Audio Art Authentication and Classification with Wavelet Statistics

Joel Martin *jrm2107@columbia.edu*
Columbia University
Department of Electrical Engineering
New York, NY, 10027, USA

Abstract

An experimental computation technique for audio art authentication is presented. Specifically, the computational techniques used by painting/drawings art authentication are transformed from two-dimensional (image) into one-dimensional (audio) methods. The statistical model consists of first and higher-order wavelet statistics. Classification is performed with a multi-dimensional scaled 3D visual model. The results from the analyses of music/silence discrimination, audio art authentication, genre classification, and audio fingerprinting are demonstrated.

Keywords: Audio Classification, Feature Extraction, Musical Genre Classification, Wavelets.

1. INTRODUCTION

The detection of forgery and authenticity of fine art paintings has been in demand for centuries. In the past, the intuitive trained eye of an art professional was relied upon to detect fakes with good, relative success. However, the digital age has spawned computer-imaging techniques that use advanced statistical algorithms for art content analysis.

In particular, the intriguing work of Lyu, Rockmore, and Farid [1] of Dartmouth University has propelled the art authentication industry into the 21st century. Various works of Flemish artist Pieter Bruegel the Elder and Italian painter Pietro di Cristoforo Vannucci were analyzed for authenticity using statistical wavelet techniques. The results were quite remarkable, discerning forgeries and the works of assistants in the paintings with fairly accurately.

These techniques could also be used to distinguish music authenticity and distinguish between musical genres. Although a fake Beethoven recording cannot be purchased on the internet (since Beethoven did not record any music), the idea is still intriguing. For example, statistical components of an audio file could be used to describe the finger-picking style of a folk guitar artist or the vocal-trends of a pop singer. These results could be used for commercial authentication, classification purposes, or simply as musical theory analysis for the aspiring musician.

2. RELATED WORK

The demand for audio classification has initiated some very innovative research. Speech/music discrimination [2], audio violence detection [3], audio search and retrieval systems [4], and music genre classification [5] are a few examples of audio content analysis research. The basic principle behind each topic is the pertinent feature extraction from the audio data, i.e. extracting an N-length vector that sufficiently summarizes the content of the audio data.

The differentiations between music, speech, and silence/environmental sounds have been proven fairly effective by using a variety of speech-processing techniques [6]. However, the classification of musical genres has been quite tedious [7]. Previous research has used a variety of digital musical analysis. Digital MIDI data was used to classify song melodies [8]. Irish, German, and

Austrian folk music have been distinguished using hidden Markov models [9]. Even neural networks have been used to differentiate between pop and classical music [10].

Music experts agree that wavelet-statistics is an excellent method for audio analysis, because music is accurately described using a combination of wavelets [11]. Lambrou, Kudumakis, Speller, Sandler, and Linney used wavelet-statistics for genre classification for rock, piano, and jazz styles. First order statistics, such as mean, variance, skewness, and kurtosis; and second order statistics, such as angular second moment, correlation, and entropy, were used to build a feature vector. Finally, classification produced excellent accuracy, using either the minimum distance classifier or the least squares minimum distance classifier [12].

This paper attempts to use wavelet statistics for genre classification, music/silence differentiation, audio art authentication, and audio fingerprinting. The same wavelet-based techniques used by two-dimensional (image) painting/drawing analysis are transformed into one-dimensional (audio) methods. As stated, the feature extraction is done using wavelet-statistics, and classification is performed using multi-dimensional scaling [1].

3. FEATURE SELECTION

The music data is conditioned before the feature vectors are extracted. If needed, a digital filter is used to extract the frequency range of a specific musical instrument. For example, the guitar frequencies can be extracted from other instruments in a five-piece band. Then the data is read from wave files and auto-scaled to fill the full intensity range (0,255).

The pertinent musical features are extracted after decomposing the audio file into basis wavelets. Each audio file is transformed by using a five-level, wavelet-like decomposition. As illustrated in Fig, 1, this decomposition splits the frequency space into multiple scales and two orientations, whereas a two-dimensional image would be split into four orientations (lowpass, vertical, horizontal, and diagonal subbands) [1]. Subsequent scales are created by subsampling the lowpass basis by a factor of two and recursively filtering. The highpass subbands at scale i = 1,…,n are denoted as $D_i(x,y)$. Shown in Fig. 2, is a three-level decomposition of an audio sample.

The feature vector is composed of two parts. The first part is a statistical model composed of the mean, variance, skewness, and kurtosis of the highpass subband coefficients at each scale i = 1,…, n – 2. However, whereas the feature vector for a 2D image results in 12(n – 2) values, the 1D audio vector consists of 4(n – 2) values [1].

The second half of the feature vector consists of 4(n – 2) statistical values based on the errors of an optimal linear predictor of coefficient magnitude. For the highpass subband $D_i(x,y)$, a linear predictor for the magnitude of these coefficients in a subset of all possible neighbors may be given by (1).

$$\begin{aligned}\left|D_i(x)\right| = {} & w_1\left|D_i(x)\right| + w_2\left|D_i(x+1)\right| + \\ & w_3\left|D_{i-1}(\tfrac{x}{2}-1)\right| + w_4\left|D_{i-1}(\tfrac{x}{2})\right| + \\ & w_5\left|D_{i-1}(\tfrac{x}{2}+1)\right| + w_6\left|D_{i-2}(\tfrac{x}{4}-1)\right| + \\ & w_7\left|D_{i-2}(\tfrac{x}{4})\right| + w_8\left|D_{i-2}(\tfrac{x}{4}+1)\right|\end{aligned} \quad (1)$$

The linear predictor takes the form (2), in which the neighbors are arranged in a matrix Q. Next, the predictors $\vec{w}$ are found with (3). Lastly, the log error in the linear predictors is calculated from (4).

$$\vec{D} = Q\vec{w}, \quad (2)$$

$$\vec{w} = (Q^T Q)^{-1} Q^T \vec{D} . \qquad (3)$$

$$\vec{E}_D = \log_2(\vec{D}) - \log_2(|Q\vec{w}|) . \qquad (4)$$

The mean, variance, skewness, and kurtosis of the highpass predictor log error is found at each scale i = 1,…, n – 2; resulting in 4(n – 2) values [1].

Combining both sets of elements with a five-level frequency-domain decomposition projects a 24-length feature vector. Our assumption is that the wavelet domain statistics are indicative to the inherent characteristics of the musician. For example, the finger-picking style or favorite notes of a musician may be extracted from a frequency-based analysis. This assumption proved to be suitable for the painting art analysis of Lyu, Rockmore, and Farid [1].

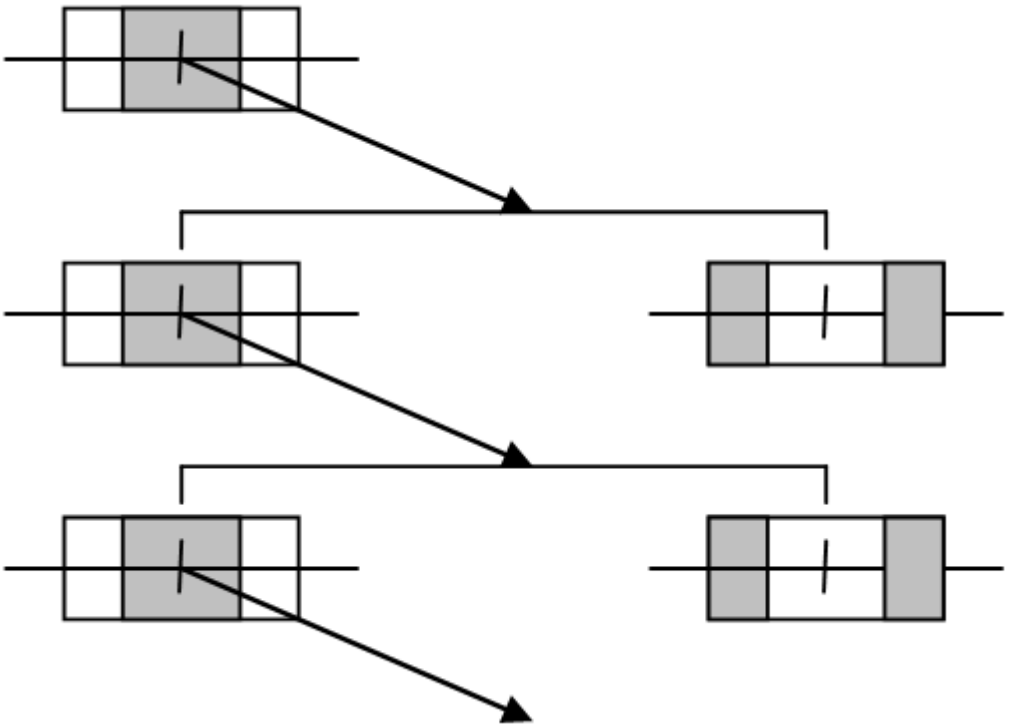

FIGURE 1: Multi-scale and orientation decomposition of frequency space for 1-dimensional. Shown, from top to bottom, are levels 0, 1, and 2, and from left to right, are the lowpass and highpass subbands.

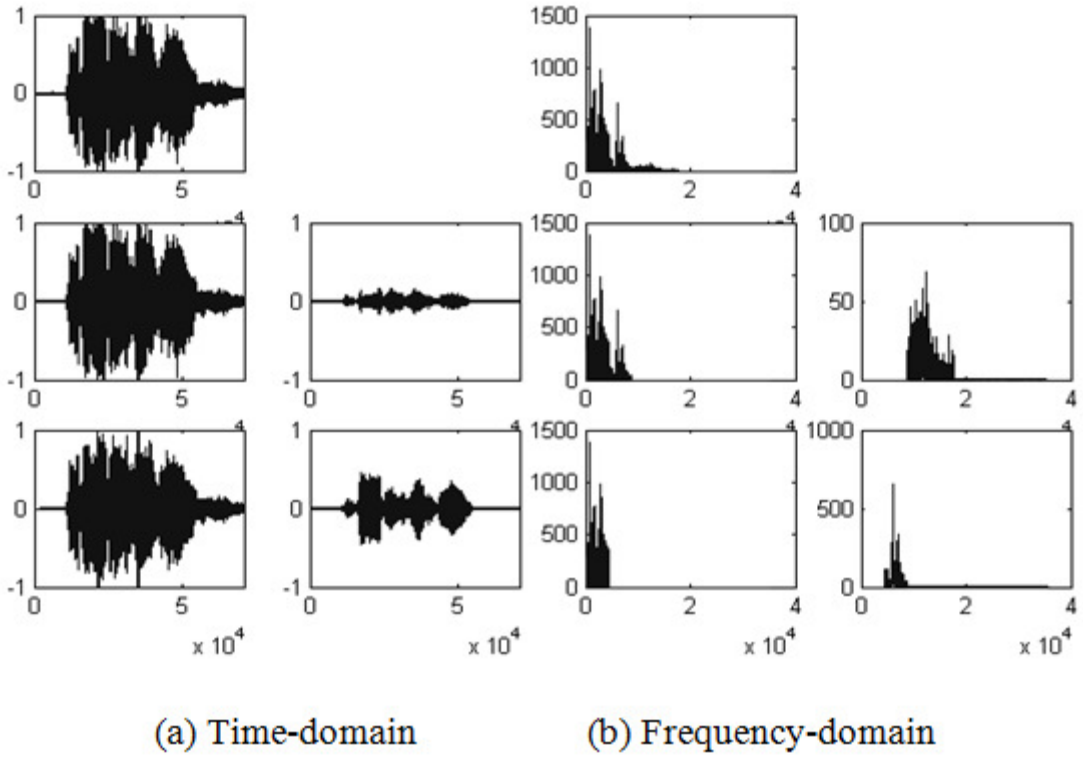

FIGURE 2: Time-domain (a) and frequency-domain (b) decomposition of a 16kpbs wav. Shown from top to bottom, are levels 0, 1, and 2, and from left to right are the lowpass and highpass subbands.

4. CLASSIFICATION

The 24-length feature vector is computed for each of the N-songs. Next, the Hausdorff Distance is computed between all sets of feature vectors to search for similarities between the audio samples [1]. Finally, the resulting NxN distance matrix is transformed into 3D space using classical multidimensional scaling [13]. A visual 3D model can show the clustering and separation between the audio files.

5. EXPERIMENTAL RESULTS

5.1 Music/Silence Discrimination

Three sets of analysis were done. The first consisted of a set of three song recordings from a musician and three sets of silence recordings. All files were 16 kHz, 8 bit wave files.

The Hausdorff Distance between all sets of vectors produced 100% accurate results. That is, the smallest distances (besides 0) were accurate in identifying the author of the content (i.e., whether the data correlates with the musician or the silence). The multi-dimensional scaled (MDS) vectors were plotted in a 3D domain, shown in Fig. 3a.

The silence recordings clustered very closely together, while the song vectors were sparsely located. The distances from songs A,B,C from silence 1,2,3 were all near ~1.15. The simple plot illustrates the apparent differences between the two sets. The scattering of the musical data could have been due to low sound quality. Therefore, a higher bit rate should be used during recording.

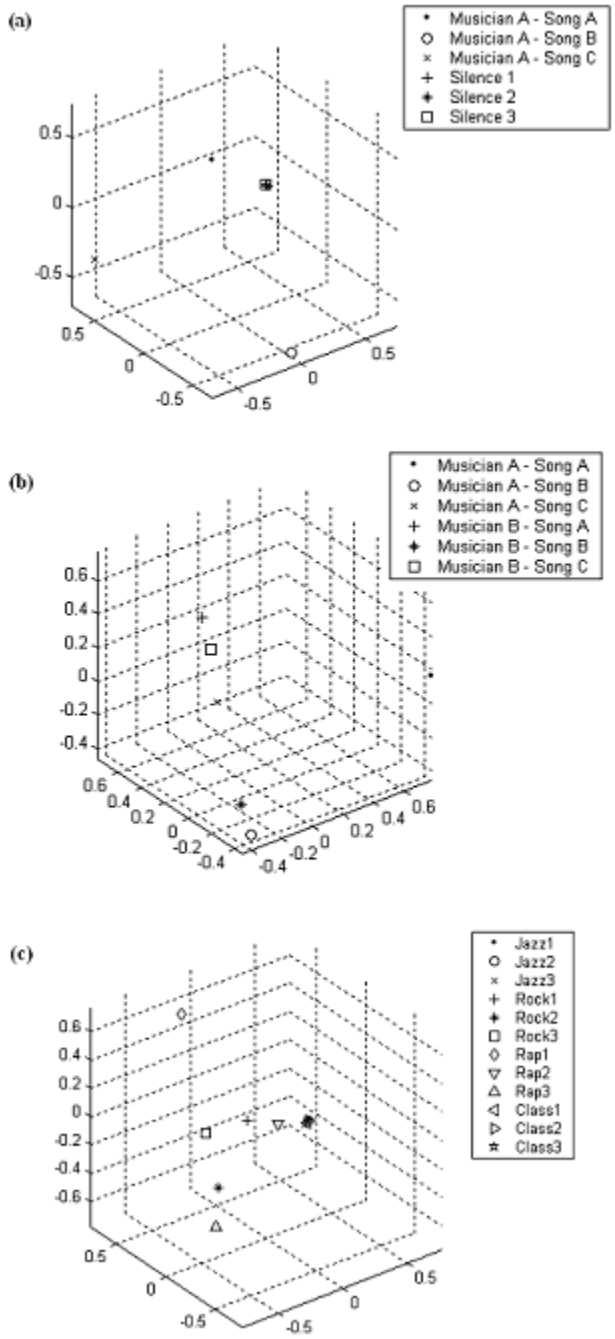

FIGURE 3: 3D MDS for (a) Music/Silence, (b) Authentication, (c) Genre Classification.

5.2 Audio Art Authentication

The second analysis consisted of three songs recorded by two different rock artists, a respected artist and an amateur (cover band) artist. All data files were 44.1 kHz, 16 bit wave files. The guitar signals of the files were extracted by band-passing the data from 80 to 5 kHz.

The Hausdorff Distance was computed between all sets of vectors, and produced 50% accuracy as to identifying the correct musician. Fig. 3b shows the multi-dimensional scaled (MDS) vectors. Musician B had songs A and C cluster together, but most of the results were sparsely located. The distances of Musician A (songs A,B,C) and Musician B (songs A, B) from Musician B (song

C) were 1.19, 0.98, 1.22, 0.17, and 0.82 respectively. Thus, the songs for Musician B were indeed closer together, while Musician A's vectors remained scattered.

The less-than-satisfactory results for the second experiment may have been due to a range of factors. For example, the professional recording may have undergone a compression technique that would have altered the frequency spectrum of the wave file.

5.3 Genre Identification

The third analysis consisted of three songs from each of the genres: jazz, rock, rap, and classical. Again, all files were 44.1kHz, 16-bit wave files. For this analysis, no bandpass filtering was performed on any of the song files.

The Hausdorff distances resulted in only 33% accurate genre classification. However, the 3D plot in Fig. 3c shows an interesting outcome. All three of the jazz and classical songs clustered very closely together, while the rock and rap songs were scattered in one primary direction. The distance of the rock 1,2,3 and rap 1,2,3 recordings from the cluster were approximately 0.48, 0.85, 0.78, 1.05, 0.26, and 1.05, respectively. The jazz and classical songs were all within ~0.06 distance from each other.

These results are quite remarkable. That is, the jazz and classical songs produced very similar wavelet domain statistics, while the rock and rap songs were independent of each other. The similar instrumentation and timing arrangements of jazz and classical music may have caused the clustering. On the other hand, the distortion and spontaneity of rap and rock music may have caused the scattering. Nevertheless, the classification of jazz and classical music from rock and rap music is quite apparent from the results.

5.4 Audio Fingerprinting

The final analysis consisted of a variety of eight pop or rock songs performed by nineteen different artists. The Hausdorff distance was computed to test whether the similar songs correlated together; e.g., the Hausdorff distance between four artists performing the same song. The results, shown in Fig. 4, were mediocre. In the graph, each song is represented as a shape. As can be seen, almost all of the shapes clustered together.

Next, 10% white noise distortion was added to each song, and the Hausdorff distance was computed between each of the undistorted and distorted songs. The fingerprint should match the distorted song with its correct undistorted version. The results were better, matching five of the eight songs - 63% accuracy. The distortion in each of the five songs was very noticeable, but did not alter the fingerprint enough to disassociate it with its original form.

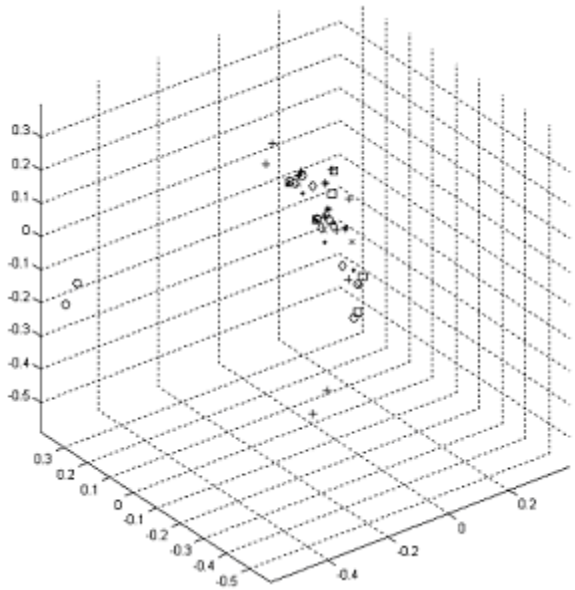

FIGURE 4: 3D MDS for Audio Fingerprinting

6. CONCLUDING REMARKS

A wavelet-based digital technique for audio art identification was presented. The presumption of the technique was that the statistical features of the frequency domain provide sufficient data to identify and classify audio art data. As shown in the paper, the technique was very effective in classifying music from silence. Sufficient results were also found for classifying two sets from a respectable band and a corresponding amateur (cover) band. Additionally, the fingerprinting proved 65% correct for 10% added distortion. These results could be used for commercial authentication purposes or simply as musical theory analysis for the aspiring musician.

7. ACKNOWLEDGEMENT

Songs by Led Zeppelin and tribute band Led-By-Zeppelin were used for the authentication test. Songs by Miles Davis, Eqsuivel, Angelo Badalamenti, Bob Dylan, The Rolling Stones, Snoop Dogg, Ice Cube, Ugly Duckling, Erik Satie, and Felix Mendelssohn were used for the genre classification test. Songs by the Beatles, Ray Charles, James Taylor, Boys 2 Men, Pink Floyd, Creedance Clearwater Revival, Boonie Tyler, the Animals, Madonna, Kylie Minoge, Tina Turner, Grand Funk Railroad, Bob Dylan, Bob Marley, and Eric Clapton were used for the audio fingerprinting test.

8. REFERENCES

1. S. Lyu, D. Rockmore, H. Farid. “A Digital Technique for Art Authentication.” Proc Natl Acad Sci USA. 101(49): 17006-10. 2004.

2. E. S. Parris, M. J. Carey, H. Lloyd-Thomas, “A comparison of features for speech, music discrimination.” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. 1:149-152. 1999.

3. S. Pfeiffer, S. Fischer, W. Effelsberg, “Automatic audio content analysis.” Proc. of 4th ACM Multimedia Conference. 1:21-30. 1996.

4. E. Wold, T. Blum, D. Keislar, J. Wheaton. “Content-Based Classification, search, and Retrieval of Audio.” IEEE Multimedia. 3(3):27-36. 1996.

5. G. Tzanetakis, P. Cook. “Music genre classification of audio signals.” IEEE Transactions on Speech and Audio Processing. 10(5):293-302. 2002.

6. L. Lu, H. Jiang, H. J. Zhang. “A robust audio classification and segmentation method.” Proc. 9th ACM Int. Conf. Multimedia. 1:203--211. 2001.

7. T. Li, G. Tzanetakis. “Factors in automatic musical genre classification of audio signals.” Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2003.

8. M. Shan, F. Kuo. “Music style mining and classification by melody.” Proc. IEEE International Conference on Multimedia and Expo. 1:97-100. 2002.

9. W. Chai, B. Vercoe. “Folk music classification using hidden Markov models.” Proc International Conference on Artificial Intelligence. 2001.

10. B. Matityaho, M. Furst. “Neural network based model for classification of music type.” Proc. 18th Conv. Electrical and Electronic Engineers in Israel. 4.3.4/1-5. 1995.

11. L. Endelt, A. Cour-Harbo, “Wavelets for sparse representation of music.” Proc of Wedelmusic. 2004.

12. T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, A. Linney. "Classification of audio signals using statistical features on time and wavelet transform domains.” Proc. IEEE ICASSP. 6:3621-24. 1998

13. T. Cox, M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994