ECG Analysis based feature extraction using Wavelet Transform for Music Genre Classification

Pranav Vijaya Kumar Rao, Vishwas Nagesh Moolimani

I. ABSTRACT

Music genre is arguably one of the most important and discriminative information for music and audio content. Music genre recognition based on Fourier Transform using MFCCs and Spectrograms have been successfully explored over the years. On the other hand, a variety of wavelet transform techniques have been used with promising results in the field of bio-signal processing and analysis. We discuss the application of wavelet transform based techniques to music signals to help generate better discriminative feature sets. We run a comprehensive evaluation of the conventional feature extraction methods with the proposed wavelet-based methods using several popular classifiers and verify that the proposed methods achieve near state-of-the-art performance on the GTZAN benchmark dataset.

II. INTRODUCTION

With the rapid development of multimedia technology, the amount of online music has accumulated so dramatically that structuring large-scale music is becoming a fundamental problem. Music information retrieval (MIR) is one of those studies, which provides a significative attempt to deal with music data. Genre classification is a basic and essential tool for MIR to analyze and process music information.

To aid classification, discriminative feature extraction is an important and necessary task. Conventionally, acoustic features such as MFCC, spectral coefficients [1] and spectrogram transformation [2] has been used for classification. Most of these techniques employ frequency analysis of the audio signal using Fourier Transform (FT). There are a few drawbacks of using the Fourier transform which can be superseded by using the Wavelet transform. Music signals are highly non-stationary with a wide dynamic range of multiple frequency components. Fourier transform only works well on stationary signals. Wavelet transform, on the other hand, performs multi-scale frequency analysis with high frequency resolution in small time range and high time resolution in low frequencies as can be seen in Figure 1.

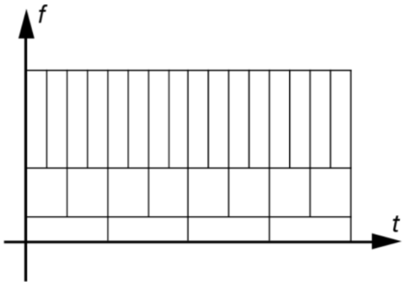

Fig. 1. Time resolution vs. Frequency resolution for Continuous Wavelet Transform

Over the past few years, wavelet-based feature extraction has been successfully explored for ECG signal analysis and classification. [3] uses wavelet packet decomposition for automatic beat classification in ECG. Discrete wavelet transform (DWT) sub-bands after dimensionality reduction [4] were used as features for arrhythmia classification. Very recently, discrete wavelet transform (DWT) coupled with 1-D hexadecimal local pattern (1D-HLP) [5] provided state-of-the-art performance for arrythmia classification. Taking inspiration from these results, we extend wavelet-based feature extraction techniques to the problem of music genre classification.

In this work, we present four concepts, namely – Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT), Dual-Tree Complex Wavelet Transform (DTCWT) and Wavelet Scattering Transform, as prospective feature extractors for music signals. Based on the results, we verify that the proposed methods achieve near state-of-the-art performance on the GTZAN benchmark dataset.

III. CONVENTIONAL APPROACHES

Traditionally, all the feature extraction techniques for music and audio signals have heavily relied on Fourier Transform and its variants. Two of the most common methods are discussed and will be further used for comparison with proposed approaches.

A. *Conventional 1D feature extraction using Fourier Transform*

Features are extracted by performing 1-D Fourier Transform on audio signals (as shown in Table I). In this section, we explain all the FT based features that are popularly used for music retrieval applications. Zero crossing rate is the rate at which the sign of the signal changes (positive to

Submitted in partial fulfillment of the final project for ECE 251C.
Vishwas Nagesh Moolimani, UCSD. Pranav Vijaya Kumar Rao, UCSD.

negative or vice versa). Spectral Centroid indicates where the "center of mass" of the signal is located. This is done by estimating the weighted mean of frequencies present in the audio track. Spectral bandwidth is the magnitude at frequency bins and is dependent on Spectral centroid. Spectral roll-off represents the frequency below which a specified percentage of the total spectral energy lies. Mel-Frequency Cepstral Coefficients (MFCCs) are a small set of features that describe the overall shape of the spectral envelope. Chroma frequencies divides the entire spectrum in to 12 bins representing 12 different semitones (or chroma) of the music octave.

TABLE I

1-D FEATURE SET USING FOURIER TRANSFORM
MFCCs
Spectral Centroid
Chroma frequencies
Spectral roll-off
Root Mean Square
Zero Crossing Rate
Spectral Bandwidth

Some of these features are very useful to differentiate different genres of music. For example, spectral centroid for blues song will lie somewhere near the middle of the spectrum while that for a metal song would be toward the end. [6] gives a comprehensive study of FT based features for audio signals.

In this work, we propose 2 wavelet based 1-D feature extractors, DWT and DTCWT to overcome the shortcomings of Fourier Transform. The classification performance of the proposed approaches is compared with the FT approach.

B. Image based feature extraction using Mel Spectrogram

Short Time Fourier Transform (STFT) is Fourier transform applied to a signal in overlapping/non-overlapping time windows. STFT provides the time-localized frequency information for situations in which frequency components of the signal vary over time.

Spectrogram is a 2D representation of the spectrum of frequencies of a signal as it varies over time using STFT. Spectrograms have many variants such as Mel-spectrogram, ERB-spectrogram, log-spectrogram based on the type of non-linearity applied to the frequency scale. All the variants provide useful information about specific spectrums in a signal. Hence, spectrograms can be used as high-level features to perform analysis on various kinds of signals. The most commonly used spectrogram for music or audio signals is the Mel-spectrogram. Spectrograms and Mel-spectrograms are inputted to a CNN which learns kernels to extract the best features which will be further used for classification.

[7] compares spectrogram with MFCC features and conclude that the spectrogram contains more details of music components such as pitch, flux, etc. [8] also proposes the spectrogram approach as it considers a zoning mechanism to perform local feature extraction. [9] introduced the network structure with '1D-CNN' to process spectrograms in music classification.

In this work, we propose alternative feature extraction techniques using continuous and discrete wavelet transform. The intuition is that fixed size windows are not appropriate for frequency analysis of music signals. Hence, variable size windows offered by wavelet transforms are used and four different approaches to perform feature extraction are proposed in the next section.

IV. PROPOSED METHODS

To perform a comprehensive comparison of traditional Fourier feature extraction approaches with wavelet techniques, we propose four different methods of feature extraction using various forms of continuous and discrete wavelet transforms:

A. Discrete Wavelet Transform (DWT) [10]

As mentioned before, there are two types of wavelet transforms; Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). Mathematically, CWT is defined by the following equation

$$X_w(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \qquad (1)$$

where $\psi(t)$ is the continuous mother wavelet which gets scaled by a factor of a and translated by a factor of b.

The main difference for DWT in (1) is that it uses discrete values for the scale and translation factors. Moreover, scale factor increases in powers of two and translation factor increases by integer values. It is also important to note that DWT is only discrete in scale and translation domain, not in time-domain.

Due to the constraints of scale, we can say that in practice, DWT is always implemented as a filter-bank. Filter banks are an efficient way of splitting a signal into several frequency sub-bands. This means that it is implemented as a cascade of high-pass and low-pass filters.

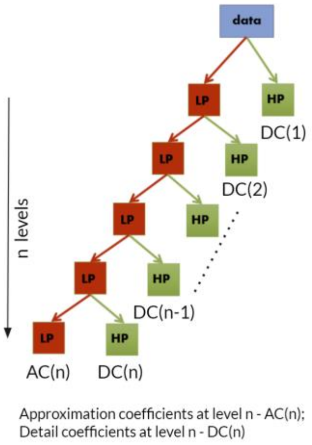
data
LP
HP
DC(1)
LP
HP
DC(2)
LP
HP
LP
HP
DC(n-1)
LP
HP
AC(n)
DC(n)
n levels
Approximation coefficients at level n - AC(n);
Detail coefficients at level n - DC(n)

Fig. 2. Schematic representation of DWT

Figure 2 shows the filter design used in DWT. There are n levels of decomposition where n is the maximum decomposition level which is reached when the number of samples in our samples in our signal is smaller than length of the wavelet filter. At each level, scale factor increases by 2 (down-sampled by 2) and frequency factor decreases by 2.

DWT returns 2 sets of coefficients, approximation coefficients (output of low pass filter) and detail coefficients (output of high pass filter). Coefficient set is determined by concatenating approximation and details coefficients as shown in Table II.

TABLE II
ELEMENTS OF DWT COEFFICIENT SET

[AC(n), DC(n), DC(n-1), DC(n-2), …, DC(2), DC(1)]

Feature set is computed by extracting statistical metrics from the coefficient set (obtained in Table II). The feature set (shown in Table III) is then passed through machine learning models to estimate the performance of classification.

TABLE III
STATISTICAL METRICS EXTRACTED FROM COEFFICIENT SET

FEATURE SET
Entropy
5th percentile of data
25th percentile of data
75th percentile of data
95th percentile of data
Median
Mean
Standard deviation
Variance
Root mean square
Number of zero crossings
Number of mean crossings

B. *Extracting and Down sampling Scalogram Image Features [10]*

The scalogram is the absolute value of the continuous wavelet transform (CWT) of a signal, plotted as a function of time and scale, where scale refers to inverse frequency. Scalogram can be useful for real world signals with multiple scales and hence are used for feature extraction of audio signals.

Computing the scalogram of an audio signal is computationally intensive and produces scale x sample size coefficients which can be large for music signals. Hence, the scalogram coefficients need to be critically down-sampled to create a 2D image feature set. Two methods of down-sampling are employed – max pooling and average pooling.

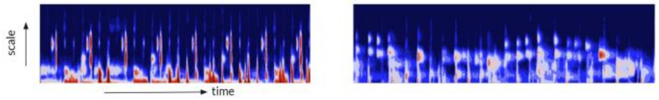

Fig. 3. 2D scalogram image features for two different genres – ‘pop’ and ‘jazz’, shows the differences between the two

After down-sampling, each audio signal is converted into a 2D scalogram image as shown in Figure 3, which act as inputs to a CNN model as given by the pipeline in Figure 4. The custom defined CNN model, as shown in Figure 5, learns appropriate which kernels which extract high level features useful for music genre classification.

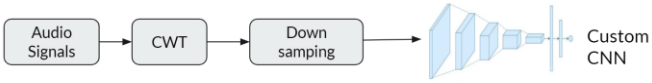

Fig. 4. Complete pipeline of the proposed method B

Two different Convolution models are used based on 1D convolution and conventional 2D convolution. From Table VI, the test accuracy on scalograms give promising results but the values are lesser than the FFT based spectrogram approach. The lower accuracy can be attributed to the excessive need for down sampling.

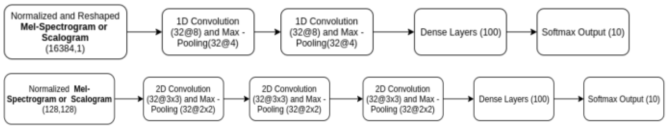
Normalized and Reshaped Mel-Spectrogram or Scalogram (16384,1)
1D Convolution (32@8) and Max - Pooling(32@4)
1D Convolution (32@8) and Max - Pooling(32@4)
Dense Layers (100)
Softmax Output (10)
Normalized Mel-Spectrogram or Scalogram (128,128)
2D Convolution (32@3x3) and Max - Pooling (32@2x2)
2D Convolution (32@3x3) and Max - Pooling (32@2x2)
2D Convolution (32@3x3) and Max - Pooling (32@2x2)
Dense Layers (100)
Softmax Output (10)

Fig. 5. Proposed CNN architecture for scalogram feature extraction

C. *Dual-Tree Complex Wavelet Transform (DTCWT) [11]*

The dual-tree complex wavelet transform is a relatively recent enhancement to DWT. For complex modulated signals like audio, DWT encounters few shortcomings; oscillations, shift variance, aliasing and lack of directionality.

Since wavelets are bandpass functions, the wavelet coefficients tend to oscillate between positive and negative values around singularities. This considerably complicates wavelet-based processing. Additionally, a small shift of the signal greatly perturbs the wavelet coefficient oscillation pattern around singularities. The wide spacing of the wavelet coefficient samples, or the fact that wavelet coefficients are computed via iterative discrete-time down-sampling operations interspersed with nonideal low-pass and high-pass filters, resulting in substantial aliasing. Finally, while Fourier sinusoids in higher dimensions correspond to higher dimensional plane waves, the standard tensor product construction of M-D wavelets produces a checkerboard pattern that is simultaneously oriented along several directions.

Fourier transform does not suffer from these problems since it is based on complex-valued oscillating sinusoids. Inspired by Fourier representation, complex wavelet transforms are designed where the real component $\psi_r(t)$ and the imaginary component $\psi_i(t)$ form Hilbert transform pair.

DTCWT is an effective approach for implementing complex wavelet transform (shown in Figure 6). It employs two real DWTs, the first DWT gives the real part of the transform while the second DWT gives the imaginary part. Analysis and synthesis filter banks are used to implement the DTCWT.

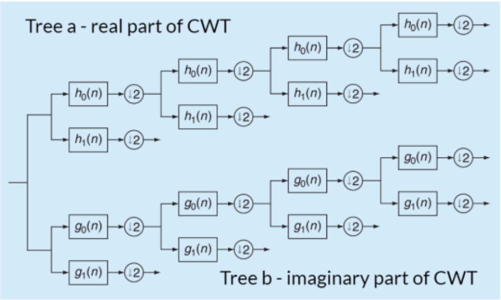

Fig. 6. Analysis FB for DTCWT

The two real DWTs use two different set of filters, with each satisfying the PR conditions. The two sets of filters are jointly designed so that the overall transform is approximately analytic. Let $\mathrm{h}_0(\mathrm{n})$, $h_1(n)$ denote the low-pass/high-pass filter pair for the upper FB and let $g_0(n)$, $g_1(n)$ denote the low-pass/high-pass filter pair for the lower FB.

D. *Wavelet Scattering Transform [12]*

Proposed approach IVB is limited as there is loss of information when reducing the huge set of CWT coefficients to image size features. To combat this loss, a new method of Wavelet Scattering transform is proposed. Wavelet scattering employs a filter-bank approach to derive low variance features from real valued time series signals based on scalogram coefficients. The scattering framework uses predefined wavelet and scaling filters to define scattering coefficients in one shot without any learning.

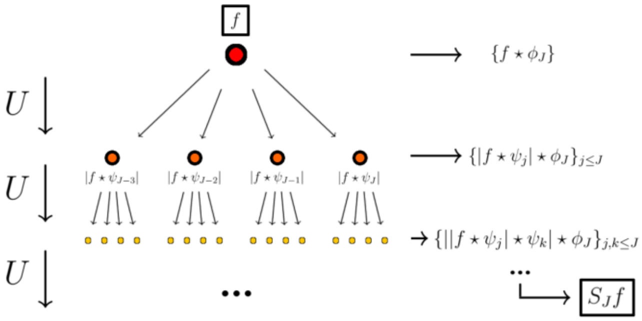

Fig. 7. Schematic layout of the scattering transform's computation

`

Figure 7 is a tree representation of scattering transform computation. At every level the signal *f* is convolved with the respective wavelet filter ψ_j. $|f \star \psi_j|$ at each node gives the scalogram coefficient and $\{|f \star \psi_j| \star \phi_J\}$ gives the scattering coefficient at every level of the tree.

The basic building block wavelet used is the Morlet wavelet. Morlet wavelet is used because it can capture short bursts of repeating and alternating music notes. The scaling filter used is a general gaussian low pass filter. An important characteristic of wavelet scattering is, the derived features are insensitive to translations defined on the invariance scale and are continuous with respect to deformations. The support of the scaling function determines the size of invariance scale and the support of the wavelet filter is also restricted to the same. Figure 8 represents the support of both the filters for an invariance scale = 0.5s

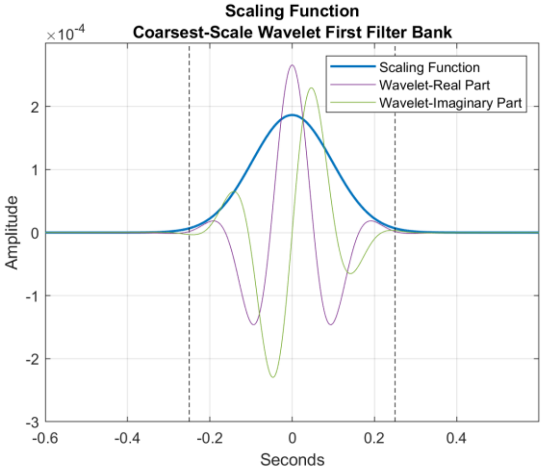

Fig. 8. Support of scaling and wavelet function

To produce scattering coefficients, a 2-level filter-bank is realized with Q factor of 8 for level 1 and 1 for level 2. Q factor specifies the number of filters per octave in the respective level. The feature set composed of scattering coefficients is then passed through machine learning classifiers to estimate performance characteristics.

V. Comparison

A. *Dataset*

Our proposed methods are experimented on the benchmark GTZAN [13] dataset. The GTZAN dataset consists of 1000 audio tracks of 30 seconds duration. All tracks are 22,050Hz, Mono 16-bit audio files in .wav format. The GTZAN dataset has been widely used in many studies with the aim of music genre classification. It was collected and proposed in [14]. The genre labels and numbers of corresponding genres are given in Table IV. The test – train split used is 80 % training and 20% testing.

TABLE IV
GTZAN DATASET SPECIFICATIONS

Genre	Number of Tracks
Classic	100
Jazz	100
Blues	100
Metal	100
Pop	100
Rock	100
Country	100
Disco	100
Hip-hop	100
Reggae	100

B. *Results*

In this section, we compare the performance of the proposed wavelet-based feature extraction techniques with their conventional Fourier Transform counterparts.

For DWT (proposed in Section IVA), the extracted feature set is test on various classification algorithms like Random Forest, Gradient boosting, Kernel SVM, etc. However, since Random Forest (RF) and Gradient boosting (GB) performed the best, we restrict the depiction of results to only these two algorithms. We experiment with different number of estimators (1500, 2000, 4000) for each algorithm. Additionally, this algorithm was experimented with different orders of Daubechies wavelet (upto db20). Daubechies wavelet is used as it was empirically found to perform the best for many applications. Table V contains the comparison of top 5 best performing models using DWT feature extractors. For example, db12_GB4000 implies that choice of wavelet is db12 and gradient boosting with 4000 estimators is used as the classifier. As compared to 1-D FT performance (in Table IV), we can see that DTCWT brings in 14.5% increase in Test accuracy.

TABLE V
DWT PERFORMANCE COMPARISON

FEATURE EXTRACTION	TEST ACCURACY (%)
db12_GB4000	**81.5**
db17_GB2000	78
db5_RF2000	78
db19_RF2000	78
db9_RF4000	77.5

For the Scalogram approach (proposed in Section IVB), the scalogram image is obtained after max-pool down-sampling of the CWT coefficients. The resultant scalogram images are passed through 2 different architectures of CNN as shows in Figure 5. Table VI contains the performance results of the CNN models using the scalogram features.

TABLE VI
SCALOGRAM PERFORMANCE COMPARISON

FEATURE EXTRACTION	TEST ACCURACY (%)
1D – Convolution Model	57
2D – Convolution Model	**60**

For DTCWT (proposed in Section IVC), the extracted feature set was tested on the same set of classifiers mentioned in the context of DWT. For all audio signals, we perform 17 levels of decomposition to extract the wavelet coefficients. Table VII contains the comparison of different bi-orthogonal and qshift filters. The exact filters used are mentioned in the Appendix. As compared to 1-D FT performance (in Table IV), we can see that DTCWT brings in 18% increase in Test accuracy.

TABLE VII
DTCWT PERFORMANCE COMPARISON

FEATURE EXTRACTION	TEST ACCURACY (%)
14_GB4000	**85**
14_GB2000	84
14_GB1500	84
15_RF2000	83.5
24_GB4000	83

For the Wavelet Scattering (proposed in Section IVD), the hyperparameters for extracting different feature sets are invariance scale (IS) in seconds (s), level of decomposition, number of filters in each level. [8 1] represents a 2-level decomposition which 8 filters per octave in the respective level. Table VIII contains the performance result of a few variants of wavelet scattering with the best combination.

TABLE VIII
WAVELET SCATTERING PERFORMANCE COMPARISON

FEATURE EXTRACTION	TEST ACCURACY (%)
IS_0.5 - [8 2 1]	86.5
IS_1 - [8 1]	84
IS_0.5 - [8 1]	**88**

Table IX shows the best performance for each of the proposed wavelet-based feature extractors and compares it with conventional FT techniques. We can conclude that Wavelet-based feature extractors perform significantly better.

TABLE IX
SUMMARY OF RESULTS

FEATURE EXTRACTION	TEST ACCURACY (%)
1-D FT	67
Mel Spectrogram	73
DWT	**81.5**
Scalogram	**60**
DTCWT	**85**
Wavelet Scattering	**88**

VI. CONCLUSION

In this project, we propose four variations of feature extraction using wavelet-based transforms for music genre classification. The proposed methods were directly inspired from the successful wavelet-based feature extraction models used in bio-signal processing and analysis. We perform a comprehensive comparison of these methods with the conventional Fourier-based techniques by stating the shortcomings of the Fourier transform. Owing to the strong feature extraction capability of wavelet transforms, our proposed methods achieve near state-of-the-art performance on the benchmark GTZAN dataset.

In future, better classifiers can be designed specifically to handle multi-scale features. Deeper learning models can also be used to classify the extracted feature sets. Further exploration into designing such better classifiers would provide better results for various audio classification tasks.

REFERENCES

[1] Andén, J., & Mallat, S. (2011, October). Multiscale Scattering for Audio Classification. In ISMIR (pp. 657-662). [Online]. Available: https://www.di.ens.fr/data/publications/papers/ismir-final.pdf

[2] Costa, Y. M., Oliveira, L. S., Koericb, A. L., & Gouyon, F. (2011, June). Music genre recognition using spectrograms. In 2011 18th International Conference on Systems, Signals and Image Processing (pp. 1-4). IEEE. [Online]. Available: http://www.ppgia.pucpr.br/~alekoe/Papers/ALEKOE-IWSSIP2011.pdf

[3] Faziludeen, S., & Sabiq, P. V. (2013, April). ECG beat classification using wavelets and SVM. In 2013 IEEE Conference on Information & Communication Technologies (pp. 815-818). IEEE. [Online]. Available: https://www.researchgate.net/profile/Sabiq_Pv/publication/261355238_ECG_beat_classification_using_wavelets_and_SVM/links/546c43cf0cf2397f7831d1b2/ECG-beat-classification-using-wavelets-and-SVM.pdf

[4] Martis, R. J., Acharya, U. R., & Min, L. C. (2013). ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. Biomedical Signal Processing and Control, 8(5), 437-448. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809413000062

[5] Tuncer, T., Dogan, S., Pławiak, P., & Acharya, U. R. (2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. Knowledge-Based Systems, 186, 104923. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0950705119303727

[6] Lartillot, O., & Toiviainen, P. (2007, September). A Matlab toolbox for musical feature extraction from audio. In International conference on digital audio effects (pp. 237-244). [Online]. Available: http://www.academia.edu/download/3248943/p237.pdf

[7] Kong, Q., Feng, X., & Li, Y. (2014). Music genre classification using convolutional neural network. In Proc. Int. Soc. Music Inform. Retrieval (ISMIR). [Online]. Available: https://pdfs.semanticscholar.org/5477/9685b293a5afa2cd3107f479bf4503a99e82.pdf

[8] Costa, Y. M., Oliveira, L. S., Koericb, A. L., & Gouyon, F. (2011, June). Music genre recognition using spectrograms. In 2011 18th International Conference on Systems, Signals and Image Processing (pp. 1-4). IEEE. [Online]. Available: http://www.din.uem.br/yandre/iwssip_2011.pdf

[9] Bian, W., Wang, J., Zhuang, B., Yang, J., Wang, S., & Xiao, J. (2019, August). Audio-Based Music Classification with DenseNet and Data Augmentation. In Pacific Rim International Conference on Artificial Intelligence (pp. 56-65). Springer, Cham. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1906/1906.11620.pdf

[10] Ahmet Taspinar, "A guide for using the wavelet transform in machine learning", unpublished. [Online]. Available: http://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/

[11] Selesnick, I. W., Baraniuk, R. G., & Kingsbury, N. G. (2005). The dual-tree complex wavelet transform. IEEE signal processing magazine, 22(6), 123-151. [Online]. Available: https://scholarship.rice.edu/bitstream/handle/1911/20355/Sel2005Nov1TheDualTre.PDF

[12] Andén and Lostanlen, "Music Genre Classification using Wavelet Time Scattering". MATLAB. [Online]. Available: https://www.mathworks.com/help/signal/examples/music-genre-classification-using-wavelet-scattering.html

[13] *GTZAN Genre Collection*. http://marsyas.info/downloads/datasets.html

[14] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1021072/

Appendix

As mentioned before, different types of bi-orthogonal and qShift filters are used for DTCWT feature extraction (as shown in Tables X and XI). The feature extraction information shown in Table VI for DTCWT performance comparison depicts the filter index used. For example, 14_GB4000 uses 1st filter from Table X and 4th filter from Table XI. These filters are not explained in detail as they are out of scope for our project.

TABLE X
TYPES OF BIORTHOGONAL FILTERS

Filter Index	Filter Name
1	Antonini 9,7 tap filters
2	LeGall 5,3 tap filters
3	Near-Symmetric 5,7 tap filters
4	Near-Symmetric 13,19 tap filters

TABLE XI
TYPES OF QSHIFT FILTERS

Filter Index	Filter Name
1	Q-shift 10,10 tap filters (only 6,6 non-zero taps)
2	Q-shift 10,10 tap filters
3	Q-shift 14,14 tap filters
4	Q-shift 16,16 tap filters
5	Q-shift 18,18 tap filters