International Journal of Digital Information and Wireless Communications (IJDIWC) 2(4): 66-78
The Society of Digital Information and Wireless Communications, 2012 (ISSN: 2225X-658X)

NNET BASED AUDIO CONTENT CLASSIFICATION AND INDEXING SYSTEM

Muhammad M. Al-Maathidi[1], Francis F. Li[2]

Acoustics Research Centre
School of Computing Science and Engineering. University of Salford
Salford, Greater Manchester, M5 4WT, UK
[1] e-mail: M.M.Abd@edu.salford.ac.uk
[2] e-mail: F.F.Li@salford.ac.uk

ABSTRACT

Rapid advancement in computers and internet technology has led large volume of multimedia files. The archiving and digitization of the old media contents also contributes to the growth of the digital library. The usefulness of these collections is largely dependent upon the availability, information retrieval, and search tools. Soundtracks are information-rich and there is a lot of related information can be extracted from them; that enabling metadata generation and semantic search. This paper proposes a machine learning system based on neural network; the system use set of audio features that belong to three different feature spaces as a training/testing features; this regime will presents a generic structure of a pre-processing stage for automated metadata generation.

KEYWORDS

Audio classification, multimedia, indexing, metadata, content descriptor, information retrieval, machine learning, feature space, machine audition, noise robustness.

1 INTRODUCTION

The development in computers technology has resulted in a large number of media files, such as broadcasted radio and television programs, recorded meetings, and voice mails. To extract information from the massive archive of multimedia files, effective search and indexing tools are vitally important. On the other hand the manual indexing is not practical, time-consuming, and individual related. In This paper proposes a high level architecture system that can automatically generate metadata that describe the content of the signals using purposely chosen feature spaces and machine learning regimes. Essentially audio signals are segmented into speech, music and other sound events. These three types of audio signals can be analyzed subsequently using many existing machine audition tools to give text information about the soundtrack.

This is an active research field in which many results have been published and these covered diverse topics. Some research focused on audio segmentation into pre-defined classes such as news, music, advertisement cartoon, and movies [1, 2]. While others emphasized on content types, for example silence, speech, and music [2, 3]. Moreover, other research focused on sport game segmentation such as football, basketball, tennis, hockey, Ping-Pong, and badminton [4]. Others concentrated on acoustic homogenousity and self-similarity [5, 6]. In addition, some music research placed an emphasis on different areas like music structure segmentation [7], and instrument type [8]. In speech field research, the focus are on speaker clustering and identification [9, 10].On the other hand, speech recognition and keyword spotting systems achieve notable results; there are some commercial products

such as Dragon Naturally Speaking software in addition to smartphone operating systems that support natural language speech recognition such as Android and Apple iOS.

The current research in this field has encountered several limitations and weaknesses:

1. The available classification systems designed to handle specific class sets and cannot be generalized.
2. Results from different research cannot be compared due to the use of different training/testing samples to evaluate the proposed systems.
3. Most research utilizes specific audio content or an idolized sample set in system evaluation. This makes it hard to compare different techniques and results.
4. No research attention was given to the detection of an overlapped classes "combination of different classes" but this is an important case if we wanted to handle a real life audio content.
5. Finally, despite all the current research in the field, there is no available "neither free nor commercial" system that claims the ability to search the audio files content even though it is highly required.

This paper aims to identify the optimal combination between audio feature space and machine learning techniques in order to achieve the most accurate classification and indexing results. In this work we will try to address some of the above shortcoming by utilizing the best set of features and the most promising machine learning algorithms and test this combination on a same set of samples in order to identify the best system component to build the classification system. Also the trained classification system will be tested to check its noise robustness.

2 A HIGH LEVEL AUDIO CLASSIFICATION SYSTEM ARCHITECTURE

This research is a step toward an automated high level architecture, media indexing and archiving system. This system aims to identify the class of the input audio file, the following three classes will be utilize: speech, music, and other the event sound. After identifying the content that belongs to each class, the system can re-deploy many existing algorithms to generate content related texts to each one of these classes, by utilizing text mining semantic analysis algorithms. Subsequently, the generated metadata can be combined with text search engines to enable audio searching; also the result can be integrated with MPEG-7 audio description standard. Such a system is urgently needed in order to utilize the content-based audio retrieval and audio search engines.

The proposed system architecture is illustrated figure 1.

The metadata generation system will explore each entity in the audio database and generate a text indexes and keyword that describes the content of the audio file, the text indexes will be stored in a database. This will enable the system to accept text quires, this the text will be fed to text search engine, the search engine will search for the best matching entities in the text indexes database and retrieve their corresponding audio files to be delivered to the user. The user Also can submit an audio queries "query by example," in this case the metadata generation system will take the input audio file and generate a text keywords, then these keyword can be treated d as a text query.

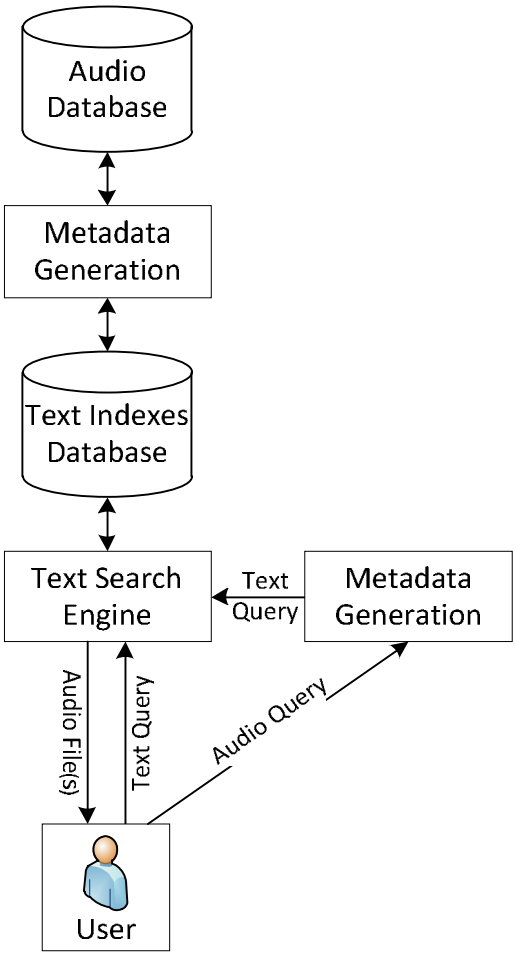


Figure 1: The proposed system architecture is illustrated.

3 AUDIO FEATURE SPACES

Feature extraction is an important step in the audio classification system; an efficient features set should be able to capture the most significant audio properties of the distinct class; it should be reasonably robust against environmental noises and powerful enough to discriminate between various audio classes.

3.1 Temporal Domain

This is the native domain for the audio signal in which the signal is represented over time. The temporal features are fast to extract and it has been used successfully as a classifying feature [11]. However it might fail to differentiate between mixed type audio data.

The following time domain features selected to be tested in the proposed system.

Zero Crossing (ZC)

The ZC is a basic property of an audio signal, it can be found by calculating the times that the audio signal crosses the zero axes. The ZC feature has been used as an audio signal classification feature in particular for voiced/unvoiced speech discrimination, speech/music discrimination [12], and for music genre classification [13, 14], but the ZC will fail in the case of mixed classes audio data, also it will fail in some genres of music[2].

RMS (RMS)

Approximate the loudness of an audio signal[15, 16].

3.2 Frequency Domain "Spectral Domain"

In the frequency domain audio signal represented by its spectral distribution features characterize the short-time spectrum. The following frequency domain features are selected to be tested in the proposed system:

Spectral Entropy (SE)

The SE returns the relative Shannon entropy of the spectrum.

Spectral Rolloff Frequency (SL)

The SRF is the frequency that below which 85% of the spectrum magnitude is concentrated [13]. Other references use different ratios to approximate the SL.

Brightness (BR)

Measures the amount of the high-frequency content in an audio signal; this is done by measuring the amount of energy above the cut-off frequency. The increment in loudness also increases the amount of high spectrum

content of a signal thus making a sound brighter.

Roughness (RF)

Roughness is a way to describe the pleasantness reduction in hearing, the total roughness estimation is achieved by finding all the peaks and taking the average all the dissonance between all the possible pairs [17]

Irregularity (IR)

The spectrum irregularity measures the degree of variation in successive spectrum peaks; it can be approximated by finding the square difference in amplitude between adjacent partials [17]

Spectral Flux (SF)

The SF is the value of average variation in a signal spectrum between adjacent frames. It measures the local spectral change. Speech SF values is higher than music SF values and the environment sounds have the lowest SF values. Also, the environmental sounds changes dramatically between successive frames [18]. SF has been used for Speech/Music Discrimination in [12, 14].

Spectral Centroid (SC)

The SC is used to characterize the spectrum. It is calculated as the weighted average of the discrete frequencies present in the signal, as determined by DFT. It is has been designed to discriminate between different musical instrument timbres [16].

Audio Spectrum Centroid (ASC)

The ASC measures the center of gravity of a log-frequency power spectrum. The ASC provides information on the power spectrum shape, it indicates whether a power spectrum is dominated by high or low frequencies, and also it gives an approximation of the signal perceptual sharpness. The log-frequency scaling utilized to approximate the frequency perception of the human auditory system[16].

Audio Spectrum Spread (ASS)

The ASS measures the spectral shape, it describes the spectrum around its centroid, so a low ASS value means that the spectrum power might be concentrated around the centroid, while a high value means the spectrum power might distributed across a wider range of frequencies. It is designed specifically to help differentiation between noise-like and tonal sounds[16].

3.3 Cepstral Domain

The cepstrum concept was introduced for the first time for by Bogert et. al. [19]. It is the result of taking the Fourier transform of the logarithm of the magnitude of the spectrum.

$$Pitch = 1127.0148 \ log\left(1+\frac{f\,(Hz)}{700}\right) (1)$$

The second Fourier transform can be replaced by the IDFT, DCT, and IDCT. But because the DCT decorrelates the data better than the DFT, it is often preferred[20].

The cepstrum features is a good technique for separating the components of complex signals that are made up of several different but simultaneous elements combined together such as speech features.

Mel Frequency Cepstrum Coefficients (MFCCs)

The Mel-Frequency Cestrum Coefficients (MFCCs) is a perceptually motivated representation that defines as a short window cepstrum of a signal. MFCC is the most popular cepstrum-based audio feature; it represents an excellent feature vector for both speech and music signals [16]. The non-linear Mel-frequency scale was developed to approximate the responses of the human auditory system. The Mel is a unit of pitch that

has been judged by listeners to be equally spaced. To convert from frequency f in hertz to the equivalent Mel frequency achieved by using the following equation:

$$Pitch = 1127.0148 \ log\left(1+\frac{f\,(Hz)}{700}\right) (2)$$

Figure 2 illustrate the mal-spaced filter bank for frequency range from 0 to 8kHz.

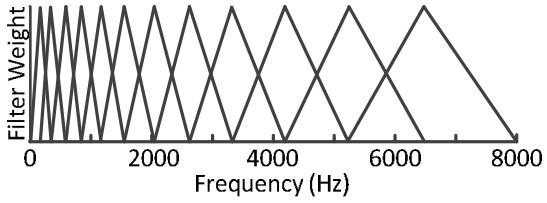


Figure 2: Mel Filter Banks.

The MFCC is proven to be beneficial in the field of audio classification, it is mostly used as a main feature in audio classification in [8, 10, 21-24]. Therefore MFCC selected to be the main classification feature.

MFCC Vector Calculation

MFCC is a short term spectral based feature [25]. Therefore, the first step to calculate MFCC is to segment the audio file into overlapping small windows typically with a size of (20ms-40ms). To overcome windowing edge effect, a Hamming window is applied on each frame[26]. Below is the steps followed to extract the MFCC.

1. Finding window spectrum: the spectrum of each window is calculated by:
   a. Applying Fast Fourier Transform.
   b. Finding the magnitude by taking the absolute value of the spectrum.
   c. Discard the second half of the spectrum coefficients; because of the symmetry property of FFT.
2. Reducing window spectrum: the spectrum will be reduced to N values only, where N is the size of the result MFCC vector. Usually it will be from 12 to 20 and this can be achieved by:
   a. Identifying the minimum and maximum frequencies, usually the minimum frequency is set at 0Hz and the maximum frequency in speech recognition system is set for 4kHz "notice that the frequencies in Hertz".
   In our classification problem, a higher frequency range needs to be covered because we are dealing with music file content.
   b. Convert these frequencies from Hertz to Mel using equation (1), and find N+2 equally spaced centers in Mel domain that start with the first center that equals 0 and ends with N+2 that equal to maximum frequency.
   c. Convert the N+2 Centers to Hertz using equation (2).
   d. Find triangular filters for frequency centers 2 to N+1, so that the triangular filter M is started from frequency center M, centered on frequency center M+1, and ends on frequency center M+2, see figure 2.
3. Reducing spectrum: the spectrum of each frame will be multiplied by the N Mel filters; this will produce a vector of N elements, "one element for each filter" that represent the reduced spectrum.
4. Calculate the log Energies Vector: calculates the log energies of the result vector.
5. Finding the final MFCC vector: the final MFCC vector is achieved by taking the discrete cosine transform of log energy vector.

4 AUDIO CLASSIFICATION

This paper presents an automated multi audio classification system that discriminates and categorizes input audio into three classes: speech, music, and other event sound. The following is the classification process scheme:

4.1 Framing

In the framing step the input audio file will be split into overlapped frames of 40ms size and a hamming window will be applied when required.

4.2 Frame Feature Extraction

This step aims to reduce the dimension of the input audio file by extracting the unique features that can be used by the decision making system (DMS) to discriminate between frames that belong to different audio classes. Before feature extraction, the silent segments with diminished loudness are removed then the features listed in section 3 will be extracted from all the non-silent frames.

4.3 Classification System

The classification system aims to classify the input frames according to their features into to one of the following three classes: speech, music, and other event sound. The input for this step will be one feature vector for each audio frame and the output will be a number between 1 and 3 that represents one of the three different classes that the frame belongs to.

The system will use a supervised Neural Network NNet. the NNet has been successfully used as an audio classification module in [1, 24].

The supervised NNet require a training unit and a testing unit, each unit will be discussed briefly.

Training Units

The features of the manually classified test samples will be used to train the NNet. Three different NNet modules will be utilized, one module for each class, these modules will be trained to reach the result "1" for all frames that belong to the same class, and "-1" for all frame that doesn't belong to the same class.

Classification Units

The task here is to classify frames into one of the pre-defined classes; classification is accomplished by using the three trained classification modules that will take the input feature vector of the frames. Each frame will be examined by the three classification units in order to find the frame class. To check the classification accuracy, the testing audio files will be classified and the classification result will be compared with a manually identified file content. The percentage of the true classified frames wills the classification accuracy.

5 AUDIO SAMPLES DATABASE

A real-life, non-biased, high quality, miscellaneous audio database is a key point of successful system training and performance evaluation. The samples are in audio CD quality 44.1K sampling rate, and 16 bit depth. All the samples are saved in an uncompressed wave file format to make it easier and faster to manipulate. Also, this uncompressed format is used to avoid any quality degradation or information loss.

Each sample is classified manually into one of the following three main classes:

1. **Speech**: contains a pure speech with no background music or noise, including a variety of voices of males, females, kids and a group of people, voices of lectures, conversations, shouting, and narration.
2. **Music:** contains different types, genre, modes and instruments of music files.
3. **Other:** contains samples that do not fit in the previous two classes, like sounds of rain, storms, thunders, screaming,

helicopter, crashing, busy road, school yard, and many others samples.

The average length of these samples is about 15.2 seconds, and the contents of each sample are selected to be acoustically homogenous. Table 1 shows the number of samples in each class and it average length.

Table 1: The number of samples in each class.

Class	Number of Samples	Average Sample Length
Speech	115	11.6 Sec
Music	115	17.6 Sec.
Other	115	16.3 Sec

6 CLASSIFICATION SYSTEM TESTING AND EVALUATION

Testing, evaluation and a brief discussion for the result will be carried out in the following sections. These tests have been carried out using a MatLab developed system. For training and testing purposes, the samples that belong to the same class will be split into two groups: the first group will be used for training and the second group will be used for testing. A *one-against-all* classification technique will be utilized for each one of the three classes. A neural network based classification module will be utilized as a DMS. System parameters will be initiated with values that are found to be suitable "for the classification system" through the literature review and during system development. This parameter will be fixed through all the tests in order to emphasize the effect of the features selection on the classification accuracy. The focus in testing will be on the key parameter that is theoretically expected to improve the system performance.

The classifier will give one of two outputs for each frame: *positive* or *negative*. The *positive* means that the frame belongs to the classifier class and the *negative* means that the frame does not belong to the classifier class. In either case, the classifier may reach the correct decision result and therefore it will be referred to as a "*true classification*;" if it misses, the correct decision result will be referred to as "*false classification*." The four possible combinations of these results are "*positive-true*, *negative-true*, *positive-false*, *negative-false*." In system evaluation, the classification-results tables will show only the "*positive-true"* and the "*negative-true.*" Therefore, in the case of the *speech-against-all* classification, the percent of *positively-true* classified frames will show the classification accuracy of audio files with speech content, the percent of *negatively-true* classified frames will present the classification accuracy of audio files with other content like music or any class other than speech. The aim is to achieve the higher percentage for truly classified frames in the file in order to attain the optimum classification performance.

6.1 Classification System Parameters

In this section, the key system parameter will be discussed in order to determine the most suitable values that can be fixed during system testing to emphasize the effect of using different features sets on the classification accuracy.

Frame size and frame overlap

Typically, the frame size ranges between 20ms to 40ms, with a 50% overlap will be utilized to minimize signal discontinuity [16]. During test frames, size will be set to 40ms in order to preserve the low frequencies.

Minimum frame energy

This value determines the minimum energy that above which the frame will be considered as a non-silent frame, so if the frame is non-silent it can be used in NNet training and testing. This parameter has a direct effect on

training speed and classification accuracy, a value higher than zero will be selected in order to void confusing the NNet; This confusion accrue wen the NNet fed with the multiple silent frames and each one of them belong to different classes, that will increase training time and decrease the classifier accuracy.

The test results shown in table 2 held using speech-against-all classification module that utilizes MFCC as a classification feature. Different values for minimum frame energy will be tested in order select the optimum parameter value.

Table 2: The "Minimum frame energy" parameter and its effect on classification result.

Minimum Frame Energy	Positively True Classification
0%	81.8%
5%	84.0%
10%	85.8%
15%	83.3%

The value of the "minimum frame energy" parameter will set to 10% during system training and testing.

Smoothing parameter

To improving the classification result smoothing is introduced. A simple smoothing algorithm is utilized where a window of size equal to "smoothing parameter". This parameter is used to smooth the frames classification result of a file. The smoothing result will range from "-1" to "1", all the results between "-0.25" and "0.25" will be ignored. This simple approach will improve classification accuracy; also it will reduce fluctuation in classification result. Different values for smoothing parameter will be examined using *speech-against-all* classification module and choosing MFCC as a classification feature. Table 3 shows the effect of the smoothing parameter on classification accuracy.

Table 3: The "smoothing" parameter and its effect on classification result.

Smoothing Parameter	True Classification Accuracy
0	69.5%
5	77.3%
10	80.0%
15	85.5%
25	77.3%

The value of the "smoothing parameter" parameter will set to 10% during system training and testing.

6.2 Classification Features and System Evaluation

In this section, several feature combinations will be used as a classification features in order to find the best classification feature set. Because MFCC is proven to one of the best classification feature it has been used as the main feature and is supported with some other features to improve the classification accuracy. A *one-against-all* approach will be followed in order to allow future system developed to enable the overlapped class detection.

Testing result shown in three figures, each figure will show the test number, the present of positive true classified frames, and the percent of and the negative true classified frames. Table 4 show the selected feature set in each one of the 15 tests. And the classification results of *speech-against-all*, *music-against-all*, and *other-against-all* tests presented in figures 3, 4 and 5 respectively.

6.3 Classification Results Discussion

It is clear from the testing results that the MFCC is the best feature that leads to a high accurate classification results. The rest of the features can improve the results slightly if two to three of them are combined with the MFCC. At the same time, combining more features may lead to difficulties in the system training and decrease classification accuracy.

Table 4: The utilized classification features in each test.

Test No.	Classification Features	Test No.	Classification Features
1	MFCC	9	MFCC+SP
2	MFCC+ZC	10	MFCC+EN
3	MFCC+RMS	11	MFCC+FL
4	MFCC+RL	12	MFCC+ZC+RF
5	MFCC+BR	13	MFCC+ZC+IR
6	MFCC+RF	14	MFCC+RM+RF
7	MFCC+IR	15	MFCC+ZC+RM+SP
8	MFCC+SC		

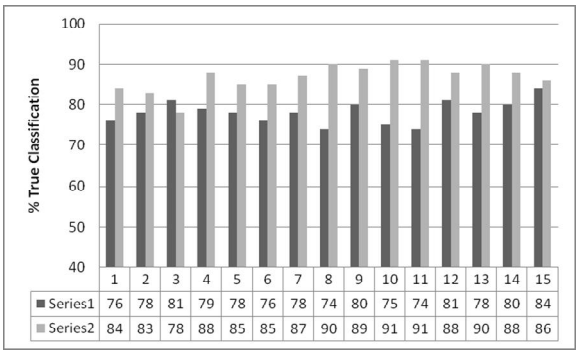


Figure 3: *Speech-against-all* classification results.

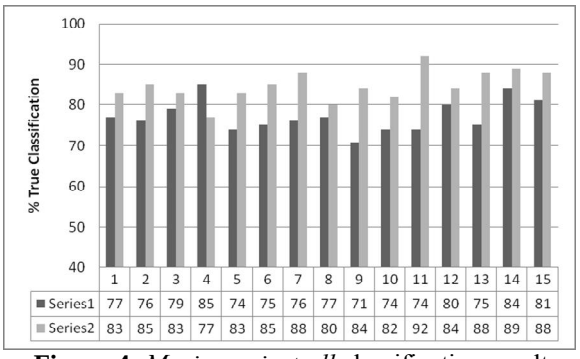


Figure 4: *Music-against-all* classification results.

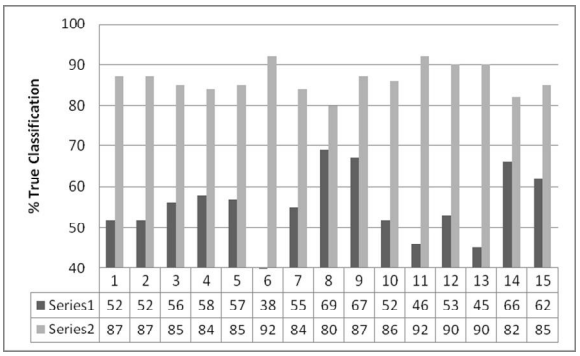


Figure 5: *Others-against-all* classification results

The first look on the result we can find that the best negative true classification accuracy was 91%, 92%, and 92% for *speech-against-all*, *music-against-all*, and *other-against-all* respectively; and best positive classification for *speech-against-all* was 84%, and for *music-against-all* was 85%, but for the *music-against-all* test the result was very low with an average equal 55.2%. But if we look to the result from another view point we can say that the NNet performed well by trying to identify the properties of both speech and music frames and classify them negatively, it has achieved the average of 86.4% of negative true accuracy. This is a good way to deal with the wide range of varieties in this class. This difficulty can be overcome by depending on the negative true classification result only and all the frames that haven not been classified as negatively true their class can be counted as other.

The overall system accuracy in all tests for all classes is 78.1%. The best achieved true classification accuracy in *speech-against-all* tests was achieved in test number 15 in which the average of positive result was 85%, for *music-against-all* it was 86.5% for positive true result achieved in test no 14, and for *other-against-all* the best achieved negative true result was 92% in tests number 6,11. This will give an overall system classification accuracy of 87.8%.

6.4 Noise Robustness System Testing

In this section the classification system will be examined against noise robustness, in order to do that three versions of the test files is created by mining the file with three noise samples, these noise samples includes brown noise, pink noise, and white noise. Table 5 shows the noise type, the Peak Signal to Noise Ratio (PSNR) for the noise sample, and the average of Signal to Noise Ratio (SNR) for each class in the test audio database after being mixed with the noise sample.

Table 5: Average SNR after adding noise to the test files.

Test Sample	Noise Sample Type	Noise Sample PSNR (dB)	Av. Speech SNR (dB)	Av. Music SNR (dB)	Av. Other SNR (dB)
1	Brown	38.0	17.0	20.3	14.8
2	Pink	46.1	24.7	28.0	22.4
3	White	42.2	21.0	24.3	18.7

Testing result shown in figures 6 to 14, these figures present the result same as figures 3 to 5. Table 6 indicates figure number, the utilized noise sample, and the examined audio classification module.

Table 6: The noise robustness tests figure details.

Mixed Noise Sample	Speech Against All	Music Against All	Other Against All
1	Figure 6	Figure 9	Figure 12
2	Figure 7	Figure 10	Figure 13
3	Figure 8	Figure 11	Figure 14

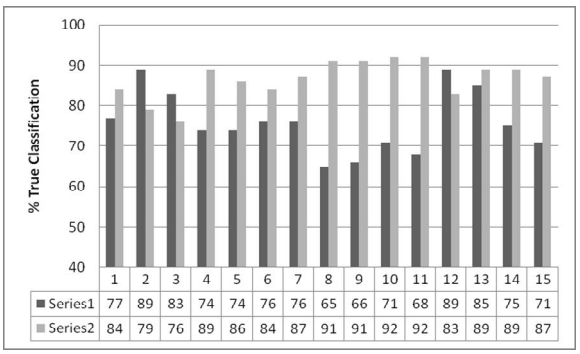


Figure 6: Results of *speech-against-all* module using noise sample 1.

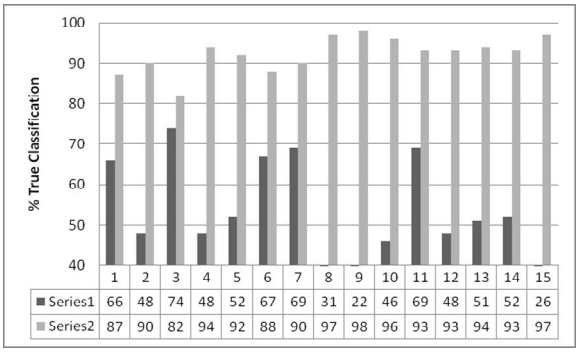


Figure 7: Results of *speech-against-all* module using noise sample 2.

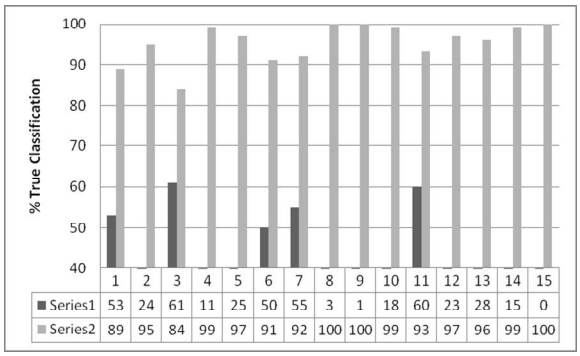


Figure 8: Results of *speech-against-all* module using noise sample 3.

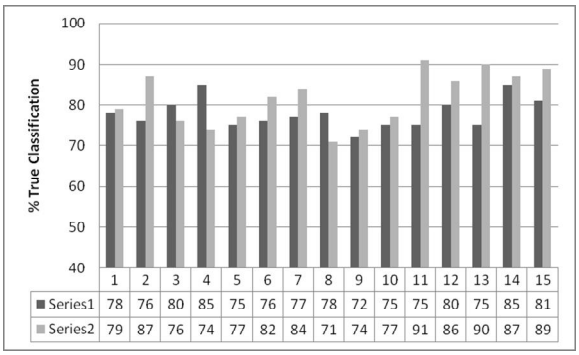


Figure 9: Results of *music-against-all* module using noise sample 1.

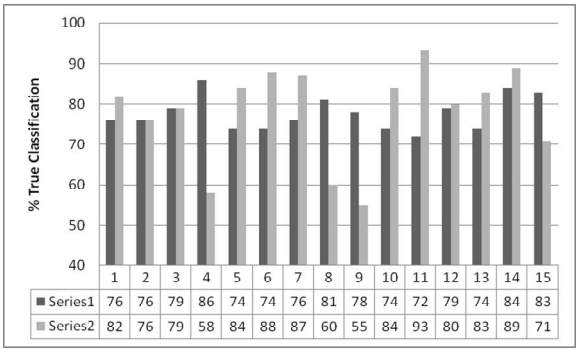


Figure 10: Results of *music-against-all* module using noise sample 2.

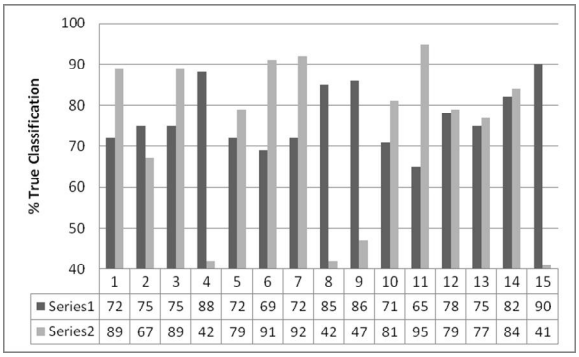
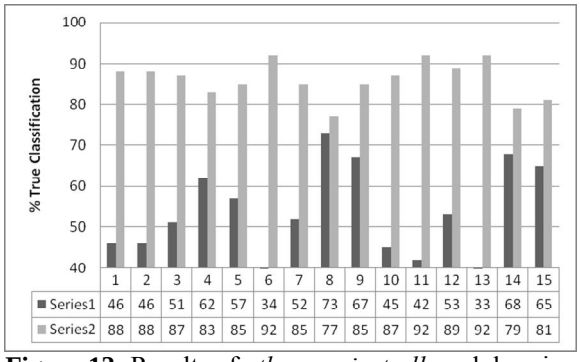
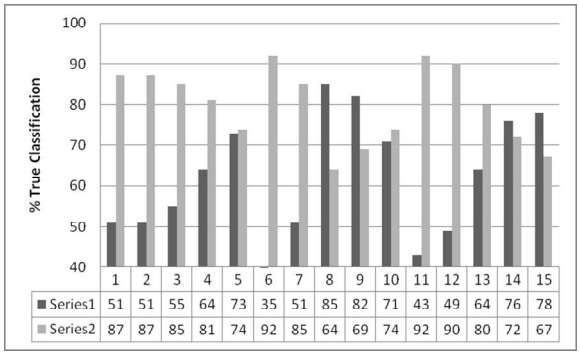


Figure 11: Results of *music-against-all* module using noise sample 3.



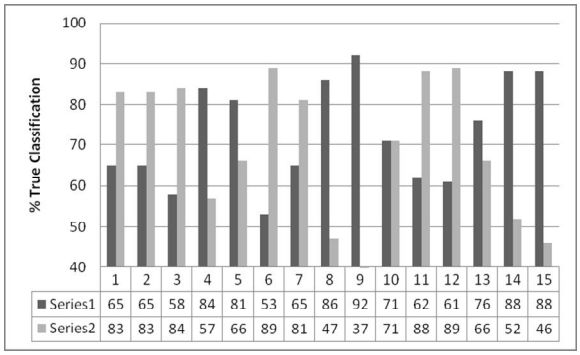
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Series1	46	46	51	62	57	34	52	73	67	45	42	53	33	68	65
Series2	88	88	87	83	85	92	85	77	85	87	92	89	92	79	81

Figure 12: Results of *other-against-all* module using noise sample 1.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Series1	51	51	55	64	73	35	51	85	82	71	43	49	64	76	78
Series2	87	87	85	81	74	92	85	64	69	74	92	90	80	72	67

Figure 13: Results of *other-against-all* module using noise sample 2.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Series1	65	65	58	84	81	53	65	86	92	71	62	61	76	88	88
Series2	83	83	84	57	66	89	81	47	37	71	88	89	66	52	46

Figure 14: Results of *other-against-all* module using noise sample 3.

6.4 Noise Robustness Result Discussion

The result will be discussed according to the introduced noise signal, for the brown noise with PSNR of 38 dB the classification result of *speech-against-all*, *music-against-all*, and *other-against-all* test results is presented in figures 6, 9, and 12 respectively. The system shows an excellent robustness against the added noise. For the pink noise with PSNR of 46.1% dB the classification result of *speech-against-all*, *music-against-all*, and *other-against-all* test results is presented in figures 7, 10, and 13 respectively. The *speech-aginst-all* system show a poor robustness especially in the positive true classification result, for the *music-against-all* test both of the positive true and negative true show a very good robustness, and for *other-against-all* the classifier reliable negative true result was very good also, such behavior is expected because the pink noise has an equal energy per octave so that the lower frequencies that contain the speech signal will have a higher noise than the other high frequency ranges. And finally for the white noise with PSNR of 42.2 dB the classification result of *speech-against-all*, *music-against-all*, and *other-against-all* test results is presented in figures 8, 11, and 14 respectively. The system show very poor robustness against the added noise, especially *speech-against-all*, and *other-against-all* tests; but the good news is that most of the ailment noises are brown or pink noises, so the system expected to have a good robustness against such amount noises.

Through all the tests the MFCC feature shows a decent robustness against the added noises.

7 CONCLUSIONS AND FUTURE WORK

This paper presents and evaluates a machine learning regime for audio content classification. Different features set were examined in and the classification results are compared; and finally the classification modules tested against three different colors noise. The tests result show that the MFCC is a good classification feature also it have a decent robustness against noise, the results showed also that adding one to three features to the MFCC can be improved classification result.

In future work, we will try to improve the work by the focus will be on the following topics:

1. Classification results can be further improved by introducing onset/offset class boundaries identification.
2. Each one of the three classes can be further classified into relevant sub classes by utilizing more dedicated algorithms that have been listed in the literature.
3. Combing the results of the three classification modules: speech, music, and others, in order to attain a mixed class classification system.
4. The utilization of the learning vector quantization algorithm that will enable the system to automatically select the best set of features in order to reach the best possible classification results.
5. Integrate system output with the MPEG-7 multimedia content description standard to enable a fast and efficient searching for the analyzed audio files.

8 REFERENCES

1 Dhanalakshmi, P., Palanivel, S., and Ramalingam, V.: 'Classification of audio signals using AANN and GMM', Applied Soft Computing, 2010

2 Panagiotakis, C., and Tziritas, G.: 'A speech/music discriminator based on RMS and zero-crossings', Multimedia, IEEE Transactions on, 2005, 7, (1), pp. 155-166

3 Pikrakis, A., Giannakopoulos, T., and Theodoridis, S.: 'A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks', Multimedia, IEEE Transactions on, 2008, 10, (5), pp. 846-857

4 Junfang, Z., Baochen, J., Li, L., and Qingwei, Z.: 'Audio Segmentation System for Sport Games', in Editor (Ed.)^(Eds.): 'Book Audio Segmentation System for Sport Games' (2010, edn.), pp. 505-508

5 Zhang, J.X., Whalley, J., and Brooks, S.: 'A two phase method for general audio segmentation', in Editor (Ed.)^(Eds.): 'Book A two phase method for general audio segmentation' (2009, edn.), pp. 626-629

6 Lie, L., and Hanjalic, A.: 'Text-Like Segmentation of General Audio for Content-Based Retrieval', Multimedia, IEEE Transactions on, 2009, 11, (4), pp. 658-669

7 Heng-Tze, C., Yi-Hsuan, Y., Yu-Ching, L., and Chen, H.H.: 'Multimodal structure segmentation and analysis of music using audio and textual information', in Editor (Ed.)^(Eds.): 'Book Multimodal structure segmentation and analysis of music using audio and textual information' (2009, edn.), pp. 1677-1680

8 Changseok, B.Y.Y., Chung; Shukran, M. A. M.; Choi, E.; Wei-Chang, Yeh: 'An intelligent classif ication algorithm for LifeLog multimedia applications', in Editor (Ed.)^(Eds.): 'Book An intelligent classif ication algorithm for LifeLog multimedia applications' (2008, edn.), pp. 558-562

9 Shao, Y., Srinivasan, S., Jin, Z., and Wang, D.: 'A computational auditory scene analysis system for speech segregation and robust speech recognition', Computer Speech & Language, 2008, 24, (1), pp. 77-93

10 Liu, S.-C., Bi, J., Jia, Z.-Q., Chen, R., Chen, J., and Zhou, M.-M.: 'Automatic Audio Classification and Speaker Identification for Video Content Analysis', 2007, pp. 91-96

11 Srinivasan, S., Petkovic, D., and Ponceleon, D.: 'Towards robust features for classifying audio in the CueVideo system'. Proc. Proceedings of the seventh ACM international conference on Multimedia (Part 1), Orlando, Florida, United States1999 pp. Pages

12 Scheirer, E., and Slaney, M.: 'Construction and evaluation of a robust multifeature speech/music discriminator', in Editor (Ed.)^(Eds.): 'Book Construction and evaluation of a robust multifeature speech/music discriminator' (1997, edn.), pp. 1331-1334 vol.1332

13 Tzanetakis, G., and Cook, P.: 'Musical genre classification of audio signals', Speech and Audio Processing, IEEE Transactions on, 2002, 10, (5), pp. 293-302

14 burred, l.: 'a hierarchical approach to automatic musical genre classification', 6th Int. conference on digital audio effects (DAFx-03), London, UK, September 8-11, 2003, 2003

15 Wold, E., Blum, T., Keislar, D., and Wheaton, J.: 'Content-Based Classification, Search, and Retrieval of Audio', IEEE MultiMedia, 1996, 3, (3), pp. 27-36

16 Hyoung-Gook Kim, N.M., Thomas Sikora: 'MPEG-7 Audio and Beyond', 2005

17 Skowronski, M.D., and Harris, J.G.: 'Improving the filter bank of a classic speech feature extraction algorithm', in Editor (Ed.)^(Eds.): 'Book Improving the filter bank of a classic speech feature extraction algorithm' (2003, edn.), pp. IV-281-IV-284 vol.284

18 Lie, L., Hong-Jiang, Z., and Hao, J.: 'Content analysis for audio classification and segmentation', Speech and Audio Processing, IEEE Transactions on, 2002, 10, (7), pp. 504-516

19 B. P. Bogert, M.J.R.H., and J. W. Tukey: 'B. P. Bogert, M. J. R. Healy, and J. W. Tukey', Proceedings of the Symposium on Time Series Analysis 1963, The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking, pp. Chapter 15, 209-243

20 Mitrović, D., Zeppelzauer, M., and Breiteneder, C.: 'Features for Content-Based Audio Retrieval', in Marvin, V.Z. (Ed.): 'Advances in Computers' (Elsevier, 2010), pp. 71-150
21 Shirazi, J., Ghaemmaghami, S., and Razzazi, F.: 'Improvements in audio classification based on sinusoidal modeling', in Editor (Ed.)^(Eds.): 'Book Improvements in audio classification based on sinusoidal modeling' (2008, edn.), pp. 1485-1488
22 Wei, C., and Champagne, B.: 'A Noise-Robust FFT-Based Auditory Spectrum With Application in Audio Classification', Audio, Speech, and Language Processing, IEEE Transactions on, 2008, 16, (1), pp. 137-150
23 Chu, S., Narayanan, S., and Kuo, C.C.J.: 'Environmental Sound Recognition With Time & Frequency Audio Features', Audio, Speech, and Language Processing, IEEE Transactions on, 2009, 17, (6), pp. 1142-1158
24 Dhanalakshmi, P., Palanivel, S., and Ramalingam, V.: 'Classification of audio signals using SVM and RBFNN', Expert Systems with Applications, 2009, 36, (3), pp. 6069-6075
25 Logan, B.: 'Mel Frequency Cepstral Coefficients for Music Modeling', International Symposium on Music Information Retrieval (ISMIR), 2000
26 Wei, H., Cheong-Fat, C., Chiu-Sing, C., and Kong-Pang, P.: 'An efficient MFCC extraction method in speech recognition', in Editor (Ed.)^(Eds.): 'Book An efficient MFCC extraction method in speech recognition' (2006, edn.), pp. 4 pp.