training dataset is divided into 70% and 30% for training and testing respectively. The number of frames consists of training and the testing dataset is given in Table 1 and Table 2. The length of the frame is 5s.

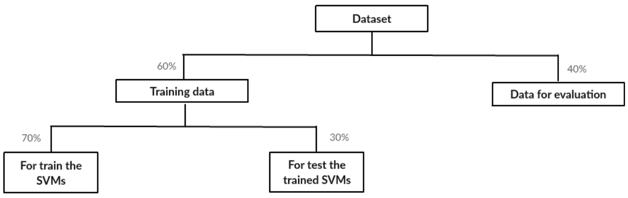


Figure 1: Dataset Partition

Table 2: Number of frames in the training dataset

Class	Number of frames
News	495
Conversations	498
Advertisements	488
Drama	508
Religious Programs	501

Table 2: Number of frames in the testing dataset

Class	Number of frames
News	320
Conversations	325
Advertisements	316
Drama	316
Religious Programs	323

A quantitative interpretation of audio data is required for the analysis to identify the most suitable features for distinguishing each class separately. This research specifically focuses on the analysis of time series and frequency series of audio signals. Figure 2 depicts the design of the proposed approach.

A. *Feature Analysis*

Features are used to capture the measurable information in the dataset. In a classification, identifying the most appropriate features is essential for differentiate one class from another. In order to select the appropriate features, the dataset should be thoroughly analyzed. Here, altogether 34 features used in the most recent and relevant past study [15] are analyzed. These features belong to time domain features, frequency domain features, cepstral features, and chroma features.

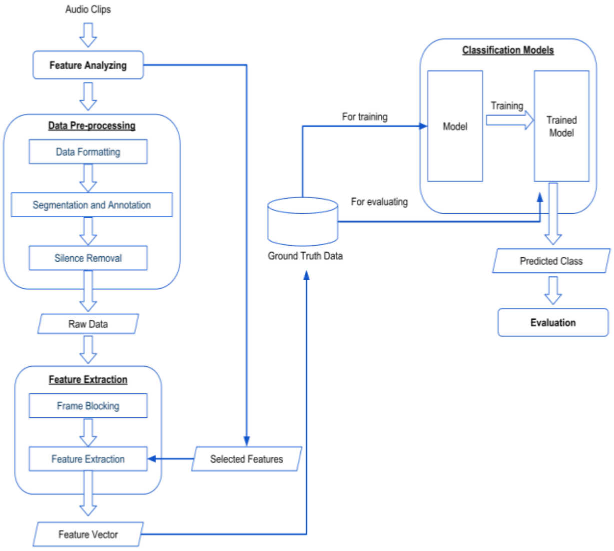


Figure 2: Design Overview

The novelty of the research is that, rather than feeding all the features together, specific sets of features are fed separately into each class. This assists to avoid the input of unnecessary features, reduce dimensions, and make classification faster and more accurate. Since a multi-class SVM holds multiple binary SVMs, one of the advantages of using a multi-class SVM is that it can input specific features for each binary SVM separately.

Since this research compare the performance of two types of multi-class SVM models, the feature selection carried out separately for both multi-class SVM models. As illustrates in Table 3, binary SVM models used to construct multi-class SVM models are trained to classify different class pairs. Therefore, for each binary SVM model, the features are identified by the class pairs that are to be classified.

Table 3: Binary classifiers of two multi-class SVMs

Multi-class SVM	Binary classifiers	Identical classes
One Vs One	SVM 1	News Vs. Advertisements
	SVM 2	News Vs. Conversations
	SVM 3	News Vs. Radio drama
	SVM 4	News Vs. Religious program
	SVM 5	Advertisements Vs. Conversations
	SVM 6	Advertisements Vs. Radio drama
	SVM 7	Advertisements Vs. Religious program

Multi-class SVM	Binary classifiers	Identical classes
	SVM 8	Conversations Vs. Radio drama
	SVM 9	Conversations Vs. Religious program
	SVM 10	Radio drama Vs. Religious program
One V s All	SVM 1	News Vs. others
	SVM 2	Advertisements Vs. others
	SVM 3	Conversations Vs. others
	SVM 4	Radio drama Vs. others
	SVM 5	Religious program Vs. others

1) *Feature Analysis: One Vs. One*

According to Table 3, the best features are analyzed to distinguish ten class pairs by looking at the spectrums. For that purpose, with the class pairs, 10 audio clips are prepared as in Figure 3, where one class is 15 minutes long.

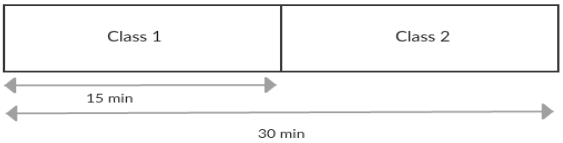


Figure 3: Audio clip structure designed for analyze features of two classes

By observing spectrums of each pair, 24 of the 34 features are selected by eliminating the features that do not show a spectrum discrimination pattern for any class pair. As an example, Figure 4 shows the spectrum of energy feature which is selected to distinguish advertisements and religious programs.

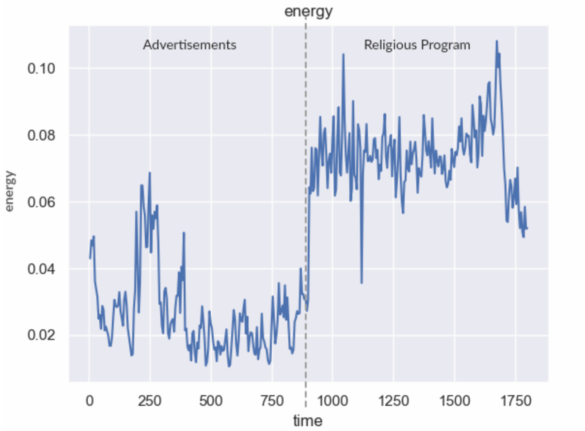


Figure 4: Energy Feature spectrum for advertisements Vs. religious programs

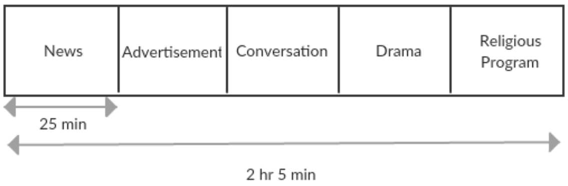


Figure 5: Audio clip structure designed for analyze features of one class from rest

Then ranked the selected features by calculating the score of the feature importance. Feature importance is scored using a tree-based classifier, which provides a measurement of the relevance of a feature towards the output variable. It is an inbuilt class provided by the scikit-learn library. Table 4 shows the selected features for One Vs. One model with the ranks.

2) *Feature Analysis: One Vs. All*

As shown in Table 3, One Vs. All method required only five binary classifiers because it constructs the number of binary SVMs equal to the number of classes. Each classifier allocates for the classification of one class. Here, the relevant features for distinguishing a class from the rest were analyzed. For that, 2 hours and 5 minutes long audio clip are prepared as Figure 5, which includes all classes of 25 minutes per class. Figure 6 shows a pattern obtained from the frequency spectrum of Spectral Entropy against the five classes. Using this 2 hours and 5 minutes lengthened sample audio clip, 34 features are analyzed and 24 features are selected as shown in Table 5.

Table 4: Selected features for One Vs. One model with ranks

Features	Binary SVMs in One Vs. One									
	1	2	3	4	5	6	7	8	9	10
ZCR										7
Energy			7	3			4		1	3
Energy entropy	1					3			5	4
Spectral centroid		2	3	4	2	1	3			6
Spectral spread			4	1		6	1		2	1
Spectral entropy		1			1	4		1		
Spectral flux					6					
Spectral roll off		4			4			3		
MFCC 1									7	
MFCC 2		3			5					
MFCC 3				5			2		4	2
MFCC 4	3	5	1			2		4	8	
MFCC 5			6			7				
MFCC 6			6							
MFCC 7						5	5			
MFCC 8	6					7		6		8
MFCC 9	2		2	2			6		3	
MFCC 10		7			3		7			
MFCC 11				6					6	5
MFCC 12			5					2		
MFCC 13								5		9
Chroma vector 1										
Chroma vector 2	5									
Chroma vector 3-9										
Chroma vector 10	7									
Chroma vector 11	4									
Chroma vector 12										
Chroma std										

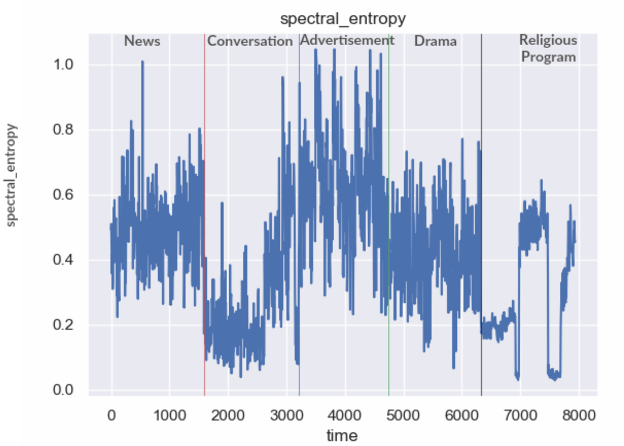


Figure 6: Spectral Entropy for five classes

Table 5: Selected features for One Vs. All model with ranks

Features	**Binary SVMs in One Vs. All**				
	1	**2**	**3**	**4**	**5**
ZCR			4		7
Energy	5				2
Energy entropy		8	13		8
Spectral centroid			2		5
Spectral spread			5		1
Spectral entropy		4	1		
Spectral flux			12		9
Spectral roll off		5	3		
MFCC 1		2			
MFCC 2		1	6		
MFCC 3			7	4	3
MFCC 4	1			3	
MFCC 5		10	9		
MFCC 6					10
MFCC 7			10		11
MFCC 8		9		2	
MFCC 9	2	7		6	4
MFCC 10		6	8		
MFCC 11	3			7	6
MFCC 12		3		1	
MFCC 13				5	
Chroma vector 1-2					
Chroma vector 3	4				
Chroma vector 4-6					
Chroma vector 7			11		
Chroma vector 8-10					
Chroma vector 11	6				
Chroma vector 12					
Chroma std					

B. *Data Pre-processing*

Data pre-processing helps to create raw data from audio files in a consistent way before extracting the features. As in Figure 2,

- In the data formatting stage, the data files convert to the .wav file format. Then the monophonic channel is chosen as the channel type and 44.1 kHz is selected as the sample rate [34].
- In the data annotation, manually listen to the audio clips using "Audacity" tool, segment them into different classes and label with relevant class labels.
- Then, remove silence only from news and conversations. The reason for remove silence from news and conversations will be described in section IV.

C. *Feature Extraction*

In audio classification, feature extraction is the most important component. Frame blocking before extracting the features. The audio waves are framed into 5s of blocks. Then extract the best subset of features from selected features. Extracted features are expressed as feature vectors. Table 6 lists all the features that are extracted in both One Vs. One and One Vs. All.

D. *Classification*

As stated in the related work, the multi-class SVM classifier is selected as the best approach which fits into our problem domain. To acquire higher performance, One Vs. One and One Vs. All multi-class SVM models are parallelly implemented and evaluated.

Table 6: Extracted features for each binary SVM

Multi -class SVM	**Binary SVMs**	**No. of features**	**Identical classes**
One Vs One	SVM 1	7	Chroma2, Chroma10, Chroma11, Energy entropy, MFCC4, MFCC8, MFCC9
	SVM 2	7	MFCC2, MFCC4, MFCC5, MFCC 10, Spectral centroid, Spectral entropy, Spectral rolloff
	SVM 3	7	Energy, MFCC4, MFCC6, MFCC 9, MFCC12, Spectral centroid, Spectral spread
	SVM 4	6	Energy, MFCC3, MFCC9, MFCC11, Spectral centroid, Spectral spread
	SVM 5	7	MFCC2, MFCC5, MFCC10, Spectral centroid, Spectral entropy, Spectral flux, , Spectral rolloff
	SVM 6	7	Energy entropy, MFCC4, MFCC7, MFCC 8,Spectral centroid, Spectral spread, Spectral entropy
	SVM 7	7	Energy, MFCC3, MFCC7, MFCC9, MFCC10, Spectral spread, Spectral centroid
	SVM 8	6	MFCC4, MFCC8, MFCC12, MFCC13, Spectral entropy, Spectral rolloff
	SVM 9	8	Energy, Energy entropy, MFCC1, MFCC3, MFCC4, MFCC9, MFCC11, Spectral spread
	SVM 10	9	Energy, Energy entropy, MFCC3, MFCC 11, MFCC13, MFCC8, Spectral centroid, Spectral spread, ZCR
One Vs All	SVM 1	6	Chroma3, Chroma11, Energy, MFCC4, MFCC9, MFCC11
	SVM 2	10	Energy entropy, MFCC1, MFCC2, MFCC 5, MFCC8, MFCC9, MFCC10, MFCC12, Spectral entropy, Spectral rolloff
	SVM 3	13	Chroma7, Energy entropy, MFCC2, MFCC 3, MFCC5, MFCC7, MFCC 10, Spectral entropy, Spectral centroid, Spectral flux, Spectral rolloff, Spectral spread, ZCR
	SVM 4	7	MFCC3, MFCC4, MFCC8, MFCC9, MFCC11, MFCC12, MFCC13
	SVM 5	11	Energy, Energy entropy, MFCC3, MFCC 6, MFCC7, MFCC9, MFCC11, Spectral flux, Spectral centroid, Spectral spread , ZCR

1) *One Vs. One model*

In this approach, N(N-1)/2 number of binary SVMs are implemented to classify N number of classes. Therefore, we design ten binary SVMs where each SVM classifies a pair of classes as shown in Table 3. SVM classifies ith and jth classes for a data point D = (x_t, y_t) as follows,

$$if\ (w^{ij})^T\ \varphi(x_t) +\ b^{ij}\ \geq 1 -\ {\varepsilon_t}^{ij};\ y_t = class\ i \qquad (1)$$

$$if\ (w^{ij})^T\ \varphi(x_t) +\ b^{ij}\ \leq -1 +\ {\varepsilon_t}^{ij};\ y_t = class\ j \qquad (2)$$

$(w^{ij})^T\ \varphi(x_t) +\ b^{ij}$ called as decision boundary where w^{ij} is the weight vector, x_t is the input vector, b^{ij} is the bias, and data x_t is mapped to a higher dimensional space by the function φ. The motivation behind the SVM is maximizing the decision boundary between two classes. The maximized decision boundary for *ith* and *jth* classes acquired by minimizing the magnitude of w^{ij}. Hence, to find the maximum margin, the magnitude of w^{ij} should be minimized as in the Equation (3). When the data is non-linearly separable, $C\sum_t {\varepsilon_t}^{ij}$ is introduced as the penalty terms to reduce the number of training errors.

$$min_{w^{ij},b^{ij},\varepsilon^{ij}} \frac{1}{2}\left(w^{ij}\right)^T\left(w^{ij}\right) + C\sum_t {\varepsilon_t}^{ij} \qquad (3)$$

One Vs. One model builds ten binary SVMs to classify five classes. Since there are ten decision boundaries, the predicted class for a particular data point is identified using a voting strategy called “Max Winning” strategy. If the decision boundary says the data point belongs to *ith* class, then vote for the *ith* class. Otherwise, vote for the *jth* class. Then the class with the maximum votes is taken as the predicted class.

2) *One Vs. All model*

One Vs. All method constructs N number of binary SVMs where it has N number of classes to classify. Therefore, five SVM classifiers are designed as shown in Table 3. Each SVM is trained with the whole dataset where the data belongs to *ith* class with positive labels and remain of the data with negative labels. An SVM solves data point D = (x_t, y_t) for *ith* class according to the following equations Equation (4) and Equation (5).

$$if\ (w^{i})^T\ \varphi(x_t) +\ b^{i}\ \geq 1 -\ {\varepsilon_t}^{i};\ y_t = class\ i \qquad (4)$$

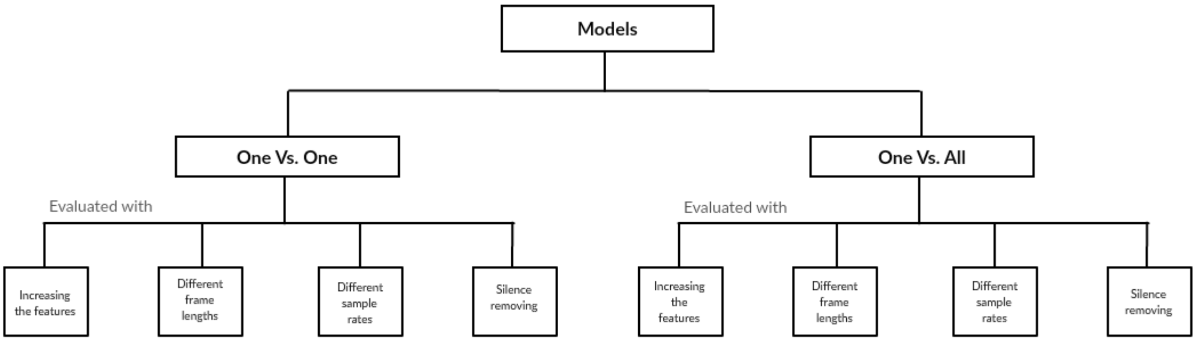
$$if\ (w^{i})^T\ \varphi(x_t) +\ b^{i} \leq -1 +\ {\varepsilon_t}^{i};\ y_t \neq class\ i \qquad (5)$$

To find the maximum margin, the magnitude of w^i should be minimized as in the Equation (6) where C is the constant used to reduce training error.

$$min_{w^{i},b^{i},\varepsilon^{i}} \frac{1}{2}\left(w^{i}\right)^T\left(w^{i}\right) + C\sum_t {\varepsilon_t}^{i} \qquad (6)$$

One Vs. All model implements five binary SVMs to classify each class individually. After training five classifiers, the class of a data point x is predicted by finding the decision boundary which has the maximum value. Equation (7) gives the prediction function for data point x.

$$class\ of\ x = argmax\left(\left(w^{i}\right)^T\ \varphi(x_t) +\ b^{i}\right) \qquad (7)$$



In these experiments, while changing one parameter other parameters are keeping as constants.

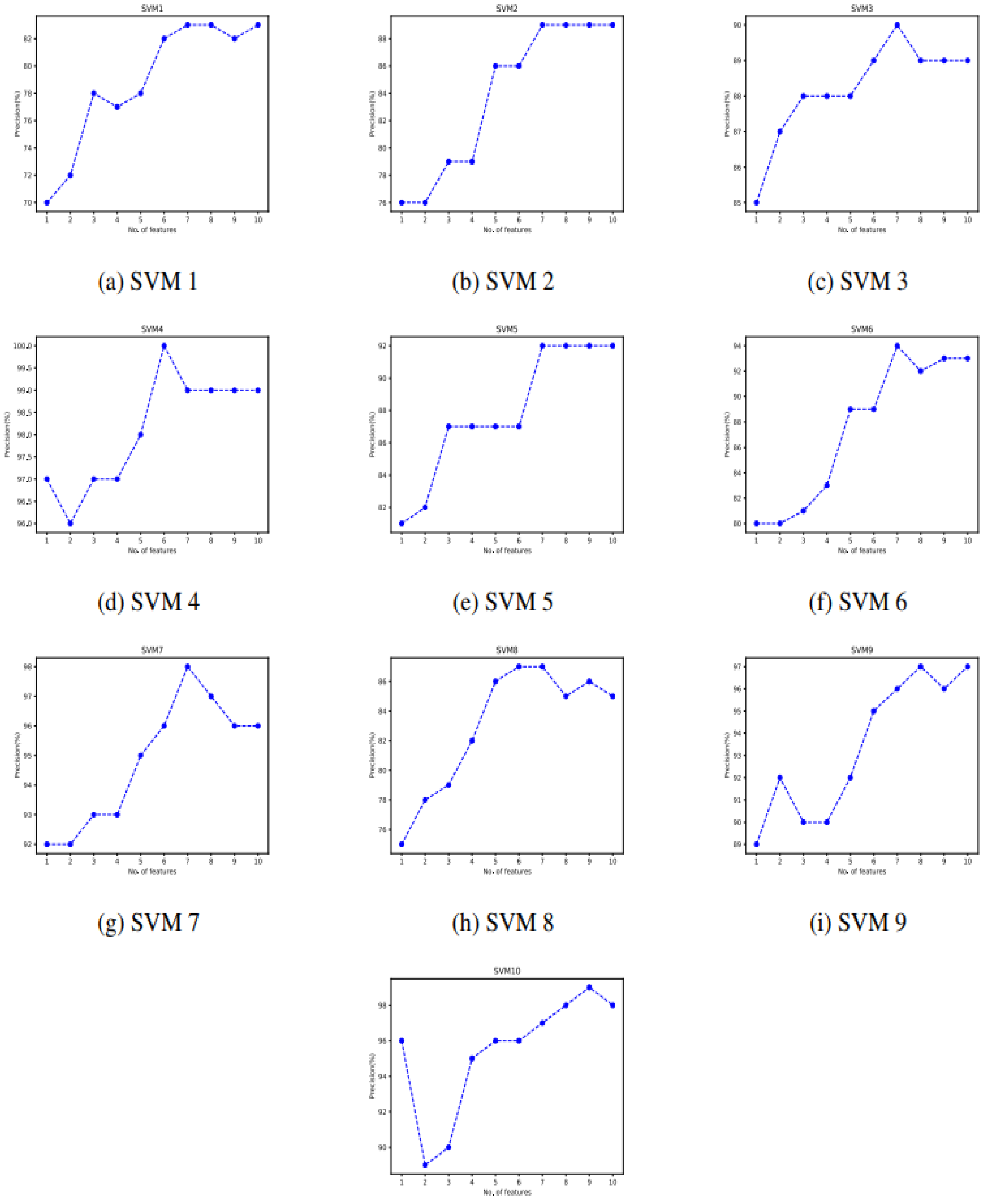
- K value = 10
- Frame lenght =5s
- Sample rate = 44100 Hz
- With silence remove of news and conversations

Figure 7: Evaluation Plan

E. *Evaluation*

In evaluation, the performance of One Vs. One and One Vs. All multi-class SVM models are evaluated. Ground truth data is required to evaluate the accuracies of the two models. 40% of the total data that is never in the training set is taken as the ground truth data. The ground truth data contains 28 minutes long audio recordings of each class (news, conversations, advertisements, drama, and religious programs). “Audacity” tool is used to annotate the ground truth data.

The models are evaluated under different criteria as depicted in Figure 7. By increasing the features, the performances of the models are evaluated. Additionally, the two models are evaluated using selected frame lengths, selected sample rates, before silence removal and after silence removal. The performances of the two models are presented using graphs and confusion matrices. Necessary



(a) SVM 1

(b) SVM 2

(c) SVM 3

(d) SVM 4

(e) SVM 5

(f) SVM 6

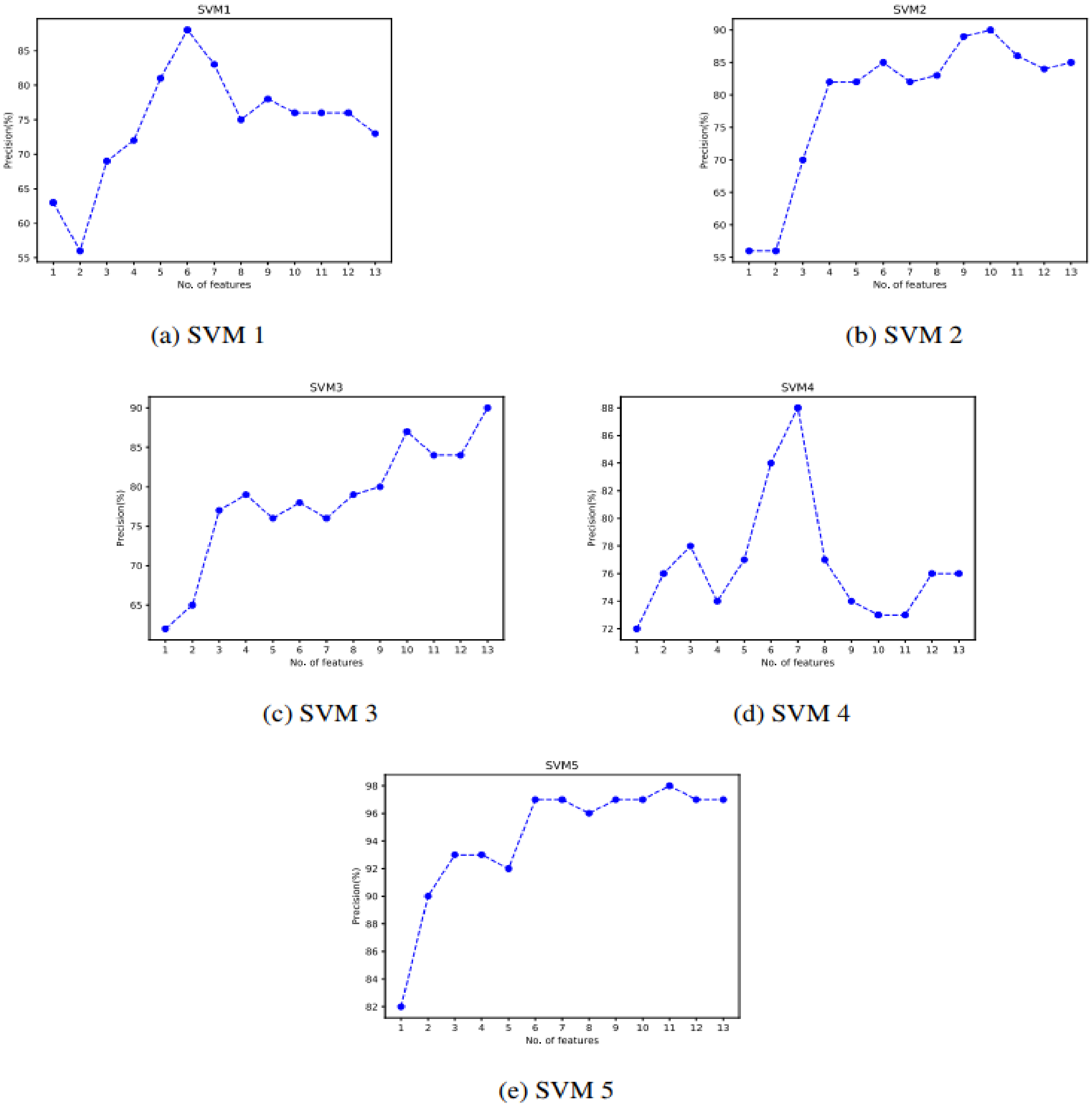
(g) SVM 7

(h) SVM 8

(i) SVM 9

(j) SVM 10

Figure 8: Changing of precision when increasing features of binary SVMs in One Vs. One model



(a) SVM 1

(b) SVM 2

(c) SVM 3

(d) SVM 4

(e) SVM 5

Figure 9: : Changing of precision when increasing features of binary SVMs in One Vs. All model

refinements for both models are made based on the evaluation results. The precision value is taken to measure performance.

IV. EXPERIMENTS AND RESULTS

Models were initially evaluated using data that were not used in the training phase. All the experiments mentioned in section III.E were repeated for five times and the average of the results was calculated to determine the overall performance of the system. According to the results, necessary refinements were done to the classification model.

A. *Increase the number of features*

This method is used to prevent "diminishing returns". First, features selected for each SVM are ranked according to the score value of feature importance. Thereafter, many rounds of experiments were conducted by increase the input feature count in order to determine the optimal feature set. Figure 8 and Figure 9 illustrate the obtained results for One Vs. One and One Vs. All models. These figures indicate less number of features can achieve the highest precision value. The minimal, required subset of features for each SVM is selected through this process. The obtained features are listed in Table 6.

B. *Different frame sizes*

Selecting a correct frame size to extract the features is essential when comes to a classification problem. In the closest work to this research done by C. Weeratunga et al [15] has used 2.5s as the frame size. Other than that, the works that are found in the literature have used different frame sizes such as 25ms, 0.25s, and 0.3s etc. Therefore, the model is evaluated with respect to different frame sizes and reports the results for the chosen frame sizes 0.25s, 2.5s, 4s, and 5s. Increasing the frame size more than 5s is impossible in this case since some of the data segments in the dataset has the length in between 5s and 6s. When changing the frame size, the rest of the model's parameters such as K value and sample rate were kept constant. The obtained results are shown in

Figure 10. Length of 5s frame is selected as the best frame size.

C. *Different sample rates*

FM radio channels have a bandwidth of 15 kHz approximately. Bandwidth is the difference between the highest and lowest frequencies carried in an audio stream. According to Nyquist Shannon theorem, the highest frequency is half of the sample rate. Practically, the highest frequency for a radio stream is in between 22050 Hz - 20000 Hz because the highest audible frequency of a human is 20000 Hz [34]. Thus, logically the best sample rate for our study is 44100 Hz. In addition to that, 16000 Hz and 22050 Hz were also used as the sample rates in previous works related to radio broadcasting classification. Weeratunga et al [15] proposed 22050 Hz as the sample rate, John Saunders [16] and Vavrek, J. et al [20] proposed 16000 Hz as the sample rates for their studies. Therefore, we evaluate the models with sample rates 16000 Hz and 22050 Hz, and 44100 Hz to find the most reliable sample rate for the research. Figure 11 illustrates the changing of performance against different sampling rates for both One Vs. One and One Vs. All models.

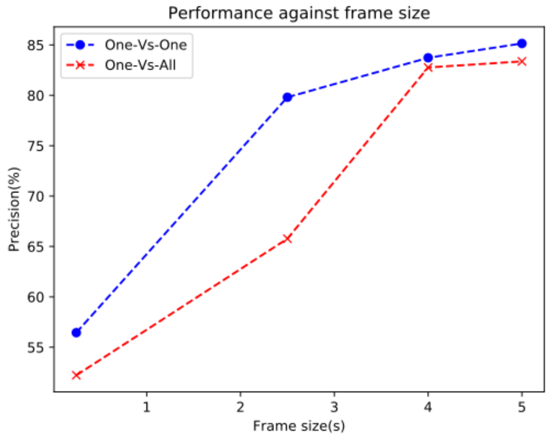


Figure 10: Variation of precision against frame size

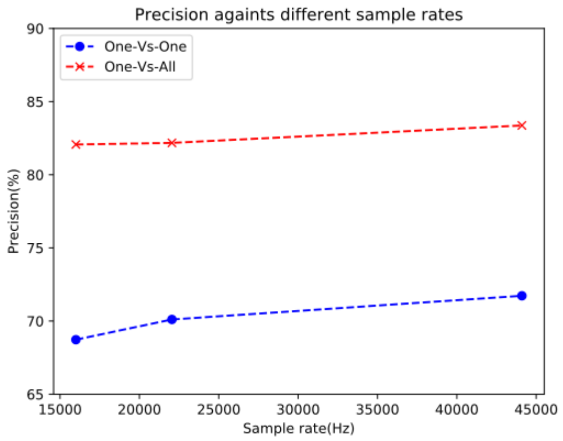


Figure 11: Variation of precision against sample rate

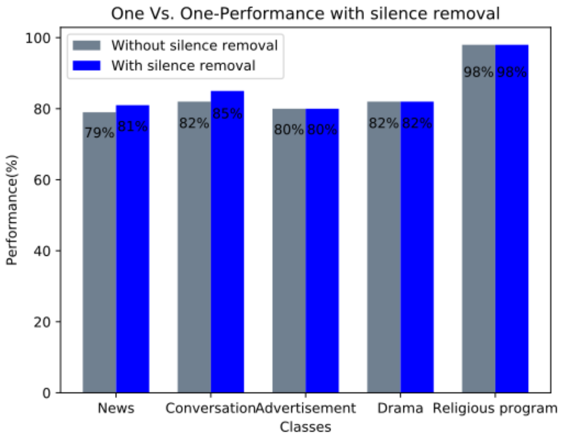


Figure 12: Variation of the precision values of One Vs. One against Silence removal

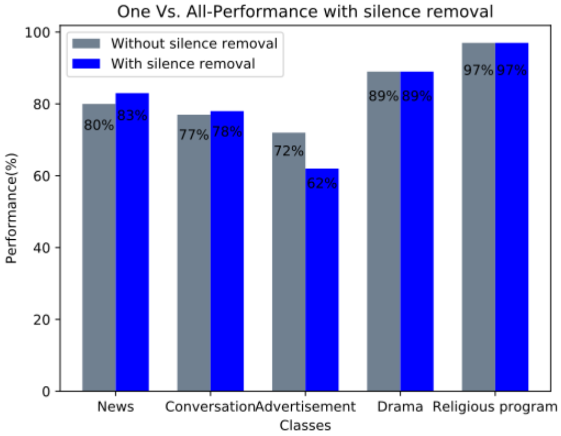


Figure 13: Variation of the precision values of One Vs. All against Silence removal

D. *With silence removing*

In the data pre-processing phase, silence removal is done only to news and conversation contents because they have long silence periods within an audio clip. For the evaluating purposes, the model is evaluated without silence removal and with silence removal from all classes. Figure 12 and Figure 13 show the changing of the performances of the classification without silence removal and with silence removal in One Vs. One model and One Vs. All model respectively. As shown in figures, after the removal of the silence from all data, performances increased only in news and conversations. Therefore, we decided to remove silence only from news and conversations.

V. EVALUATION

One of the main aspects of this research is to select the optimal subset of features for classification. Table 7 and Table 8 provide the training accuracy (precision value) of each SVM when using all features and the selected subset of features. This indicates that the overall performance of the models increases with the optimal subset of features.

Table 7: Training accuracies with all the features and the optimal subset of features of One Vs. One model

Models	**Accuracy using all the features (Precision)**	**Accuracy using the optimal subset of features (Precision)**
news/ advertisement	85%	87%
news/ conversation	88%	92%
news/ drama	90%	92%
news/ religious program	99%	100%
advertisement/ conversation	89%	92%
advertisement/ drama	91%	93%
advertisement/ religious program	98%	99%
conversation/ drama	84%	87%
conversation/ religious program	95%	95%
drama/ religious program	97%	99%
One Vs. One	**81%**	**85%**

Table 8: Training accuracies with all the features and the optimal subset of features of One Vs. All model

Models	**Accuracy using all the features**	**Accuracy using the optimal subset of features**
news/ other	76%	82%
conversation/ other	82%	92%
advertisement/ other	86%	91%
drama/ other	79%	82%
religious program/ other	93%	96%
One Vs. All	**78%**	**83%**

After applying the optimal features with selected parameters, the performance of One Vs. One and One Vs. All models show respectively in Table 9 and Table 10.

Table 9: Confusion matrix of One Vs. One model

		Predicted Class					Support	Precision
		news	conv	advert	drama	relprog		
True Class	news	**260**	8	43	6	0	317	81%
	conv	25	**247**	14	33	6	325	85%
	advert	28	12	**242**	22	3	307	80%
	drama	7	13	6	**290**	0	316	82%
	relpog	0	7	0	2	**314**	323	98%
Overall results								85.20%

Table 10: Confusion matrix of One Vs. All model

		Predicted Class					Support	Precision
		news	conv	advert	drama	relprog		
True Class	news	**265**	9	38	5	0	317	73%
	conv	38	**247**	4	34	2	325	85%
	advert	44	11	**198**	51	3	307	80%
	drama	13	12	6	**283**	2	316	76%
	relpog	1	11	0	0	**311**	323	98%
Overall results								83.37%

When considering the results of both models, a considerable amount of news frames has misclassified as advertisements, while the conversations and advertisements have misclassified as news and drama. The religious program has classified better than others. Even though the news reading can be considered as monotonic, in some scenarios conversations and drama also have a monotonic nature as news. Even in the advertisements show a monotonic nature after removal of music and jingles. Therefore, the monotonic nature of news, conversations, advertisements, and drama might be the reason for these misclassifications. Table 9 and Table 10 shows the obtained accuracies of each binary SVM in both models.

As included in Table 11, even though ten binary SVMs trained with high training accuracies, after combining the ten models together the accuracy of the alliance is degraded. The reason might be the ``Max Winning" strategy that uses to predict the classes in One Vs. One. In ``Max winning" strategy when computing the mode, if the maximum number of votes is equal to two classes, then it outputs only one class which appears first in the array. This might be caused to degrade the final result of One Vs. One SVM. In our training dataset, approximately 13% of data has faced this issue when the frame length is 0.25s. But when we increase the frame length, the proportion of identical classes decreased to 4% as shown in Figure 14.

Table 11: Overall results of One Vs. One model

Models	**Accuracy**	**Precision**	**Recall**	**F1-score**
news/ advertisement	87%	87%	87%	87%
news/ conversation	90%	92%	91%	91%
news/ drama	92%	92%	92%	92%
news/ religious program	99%	100%	100%	100%
advertisement/ conversation	91%	92%	92%	92%
advertisement/ drama	94%	93%	94%	94%
advertisement/ religious program	99%	99%	99%	99%
conversation/ drama	85%	87%	86%	87%
conversation/ religious program	95%	95%	95%	95%
drama/ religious program	99%	99%	99%	91%
One Vs. One	**85%**	**85%**	**85%**	**85%**

Table 12: Overall results of One Vs. All model

Models	**Accuracy**	**Precision**	**Recall**	**F1-score**
news/ other	86%	82%	75%	78%
conversation/ other	91%	92%	78%	84%
advertisement/ other	91%	91%	79%	85%
drama/ other	88%	82%	81%	82%
religious program/ other	98%	96%	97%	97%
One Vs. All	**83%**	**83%**	**83%**	**83%**

As depicted in Table 12, One Vs. All model shows less accuracy than One Vs. One model. One disadvantage of One

Vs. All model is when analyzing the features, each binary SVM in One Vs. All model must look at the features as one class against the four classes. Therefore, to choose the best, it is difficult to observe the distinguishing features for one class versus four. But when analyzing features for One Vs. One model, features should be analyzed against two classes each time whenever it is easy to identify a pattern that distinguishes two classes. Another drawback of One Vs. All model is that it takes high training time since each binary SVM classifier in One Vs. All model requires a complete dataset for individual training. But the binary SVM classifiers in One Vs. One model takes less training time compared to the binary SVMs in One Vs. All model as it requires data only from two classes for training.

According to Table 9 and Table 10, the One Vs. One model achieved 85% of overall precision and the One Vs. All model achieved 83% of overall precision. The obtained results of this study guide us to find the most suitable multi-class SVM for this problem domain. According to the performance of these two models, the One Vs. One model with 85% of precision, is chosen as the most appropriate model for this research.

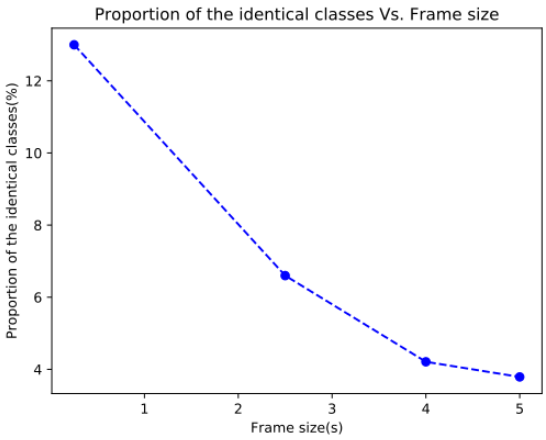


Figure 14: Proportion of the identical classes against the frame length

VI. CONCLUSION AND FUTURE WORK

The main aim of this research is to identify voice dominant content categories to automate the radio broadcasting context in Sri Lanka. For that, a multi-class SVM was proposed. Multi-class SVM was built using two conventional ways, "One Vs. One" and "One Vs. All" and compared the performance to find the best model for this domain. The novelty of this approach is that instead of feeding all the features once, only selected features were fed separately to each classifier in the model.

The performance of these two models was evaluated under different criteria. One Vs. One model successfully classified the pre-defined content categories with the accuracies of 81% for news, 85% for conversations 80% for advertisements, 82% for drama and 98% for religious programs. The One Vs. All model successfully classified the categories with the accuracies of 73% for news, 85% for conversations, 80% for advertisements, 76% for drama and 98% for religious programs. The final overall accuracies of the One Vs. One and One Vs. All models are 85% and 83% respectively. Moreover, this proposed methodology is able to increase the classification accuracy of news contents to 81% and the accuracy of the existing methodology [15] was 41%.

The major limitation of this research is that the model is trained and tested only for "SLBC" radio FM channel. However, this study creates a platform to further generalize this model to all Sri Lankan FM channels. Another limitation is that restricts the data of each class to 1 hour and 10 minutes. The reason is that religious programs were unable to provide data for more than 1 hour and 10 minutes long. Therefore, the data of the rest of the classes were also limited to 1 hour and 10 minutes to avoid the proportional bias in the dataset. Therefore, using more training data for classification to improve performance is a good choice. In addition, identifying the most prominent features yields more accurate results.

REFERENCES

[1] C. R. S. Celebrating Radio: Statistics / World Radio Day 2015, 2018. [Online].Available:http://www.diamundialradio.org/2015/en/content/celebrating-radiostatistics.html

[2] 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 (2018): Radio Broadcast Monitoring to Ensure Copyright Ownership. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 11(1)

[7] Dhanith Chaturanga and Lakshman Jayaratne (2013): 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

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

[9] Rajitha Amarasinghe and Lakshman Jayaratne (2016): 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

[10] Rajitha Peiris and Lakshman Jayaratne (2016): 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

[11] Tharika Madurapperuma, Gothami Abayawickrama, Nesara Dissanayake, Viraj B. Wijesuriya and K. L. Jayaratne (2017): 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.

[12] Rajitha Peiris and Lakshman Jayaratne (2016): 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.

[13] M. G. Viraj Lakshitha and K. L. Jayaratne (2016): 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.

[14] 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.

[15] C.O.B. Weerathunga, P.V.K.G. Gunawardena and K.L. Jayaratne (2018): Classification of Public Radio Broadcast Context for Onset Detection. *European Journal of Computer Science and Information Technology (EJCSIT) by European Centre for Research Training and Development UK*, 7(6):1-22, Published by ECRTD – UK, ISSN2054 – 0957 print 2054 – 0965 Online, www.eajournals.org, 13 Duncan Rd, Gillingham Kent ME7 4 LA, UK.

[16] J. Saunders, "Real-time discrimination of broadcast speech/music," in *icassp*. IEEE, 1996, pp. 993–996.

[17] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *AAAI/IAAI*, 2000, pp. 679–684.

[18] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 81–84.

[19] S. G. Koolagudi, S. Sridhar, N. Elango, K. Kumar, and F. Afroz, "Advertisement detection in commercial radio channels," in *Industrial and Information Systems (ICIIS), 2015 IEEE 10th International Conference on*. IEEE, 2015, pp. 272–277.

[20] J. Vavrek, E. Vozarikov ´ a, M. Pleva, and J. Juh ´ ar, "Broadcast news audio classification using ´ svm binary trees," in *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*. IEEE, 2012, pp. 469–473.

[21] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 203–211.

[22] M. Khan, W. G. Al-Khatib, and M. Moinuddin, "Automatic classification of speech and music using neural networks," in *Proceedings of the 2nd ACM international workshop on Multimedia databases*. ACM, 2004, pp. 94–99.

[23] Z. Kons, O. Toledo Ronen, and M. Carmel, "Audio event classification using deep neural networks." in *Interspeech*, 2013, pp. 1482–1486.

[24] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[25] C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transaction onNeural Networks13 (2)*, pp. 415–425, 2002.

[26] F. Aurino, M. Folla, F. Gargiulo, V. Moscato, A. Picariello, and C. Sansone, "One-class svm based approach for detecting anomalous audio events," in *Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on*. IEEE, 2014, pp. 145–151.

[27] A. Bouril, D. Aleinikava, M. S. Guillem, and G. M. Mirsky, "Automated classification of normal and abnormal heart sounds using support vector machines," in *Computing in Cardiology Conference (CinC), 2016*. IEEE, 2016, pp. 549–552.

[28] S. E. Kuc¸ ¨ ukbay and M. Sert, "Audio-based event detection in office live environments using ¨ optimized mfcc-svm approach," in *Semantic Computing (ICSC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 475–480.

[29] I. Mart´ın-Morato, M. Cobos, and F. J. Ferri, "A case study on feature sensitivity for audio ´ event classification using support vector machines," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016, pp. 1–6.

[30] J. C. Wang, J. F. Wang, C. B. Lin, K.-T. Jian, and W. Kuok, "Content-based audio classification using support vector machines and independent component analysis," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4. IEEE, 2006, pp. 157–160.

[31] L. Lu, S. Z. Li, and H. J. Zhang, "Content-based audio segmentation using support vector machines," in *Proc. ICME*, vol. 1, 2001, pp. 749–752.

[32] Y. Zhu, Z. Ming, and Q. Huang, "Automatic audio genre classification based on support vector machine," in *Natural Computation, 2007. ICNC 2007. Third International Conference on*, vol. 1. IEEE, 2007, pp. 517–521.

[33] B. Kijsirikul and N. Ussivakul, "Multiclass support vector machines using adaptive directed acyclic graph," in *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, vol. 1. IEEE, 2002, pp. 980–985.

[34] "Sample rates - audacity manual," https://manual.audacityteam.org/man/sample rates.html, (Accessed on 12/22/2018).

[35] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.