CLASSIFICATION OF MUSIC INSTRUMENTS USING WAVELET-BASED TIME-SCALE FEATURES

Farbod Hosseyndoust Foomany[1] and Karthikeyan Umapathy[2]
Ryerson University[1,2]
ffoomany@rnet.ryerson.ca[1], karthi@ee.ryerson.ca[2]

ABSTRACT

Separation of sounds from different sources plays a significant role in success of auditory scene analysis and multimedia content recognition. In this paper, we propose wavelet-based features for discrimination of signals from various music instruments. One hundred and fifty-two music segments from thirteen different instruments were selected from a public music database (Universitat Pompeu Fabra). We performed automatic instrument classification of segments from 13 instruments using selected wavelet features which resulted in accuracy as high as 85%. The wavelet features, along with the considerations suggested and elaborated on here, while are successful for solving the problem at hand, could be applied to many signal processing problems in other domains.

Index Terms— Wavelet Analysis, Music Instruments, Feature Extraction, Pattern Classification.

1. INTRODUCTION

Automatic classification of events in audio channels of multimedia streams has numerous applications in multimedia content recognition and retrieval. Lew et al. (2006) [12] specified two requirements for multimedia information retrieval systems: capability of searching for a particular item and facilitating the process of browsing/summarizing the collections. A majority of the existing works, in this area, is on multi-group audio and speech classifications [7, 18] with only a few works focusing exclusively on music instrument classification.

Music sounds are described by four different perceptual attributes: pitch, loudness, duration and timbre [16]. Timbre refers to the perceptual qualities of sounds or, what they sound like [8]. Herrera-Boyer et al. (2003) [9] in a review of works on music separation specified that in the timbre space analyzed by Grey and Gordon (1978) [6] only one dimension (spectral centroid) significantly correlated with the perceptual dimension. Since timbre, is not a mathematically defined quality, some studies have aimed at correlating it with well defined features [11]. Others, ignoring the perceptual qualities, have pursued the goal of devising features that optimally discriminate between music instruments [14].

In the realm of wavelet analysis for music classification, studies have been focused on extraction of characteristics such as instantaneous frequency, pitch and sub-band energies from wavelet coefficients. In an earlier work, Delprat et al. (1992) [4] had succeeded to employ continuous wavelet coefficients and Gabor coefficients for extracting instantaneous frequency. Similarly Dai et al. (2008) [3] proposed a method for estimation of instantaneous frequency by wavelet ridge extraction. Kostek and Czyzewski (2001) [10] were among the people who used wavelet sub-band analysis for music classification. Ozbek et al. (2012) [15] reviewed the previous works on spectral and wavelet music classification and contended that "features based on parametrizing the wavelet sub-band energies have been found efficient but they lack the time domain information". Wavelet transform provides powerful time-frequency analysis tools for various applications in which temporal characteristics of non-stationary signals are studied [2, 13]. Here, we use wavelet features to capture temporal and spectral qualities of the music signal simultaneously. As mentioned earlier, spectral centroid is shown to provide some discrimination power in music classification. Nevertheless the centroid frequency varies even for a particular sound of an instrument, e.g. striking mallet [8]. Therefore we need to model the temporal evolution of the features. This conforms to the suggestions of Ozbek et al. (2012) [15].

By using the continuous wavelet transform, not only can we capture the qualities related to center frequency (and dominant scale), but we can characterize the evolution of these traits in time. We propose wavelet features that capture three categories of traits in a signal: 1) Wavelet-based bandwidth characteristics 2) Dominant wavelet scale (resembling centroid frequency) and 3) Wavelet-based temporal variation of dominant scale. We believe that the proposed features capture meaningful traits in signal that could be useful in many signal processing applications.

2. METHOD

2.1. Database and Pre-processing

The database used in this study is from Music Technology Group at Universitat Pompeu Fabra [19]. Thirteen pieces of music were chosen from various mixtures so that they represent a range of music instruments. The signals were preprocessed in four steps of re-sampling, silence removal, filtering, and splitting into segments.

The piece from each instrument was split into segments of 1.4 seconds (with no overlap) after silence removal and filtering. All signals were filtered with a Butterworth filter of 30-7000Hz and all the signals were re-sampled to 16KHz. Total of 152 segments were obtained after these steps in the database. The instrument-set comprised of 7 different guitar styles (the difference between the sounds of clean guitar and slag-distorted guitar was very minimal), 2 styles of drum and 4 other instruments: piano, cello, pad3, and harmonica.

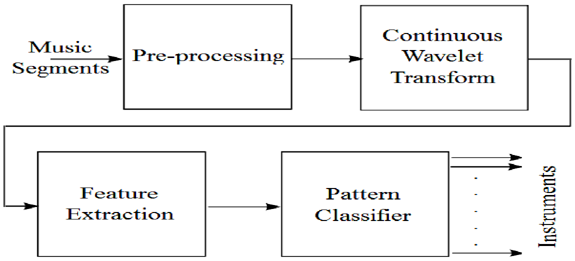

Fig. 1. Block diagram of the proposed method.

2.2. Wavelet Features

The continuous wavelet transform (CWT) of the signal *x(t)* could be shown as *T(a,b)* where *a* and *b* are scale and shift respectively and $\varphi(t)$ is the wavelet basis function [2]:

$$T_x(a,b)=1/\sqrt{a}\int_{-\infty}^{+\infty}x(t)\varphi(\frac{t-b}{a})dt \qquad (2)$$

We can show the contribution of each scale to the total energy by this equation (Addison, 2005) [1]:

$$E=\frac{1}{C_g}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left|T_x(a,b)\right|^2/a^2dadb=\int_{-\infty}^{+\infty}\left|x(t)\right|^2 \qquad (3)$$

in which *Cg* is admissibility constant for the wavelet basis function and could be ignored for our relative features.

We extract the following features from CWT of the signal in line with our motivation of capturing the three traits in the signal:

1. Scale Distribution Width (SDW): Scale distribution width is the difference between the scales, at which the CWT of the signal reaches its half (or a predefined portion of) its maximum energy per scale. The energy per scale in this sense is the averaged energy for a particular scale, over all the shift values (in the frame). The averaging process and how SDW is calculated is demonstrated in Fig. 2. The idea behind using this feature is to represent the organization of signal around a dominant scale. This feature has been employed in the literature in the context of analysis of electrocardiogram signal by Umapathy et al. (2011) [17].

Due to the non-linear relation between scales/frequency, SDW feature is influenced by the variation in the location of the dominant scale. However this non-linear relation also provides possibilities to derive many variations of SDW. It could be seen from wavelet equation (2) that for a particular scale, *Ta* is the convolution of signal with $\varphi(\frac{-t}{a})/\sqrt{a}$. Therefore an inverse relation exists between the scale and the frequency of the wavelets (*a1/a2=f2/f1* for two pairs of scale(*a*) and frequency*(f)*). Sequential choice of scales (for example from 1 to 150), as shown in Fig. 3 gives a poor resolution for higher frequencies. This could be compensated, by exponential choice of scales as we suggest. It is demonstrated in the bottom panel.

It is notable that while choice of scales affects the resolution, hence, the accuracy of calculations, it will not influence the value of SDW (difference between the scales for which the energy is half its maximum) and is still influenced by the variation in the location of the dominant scale.

To compensate for this, phenomenon, we suggest log-SDW.

2. Log Width (log-SDW): If *a1* and *a2* are the scales at which CWT energy reaches its half maximum, log-SDW is defined as:

$$\log\text{-}SDW=\log(a1)-\log(a2) \qquad (4)$$

Based on the above discussion it could be seen that log-SDW serves as a bandwidth measure in log-frequency domain (*log(a1/a2)=-log(f1/f2)*).

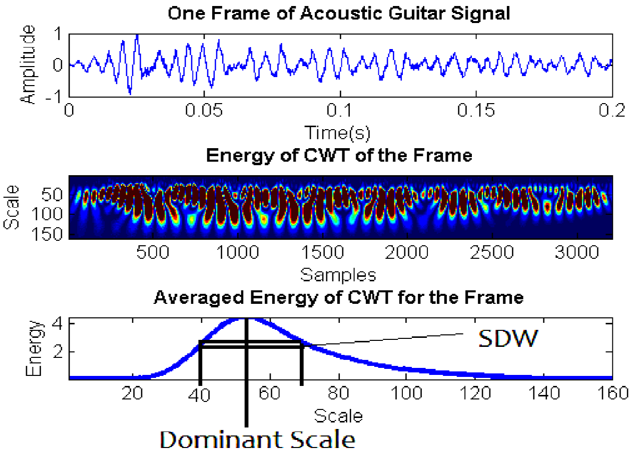

Fig. 2. A frame of audio signal from acoustic guitar (top) energy of CWT coefficients (middle) and demonstration of how SDW and dominant scales are calculated (bottom)

3. Dominant Scale: This is the scale at which the energy of CWT coefficients (averaged over frame) is maximum.

4. Time Variance of Dominant Scale (TVDS): The need for wavelet temporal features was highlighted in the introduction section. The features introduced up to now, are based on averaging over frame and do not seize the advantages of wavelet transform in time domain. To solve this problem, we suggest the TVDS which is the standard deviation (STD) of values of dominant scale at each sample (shift). Contrary to SDW and dominant scale, for this

feature, we do not average the energy over the frame. Instead we calculate the dominant scale, at each shift, and calculate the standard deviation of the scales for the frame.

5. Wavelet Mean of Inverse Scale (WMIS): Based on the inverse relation between the frequency and scale, we extend the idea of centroid frequency to the continuous wavelet domain by defining WMIS as:

$$WMIS = \sum_{b}\sum_{a}\frac{|T_x(a,b)|}{\sum_{a}|T_x(a,b)|}*(1/a) \qquad (5)$$

in which *a* and *b* are scale and shift respectively and *T* is the continuous wavelet transform as defined earlier. WMIS is calculated for each frame, but takes into account all shift values (*b*).

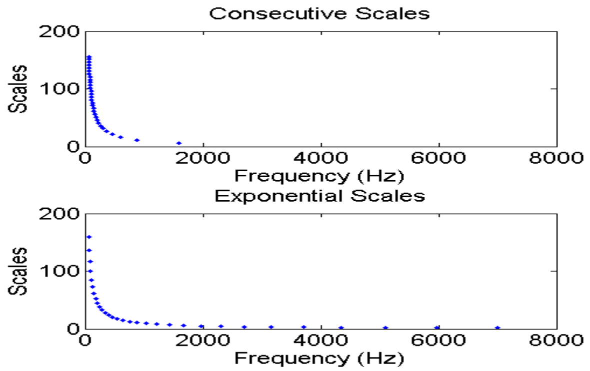

Fig. 3. Frequency-scale relation and effect of choice of scales in wavelet analysis: Top panel shows the scales from 1 to 150 and their frequency correspondence. Bottom panel demonstrates the scales scattered exponentially in the same range (giving a better frequency resolution for high frequencies)

2.3. Classification and Evaluation Methods

For each segment the signal was split into the overlapping frames of 0.2 seconds, and features were calculated on these frames (after proper hamming windowing). The value assigned to the segment was the median of the feature values of the frames. In this study for SDW variations and dominant scale, we discarded the frames in which the maximum averaged energy is less that 80% of its maximum for all the frames in the segment. Gaussian wavelet of order 6 was used as mother wavelet. For all classification tasks we employed linear discrimination analysis (LDA) classifiers with leave-one-out cross-validation method (Duda et al. 2001) [5] as a simple classifier to analyze and compare the power of features (rather than classifiers).

3. RESULTS AND DISCUSSIONS

Table 1 summarizes the results for selected feature sets. Box-plots for 6 feature sets are plotted in Fig. 4. CF and frequency Bandwidth are displayed in addition to 4 proposed wavelet features for comparison purposes. Frequency bandwidth was computed from the power spectrum of the signals. For the last combination of three features (log-SDW +TVDS-S+ WMIS), most classes had accuracy of 100% (except instruments 3, 8, 9, 11 and 13).

Table 1. Results of single-level instrument classification

Feature-Set	Classification Accuracy
CF	49.3%
Bandwidth	48.7%
CF + Bandwidth	71.7%
SDW	47.4%
TVDS-S	50.0%
WMIS	65.8%
log-SDW	13.2%
TVDS+ WMIS	82.9%
SDW+TVDS+ WMIS	80.3%
log-SDW +TVDS+ WMIS	85.5%

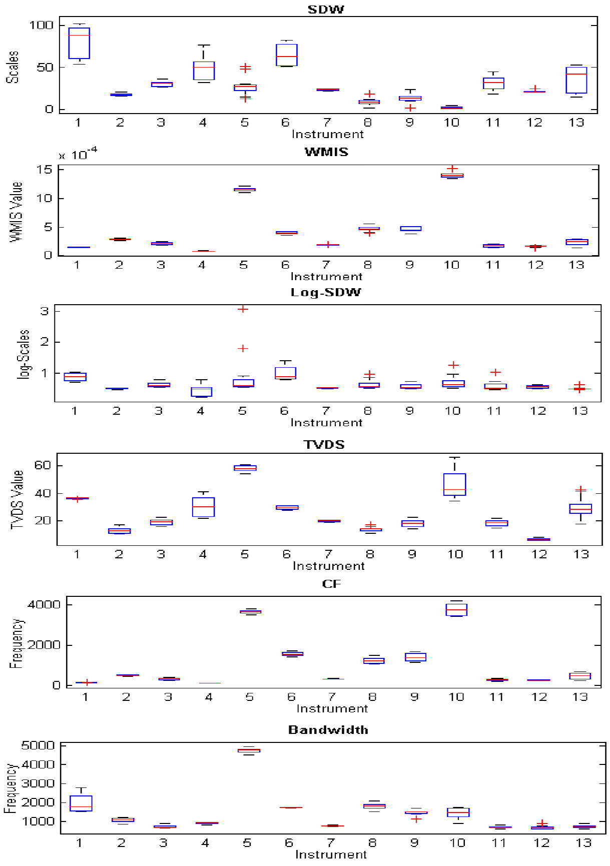

Fig. 4. Box-plots of the feature values for 13 instruments (the y-axis shows the values and x-axis shows the groups/instruments)

Analyzing the results in Tables 1 and revisiting our motivation of characterizing music instruments with 3 simple signal traits in wavelet domain (Bandwidth i.e. richness of content; dominant scale, which is pitch related, and temporal variation of dominant scale which is rhythm related) we observe that the proposed features did yield good results. It is interesting to note that the combination of

TVDS and WMIS performed well. TVDS captured the temporal variation in the wavelet domain and justified the literature in the need for such features while WMIS is a good alternative to CF. Non-linear scale frequency relation and choice of wavelets added flexibility to the WMIS feature. TVDS captures a different trait of signal (compared to centroid frequency and bandwidth), which is only obtainable through wavelet analysis. TVDS along with WMIS provided the best results with SDW and log-SDW in two different schemes of classifications. Log-SDW provided additional information in the direct 13-groups classification where SDW did not help the feature combinations.

4. CONCLUSION

Although various spectral/cepstral and time-frequency features exist in the literature for audio classification, this work attempted to present five simple wavelet features capturing meaningful signal traits for classifying music instruments. We accommodated the highlighted need for extracting temporal variations in wavelet domain as well. We tested these feature on 152 music segments from 13 different music instruments, from a public music database. The results and discussions not only demonstrated the potential of the proposed features in this particular application but highlighted the possibility of using time-scale and time-frequency features which have meaningful association with music signal traits.

5. REFERENCES

[1] P. S. Addison, “Wavelet transforms and the ECG: a review,” Physiological measurement, vol. 26, p. R155, 2005.

[2] P. S. Addison, J. Walker, and R. C. Guido, “Time–frequency analysis of biosignals,” Engineering in Medicine and Biology Magazine, IEEE, vol. 28, no. 5, pp. 14–29, 2009.

[3] Y. Dai, Q. Ma, and W. Tang, “Efficient wavelet ridge extraction method for asymptotic signal analysis,” Review of Scientific Instruments, vol. 79, no. 12, pp. 124703–124703, 2008.

[4] N. Delprat, B. Escudie, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani, “Asymptotic wavelet and Gabor analysis: extraction of instantaneous frequencies,” Information Theory, IEEE Transactions on, vol. 38, no. 2, pp. 644–664, 1992.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification. 2nd,” Edition. New York, 2001.

[6] J. M. Grey and J. W. Gordon, “Perceptual effects of spectral modifications on musical timbres,” The Journal of the Acoustical Society of America, vol. 63, p. 1493, 1978.

[7] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” IEEE Trans. Neural Netw., vol. 14, no. 1, pp. 209–215, Jan. 2003.

[8] S. Handel, “Timbre Perception and Auditory Object Identification,” In Moore (ed.) Hearing. New York, Academic Press., 1995.

[9] P. Herrera-Boyer, G. Peeters, and S. Dubnov, “Automatic classification of musical instrument sounds,” Journal of New Music Research, vol. 32, no. 1, pp. 3–21, 2003.

[10] B. Kostek and A. Czyzewski, “Representing musical instrument sounds for their automatic classification,” J. Audio Eng. Soc, vol. 49, no. 9, 2001.

[11] J. Krimphoff, S. McAdams, and S. Winsberg, “Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique,” Le Journal de Physique IV, vol. 4, no. C5, pp. 5–5, 1994.

[12] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-based multimedia information retrieval: State of the art and challenges,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 2, no. 1, pp. 1–19, 2006.

[13] S. G. Mallat, A wavelet tour of signal processing. Academic Pr, 1999.

[14] J. Marques and P. J. Moreno, “A study of musical instrument classification using Gaussian mixture models and support vector machines,” Cambridge Research Laboratory Technical Report Series CRL, vol. 4, 1999.

[15] M. E. Ozbek, N. Ozkurt, and F. A. Savaci, “Wavelet ridges for musical instrument classification,” Journal of Intelligent Information Systems, vol. 38, no. 1, pp. 241–256, 2012.

[16] G. Peeters, S. McAdams, and P. Herrera, “Instrument sound description in the context of MPEG-7,” in Proceedings of the 2000 International Computer Music Conference, 2000, pp. 166–169.

[17] K. Umapathy, F. H. Foomany, P. Dorian, T. Farid, G. Sivagangabalan, K. Nair, S. Masse, S. Krishnan, and K. Nanthakumar, “Real-time electrogram analysis for monitoring coronary blood flow during human ventricular fibrillation: Implications for CPR,” Heart Rhythm, vol. 8, no. 5, pp. 740–749, 2011.

[18] K. Umapathy, S. Krishnan, and R. K. Rao, “Audio signal feature extraction and classification using local discriminant bases,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 4, pp. 1236–1246, 2007.

[19] M. Vinyes, MTG MASS database, 'http://www.mtg.upf.edu/static/mass/resources' [Subsets include: (Artist: Sargon, Song: Silenci, Kcleta Studios), (Artist: Bearlin (I. Calvo and J. Rabascall), Producer: Sergi Vila & Bearlin), ("Que Pena / Tanto Faz" by T. Curvemusic), ("Remember the Name" by F. Minor, Warner Bros. Records), ("Ana" by Vieux Farka Touré) ,("TV on" by Kismet, Studio Moskou)].