Two Kinds of Audio Feature Selection Algorithms in Wavelet Domain

De Li, DaYou Jiang
Department of Computer Science
Yanbian University
Yanji, China
Email: leader1223@ybu.edu.cn
ybdxgxy13529@163.com

Jongweon Kim*
Department of Contents and Copyright
Sangmyung University
Seoul, Korea
Email:jwkim@smu.ac.kr

Abstract—**In this paper, we propose two new audio feature selection algorithms based on Discrete Wavelet Transform (DWT) domain. One of them is combined with Discrete Cosine Transform (DCT), the other one is combined with Mel Frequency Cepstral Coefficients (MFCCs). First, we introduce two different audio selection algorithms; second, we use the audio attack experiments to verify reliability of those algorithms. The tests show that the DWT-DCT algorithm has better stability under most of audio attacks, but not is robust to amplify attack for electronic music, while DWT-MFCCs algorithm is especially stable under amplify attack, but not is robust to mp3 compression attacks.**

Keywords-Audio feature selection; DWT; DCT; MFCCs

I. Introduction

In recent years, several front-ends have been proposed in the field of audio feature extraction. Some of them are based on short-term features, such as Fast Fourier Transform coefficients (FFTC) [1], DWT coefficients [2], DCT coefficients [3], MFCCs [4], real cepstral coefficients (RECC) [5], log filterbank energies [6], Perceptual linear Prediction (PLP) [7], log-energy, spectral flux, zero-crossing rate (ZCR) [8] and fundamental entropy. Others are based on the application of different temporal integration techniques over these short-term features. A. Meng [9] proposed a multivariate auto-regressive feature model which gives two different feature sets, the diagonal auto-regressive and multivariate auto-regressive features. P. Ruvolo [10] extended his previous work on Spectro-temporal box filters (STBFs) by proposing a hierarchical approach to combine features at multiple time scales. A. Suhaib [11] proposed a new method for audio feature extraction which is by using Probability Distribution Function (PDF). The PDF is a statistical method which is usually used as one of the processes in complex feature extraction methods such as Gaussian Mixture Models (GMM) and Principle Component Analysis (PCA). X. Y. Zhang [12] proposed a new time-frequency audio feature extraction scheme, in which features are decomposed from a frequency-time-scale-tensor. The tensor, derived from a weight vector and a Gabor dictionary in sparse coding, represents the frequency, time centre and scale of transient time-frequency components with different dimensions. I. Vatolkin [13] proposed an approach on how evolutionary multi-objective feature selection can be applied for a systematic maximisation of interpretability without a limitation to the usage of only interpretable features. In the experiments, 636 relevant low-level audio features and 566 high-level audio features were used. H. Muthusamy [14] proposed the particle swarm optimization based clustering (PSOC) and wrapper based particle swarm optimization (WPSO) to enhance the discreming ability of the features and to select the discriminating features respectively. In the experiments, MFCCs, linear predictive cepstral coefficients (LPCCs), PLP features, gammatone filter outputs, timbral texture features, stationary wavelet transform based timbral texture features and relative wavelet packet energy and entropy features were extracted.

In this paper, we proposed two audio selection algorithms in wavelet transform domain, which combined DCT and MFCCs. They have good stability to a series of audio attacks. Features of audio music are expressed as binary image after combined transformation.

The details of proposed algorithms will be addressed in following sections. In the next Section 2, we study MFCCs and DWT. Then Section 3 explains the proposed two different audio selection algorithms, the DWT-MFCCs and DWT-DCT. Section 4 shows the results of the audio attacks experiments with respect to two proposed algorithms to demonstrate the performance and stability of those algorithms. And finally, Section 5 gives the conclusions and future works.

II. Background and Related Work

A. DWT

DWT represents an analog signal in the time-frequency domain with Sine and Cosine functions and the coefficients are calculated by using Mallat's pyramid algorithm [15]. The general procedure for DWT is illustrated in Figure.1. DWT decomposes a signal to approximation coefficients and detail coefficients by applying low-pass and high-pass filters respectively. The detail coefficients can be sent to another set of filters for further decomposition. The filterbank implementation of wavelets can be interpreted as computing the wavelet coefficients of a discrete set of child wavelets for a given mother wavelet ψ (t). In the case of the discrete wavelet transform, the mother wavelet is shifted and scaled by powers of two.

$$\psi_{j,k}(t)=\frac{1}{2^j}\psi(\frac{t-k2^j}{2^j}) \qquad (1)$$

Where *j* is the scale parameter and *k* is the shift parameter,

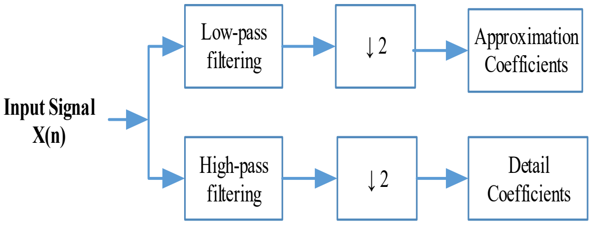

Figure 1. The general procedure for DWT

B. DCT

The DCT is a real transform that has great advantages in energy compaction. The DCT is actually shift variant, due to its cosine functions. Its definition for spectral components $f_y(k)$ is

$$f_y(k)=\begin{cases}\sqrt{\frac{1}{N}}\sum_{n=0}^{N-1}x(n)\cos\frac{k(2n+1)\pi}{2N} & k=0\\ \sqrt{\frac{2}{N}}\sum_{n=0}^{N-1}x(n)\cos\frac{k(2n+1)\pi}{2N} & k=1,2,\dots\end{cases} \qquad (2)$$

Where x(n) is a frame sequence of audio signal, the length of which is N. The $f_y(k)$ is the coefficients sequence obtained by applying DCT.

C. Algorithm of MFCCs feature selection

MFCCs were firstly introduced by Davis and Mermelstein in the 1980's, which is one of the most popular features for speech recognition [16]. In the past few decades, MFCC became popular parameterization method that has been developed and has been widely used in the speech technology field. MFCC analysis is similar to cepstral analysis and yet the frequency is warped in accordance with Mel-scale. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum.

The MFCC features for a segment of music file are computed using the following procedures:

Step1: The host audio clip is loaded as signal s(n). Taking it through a preemphasis filter, after through high-pass filter, the signal s'(n) = s(n) - a × s(n-1). Set the variable a as 0.95.

Step 2: Devide the signal into short frames si(n) with frame duration as F_d and frame step as F_s. In our scheme, we set the F_d as 256 and F_s as 128, where i=1,2,⋯,Nf, and Nf is the number of frames .

Step 3: Take FFT to the signal and calculate a periodogram spectral estimate of the pow spectrum pi(k).

Step 4: Apply the mel-frequence cepstral coefficients to the pow spectral, sum the energy in each filter. In our scheme, we set the filter number as 24.

Step 5: Take the Discrete Cosine Transform (DCT) of the logarithm of all filterbank energies.

Step 6: Keep DCT coefficients 6-9, discard the rest.

Step 7: Calculate two order difference coefficients of the DCT coefficients 6-9 and calculate the mean value of parameters in each frame.

Step 8: Choose the middle 1025*2 values F_{dc} from frames number 1024 then set them number from 1 to 2050. Compute the mean value of the neighbouring two values by the following formula:

$$Fp(i)=(F_{dc}(2*i-1)+F_{dc}(2*i))/2 \quad i=1,2,\dots \qquad (3)$$

Step 9: Apply binarization algorithm by the following formula:

$$Fp(i)=\begin{cases}1 & Fp(i)>Fp(i+1) \quad i=1,2,\dots\\ 0 & otherwise\end{cases} \qquad (4)$$

III. TWO DIFFERENT AUDIO FEATURE SELECTION ALGORITHMS

In this paper, we proposed two audio feature selection algorithms. The first one DWT-DCT is based on dual domains, while the second one is using MFCCs, which indicates the short time scale features. The feature values are finally expressed as binary image with size 32 × 32. So the audio features are represented using 1024 points.

A. DWT-DCT

The main characteristic of DWT is multi-resolution. After taking DWT to audio signal, the audio signal is decomposed into time domain and frequency domain by different scales corresponding to different frequency ranges. The approximation components indicating the low frequency components of the signal can effectively resist various attacks. The DCT has the property of decorrelation and the DC coefficient of DCT has good stability. The proposed audio feature selection algorithm based on DWT-DCT is shown in Figure 2.

The procedure of the proposed DWT-DCT feature selection algorithm is described as follows:

Step1: The host audio clip is loaded as signal s(n). Divide the signal into short frames $s_i(n)$. Each frame has 512 points, where i=1,2,⋯,Nf, and Nf is the number of frames.

Step 2: Devide each short signal $s_i(n)$ into 4 shorter frames $s_{ip}(n)$ with frame duration as F_d. where ip=1,2,3 and 4.

Step 3: Perform 2-level DWT with 1-coefficients Daubechies wavelet (Db1) to each shorter frames $s_{ip}(n)$.

Step 4: Apply 1-D DCT to approximation components, then take the DC coefficient.

Step 5: Obtain the mean value of 4 DC coefficients as the feature F_c of each frame.

Step 6: Transform F_c into integer by the following formula:

$$Fc=\begin{cases}\lfloor Fc\rfloor & ,if\ Fc-\lfloor Fc\rfloor<0.5\\ \lceil Fc\rceil & otherwise\end{cases} \qquad (5)$$

Where $\lfloor\ \rfloor$ and $\lceil\ \rceil$ are the ceil function and floor function.

Step 7: Apply binarization algorithm by the following formula:

$$Fc = \begin{cases} 1 & if \ \mathrm{mod}(Fc, 2) == 0 \\ 0 & otherwise \end{cases} \qquad (6)$$

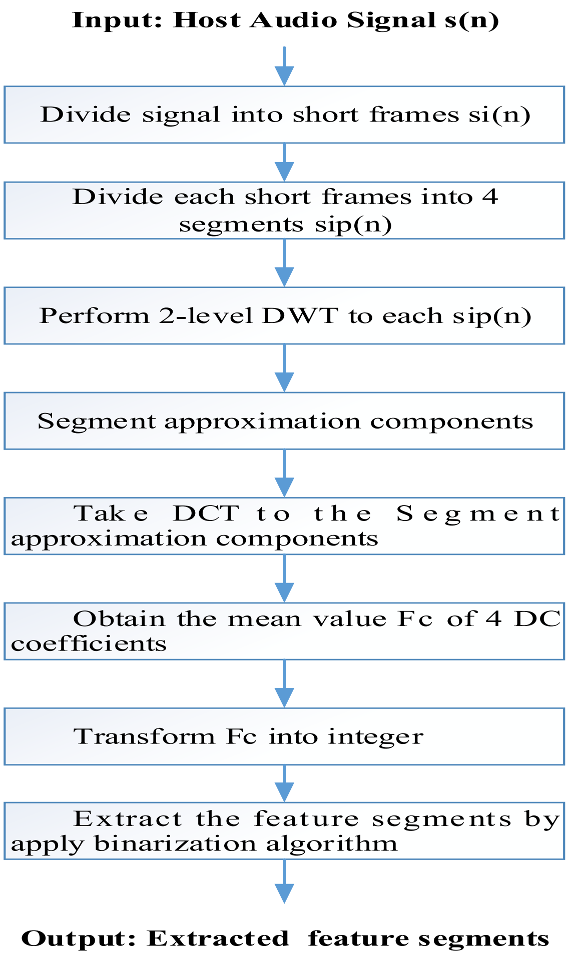
Input: Host Audio Signal s(n)
Divide signal into short frames si(n)
Divide each short frames into 4 segments sip(n)
Perform 2-level DWT to each sip(n)
Segment approximation components
Take DCT to the Segment approximation components
Obtain the mean value Fc of 4 DC coefficients
Transform Fc into integer
Extract the feature segments by apply binarization algorithm
Output: Extracted feature segments

Figure 2. Follow chart of the DWT-DCT algorithm

B. *DWT-MFCCs*

The design inspiration of DWT-MFCCs algorithm mainly comes from MFCCs. The low frequency components of the signal in DWT domain have higher robustness against many attacks. And the middle coefficients of MFCCs are stable. So, we can combine the DWT with MFCCs to select the feature of audio signal. The procedures of DWT-MFCCs algorithm shown in Figure 3 is virtually the same as MFCCs. The minor difference between them is that the former should perform 1-level DWT transformation on signal audio to get the approximation components for the next procedures.

IV. AUDIO ATTACK TESTS

To test our attacks,we have chosen three sound files with variant characteristics:

Electronic Music (see Figure 4).

“I am ready” : Pop music by Bryan Adams (see Figure 5).

“Mark the knife” : Jazz music by Westlife (see Figure 6).

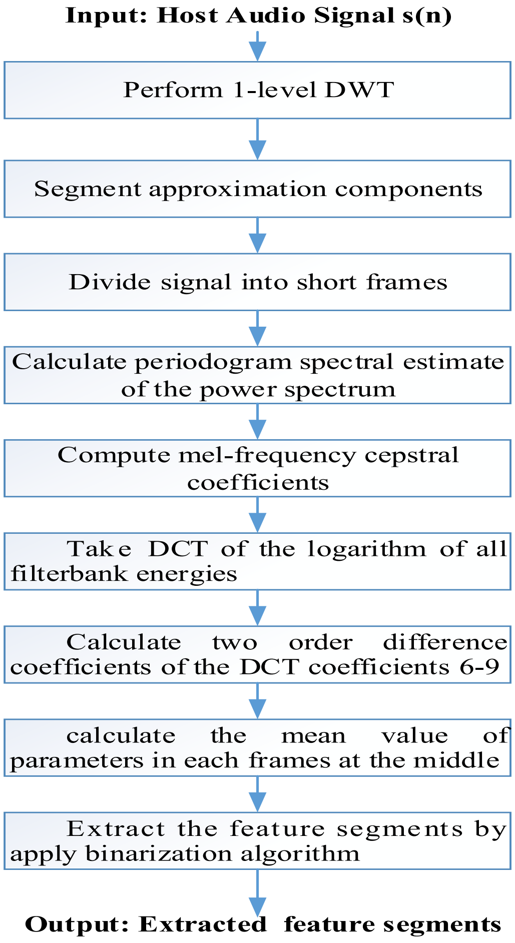
Input: Host Audio Signal s(n)
Perform 1-level DWT
Segment approximation components
Divide signal into short frames
Calculate periodogram spectral estimate of the power spectrum
Compute mel-frequency cepstral coefficients
Take DCT of the logarithm of all filterbank energies
Calculate two order difference coefficients of the DCT coefficients 6-9
calculate the mean value of parameters in each frames at the middle
Extract the feature segments by apply binarization algorithm
Output: Extracted feature segments

Figure.3. Follow chart of the DWT-MFCCs algorithm

Figure 4. Electronic music

Figure 5. Pop music

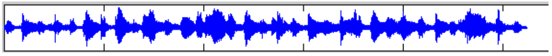

Figure 6. Jazz music

A. *Attacks type*

All of the audio signals in the test are music with 16bits/sample, 44100Hz sample rates, and 12 seconds. In order to illustrate the robustness of the algorithm, common signal attacks and audio stir-mark attacks are used. The following attacks are chosen:

Common signal attacks:

Noise addition: White noise with 20 dB of the power is added.

Delay: A delayed copy of the original is added to it. Set the delay time of 50 ms and a decay of 10%.

Echo: An echo signal with a delay of 50 ms and a decay of 10% is considered.

Re-quantization: Re-quantization of a 16-bit audio signal to 8-bit and back to 16-bit. Re-quantization of a 16-bit audio signal to 32-bit and back to 16-bit.

Re-sampling: the original audio signal is down sampled to 22050Hz and the up sampled back to 44100Hz.

MPEG compression: The Adobe Audition 3.0 was used to perform coding and decoding with bit rates 128Kbit and 64Kbit.

Audio Stir-mark attacks:

Noise addition: add white noise addbrumm_100.

Low-pass filtering: the resistor capacitor circuit (RC) low-pass filter with cutoff frequency of 11025 Hz is applied.

Addsinus: Add Sinus attack with frequency in 900 Hz and amplitude as the value of 1300.

Amplify: set the volume down to 50%.

In the study, reliability was measured as the bit error rate (BER) of the extracted feature, and its definition is:

$$BER = \frac{BE}{NL} \times 100\% \qquad (6)$$

Where BE and NL are respectively the number of erroneously detected bits and the gross feature bits.

B. Experimental results

Tables 1 and 2 show the results under the audio attacks mentioned above.

TABLE I. ATTACKS RESULTS FOR DWT-DCT ALGORITHM (BER)

attacks	*Pop*	*Electronic*	*Jazz*
addnoise (20dB)	0	0.0127	0.001
addbrumm_100	0.001	0.0059	0.0029
addsinus	0	0.0166	0.002
amplify (0.5)	0.0068	0.1172	0.0301
echo_50_10%	0.001	0.0176	0.0059
delay_50_10%	0	0.042	0.0273
rc_lowpass	0	0.001	0.001
Requantization 8	0	0.0029	0
Requantization 32	0	0	0
resampling	0	0	0
mp3 128kbps	0.0088	0.2041	0.0447
mp3 64kbps	0.0088	0.2041	0.0457

As shown in Table 1, since that the DC values have nice stability, the DWT-DCT algorithm resists most attacks, especially under add noise, add sinus, requantization and re-sampling attack. After mp3 compression, the feature values didn't change a lot for Pop music and Jazz music. Only for Electronic music, the robustness under amplify attack and mp3 compression attacks is lower than the other music.

TABLE II. ATTACKS RESULTS FOR DWT-MFCCS ALGORITHM (BER)

attacks	*Pop*	*Electronic*	*Jazz*
addnoise (20dB)	0.0264	0.0417	0.0296
addbrumm_100	0.0176	0.0078	0.0269
addsinus	0.008	0.032	0.0164
amplify (0.5)	0	0	0
echo_50_10%	0.0352	0.0521	0.0449
delay_50_10%	0.0469	0.0703	0.0623
rc_lowpass	0	0	0.002
Requantization 8	0.0137	0.0234	0.0273
Requantization 32	0	0	0
resampling	0	0	0
mp3 128kbps	0.3554	0.5344	0.3961
mp3 64kbps	0.3554	0.5344	0.3971

As shown in Table 2, the DWT-MFCCs algorithm is robust to most attacks, especially under amplify, rc-low-pass filtering and re-sampling attack. The limitation is that it cannot resist to MP3 compression. Because compression and reductive process, though the format of audio changed back to .wav, the length of the signal has been changed.

Comparing the experimental results of DWT-MFCCs algorithm with the DWT-DCT algorithm, we can find that the DWT-DCT algorithm is generally much better under these attacks. But for audio segment like the electronic music tested in the paper, the audio feature selection algorithm using the DWT-MFCCs algorithm can be more robust under amplify attack.

V. CONCLUSIONS

We proposed two audio feature selection algorithms based on wavelet domain. To verify the reliability of the algorithms, three different audio music signals were tested under a series of attacks including common signal attacks and audio stir-mark attacks. The experimental results show that the DWT-DCT algorithm has better stability than DWT-MFCCs to most attacks, and the DWT-MFCCs is more robust under amplifying attack.

The proposed algorithms have desynchronization problems and the audio features will have a big change after the desynchronization attacks. The future work will be focused on solving the problem of synchronization.

ACKNOWLEDGEMENTS

This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2015.

REFERENCES

[1] D. Megias, J. Serra-Ruiz, and M. Fallahpour. “Efficient self-synchronized blind audio watermarking system based on time domain and FFT amplitude modification”.Signal processing vol.90, 2010, pp.3078-3092.

[2] S. G. Mallat. “A wavelet tour of signal processing”. Academic, New York. 1999

[3] H. T, Hu and L. Y. Hsu. “Robust, transparent and high-capacity audio watermarking in DCT domain”. Signal Processing vol.109, 2015, pp.226-235.

[4] L. C. Jimmy and G. A. Ascensiὀn. “Feature extraction based on the high-pass filtering of audio signals for Acoustic Event Classification”. Computer Speech and Language vol.30, 2015, pp.32-42.

[5] B. Gold and N. Morgan. “Speech and audio signal processing: Processing and Perception of Speech and Music.Wiley”.2000

[6] X. Zhuang, et al. “Real-world acoustic event detection”. Pattern Recognition Letters. vol.31, 2010, pp.1543-1551.

[7] J. Portel, et al. “Non speech audio event detection”. In: IEEE Int. Conf. On Acoustics, Speech, and Signal Processing (ICASSP), 1973-1976.

[8] A. Temko and C. Nadeu. “Classification of acoustic events using SVM-based clustering schemes”. Pattern Recognition vol.39, 2006, pp.684-694.

[9] A . Meng, P. Ahrendt and J. Larsen. “Temporal feature integration for music genre classification”. IEEE Trans. Audio Speech Language Process. vol.15, 2007, pp.1654-1664.

[10] P. Ruvolo, I. Fasel, and J. R. Movellan. “A learning approach to hierarchical feature selection and aggregation for audio classification”. Pattern Recognition Letters. vol.31, 2010, pp.1535-1542.

[11] A. Suhaib, et al. “Audio feature extraction using probability distribution function”. AIP Conf. Proc. Penang, Malaysia, May 28-30, 2014.

[12] X. Y. Zhang, et al. “Time–frequency audio feature extraction based on tensor representation of sparse coding”. Electronic Letters. Vol.51, no.2, 2015, pp.131-132.

[13] I. Vatolkin, et al. “Interpretability of Music Classification as a Criterion for Evolutionary Multi-objective Feature Selection”.4th International Conference, EvoMUSART 2015, Copenhagen, Denmark, April 8-10, 2015, Proceedings, pp.236-243.

[14] H. Muthusamy, et al. “Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals”. PLoS One. vol.10, no.3, 2015, pp.1-20.

[15] S. Mallat. “A theory for multi-resolution signal decomposition: the wavelet representation”. IEEE Trans. Pattern Anal.vol.11, no.7, 1989, pp.674-693.

[16] P. Beyerlein, et al. “Large vocabulary cretinous speech recognition of broadcast news- the Philip/RWTH approach”, speech Commun.2002.