Visualizing Music and Audio using Self-Similarity

Jonathan Foote
FX Palo Alto Laboratory, Inc.
3400 Hillview Ave., Building 4
Palo Alto, CA 94304 USA
+1 (650) 813-7071

foote@pal.xerox.com

1. ABSTRACT

This paper presents a novel approach to visualizing the time structure of music and audio. The acoustic similarity between any two instants of an audio recording is displayed in a 2D representation, allowing identification of structural and rhythmic characteristics. Examples are presented for classical and popular music. Applications include content-based analysis and segmentation, as well as tempo and structure extraction.

1.1 Keywords

music visualization, audio analysis, audio similarity

2. INTRODUCTION

There has been considerable interest in making music visible. Most approaches quantitatively render the time and/or frequency content of the audio signal, using methods such as the oscillograph and sound spectrograph [1], [2]. Other visualizations are derived from note-based or score-like representation of music, typically from MIDI note events [3].

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, the second chorus sounds like the first. On a shorter time scale, successive bars are often repetitive, especially in popular music. This paper presents a novel method of visualizing music by its acoustic similarity or dissimilarity in time, rather than absolute acoustic characteristics. Self-similarity is visualized in a two-dimensional representation of time, such as Figure 1.

An audio file is represented as a square. Time runs from left to right as well as from bottom to top. Thus the bottom left corner of the square corresponds to the beginning of the piece, while the top right corresponds to the end. 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 the 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 pattern: a note repeated twice will give 4 bright areas at the corner of a square. The lower left area corresponds to the first instance of the note while the upper right region corresponds to the second. The two remaining 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 between repetitions.

ACM Multimedia '99 10/99 Orlando, FL, USA

2.1 Audio parameterization

To calculate the similarity between two audio "instants," they are first parameterized using the short-time fourier transform resulting a spectrogram. Figures 5 and 6 use this representation. Alternatively, Figures 1 and 7 were constructed from a Mel-frequency cepstral representation, including an energy term. The short paper format does not permit a fuller discussion; the interested reader is referred to [4] for further details.

2.2 Similarity Measure

Given two feature vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ 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 our features), the product can be normalized to give the cosine distance between the vectors:

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

Because windows, hence feature vectors, occur at a rate much faster than typical musical events, a better similarity measure S can be obtained by computing the vector correlation over a window w. Thus

$$S_w(i,j) \equiv \frac{1}{w}\sum_{k=0}^{w-1} s(i+k, j+k)$$

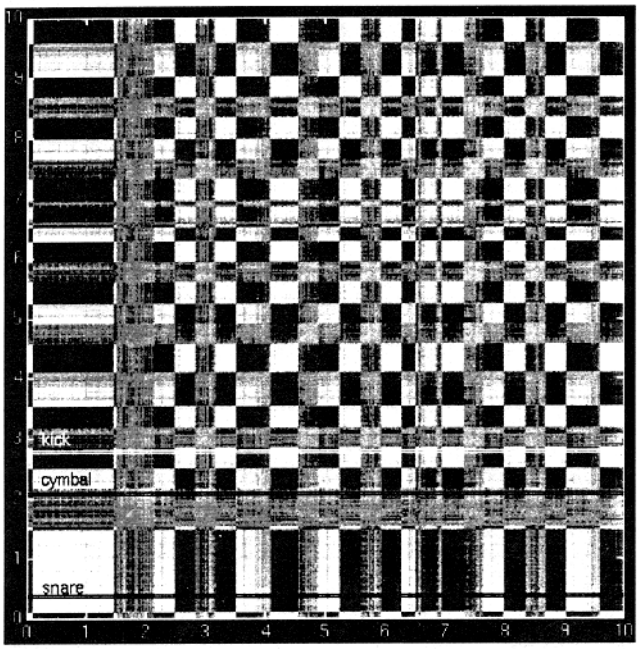

Figure 1. Self-similarity visualization of drum pattern

This also captures the time dependence of the vectors, and serves as the similarity metric used for the images in this paper.

2.3 Visualization Method

To visualize an audio file, an image is constructed so that each pixel at location *i*, *j* is given a grayscale value proportional to the similarity measure. Below are some examples; the time scales are seconds. For reasons of resolution and space most images are from small excerpts of longer works.

2.4 "Drum Solo" Example

Figure 1 is a sampled "drum solo," taken from an audio test CD. The solo starts with a snare drum roll, followed by a syncopated alternation of kick and snare hits and cymbal accents. The alternation of instruments is particularly visible in this Figure. For example, the 2 x 2 "checkerboard" between the second and third seconds of the recording is a snare drum hit followed by a kick drum hit. This sequence is

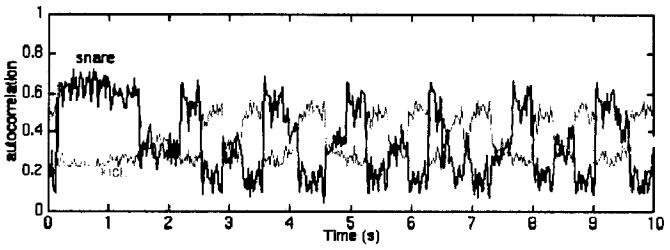

Figure 2. Graph of self-similarity vs. time

reversed (kick, then snare) between seconds 3 and 4.

Figure 2 shows the autocorrelation as a more conventional plot. Because both instrument and timing information could be automatically derived from the plot, this information could be used to generate a MIDI representation of the source music, which is in general a very difficult problem for unpitched instruments

2.5 Bach Prelude

Figure 4 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 3 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 5 shows a similar excerpt, this time from a MIDI realization using a passable piano sample and a strict tempo. Beginning silence is also visible as a square at lower left. Here, unlike the human performance, all notes have exactly the same duration and articulation. Figure 6 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. (Overlapping notes, such as the initial C, were ignored.) Clearly the structures of all three figures are highly similar, indicating that they do indeed capture the underlying structure of the music.

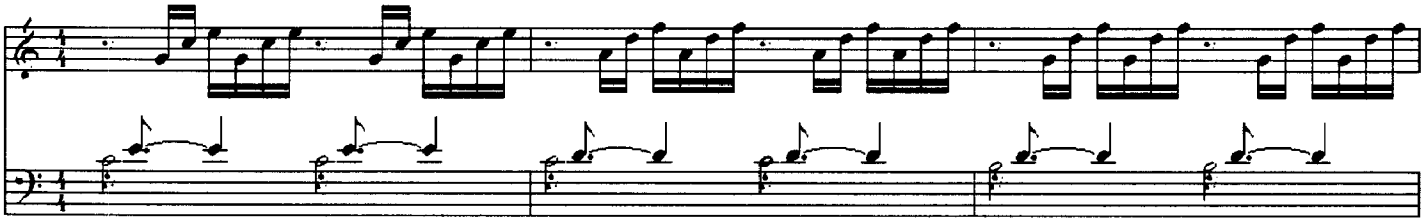

Figure 3. First bars of Bach's *Prelude No. 1 in C Major*, BVW 846, from *The Well-Tempered Clavier*

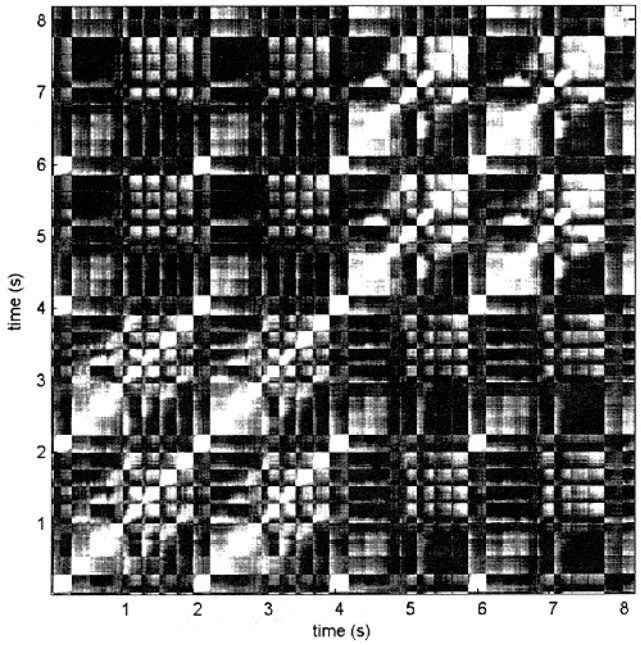

Figure 4. Self-similarity of Bach's *Prelude No. 1*

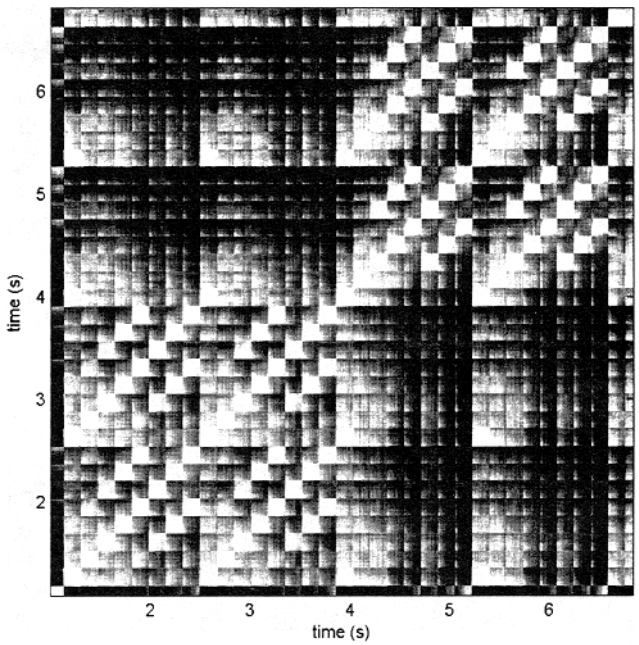

Figure 5. Self-similarity of a MIDI rendition of *Prelude No. 1*

2.6 *Day Tripper* by the Beatles.

Figure 7 shows the entire song *Day Tripper* by the Beatles. The image has been annotated to show the canonical pop song structure. Vocals in the first verse start at about 18 seconds; the 4 vocal phrases ("Got a good reason...") can be seen echoed in the second verse ("She's a big teaser...") about 20 seconds later. The chorus starts at about 30 seconds; the prominent feature at 40 seconds is the sustained "so" ("it took me *so* long/to find out") which is recapitulated halfway through the second verse at 75 seconds. Note that the "so" of the third chorus (130 seconds) is not similar to the preceding choruses as it is sung in falsetto. The signature guitar riff is particularly clear in both the introduction and its note-for-note recapitulation in the coda, and is also visible in the verses and outro, which fades out. The bar-by-bar and section-by-section periodicity are evident in the diagonal lines prevalent throughout the image.

3. Retrieval by Similarity

These visualizations show how acoustically similar passages can be located in an audio recording. Similarity can also be found across recordings as well as within a single recording. As an immediate application, this would be useful wherever known music or audio needs to be located in a longer file. For example, it would be a simple matter to find the locations of the theme music in a news broadcast, or the times that advertisements occur in a TV broadcast if the audio was previously available. In this case, the similarity measure would be computed between all

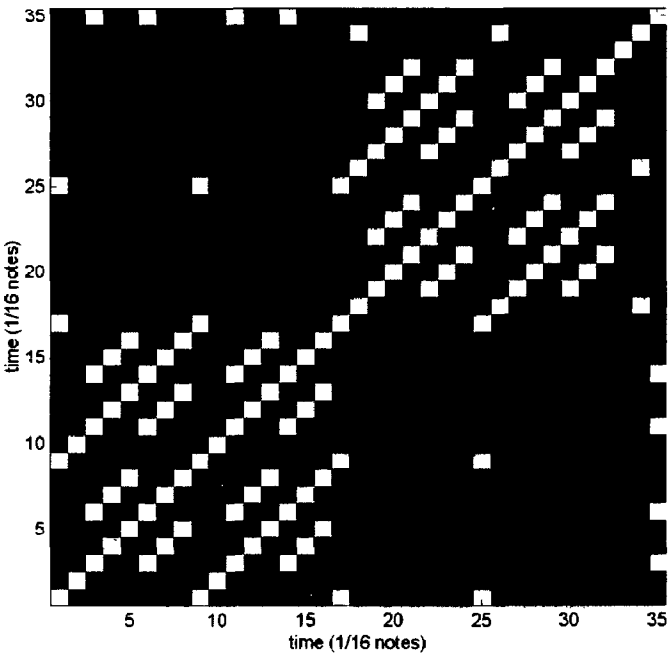

Figure 6. Self-similarity of *Prelude No. 1* computed from MIDI data only -- no acoustic information was used.

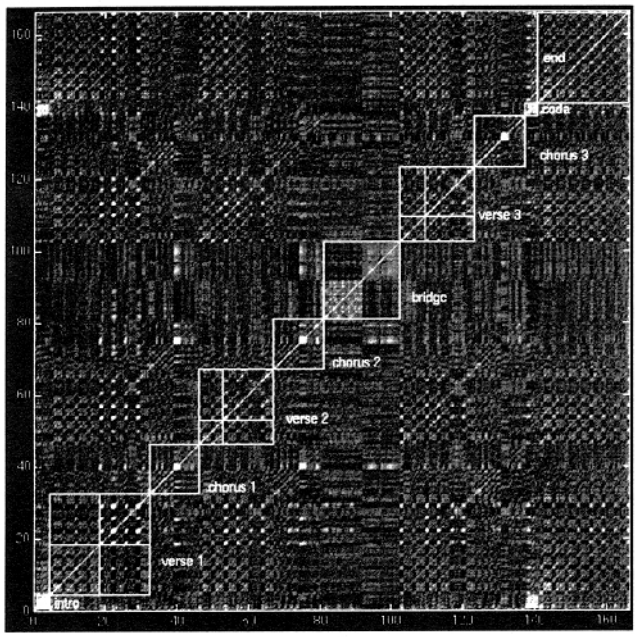

Figure 7. ***Day Tripper*** **by Lennon/McCartney**

frames of the source commercial and the TV broadcast, resulting in a rectangular similarity matrix. Advertisement onset times could be determined by thresholding the similarity matrix at some suitable value.

The structure of most music is sufficient to characterize the work. As proof by example, human experts can identify music and sound by visual structure alone. Victor Zue of MIT teaches a course in "reading" sound spectrographs. In a double-blind test, Arthur G. Lintgen of Philadelphia was able to distinguish unlabeled classical recordings by identifying the softer and louder passages visible in the LP grooves [7]. These examples indicate that the visualization method presented here might be useful for music retrieval by similarity. Not only can *acoustically* similar audio be located, but *structurally* similar audio should be straightforward to find, by comparing similarity visualizations. For example, different performances of the same score should have a similar structural visualization regardless of how or when they were performed or recorded, or indeed the instruments used.

3.1 Structure/Tempo Extraction

Unlike practically all prior work, this method characterizes self-similarity rather than specific audio attributes such as spectrum or pitch. Given the audio of a particular performance and a MIDI file representation of the same piece, as on Figures 5 and 6, it would be possible to warp the similarity matrix from the known-tempo MIDI rendition to match that of the original performance. The warping function would then serve as a tempo map, allowing the MIDI file to be played back with the tempo of the original performance. Other indications of tempo and structure could be similarly derived from the similarity map.

4. ACKNOWLEDGEMENTS

Thanks to the staff of the Institute for Systems Science (now KRDL) in Singapore. This work was funded by a William J. Fulbright Fellowship administered by the Committee for the International Exchange of Scholars.

5. REFERENCES

[1] Potter Ralph K., George A. Kopp, Harriet C. Green, *Visible Speech*, D. Van Nostrand Co., NY, 1947

[2] Koenig, Walter K., H.K. Dunn, L.Y. Lacey, "The Sound Spectrograph," in *JASA*, Vol. 18, p. 19-49.

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

[4] Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993

[5] Foote, Jonathan. "Content-Based Retrieval of Music and Audio," in C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, Vol. 3229, pp. 138-147, 1997.

[6] Carey, M. J., et al., "A Comparision of Features for Speech and Music Discrimination," in *Proc. ICASSP '99*, vol. 1, pp. 149-152, IEEE, Phoenix AZ 1999

[7] Johnson, P., "sci.skeptic FAQ," Section 0.6.2, http://www.faqs.org/faqs/skeptic-faq/[1], 1999

[1] The author would appreciate any pointers to more authoritative references.