Adaptive Window Processing Size in Music Genre Classification Using Neural Network

Md Sah Hj Salam, Eo Shyan Ling, Lee Lay Cheng, and Noor Aina Zaidan

Abstract— Music genres classification has become an indispensable and important tool for organizing a large collection of music database. Music signal consists of huge data which leads to difficulty in getting good features prior to training and recognition. The traditional features extraction was based on fixed window size assuming that the signal is stationary within short period of time. Nevertheless, since information in music is not the same at different point, fixed window may not able to accurately extract important features. Therefore, this work reports a study in an attempt to classify music genre based on comparison between fixed and adaptive non-fixed window features extraction. Three music genres are tested in this work are classical, pop and rock. These music collections are extracted using MFCC features via both fixed and non-fixed window sized. These data are then trained and classify using MLP Neural Network. The result shows that features extraction based on adaptive non-fixed window sized perform better than fixed sized window processing. The best result for fixed window sized is 78% while with adaptive non-fixed window sized is 94%.

Index Terms— Music Genre Classification, Neural Network, Adaptive window processing, MFCC

1 Introduction

With the increasing popularity of downloading digital music through the Internet, the efficiency and accuracy of automatic music information processing have been an extremely important issue in organizing increasing collection of music data. Music genres classification has become an indispensable and important tool for the reason and considered to be a cornerstone in the research area of Music Information Retrieval (MIR). It is thought that MIR will be a key element in the processing, searching and retrieval of digital music in the near future. Nevertheless, searching and organizing the music based on their genre is still a challenging task.

Human perception of a continuous sound, such as a note from a musical instrument, is often divided into three parts: loudness, pitch, and timbre [1]. An interesting research on human classification of musical genres was proposed by Lippens *et al*. [2]. They performed a comparison test between automatic and human manual genre classification using two different datasets. Results on this research shows that human outperform machine by more than 20% for both data sets. Nevertheless, there is often only limited agreement can be achieved among humans when comes to classifying the music genres. It is due to each individual has a different knowledge and understanding on how to classify a given music. Moreover, very few genres have clear definitions as there is often significant overlap between genres, and music might belong to multiple genres. This manually process is time-consuming and expensive and therefore requires machine assistance in classifying the music genre.

This work proposed an approach of music features extraction based on adaptive windowing process. The traditional approach based on fixed processing window regards all information in the music signal as the same while the proposed window processing assumes that signal with more activities will have more information and thus should have smaller window processing so that more information can be extracted. On the other hand, signal with less activity is considered to have less information and thus bigger window size is used.

This report starts with giving reviews of previous work and follows by methodology, experimental set up, result and discussion. The report ends with over all conclusion of the work.

2 Literature Review

There were many attempts conducted in finding a good algorithm for music genre classification. Most of the previous works proceeded in two processing steps: feature extraction and classification [3]. The first one is cutting the musical signal into frames and computes the feature vectors of low-level descriptors such as timbre and rhythm. The second step, music genre classification is achieved by applying the machine learning algorithm to the feature vectors. In this phase, the patterns representing genre class are trained beforehand [4].

- *Md Sah Hj Salam is with the UTM VicubeLab, UniversitiTeknologi Malaysia, Malaysia, Johore 81310. E-mail: sah@ utm.my*
- *Eo Shyan Ling is with the Computer Science Department, Universiti Teknologi Malaysia, Malaysia, Johore 81310.*
- *Lee Lay Cheng is with the Computer Science Department, Universiti Teknologi Malaysia, Malaysia, Johore 81310.*
- *Noor Aina binti Zaidan is with the Computer Science Department, UniversitiTeknologi Malaysia, Malaysia, Johore 81310. E-mail: ainazaidan@ gmail.com*

In 1996, among the earliest approaches for automatic music genre classification was proposed by Wold et al [5]. Although they were not exactly focusing to music genres but more on general sound classes from animals, music instruments, speech and machines, the work attempted to extract the loudness, pitch, brightness and bandwidth from the signal. The average, variance, and autocorrelation of those features were computed and then were statistically plotted over the whole sound clip. Gaussian classifier was applied for the classification purpose.

Dannenberg et al. [6] proposed a machine learning approach to build classifiers and 13 low-level features were extracted from MIDI music to recognize the music styles. They recorded 25 examples each of 8 different styles and used three different supervised classifiers: Bayesian Classifier, Linear Classifier and Neural Networks to train the data. The dataset in the experiment was divided into two where 4/5 of the data were used to train the classifier and 1/5 of the data is used to test the classification. Results indicated that using 8 musical styles had the overall accuracies of 70% to 90%

Soltau et al. [7] suggested a new idea to represent temporal structures of input signal. They had proposed a new architecture Explicit Time Modelling with Neural Network (ETM-NN), where this architecture used statistical analysis of temporal structure to provide some new features to the whole network. Instead of considering its output, a MLP was trained to recognize music genres and the activation of its hidden neurons was considered as a compact representation of the input feature vector. Each hidden neuron was considered as the abstract musical event which there was not necessarily related to an actual musical representation. To build one single feature vector which was fed to a second network that implements the final decision about the genre of musical piece, the sequence of abstract events over time was then analysed to determine the music class.

Li *et al.* [8] proposed a new features extraction method for music genre classification: Daubechies Wavelet Coefficient Histogram, *DWCH* to capture the local and global information of music signals simultaneously. Through this research, the authors proved that *DWCH* significantly improves the accuracy of music genre classification where they compared the classification rate of *DWCH* features, MFCC features, Fast Fourier Transform (FFT) features, beat, and pitch with different combination of classifiers. Apparently, the classification rate on *DWCH* features for every type of classifiers (SVM, GMM, Linear Discriminant Analysis **(**LDA), KNN) turned out to be highest compared to the other features.

Ahrendt *et al.* [9] had explored with special emphasis on the decision time horizon and ranking of tapped-delay-line short-time features. The authors implemented MFCC, Linear Prediction Coefficients (LPC), Zero-Crossing Rate (ZCR), and MPEG-7 features. In order to get the most suitable features, consensus sensitivity analysis is applied. A GMM classifier and a Neural Network (NN) classifier are used. Through the research, the authors concluded that consensus ranking of features sensitivities enabled the selection of the most salient features. MFCC, LPC and ZCR showed to be most relevant, whereas MPEG-7 features showed less consistent relevance. As a result, with only the 10 best features, 70% classification accuracy was obtained using a 5s decision time horizon.

Scott [10] tested the performance of the music classification system by using NN. The four genres used were rock, classical, soul/R&B, and country and western. As a result, genre classification was performed at a success rate of 94.8%, with classical music being classified the most successfully, 96.7%, and also country and western, soul/R&B, as well as rock music being classified the least successfully at success rates of 91.0%, 93.1%, and 93.3%.

Apart from that, there are researches conducted in automatic music genre classification. Pye [11] extracted the musical features by using Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) were used as classifier to classify 6 music genres which are blues, easy listening, classical, opera, dance (techno) and indie rock. In this experiment, Gaussian Mixture Model (GMM) and Tree-based Vector Quantization (TreeQ) were investigated. The comparison of GMM and TreeQ techniques is shown at their work, where the best accuracy (92%) for the performance of music genre classification was obtained by using GMM with MFCC [11].

In general, there are many approaches conducted in the problem of music genre classification using different features and classification methods. However, most of them did extraction via the regular fixed window processing size. A different approach in doing extraction via adaptive window is explained in next section as an attempt to music genre classification.

3 METHODOLOGY

This work compared two approaches in extracting music features which are the fixed window and adaptive window. Normally, 256-sample, 512-sample, 1024-sample or 2048-sample windows are most often used for audio applications, depending on the time versus frequency resolution priorities of the application and the sampling rate [12]. A general schematic process is shown in figure 3.1 below.

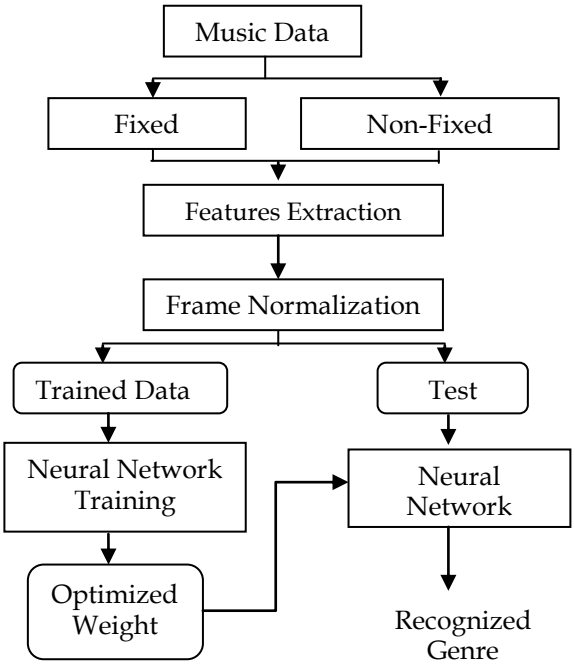

Figure 3.1: Process flow

There were 27 songs represent three music genre which are classic, rock and pop. The songs were down sampled to 22.05 kHz. Small parts of the songs in the size of 17 seconds were segmented to represent the class of the music genre. The segmented parts are taken from sound that significantly in representing the music genre. These small parts were used for training and testing.

3.1 Fixed Window

Raw audio signal divided into smaller windows. The whole raw audio signal will have M samples. Each windowed signal will have 2048 samples. The relationship between M, window size (2048) and N (number of windows) can be seen through this equation, M = 2048 x N

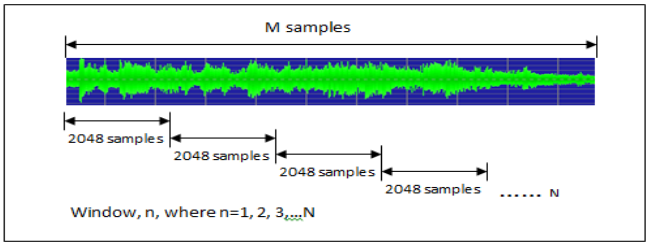

Figure 3.2: Fixed window

3.1 Adaptive Windowing

Tzanetakis *et al.* [13] proved that the classification accuracy increases significantly by the use of a window compared to the direct use of the analysis frames. In 22 kHz sampling rate, approximately 40 ms (1024-sample) or less in a window size is the best in improving classification as they provide more information than a longer window size [13]. The algorithm for dividing the signal into non-fixed processing windows was shown in Figure 3.3.

```
1.  READ signal size, buffer_size.
2.  READ signal, buffer.
3.  SET totalFrame = 0, increment = 0, frequency[100], sample[2048] = {0}
Firstly divide the signal into default size: 2048-sample window
4.  FOR window = 0 to buffer_size
      4.1  totalFrame++
      4.2  window = window + 2048
    ENDFOR
Next calculate the total of rising and falling edge (sample [i] > 0 &&
sample [i-1] < 0) of each 2048-sample window
5.  FOR frame = 0 to totalFrame
      5.1  SET subTotalFrequency = 0
      5.2  FOR i = 0 to 2048 sample
             5.2.1      sample[i] = buffer [i+increment]
             5.2.2      IF sample [i] > 0 && sample [i-1] < 0
                          5.2.2.1   subTotalFrequency++
                        ENDIF
             5.2.3      i++
           ENDFOR
      5.3  frequency[frame] = subTotalFrequency
      5.4  increment += 2048
      5.5  frame++
    ENDFOR
Get the average of the rising and falling edge for the whole signal
6.  SET totalFrequency = 0
7.  FOR j = 0 to totalFrame
      7.1  totalFrequency += frequency[j]
      7.2  j++
    ENDFOR
8.  averageFrequency = totalFrequency / totalFrame
Further divide each 2048-sample window if the total of rising and falling
edge is more than the average before proceeding features extraction
process.
9.  FOR frame = 0 to totalFrame
      9.1  IF frequency[frame] < averageFrequency
             9.1.1      Pass 2048-sample window to features extraction
                        process
      9.2   ELSE
             9.2.1      FOR counter = 0 to 2
                          9.2.1.1   Pass 1024-sample window to features
                                    extraction process
                          9.2.1.2   counter++
                        ENDFOR
           ENDIF
    ENDFOR
```

Figure 3.3: Non-fixed windows algorithm

From the above algorithm, the signal firstly being divided into default window size which is 2048 sampling points per window. The total amount of the divided window was stored in totalFrame variable. Next step proceeded in counting the total of rising and falling edge of sampling points in each window and was stored in subTotalFrequency variable. Third, subTotalFrequency for each window was being summed up (totalFrequency) and divided with totalFrame in order to get the average of rising and falling edge of sampling points for the whole signal. Lastly, if the subTotalFrequency of a particular window is greater than the average, 2048-sample window is further divided into two 1024-sample-windows before proceed to the feature extraction process. On the contrary, for the subTotalFrequency of a particular window that is less than the average, 2048-sample

window is directly feed-forward to the feature extraction process.

4 NEURAL NETWORK

Neural Network (NN) can have any number of layers, and any numbers of nodes per layer. However, more than one hidden layer can actually degrade the NN performance rather than improve it [14]. Three layers NN were used for both fixed and non fixed window size experiment: input layer, hidden layer and output layer as illustrated in figure 4.1. Both experiment used 130 nodes for input layer that represent 10 groups of MFCC features and 3 nodes for output layer that represent 3 music genres (classical, rock and pop). The numbers of hidden nodes used in experiment are 12, 20, 130 and 133. Those numbers were tested to get the highest classification rate.

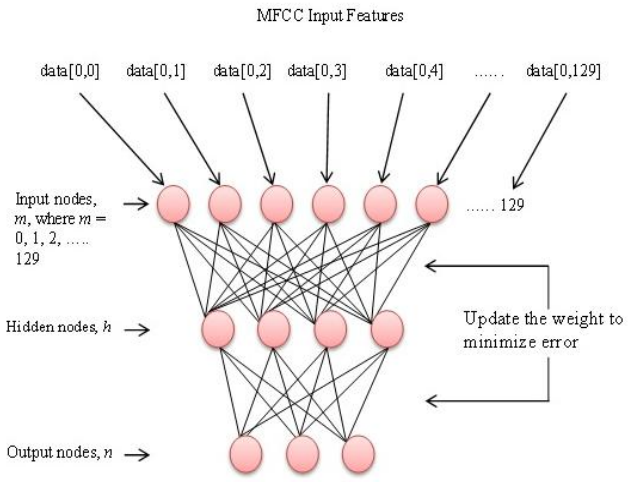

Figure 4.1: 3 layers Neural Network

Three types of experiments have been carried out to get the maximum music genre classification result: the finding of learning and momentum rate, hidden node and iteration number. These parameters control how quickly a neural net is likely to converge to a stable solution in a minimum error space.

The used of controlled variables is important for finding an appropriate pair of learning parameters. Table 4.1 shows an example of controlled variable used for the experiment.

Table 4.1: Controlled variables

Error Termination	0.0001
Number of Epoch	8000000
Network Topology	130: **h**: 3
Initial Weight Range	[-0.3: +0.3]
Learning Rate	Decided by using the result from the first phase experiments.
Momentum Rate	

The first experiment is to find a suitable pair of learning and momentum rate. Four pairs of parameters were used for the first experiment: **a** {0.1, 0.9}, **b** {0.25, 0.5}, **c** {0.5, 0.75} and **d** {1.0, 0.9} .From Table 4.2, even though **a**, **b** and **c** achieved classification rate above 90%, it can be observed that only **b** and **c** had the lower error rate compared to **a**. The convergence graph for **b** was decreasing gradually compared to **c**. Thus, for the next experiment on finding the number of hidden nodes, **b** {0.25, 0.5} was used as the constant for the learning and momentum pair.

Table 4.2: Pairs of learning and momentum rate

Variables					
Experimental		Control			
Learning Rate, η	Momentum Rate, α	Learning Iteration	Topology		
			Input Nodes	Hidden Nodes	Output Nodes
0.25	0.5	8000000	130	12	3
0.5	0.75				
1.0	0.9				
0.1	0.9				

The second experiment was about finding the suitable number of hidden nodes. The number of hidden nodes is usually about 10% the size of the input layer (Smith, 1997). Table 4.3 shows the number of hidden node used in the experiment.

Table 4.3: Number of hidden nodes

Variables					
Experimental	Control				
Hidden Nodes	Learning Rate, η	Momentum Rate, α	Learning Iteration	Topology	
				Input Nodes	Output Nodes
12 sqrt(m+n)	0.25	0.5	8000000	130	3
20 sqrt(m*n)					
130 (m)					
133 (m+n)					

5 RESULT

To get the high accuracy of music genre classification, learning rate 0.25, momentum rate 0.5 and hidden node 133 were selected to run the experiment using fixed window size. Experimental result shows that along with the increase of number of hidden nodes, the classification accuracy is increased as well. The suitable number for hidden nodes in non-fixed window size research was 20 although this value had not achieved the highest classification rate compared to 130 and 133. However, this value had the lowest error rate compared to others. Table 5.1 shows the result of music genre classification.

Table 5.1 Fixed window size - Classification accuracy (learning rate 0.25, momentum rate 0.5 and hidden nodes 133)

		Classification Accuracy (%)		
Correct Classification	**Genres**	Classical	Rock	Pop
	Classical	66.67	16.67	16.67
	Rock	0	66.67	33.34
	Pop	0	0	100

Each genre is being tested for 6 times using the data reserved for testing. For the classical genre and rock genre, it has 4 correct classifications whereas pop genre has 100% correct classification. It can be obviously seen that there is some classification confusion between rock and pop; and between classical and pop.

Table 5.2 Non-fixed window size **-** Classification accuracy (learning rate 0.25, momentum rate 0.5, hidden nodes 20)

		Classification Accuracy (%)		
Correct Classification	**Genres**	Classical	Rock	Pop
	Classical	90	0	10
	Rock	0	90	10
	Pop	30	0	70

Referring to Table 5.2, it can be seen that the best percentage lied really on the diagonal of the confusion matrix. Best scores were of 90% for classical and rock music; worse score was of 70% for pop music.

6 DISCUSSION

Learning parameters (learning rate and momentum rate) and number of hidden nodes will significantly influence the final accuracy of music genre classification whether fixed or non-fixed window size is used. During the process of finding appropriate learning parameters, it is observed that although the pair of {1.0, 0.9} has the highest classification accuracy, it is not suitable to be used as the error convergence activity is unstable which might lead to unlearned network. The experiment also indicates that if the chosen learning parameter is not suitable and the number of epoch is not enough, then it will result in an unlearned network.

It is obviously seen that along with the increase of number of hidden nodes, the classification accuracy is increased. Hence, the more interconnected weights between the layers, the better the classification accuracy.

7 CONCLUSION

It can be concluded that the extracted musical features based on fixed processing window size are informative enough for classification purpose. The processing window size needs to be changed to non-fixed where the window size is depending on the changes of audio signal. If the changes of audio signal are significant, then the size of the window must be reduced to prevent a data loss.

The research done using non fixed window sized had split non-fixedly to all the datasets where greater sampling-point changes durations were being divided into smaller windows, each containing 1024 samples; while durations that had fewer sampling-point changes, the divided window size was bigger where it contained 2048 samples. It is believed that the use of non-fixed windowing increases significantly the classification accuracy compared to the use of fixed windowing.

ACKNOWLEDGMENT

This research is supported by UTM VicubeLab at Department of Computer Graphics and Multimedia, Faculty of Computing, Universiti Teknologi Malaysia.

REFERENCES

[1] Smith, S. W., Ph.D. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. 2nd Edition. San Diego, California.: California Technical Publishing. 351-372.

[2] Lippens, S., Martens, J.P., and De Mulder, T. (2004). A comparison of human and automatic musical genre classification. *Acoustics, Speech, and Signal Processing, 2004. Proc. (ICASSP '04). IEEE International Conference on,* volume 4, iv-233-iv236.

[3]Chen, L., Wright, P., and Nejdl, W. (2009). "Improving Music Genre Classification Using Collaborative Tagging Data". *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. February 9-12. Barcelona, Spain: ACM, 84-93.

[4] Aucouturier, J. J. and Pachet, F. (2003). "Representing Musical Genre: A State of The Art". *Journal of New MusicResearch*. 32(1): 83-93.

[5] Wold, E., Blum, T., Keislar, D., and Wheaten, J. (1996). "Content-based classification, search and retrieval of audio". *IEEE Multimedia*. 3(3): 27-36.

[6] Dannenberg, R. B., Thom, B., and Watson, D. (1997). "A Machine Learning Approach to Musical Style Recognition". *Proceedings of International Computer Music Conference*, September 25-30. Thessaloniki, Greece: ICMC, 344-347.

[7] Soltau, H., Schultz, T., Westphal, M., and Waibel, A. (1998) "Recognition of Music Types". *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 12-15. Seattle, WA, USA: IEEE, 1137-1140.

[8] Li, T., Ogihara, M. and Li, Q. (2003). A comparative study on content-based music genre classification. *Proceedings of the 26th annual international ACM SIGIR*. 282-289.

[9] Ahrendt, P., Meng, A. and Larsen, J. (2004). Decision Time Horizon for Music Genre Classification using Short Time Features. *Proc. of European Signal Processing Conference (EUSIPCO)*.

[10] Scott, P. (2001). *Music Classification Using Neural Networks*. Stanford University: Project of EE37B.

[11] Pye, D. (2000). "Content-Based Methods for the Management of Digital Music". *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, June 5-9. Istanbul, Turkey: IEEE,
2437-2440.

[12] McKay, C. (2010). *Automatic Music Classification with jMIR*. McGill University: Doctoral Thesis.

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

[14] Villiers JD, and Barnard E. (1992). Backpropagation neural nets with one and two hidden layers. *IEEE Trans Neural Network*. 4: 136-41.

Dr Md Sah Hj Salam is a lecturer at Computer Science Department, Faculty of Computing, Universiti Teknologi Malaysia. He obtained his bachelor degree in Computer Science (Software Engineering) from University of Pittsburgh, PA, USA (1997), MSc (2001) and PhD(2010) in Speech processing and AI from Universiti Teknologi Malaysia. He is a member of UTM ViCubeLab Research group under Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johore. His research interests include speech segmentation and recognition, artificial intelligent and computer vision.

Eo Shyan Ling obtained her bachelor degree in Computer Science (Graphic Multimedia) from Universiti Teknologi Malaysia, Malaysia, Johore. Her research interests include speech processing and computer vision.

Lee Lay Cheng obtained her bachelor degree in Computer Science (Graphic Multimedia) from Universiti Teknologi Malaysia, Malaysia, Johore. Her research interests include speech processing and computer vision.

Noor Aina binti Zaidan obtained her bachelor degree in Computer Science (Graphic Multimedia) from Universiti Teknologi Malaysia, Malaysia, Johore. Currently, she is a Ph.D student at the Faculty of Computing, Universiti Teknologi Malaysia. Her research interests include speech processing and computer vision.