REVERBERATION FEATURES IDENTIFICATION FROM MUSIC RECORDINGS USING THE DISCRETE WAVELET TRANSFORM

Ren Gang, Mark F. Bocko, Dave Headlam

1Dept. of Electrical and Computer Engineering, Edmund A. Hajim School of Engineering and Applied Sciences, University of Rochester
2Dept. of Music Theory, Eastman School of Music, University of Rochester

ABSTRACT

This paper presents a method of extracting reverberation features from music recordings. First, we perform a short time Fourier transform to transform the audio signal into a 2D time-frequency representation in which reverberation features appear as blurring of spectral features in the time dimension. Employing image analysis methods, we may quantitatively estimate the amount of reverberation by transforming the STFT “image” to a wavelet domain where we can perform efficient edge detection and characterization. Experiments demonstrate that quantitative estimates of reverberation time extracted in this way are strongly correlated with physical measurements.

Index Terms—Reverberation, Discrete Wavelet Transform, Music Scene Analysis, Feature Extraction, Acoustic Measurement

1. INTRODUCTION

When a sound is emitted in an enclosed space an audience will perceive the combination of the direct sound and the subsequent multiple reflections and absorption of the sound by the walls as reverberation. Humans possess astonishing capabilities for perceiving and appreciating a surrounding aural space [1,2] and acoustic reverberation is a significant part of our everyday aural experience. In applications such as speech recognition or automated music transcription engineers regard reverberation as a nuisance and a great deal of the research literature is focused on how to remove or partly mitigate its effects [3]. On the other hand, for musicians centuries of tradition have led to an appreciation of musical space as an indispensable creative tool for creating music and many musical compositions are composed for specific aural spaces [4].

In the present work we describe a method to quantitatively characterize the reverberation present in musical recordings. Reverberation, both natural and artificial, is employed extensively in music recordings. Thus, reverberation features can be used to characterize various musical genres for classification or recognition purposes or they may serve as acoustic “fingerprints” that may be used in audio forensics applications, where we seek to reverse engineer a recording to reveal all the “production chain” information from the mixed-down recordings. Recent proliferation of auralization applications demands the capability of extracting reverberation features from archive recordings, this capability is of particular interest if we want to simulate some “lost” legendary music space since the physical acoustic measurements are impossible to recreate. Finally, the present method of estimating the reverberant features of a recording may contribute to methods of blind deconvolution removal of reverberation from recordings.

2. ALGORITHM DESCRIPTION

Reverberation features can readily be extracted from a wavelet domain representation of a spectrogram by image edge detection and categorization algorithms [5,6,7]. The proposed blur analysis algorithm also provides a quantitative measure of the spectrogram blur based on the statistics of edges. Quantitative measurements and descriptive features obtained from such a blur analysis can successfully depict an aural space since they are closely related to the physical measurements. Sec. 2.1 provides a brief illustration of our proposed system. Sec. 2.2 illustrates the edge detection and categorization algorithm and Sec. 2.3 illustrates the blur analysis methods.

2.1. System Structure

The system structure is composed of the following main functional modules:

Figure 1: Reverberation features are seen in a spectrographic representation as image blurring: in (a) we show the spectrogram of a single musical note in an anechoic recording; in (b) we show the spectrogram of the same part of the recording with artificial reverberation added.
(1) Short time Fourier transform (STFT): this module generates a spectrogram of the audio signal. We employ a Hamming window of 1024 sampling points with 128 points overlapping between STFT frames. The audio is recorded in 44100Hz/16-bit format.

(2) Discrete wavelet transform (DWT): the spectrogram is transformed to the wavelet domain using a 2D discrete Haar wavelet transform. Three levels of wavelet decomposition are performed to provide multi-resolution approximations and details [5]. Comparing different level of details in a wavelet transform domain provides distinctive reverberation features. Furthermore the DWT coefficients can be efficiently obtained with existing fast computation algorithms [5].

(3) Edge detection module: image edges (in time direction) in the spectrogram are representative of reverberant features. We employ the vertical detail wavelet coefficients to detect the spectrogram edges and to segment the edge areas for further categorization. The edge detection algorithm is further explained in Sec.2.2.

(4) Edge categorization: The vertical detail values in the three DWT resolution levels decompose the edge transient to reveal the shape of the edge. If the edge is sharp, vertical detail values in lower levels of the DWT are higher since they reflect the high frequency part of edge transient. For smooth edges, vertical detail values in the higher levels of the DWT are larger since they contribute to the low frequency part of the edge transient. Based on the vertical detail coefficient values at different DWT levels, we can further categorize the segmented spectrogram edges into four categories depicting four different shapes of edges. The edge categorization algorithm is given in Sec.2.2.

(5) Blur analysis: the statistical distribution of DWT coefficients depicts various levels of reverberation, e.g., one sees a greater number of smooth edges for longer reverberation time. By counting the relative frequency of occurrence of different types of edges we can obtain quantitative measurements of spectrogram blur and perform physically meaningful virtual measurements. Furthermore, the statistical distributions of different types of edges provide quantitative blur descriptors that can be utilized as feature vectors for reverberation pattern recognition. The Blur analysis algorithm is further illustrated in Sec.2.3.

2.2. Edge Detection and Categorization

The four different types of edges we encountered in spectrogram analysis and their intensity level curves are illustrated in Figure 2 [6,7]. The Dirac structure edge (D edge) separates a small area that has a distinctively different intensity level compared with its neighbors. A-step structure edge (A edge) depicts a sudden change of intensity level between two areas without intermediate transients. The G-step structure edge (G edge) depicts gradual intensity changes between two areas and the roof structure edge (R edge) depicts a relatively broad area with different intensity levels and smooth transients. If the transient smoothness of a G edge or R edge is above a certain threshold, they are further categorized as a blurred G edge (bG edge) or blurred R edge (bR edge).

![Figure 2: Four different types of edges: (a) Dirac structure edge; (b) A-step structure edge (c) roof structure edge (d) G-step structure edge.](image)

These four types of edges can be accurately detected and categorized in the wavelet domain. We use the Haar wavelet transform (HWT) to perform three levels of decomposition [3]. In each level , the spectrogram is decomposed into approximation coefficients , vertical detail coefficients , horizontal detail coefficients and diagonal detail coefficients . For comparison of the coefficient matrices of different sizes from different wavelet transform level, we map a 4 x 4 area in level 1 and a 2 x 2 area in level 2 to time-frequency location (k,l) in level 3. We define the maximum value in the level 1 and level 2 areas as the coefficient value (k,l). Also because reverberation produces image blur in the time direction, we only perform edge detection with the vertical detail coefficients. Interpreting a significant “jump” of intensity as an edge, implies that for all three wavelet decomposition levels is high. We can utilize this fact to form the following edge detection rules:

Edge Detection Rule:
if \( C_{VD,1}(k,l) > \text{Threshold E1} \), \( C_{VD,2}(k,l) > \text{Threshold E2} \) and \( C_{VD,3}(k,l) > \text{Threshold E3} \), (k,l) is an edge point.

The relative values of \( C_{VD,1}, C_{VD,2} \) and \( C_{VD,3} \) determines the edge sharpness in the time axis direction. For sharp edges, the detail coefficients in low levels are higher since it represents the high frequency components of the edge transient. We can further formulate the following edge categorization rules:

Edge Categorization Rules:
if \( C_{VD,1}(k,l) > C_{VD,2}(k,l) > C_{VD,3}(k,l) \), (k,l) is D edge or A edge;
if \( C_{VD,1}(k,l) < C_{VD,2}(k,l) < C_{VD,3}(k,l) \), (k,l) is G edge or R edge;
if \( C_{VD,2}(k,l) < C_{VD,3}(k,l) \) and \( C_{VD,2}(k,l) > C_{VD,3}(k,l) \), (k,l) is R edge;
if (k,l) is a G edge or R edge, and if \( C_{VD,1}(k,l) < \text{Threshold E4} \), (k,l) is bG edge or bR edge.

\(^{1}\) Compression schemes such as MP3 and M4A cause subtle changes of edge distribution because they exploit temporal masking and noise shaping to discard information unimportant for human auditory perception [8].
2.3. Blur Analysis for Reverberation Estimation

The first functional block in our blur analysis module is anechoic detection. Based on edge categorization results this block distinguishes anechoic recordings from reverberant recordings. This functionality is a crucial front-head for the subsequent reverberation feature extraction algorithms since we need to prevent our algorithm from extracting irrelevant reverberation features from anechoic recordings. Four scenarios must be considered for anechoic detection: (1) for clear recordings, a small number of D, A edges indicates that the recording contains anechoic components. D, A edges depict anechoic onset-offset, all of them will change into G, R edges in reverberant recordings so the decision threshold can be very low. If the spectrogram does not have a lot of bG, bR edges, we categorize it as anechoic. (2) For clear recordings with a small number of D, A edges but a large number of bG, bR edges, we categorize it as long initial-time-delay gap (ITDG) case. ITDG measures the time interval between the arrival of the direct sound and the arrival of first reflection [9]. A reverberant signal with too long of an ITDG contains a portion of anechoic signal before the reverberant signal appears. The exposed anechoic direct sound results in D, A edges. ITDG is an important architectural acoustics measurement since too long of a value for the ITDG is undesirable and makes the enclosure sound “empty” [9]. This scenario can easily be distinguished from the anechoic case by a large number of bG, br edges. (3) For noisy recordings, a large number of D, A type edges are seen. This case is well separated because we detect orders of magnitude more D, A edges than the former two cases. (4) The clear and reverberant case. Denoting $R_{DA}$ as the ratio of the number of D, A to all edges detected, $R_{GR}$ for G,R edges, $R_{bG}$ for bG,bR edges, we can form the following anechoic detection rules:

Anechoic Detection Rules:

- if $R_{DA} > \text{Threshold B1}$, the recording is noisy;
- if $R_{DA} > \text{Threshold B2}$, $R_{DA} < \text{Threshold B1}$ and $R_{bG} < \text{Threshold B3}$, the recording is anechoic;
- if $R_{DA} > \text{Threshold B2}$, $R_{bG} > \text{Threshold B3}$, the recording is reverberant with long ITDG;
- otherwise the recording is reverberant.

For the recording to be categorized as reverberated (or reverberant with long ITDG), we further perform subsequent blur analysis. The simplest and most important, architectural acoustics measurement is the reverberation time. It can be measured as the extent of the time blur in the spectrogram so we define the blur extent measure as $E_b = \frac{R_{bG}}{R_{GR}}$. As the reverberation time increases a greater proportion of G, R type edges will transform into bG, bR edges, thus $E_b$ provides a measure of the reverberation time. The performance of this measurement on recordings of known reverberation times is further evaluated in Sec.3.

Also important for characterizing an acoustic space is the reverberation time in different frequency ranges. Various different wall surfaces exhibit frequency selective absorption and scattering characteristics, thus measurement of the reverberation time at various frequencies provides a potentially detailed depiction of the enclosure. In our implementation, we analyze the spectrogram in six octave bands with central frequencies at 125, 250, 500, 1000, 2000, and 4000Hz [10] and obtain the blur extent at each band. Finer partition of the spectrogram provides better frequency resolution at the expense of long analysis frames since we then need longer duration recordings to accumulate meaningful statistics.

We also perform ITDG measurement based on the statistical distribution of edges. Long ITDG results in an increase in the number of D, A edges. Once a music recording is categorized as a long ITDG case, we can form an ITDG indicator as $\text{ITDG} = R_{DA}(R_{bG})^{3.5}$. The $(R_{bG})^{3.5}$ factor compensates for the overlap of reverberant harmonic structures covering a D, A edge: a larger $R_{bG}$ means more reverberation, and more D, A edges are covered by G, R edges and thus categorized as G, R edge in edge categorization part.

For noisy recordings we should perform de-noising algorithms [11] before the above blur analysis procedures. Also the image de-noising algorithm in [12] is especially effective in audio spectrograms since sharp edges from broadband noise have significantly different properties than edges from music onset-offset and reverberant features.

3. PERFORMANCE EVALUATION

We partitioned test music samples into analysis frames of length 4 seconds and performed reverberation feature extraction procedures as illustrated in Sec.2 to explore the effectiveness of the proposed algorithms.

First, we demonstrate the capability of the proposed anechoic detection algorithm. 200 test pieces are anechoic recordings while another 200 pieces are reverberant versions of the same recordings produced by convolving the anechoic recording with various impulse response functions (T60 reverberation times ranged from 0.10 seconds to 0.73 second). We observed no decision errors on this test set. Also the decisions between anechoic music and reverberant music with long ITDG are 98.5% accurate.

We also demonstrated the correlation of the measured blur extent and the known reverberation time. We generated reverberant sound by convolving an anechoic clarinet recording with different acoustic impulse responses. We use twelve synthetic impulse responses with different T60 values, shown in Figure 3. Each blur extent value here is calculated as the average value of blur extent measured from 20 audio frames. The standard deviation of 20 frames is marked as error bar for the sixth data point and the measurement deviation values for other data points are similar. The blur extent was observed to increase nearly linearly with reverberation time, as expected. We also tested the system with real impulse responses measured in different acoustic environments. The reverberation time of each impulse response is calculated using the method in [10]. The linear relationship between the blur extent and the reverberation time is not so evident but their strong correlation still provides a valid virtual measurement method.

We also tested the capability of the proposed method of measuring the above-defined ITDG indicator. Twelve synthetic impulse responses with different ITDG values are convolved with...
We have defined an image analysis based reverberation feature extraction algorithm that analyzes the blurs in the spectrogram to provide quantitative measures of the T60 time and ITDG. The edge detection and categorization algorithms in wavelet transform domain generate sensitive statistical edge distributions and blur analysis results. Comparing to deconvolution methods as in [13], the proposed algorithm is easier to implement when dealing with complex music scenes since it does not require detection of quiet regions for frame segmentation.

The Haar wavelet transform utilized in our paper provides the simplest implementation of statistical blur analysis. More sophisticated discrete wavelet transform have different trade-offs in edge detection capability [5] and we are investigating their performance in our system. Also, to obtain sensitive statistical measurements we need long samples of music. Although the spectrogram and DWT all have fast computation algorithms posing no perceptible computation burden, we still want to refine the sensitivity of our proposed algorithm for short analysis frames (less than 2 seconds). The resolution for short frames is important for time dependent feature extraction and also for increasing forensic sensitivity since we can shrink the analysis frame to enhance the detection of editing points in multimedia files.

5. REFERENCES