Abstract—This paper presents the technique for feature extraction to classify speech and music audio data. The combination of image processing and signal processing is used to classify audio data. There are three main steps. First, the audio data is segmented and transformed to spectrogram image and then apply image processing methods to find the salient characteristics on the spectrogram image. The next step transforms the salient spectrogram image using 2-dimensional Fourier Transform and then calculates the energy of signal at the specific frequencies to form the feature vector. Next, in classification process, Support Vector Machine is used as bi-classification tool. The method is tested on an audio database containing 510 instances with 1.5 seconds length of each. The experimental results show that the acceptable classification accuracy of our proposed technique is achieved.

Index Terms—Speech music classification, Spectrogram, Fourier Transform

I. INTRODUCTION

In the past few years, the study of speech and music classification has become an interesting research topic. There are several applications developed by using speech and music classification [1] such as, Efficient coding of multimedia sources, Automatic Speech Recognition (ASR), Automatic classification, indexing, achieving and retrieving of information from large database. The important process of those applications is to classify the audio data as either speech or music. The audio classification consists of two major parts; (i) feature extraction and, (ii) data classification. Many previous studies have been made to classify speech and music audio data, for example, Sheirer and Slaney [2] developed and evaluated different speech and music discrimination system for automatic speech recognition (ASR) to distinguish the speech segment of the soundtrack. For feature extraction process, there are two categories of effective features; Time-domain features and Frequency-domain features. Time-domain features are extracted directly from the samples of the audio signal, such as zero-crossings (ZC), root mean square (RMS) [3], Short time energy (STE), Spectrum flux [4]. Another features are Frequency-domain features represented of signal are focused on the magnitude spectral content, such as mel-frequency cepstral co-efficient (MFCC) [5], short-time fundamental frequency [6]. For feature classification process there are several classifiers have been used, such as Multilayer Neural Network [7], Support Vector Machine (SVM) [8], k-nearest neighbor [2], Hidden Markov Model (HMM) [9].

In recent survey, there is a few number of previous studies using Time-Frequency domain which considers the characteristics of audio data in time and frequency instantaneously, e.g., S. Nilufar et al. proposed the methods using visual perception of time-frequency of audio signal and combination of wavelet transform and multiple kernel learning to select the optimal subbands to discriminate audio classes [10]. There are many studies using spectrogram-based for audio classification. For instance, G. Yu and J. Slotine proposed the method to classify multiple musical instruments by treating the spectrogram as texture image with the feature extraction scheme based on time-frequency matching [11]. Y. Costa et al. compared two different textual descriptors between GLCM and LBP divided the spectrogram into several difference zones which are independently classified by different classifiers. The experimental result showed that the LBP-based archived a high overall recognition rate [12]. J. Dennis, H. D. Tran, & H. Li., proposed the classification sound event by mapping the spectrogram into a higher dimensional vector space and extracting the intensity distribution as the features [13]. It yields a robust classification accuracy in mismatch condition.

In this paper, we proposed the technique for feature extraction relying on hybrid form of spectrogram and Fourier transformation to classify speech and music audio data. There are three proposed steps. Firstly, the audio data is segmented and represented as spectrogram image using Short Time Fourier Transforms. Then, image processing methods consisting of thresholding and median filtering are used to enhance the spectrogram image. Next, a set of features are generated from spectrogram image texture analysis using local binary pattern (LBP) at the local region of spectrogram image, and combined with the feature extracted from Fourier transformation to form the an effective set of feature vectors. Next, two effective classification approaches for audio classification are investigated, i.e., Support Vector Machine (SVM) [13], [11] and Multilayer Perceptron Neural Network (MLP) [14], to measure the accuracy of feature vector extracted from our method. The benefit of SVM is that the instance data can be transformed into feature space by using a kernel function which hyperplane is a non-linear boundary.
The rest of paper is organized as follows. The proposed method is then presented in Section II. Subsequently, Experimental results and discussion are presented in Section III. Finally, the conclusion is discussed in Section IV.

II. PROPOSED METHODS

Since our method relies on the virtual representation of audio data by spectrogram, these observations have motivated us to find the feature extraction method. The overview of proposed method consists of four main steps as shown in Fig. 1: (i) preprocessing, the audio data is segmented and represented by spectrogram image, and determining the salient characteristics by thresholding and median filtering, (ii) existing features by applying texture descriptor, local binary pattern, combined with the energy of signal calculated from the transformed spectrogram image using Fourier transform at the specific frequencies to form the feature vector, and (iii) applying either SVM or MLP as a classifier. Each part is described as following Subsections.

\[ X(\tau,k) = \sum_{n=0}^{N-1} x(n) \omega(n-\tau)e^{-jn\tau} \]

where \( x(n) \) denotes the short time signal to be transformed, \( \omega(n) \) is a window function of length \( N \), \( k \) corresponds to the frequency \( f(k) = \frac{k}{N}f_s \), where \( f_s \) is the sampling frequency. Then the linear spectrogram is the squared magnitude of STFT and given as \( S_{\text{Linear}}(\tau,k) = |X(\tau,k)|^2 \). The human perception of sound is logarithmic and hence the log-spectrogram defined as \( S(\tau,k) = \log(S_{\text{Linear}}(\tau,k)) \) [10]. In our experiment, to implement the STFT in spectrogram generation, we use Hamming window function width of 23.22 ms. (512 samples) and overlapping size of 11.60 ms. (256 samples). The time-frequency matrix obtained by applying STFT is normalized into gray-scale normalized image with range scaled in [0,1] interval as follows:

\[ \hat{S}(\tau,k) = \frac{S(\tau,k) - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \]

where \( \hat{S}(\tau,k) \) denotes the normalized spectrogram image corresponding to audio data. The spectrogram generation for speech and music are shown as Figs. 2 and 3, respectively.

A. Preprocessing

1) Gray-scale Spectrogram Generation: For the first step, the audio file from dataset which is described in Section III-A, the sampling rate of 22050 Hz., audio sample size of 16 bits, a bit rate of 352 kbps, and one channel, is segmented into non-overlapping windows with 1.5 seconds length. Next, a segment is normalized with range scaled in [-1,1] interval. Each segment is transformed by applied Short Time Fourier Transform (STFT) as follows:

\[ X(\tau,k) = \sum_{n=0}^{N-1} x(n) \omega(n-\tau)e^{-jn\tau} \]
2) Binary Image Generation and Noise Reduction: To prepare the information of speech and music audio data in suitable condition, the binary image is generated by thresholding on the gray-scale spectrogram image and processing the noise reduction. The basic thresholding is given by:

\[
g(x, y) = \begin{cases} 
1 & f(x, y) > T \\ 
0 & f(x, y) \leq T 
\end{cases}
\]

where \( T \) denotes global threshold value. \( f(x, y) \) is the intensity of any point \((x, y)\) in gray-scale normalized spectrogram image obtained by (2). In this experiment, the global threshold \( T \) is set to 0.8 or intensity of gray-scale image is equal to 204. After the binary image is generated, the filtering is applied to the image to reduce noise. Median filters operate by replacing a given sample by the median of the signal value. For each binary image we use three runs of median filtering to reduce the noise occurred in image as used in a speech spectrogram segmentation algorithm [15]. The binary image for speech and music are shown as Figs. 4 and 5, respectively.

Fig. 4: Binary image of speech

Fig. 5: Binary image of music

After preprocessing process, the binary image representing the salient characteristics on spectrogram image is ready to the next step, feature extraction. In the next section, we describe some details of feature extraction by hybridization of texture feature and energy signal feature to form the feature vectors.

B. Feature Extraction

The spectrogram image obtained from the previous stage contains the distinctive patterns of the speech and music signals. Some different characteristics of speech and music signal such as the spectral of music signal contains more harmonics than that of speech signal, the music spectra change slower than speech spectra while the energy of speech signal is located at low frequency [10]. Furthermore, energy distribution of speech signal is not continuous in frequency and time axes. These observations suggest that the hybridization of spectrogram texture analysis and energy of signal using Fourier Transform will form the feature vector for speech music classification.

1) Zoning Spectrogram Image: The method called spectrogram image zoning achieved good results for music genre classification [16]. In addition, the textures of the spectrogram images of the same audio data type at the low frequencies are not different [12]. In the other word, the rationale behind zoning spectrogram is that the audio signal from the same audio data type, leads to similar intensities in the spectrogram image. Thus, in this paper, we have separated the spectrogram image to extract location information of audio data.

The speech audio data consist of many silent intervals and most of the energy is located from 0 kHz to 4 kHz, while the music audio data consist of few silent intervals, and it has continuous energy peaks for a short time and fewer frequency variations [17]. So, we take into account for separating spectrogram image from 0 kHz to 4 kHz and using texture feature extraction to obtain local feature.

Fig. 6: Zoning spectrogram image

2) Texture Feature Extraction: For extracting feature of gray-scale normalized spectrogram image from the previous stage, we take into account for Local Binary Pattern (LBP) method. The LBP was presented by Oja et al. [18] to be a model describing the texture. An important characteristic of LBP is that the difference between pixel in the window and each pixel in neighbour set of pixels. The gray value of the center pixel and the gray value of circularly neighbourhood are denoted by \( g_c \) and \( g_p (p = 0, ..., P - 1) \), respectively. The characteristics of texture are described by the joint difference distribution \( T \) obtained by the gray value of center pixel and neighbourhood pixels as follows:

\[
T \approx (s (g_0 - g_c), s (g_1 - g_c), ..., s (g_{P-1} - g_c))
\]
where
\[ s(x) = \begin{cases} 
1 & , x \geq 0 \\
0 & , x < 0 
\end{cases} \] (5)

The LBP value can be obtained by multiplying a binomial factor \( 2^p \) for each \( s(g_p - g_c) \). Finally, an image texture is transformed into a unique LBP number:
\[ \text{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \] (6)

where \( R \) denotes the distance between center pixel and neighbourhood pixel in the considered region of image.

Considering a binary code obtained from (6), Oja et al. [18] introduce a uniformity measure corresponding to the number of transitions of the binary code from 0's to 1's or vice versa. A binary code is considered to be uniform pattern if the number of transition is less than or equal to 2.

To reduce dimension of feature vectors, the uniform binary code is considered to create histogram representing the feature vectors. This descriptor called \( u \), making the LBP definition as \( \text{LBP}_{P,R}^u \). During the experiments, we choose \( P=8 \) and \( R=1 \).

### 3) 2-Dimensional Fourier Transform:
After the binarization of spectrogram image is done in the spatial domain, the binary image is converted to spectral form by the 2-D Discrete Fourier transform (DFT) described as follows:
\[ F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi(xu/M+vy/N)} , \]
\[ u \in \{0,1,...,M-1\}, \quad v \in \{0,1,...,N-1\} \] (7)

where \( f(x,y) \) is a digital image of size \( M \times N \) and \( f(x,y) \) can be obtained from \( F(u,v) \) by using the 2-D Inverse Discrete Fourier Transform (IDFT).
\[ f(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u,v) e^{j2\pi(ux/M+vy/N)} , \]
\[ u \in \{0,1,...,M-1\}, \quad v \in \{0,1,...,N-1\} \] (8)

In order to reduce the complexity to calculate equation on \( O\left((MN)^2\right) \) multiplications and additions by DFT, the Fast Fourier Transform (FFT) is performed instead of DFT. Therefore the complexity becomes \( O(MN \log_2 MN) \). Thus, we implement the FFT to obtain power spectrum or spectral density of a binary image at frequency \((u,v)\) as follows:
\[ P(u,v) = |F(u,v)|^2 \] (9)

To visualize the centered spectrum, the power spectrum image that the zero frequency \((0,0)\) is located at the center and gradually increased when approaching edges and corners is displayed.

4) Energy Signal: Considering power spectrum from 2-D Fourier transformation of speech and music audio data, as shown in Figs. 8 and 9, respectively, which are obtained from previous stage. Obviously, the vertical axes between the images are different in appearance.

These observations lead us to find the 1-D energy signal of data from speech and music transformed into power spectrum. Hence, the 1-D energy signal can be directly derived from the by 2-D Fourier power spectrum with zero horizontal frequency by following equation.
\[ X(v) = |F(0,v)|^2 \] (10)

Next, we use the Parseval’s theorem to calculate the total energy in the Frequency domain as follows:
where \(P(X)\) is the total energy of 1-D energy signal \(X\), \(X(k)\) is the Discrete Fourier Transform of a signal of length \(N\). The energy signal of speech and music audio data \(x(v)\) is shown in Fig. 10.

![Figure 10: The energy signal of speech and music at zero horizontal frequency](image)

Lastly, the input vector comprises of 60 features divided into two parts: i) 59 features obtained from the histogram described in Section B-2 and ii) 1 feature of total energy calculated by Equation (11).

**C. Classification**

From the overview of proposed method, the next stage is the classification process which each input vectors obtained from feature extraction is classified to either speech or music. To classify the input data, we use the SVM and MLP which are effective bi-classification.

1) **Support Vector Machine**: A support vector machine (SVM) is a discriminative classifier that attempts to find optimal hyperplane which separates two classes of data. Suppose a set \(S = \{(x_1, y_1), ..., (x_N, y_N)\}\) of \(\mathbb{R}^n\) is training set, where \(x_i\) is input vector, \(y_i\) is class label, \(y_i \in \{-1, 1\}\), and \(\mathbb{R}^n\) denotes \(n\)-dimensional space. To find the optimal hyperplane, the support vectors of dataset can maximize the margin, which is the distance between the hyperplane and support vectors as follows

\[
\begin{align*}
\min & \quad \frac{1}{2} ||w||^2 \\
\text{s.t.} & \quad y_i ((w^T x_i) + b) \geq 1
\end{align*}
\]

The solution to the optimization problem of SVM is given by Lagrange function as follows:

\[
L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
\]

(13)

with constraint \(\sum_{i=1}^{N} \alpha_i y_i = 0\) and \(0 \leq \alpha_i \leq C\), where \(C\) is upper bound of the Lagrange multipliers \(\alpha_i\). The kernel function \(K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle\), where \(\phi\) maps the input space into high-dimensional feature vector.

2) **Multilayer Perceptron**: The structure of Multilayer perceptron (MLP) neural network consists of three layers that are an input layer, a hidden layer, and an output layer. The input layer is composed of 60 neurons corresponding to the number of features. The number of hidden neurons depends on the value between the number of input neurons and the number of output neurons. The output layer consists of two neurons according to the number of classes. In our experiment, 10 hidden neurons are applied in the hidden layer.

The back-propagation algorithm is basically the learning algorithm for training MLP in two phases that are forward phase and backward phase. The Levenberg-Marquardt (LM) is a computationally efficient training function for back propagation neural network without having to compute the Hessian matrix. The LM algorithm derived from steepest descent method and Newton algorithm is shown as

\[
\Delta w = (J^T J + \mu I)^{-1} J^T e
\]

(14)

where \(w\) is the weight vector, \(I\) is the identity matrix, \(\mu\) is the combination coefficient, \(J\) is the Jacobian matrix containing first derivatives of the error vector, and \(e\) is error vector.

**III. EXPERIMENTAL RESULTS AND DISCUSSION**

**A. Dataset**

To evaluate the technique to classify audio data, we use audio dataset consisting of 128 files from Music Speech Collection which is separated into two categories, speech and music audio categories each of which contains 64 files. Each audio data has 30 seconds length with the sampling rate of 22050 Hz., 16 bits per sample, bit rate of 352 kbps. All of audio files are in one channel. Speech audio data set contains male and female speakers in quiet and little bit of noise. Music dataset contains a verity of music genres, e.g., pop, jazz, dance and rock.

To create training dataset, 77 files were randomly selected from speech and music dataset. Each file was randomly segmented to 10 segments with 1.5 seconds length. The total number of training data set was 770. For testing data set, 51 files were randomly selected from dataset and the total number of testing dataset was 510. In order to avoid bias, the segments from the same audio file were not in both of the training set and the test set.

**B. Classification Results**

The experiment shows the accuracy comparison of the SVM and MLP with our features. For the SVM classifier, the third-order polynomial kernel function and standard parameters have been considered. For the MLP, the number of hidden nodes is 10. The experimental result shows that depicted in Table I the our proposed feature extraction method well performs for speech and music classification. The traditional features based on ZCR and STE are compared with ours. Although the overall classification accuracy of our method are greater than that of the other, the classification accuracy for music data is greater than ours when using the MLP as a classifier. The rationale behind the slight drop of the classification accuracy for music audio data is that the spectrogram images for the music data were collected in short time and noise was also included so that some of such as images were misclassified. However, using SVM did not affect this situation.

This paper have studied the feature extraction using the characteristics of spectrogram image to classify to audio data type. However, the classification accuracy rate depending on
the some factors. For instance of the segment length of audio data, the lower segment length is more effective in multimedia retrieval [10] however, the longer segment length is more suitable for classification to gather high accuracy, this method has divided the audio data to segments of 1.5 seconds and the classification accuracy was acceptable. Therefore, the proposed method can be effectively applied to the speech and music applications.

IV. CONCLUSION

In this paper, we proposed the technique to classify speech and music audio data using the hybrid form of feature vectors by spectrogram-texture feature extraction and 2-dimensional Fourier Transformation which was used to calculate the energy of signal at specific frequency. The LBP was used to extract texture features of the spectrogram image which combined with the the total energy of the signal. The implementation on two well-known classifiers, SVM and MLP, was considered. The experimental results indicated that the acceptable classification accuracy was achieved whilst the fewer features were used to classify speech and music data.

REFERENCES


