Visualizing Musical Structure and Rhythm via Self-Similarity

Jonathan Foote and Matthew Cooper
FX Palo Alto Laboratory, Inc.
3400 Hillview Ave., Building 4
Palo Alto, CA 94304 USA
{foote, cooper}@pal.xerox.com

Abstract

This paper presents a novel approach to visualizing the time structure of musical waveforms. The acoustic similarity between any two instants of an audio recording is displayed in a static 2D representation, which makes structural and rhythmic characteristics visible. Unlike practically all prior work, this method characterizes self-similarity rather than specific audio attributes such as pitch or spectral features. Examples are presented for classical and popular music.

1. Introduction

There has been considerable interest in making music visible. Efforts include artistic attempts to realize images elicited by sound, of which the Walt Disney film *Fantasia* is perhaps the canonical example. Another approach is to quantitatively render the time and/or frequency content of the audio signal, using methods such as the oscillograph and sound spectrograph [1,2]. These are intended primarily for scientific or quantitative analysis, though artists like Mary Ellen Bute have used quantitative methods such as the cathode ray oscilloscope towards artistic ends [3]. Other visualizations are derived from note-based or score-like representation of music, typically MIDI note events [4,5].

Music is generally self-similar. With the possible exception of a few avant-garde compositions, structure and repetition is a general feature of nearly all music. That is, the coda often resembles the introduction and the second chorus sounds like the first. On a shorter time scale, successive bars are often repetitive, especially in popular music. This paper presents methods of visualizing music by its acoustic self-similarity across time, rather than by absolute acoustic characteristics. Self-similarity is visualized in a two-dimensional time representation such as Figure 1. This representation presented here is very flexible and can be used with practically any parameterization of audio. Besides audio, similar representations have been used to analyze text [7], video [8], hypertext links [9], and dynamical systems [10].

2. Similarity Analysis

An audio file is visualized as a square. Time runs from left to right as well as from bottom to top. In the square, the brightness of a point (*i*,*j*) is proportional to the audio similarity at instants *i* and *j*. Similar regions are bright while dissimilar regions are dark. Thus there is always a bright diagonal line running from bottom left to top right, because audio is always the most similar to itself at any particular time. Repetitive similarities, such as repeating notes or motifs, show up as a checkerboard patterns: a note repeated twice will give four bright areas at the corner of a square. The two regions at the off-diagonal corners are the "cross-terms" resulting from the first note's similarity to the second. Repeated themes are visible as diagonal lines parallel to, and separated from, the main diagonal by the time difference

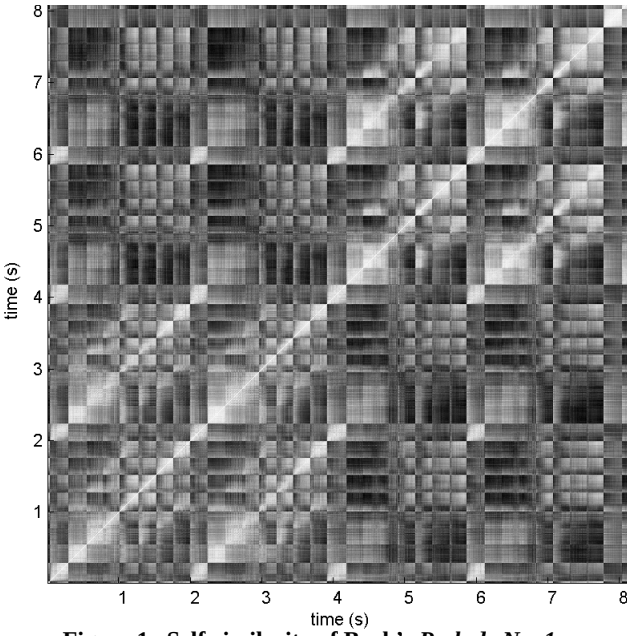

Figure 1. Self-similarity of Bach's *Prelude No. 1*

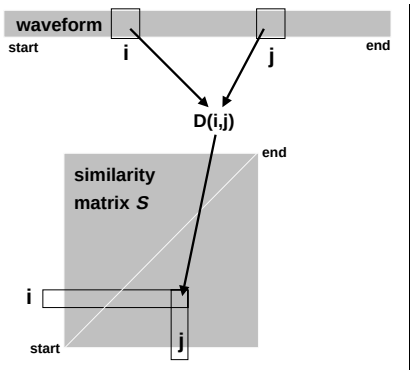

Figure 2. Distance matrix calculation

between repetitions.

To calculate the similarity between two audio "instants," they are first parameterized using the short-time Fourier transform or using Mel-frequency cepstral coefficients. Given two feature vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ derived from audio windows *i* and *j*, a simple metric of vector similarity *s* is the scalar product of the vectors. This will be large if the vectors are both large and similarly oriented. To remove the dependence on magnitude (and hence energy, given spectral features), the product can be normalized to give the cosine distance between the vectors:

$$s(i,j) \equiv \frac{\boldsymbol{v}_i \bullet \boldsymbol{v}_j}{\|\boldsymbol{v}_i\| \, \|\boldsymbol{v}_j\|}$$

To visualize an audio file, an image is constructed so that each pixel at location *i*, *j* is given a grey scale value proportional to the similarity measures described above.

Note that the actual parametrization is not crucial as long as "similar" sounds yield similar parameters. Psychoacoustically motivated parameterizations, like those described by Slaney [6], may be especially appropriate if they match the similarity judgments of human listeners.

2.1 Bach Prelude No. 1

Figure 1 shows roughly the first two bars of Bach's *Prelude No. 1 in C Major,* from *The Well-Tempered Clavier,* BVW 846. This 1963 piano performance is by Glenn Gould. The visualization makes both the structure of the piece and details of performance visible. 34 notes are visible as squares along the diagonal. The repetition time can be seen in the off-diagonal stripes parallel to the main diagonal, as well as the repeated C note at 0, 2, 4, and 6 seconds. Figure 4 shows the first three bars of the score: the repetitive nature of the piece should be clear even to those unfamiliar with musical notation. Figure l3 shows yet another similarity image of the same music, derived directly from the MIDI data. Here, no acoustic information was used. Matrix entries *(i,j)* were colored white if note *i* was the same pitch as note *j*, and left black otherwise.

2.2 Beethoven's *Fifth Symphony*

Not only can *acoustically* similar audio be located, but *structurally* similar audio should be straightforward to find, by comparing similarity matrices. For example, different performances of the same symphonic movement will have a

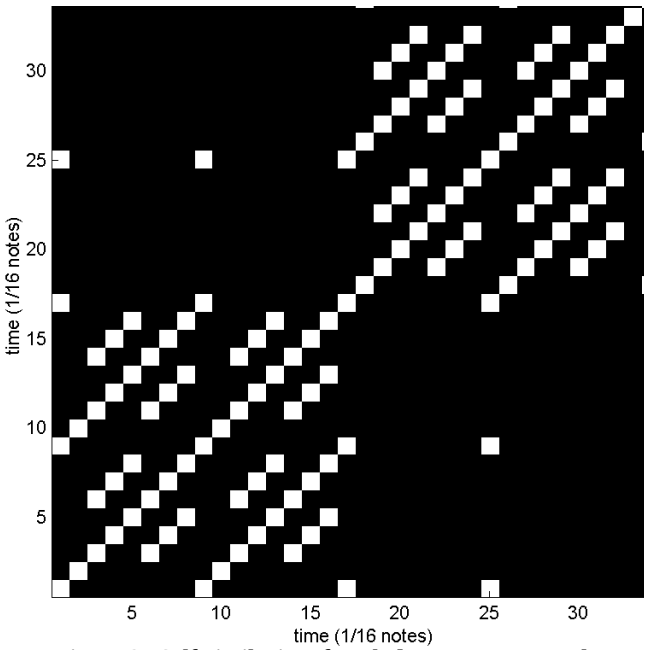

Figure 3. Self-similarity of *Prelude No. 1:* computed from MIDI note events

similar structural visualization regardless of how or when they were performed or recorded, or indeed the instruments used. Figure 5 shows the self-similarity of the entire first movement of Beethoven's *Symphony No. 5*. Two visualizations are shown, each from a different performance featuring different conductors (Herbert von Karajan and Carlos Kleiber) and orchestras (the Berlin and Vienna Philharmonics, respectively). Because the piece is more than seven minutes long, much fine detail is not observable. Each pixel represents nearly a second of music, thus the famous opening theme occurs in the in only the first few pixels. The primary visible structure is the alternation of softer string passages with louder *tutti* (all instruments playing) sections, for example the sustained climax near the end of the movement. This figure illustrates how the visualization captures both the essential structure of the piece as well as variations due to individual performances. Though similarity matrices are not directly comparable, the variation between louder and softer passages has been used as a method for similarity retrieval [14].

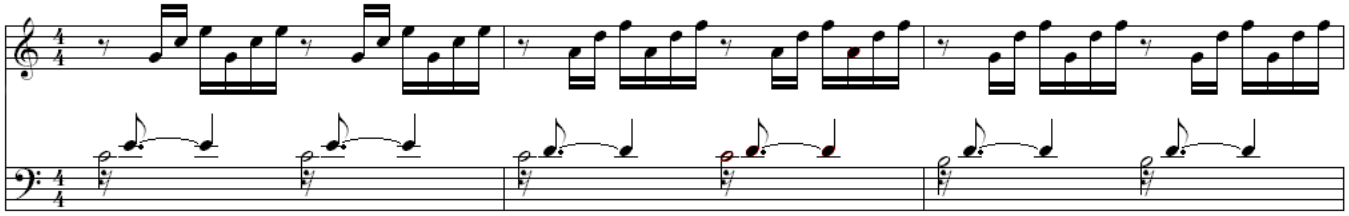

Figure 4. First bars of Bach's *Prelude No. 1 in C Major,* BVW 846, from *The Well-Tempered Clavier*

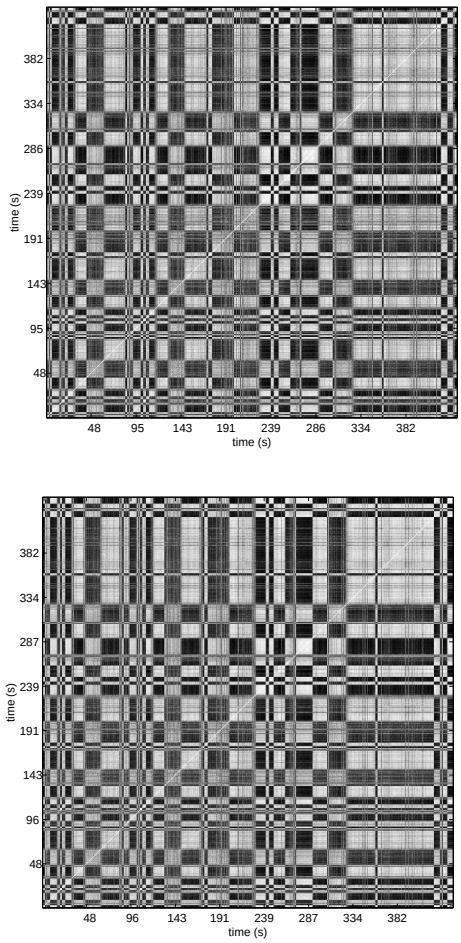

Figure 5. Self-similarity of *Symphony No. 5*. Top: von Karajan performance. Bottom: Carlos Kleiber performance.

2.3 Visualizing Musical Rhythm

Both the periodicity and relative strength of rhythmic structure can be derived from the similarity matrix. We've coined the term *beat spectrum* for a measure of self-similarity as a function of the lag [13]. Peaks in the beat spectrum at a particular lag *l* correspond to audio repetitions at that temporal rate. The beat spectrum *B(l)* can be computed from the similarity matrix using diagonal sums or autocorrelation methods. A simple estimate of the beat spectrum can be found by diagonally summing the similarity matrix ***S*** as follows:

$$B(l) \approx \sum_{k \subset R} S(k, k+l)$$

Here, $B(0)$ is simply the sum along the main diagonal over some continuous range R, $B(1)$ is the sum along the first superdiagonal, and so on. Figure 6 shows an example computed for a three-second excerpt of the Gould performance. The periodicity of each note can be clearly seen, as well as the strong eight-note periodicity of the phrase (with a sub-harmonic at 16 notes). Especially

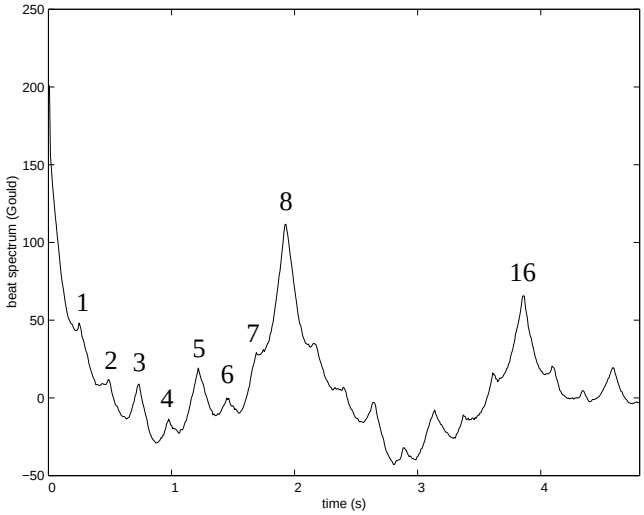

Figure 6. Beat spectrum of Gould performance from diagonal sum

interesting are the peaks at three and five notes. These comes from the three-note periodicity of the eight-note phrase: in each phrase, notes three and six, notes four and seven, and notes five and eight are identical.

A more robust estimate of the beat spectrum is the autocorrelation of ***S***:

$$B(k, l) = \sum_{i,j} S(i, j) S(i+k, j+l)$$

Because *B(k,l)* will be symmetrical, it is only necessary to sum over one variable to yield a one-dimensional result *B(l)*. This approach works surprisingly well for most kinds of musical genres, tempos, and rhythmic structures. Figure 7 shows the beat spectrum computed from the first ten seconds of the Paul Desmond jazz composition *Take 5*, performed by the Dave Brubeck Quartet. Besides being in the uncommon $5/4$ time signature, this rhythmically sophisticated work requires some interpretation. First, note that there is no obvious periodicity at the actual beat tempo (denoted by solid vertical lines in the figure). Rather, there is a marked periodicity at five beats, and a corresponding sub-harmonic at ten. Jazz aficionados know that "swing" is the subdivision of beats into non-equal periods rather than "straight" (equal) eighth-notes. The beat spectrum clearly shows that each beat is subdivided into near-perfect *triplets*. This is indicated with dotted lines spaced one-third of a beat apart between the second and third beats. A clearer visualization of "swing" would be difficult to achieve by any other means.

The beat spectrum can be analyzed to determine tempo and more subtle rhythmic characteristics. Peaks in the beat spectrum give the fundamental rhythmic periodicity [13]. Strong off-beats and syncopations can be then deduced from secondary peaks in the beat spectrum. Because the only necessary signal attribute is repetition, this approach is more robust than other approaches that look for absolute acoustic features such as energy peaks.

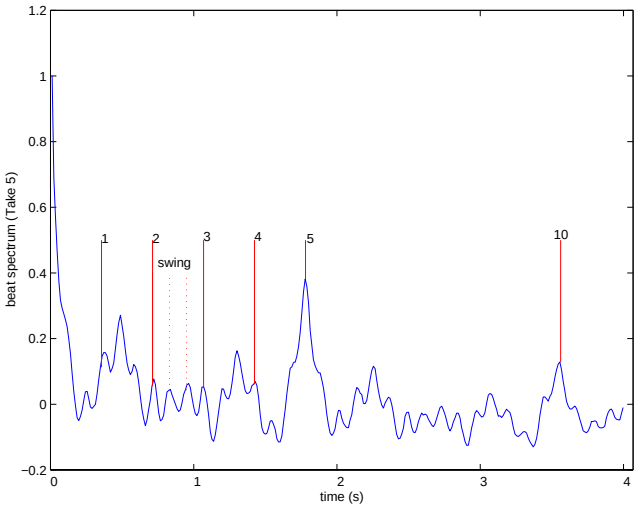

Figure 7. Beat spectrum of jazz composition *Take 5*

There is an inverse relationship between the time accuracy and the beat spectral precision. Technically, the beat spectrum is a frequency operator, and hence does not commute with a time operator. Thus beat spectral analysis, just like frequency analysis, is a trade-off between spectral and temporal resolution.

3. The Beat Spectrogram

We also introduce the *beat spectrogram* for analyzing rhythmic variation over time. Like its namesake, the beat spectrogram visualizes the beat spectrum over successive windows to show rhythmic variation over time. Time is on the *x* axis, with lag time on the *y* axis. Each pixel is colored with the scaled value of the beat spectrum at that time and lag, so that peaks are visible as brighter horizontal bars at the repetition time. Figure 8 shows the beat spectrogram of a 33-second excerpt of the Pink Floyd song *Money*. Listeners familiar with this classic-rock chestnut may know the song is primarily in the 7/4 time signature, save for the bridge (middle section), which is in 4/4. The excerpt shown starts 4 minutes and 55 seconds into the song, and clearly shows the transition from the 4/4 bridge back into the last 7/4 verse. On the left, there are strong beat spectral peaks on each beat (annotated white numbers), particularly at two and four beats (the length of a 4/4 bar), and an eight-beat subharmonic. Two beats occur in slightly less than a second, corresponding to a tempo slightly faster than 120 beats per minute (120 MM). This is followed by a short two-bar transition. Then the time signature changes to 7/4, clearly visible as a strong seven-beat peak with the absence of a four-beat component. The tempo also slows slightly, visible as a slight lengthening of the time between peaks.

4. CONCLUSION

We have presented a method of visualizing musical structure and rhythm. Unlike many other approaches, this method does not rely on detecting specific attributes like pitch or energy; rather the signal is used to model itself.

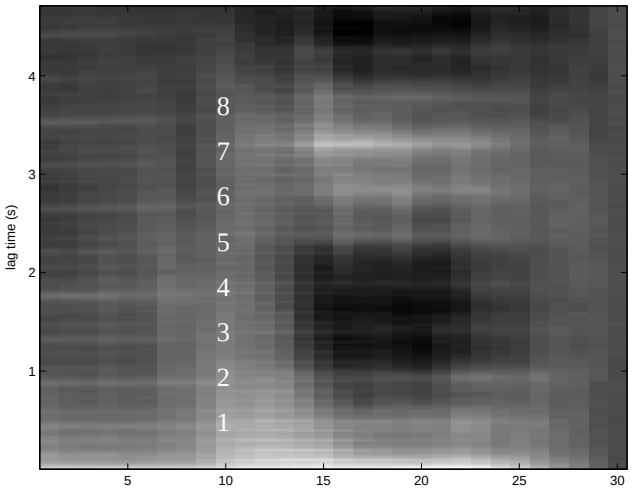

Figure 8. Beat spectrogram of Pink Floyd's *Money* (excerpt), showing transition from 4/4 to 7/4 time

5. REFERENCES

[1] Potter R., G. Kopp, and H. Green, *Visible Speech*, D. Van Nostrand Co., NY, 1947

[2] Koenig, W., H.K. Dunn, and L.Y. Lacey, "The Sound Spectrograph," in *J. Acoustical Society of America*, **18**, p. 19-49.

[3] Moritz, W., "Mary Ellen Bute: Seeing Sound," in *Animation World*, Vol. 1, No. 2 May 1996 http://www.awn.com/mag/issue1.2/articles1.2/moritz1.2.html

[4] .Smith, Sean M., and Williams, Glen, "A Visualization of Music," in *Proc. Visualization '97*, ACM, pp. 499-502, 1997

[5] Malinowski, S., "The Music Animation Machine," http://www.well.com/user/smalin/mam.html

[6] Slaney, M. (1998). "Auditory Toolbox," *Technical Report #1998-010*, Interval Research Corporation, Palo Alto, CA

[7] Church, K. and Helfman, J., "Dotplot: A Program for exploring Self-Similarity in Millions of Lines of Text and Code," in *J. American Statistical Association*, **2**(2), pp.153--174, 1993

[8] Cutler, R., and L. Davis. "Robust Periodic Motion and Motion Symmetry Detection," in *Proc. Conference on Computer Vision and Pattern Recognition* (CVPR), June 2000.

[9] Bernstein, M., et al., "Architectures for Volatile Hypertext," in Proc. Hypertext 91, pp. 243-260, December 1991.

[10] Eckman, J.P., et al., "Recurrence Plots of Dynamical Systems," in *Europhys. Lett.*, **4**(973), November 1987

[11] Scheirer, E., "Tempo and Beat Analysis of Acoustic Musical Signals," in *J. Acoust. Soc. Am*. **103**(1), Jan 1998, pp 588-601.

[12] Foote, J., "Automatic Audio Segmentation using a Measure of Audio Novelty," in *Proc. International Conference on multimedia and Expo (ICME)*, IEEE, August, 2000.

[13] Foote, J., and Uchihashi, S., "The Beat Spectrum: A New Approach to Rhythm Analysis," to appear in *Proc. International Conference on Multimedia and Expo (ICME)* IEEE, Tokyo, August 2001.

[14] Foote, J. "ARTHUR: Retrieving Orchestral Music by Long-Term Structure." In *Proc. of the International Symposium on Music Information Retrieval*, Plymouth, Massachusetts, Oct. 2000.