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 calculated and displayed as a two-dimensional representation. Similar or repeating elements are visually distinct, allowing identification of structural and rhythmic characteristics. Visualization examples are presented for orchestral, jazz, 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 measure

2. 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 attempts are primarily for scientific or quantitative analysis, (though it should be noted that the work of artists like Mary Ellen Bute [3] use quantitative methods such as the cathode ray oscilloscope towards artistic ends). Other visualizations are derived from note-based or score-like representation of music, typically from 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, the second chorus sounds like the first, and a theme is more or less similar to its variations. On a shorter time scale, successive bars are often repetitive, especially in popular music. This paper presents a novel method of visualizing the structure of music by its acoustic similarity or dissimilarity in time, rather than absolute acoustic characteristics or note events. Self-similarity is visualized in a two-dimensional representation of time. This paper presents methods[1] of displaying the acoustic self-similarity of an audio file as an image like Figure 1.

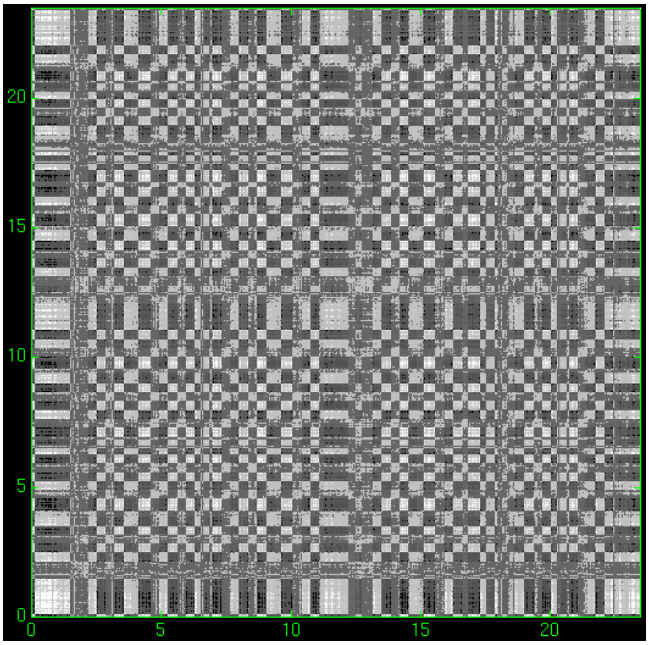

Figure 1. Self-similarity visualization of drum pattern

These images graphically depict the similarity between two time regions in an audio file. An audio file is represented as a square. Each side of the square is proportional to the length of the piece, and time runs

[1] This work was done at the Institute of Systems Science (now KRDL), affiliated with the National University of Singapore.

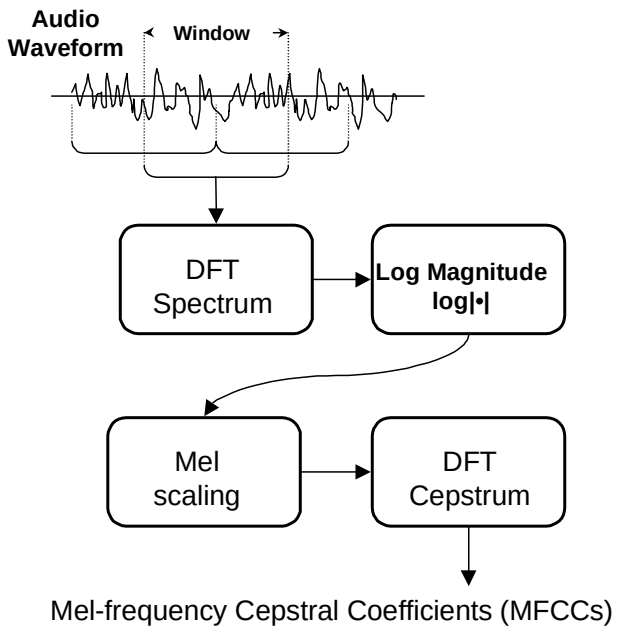

Figure 2. Acoustic processing for similarity measure

from left to right as well as from bottom to top. In the Figures, both axes are labeled with time in seconds. 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 times 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. (Technically, the autocorrelation is always a maximum at a lag of zero.) In this visualization, regions of self-similarity appear as bright regions off the diagonal, as in Figure 1. Relatively novel regions appear as dark squares. Repetitive similarity, such as repeating notes or motifs, show up as a checkerboard pattern. Long repeated themes are visible as diagonal lines parallel to and separated from the main diagonal by the time difference between repetitions.

3. Similarity Analysis

To understand a visualization like Figure 1, it helps to know how it is constructed. Consider the bottom row of pixels (or the left column; images are symmetric). This is colored by how similar the first instant of the piece is to the rest. (For the visualizations here, an "instant" is about 1/10 of a second). Thus the bottom row's halfway point is colored proportionally to the similarity between the first instant and an instant halfway through, and so forth. As we consider rows progressively higher above the bottom row, we consider instants progressively further into the piece, and compare them with the audio from start to finish across the row.

3.1 Audio parameterization

To calculate the similarity between two audio "instants," they are first parameterized into Mel-frequency cepstral coefficients (MFCCs) plus an energy term. Figure 2 shows the steps in parameterizing an audio waveform.

First, the audio is Hamming-windowed in overlapping steps. Each window is 25 mS wide and are overlapped so there are 100 windows, hence feature vectors, in a second of audio. The window width and overlap can be fine-tuned to optimize the visualizations, but the above values offer good results for most audio and are used in the examples. For each window, the log of the power spectrum is computed using a discrete Fourier transform (DFT). The log spectral coefficients are perceptually weighted by a non-linear map of the frequency scale. This operation, called Mel-scaling, emphasizes mid-frequency bands in proportion to their perceptual importance. The final stage is to further transform the Mel-weighted spectrum (using another DFT) into "cepstral" coefficients. This results in features that are reasonably dimensionally uncorrelated, thus the final DFT is a good approximation of the Karhunen-Loeve transformation of the Mel spectra. The high-order MFC coefficients are discarded, leaving the 12 lower-order MFCCs. The audio waveform is thus transformed into 13-dimensional feature vectors (12 MFC coefficients plus energy) at a 100 Hz rate.

The MFCC parameterization was originally developed for speech recognition applications, and has continually out-performed nearly all other parameterization methods. Other parameterizations such as spectral or Perceptual Linear Predictive (PLP) parameters could be used, but MFCCs result in the generally good images shown in the examples of Section 4. MFCCs have been demonstrated to work for music retrieval by similarity [6]. Furthermore, MFCCs have been shown to be better than spectral, pitch, and zero-crossing measures for discriminating between speech and music [7]. It can be objected that using MFCCs for music analysis (as opposed to speech) is "the wrong thing to do." This objection stems from the understanding that the MFCC

parameterization discards pitch information. It one sense it does this—the high-order MFCCs contain the fine harmonic structure characteristic of the driving function—but this is precisely why the MFCCs are appropriate for measuring audio similarity. A better way to characterize the MFCC transformation is as a lowpass "lifter" or frequency-domain filter. In this view, MFCCs are a smoothed representation of a sound's frequency spectrum. A single pitch in the MFCC domain is represented by roughly the envelope of the harmonics, not the harmonics themselves. Thus MFCCs will tend to match similar timbres rather than exact pitches; though single-pitched sounds will match if they are present. Having said this, it is clear from the examples of Section 4 that there may be better representations; in particular, high prominent notes appear to generate a higher similarity measure than other subjectively similar audio. Clearly, work is needed on investigating parameterizations, similarity measures, and the effect of window size on the visualizations.

3.2 Similarity Measure

The similarity measure used here is based on vector autocorrelation. Given two MFCC feature vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ derived from audio windows[1] i and j, a simple metric of vector similarity s is the scalar (dot) product of the vectors

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

This will be large if the vectors are both large and similarly oriented. 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

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

This also captures the time dependence of the vectors. To result in a high similarity score, vectors in a window must not only be similar but their sequence must be similar as well. Considering a one-dimensional example, the scalar sequence (1, 2, 3, 4, 5) has a much higher similarity score with itself than with the sequence (5, 4, 3, 2, 1). This equation serves as the similarity metric used for the images this paper.

3.3 Visualization Method

To visualize an audio file, a window with w is chosen, and the similarity measure $S(i,j)$ is calculated for all window combinations, hence time indexes i and j. Then an image is constructed so that each pixel at location i, j is given a grayscale value proportional to the similarity measure, by scaling the similarity values such that the maximum value is given the maximum brightness. Because of the rapid rate of feature vectors, it is quite possible that a long audio files will result in impracticably large images (a one minute file at a resolution of 100 vectors per second results in a 6000 x 6000-pixel image). To reduce the image size, the similarity can be averaged over short intervals, or the similarity calculated only for certain time indexes. The latter approach is taken here. Because S is already calculated over a window of size w, looking only at indexes that are an integer multiple of w reduces the image size by a that factor. Depending on the length of the audio, the examples of Section use w in the range of 5 to 10.These visualizations let us clearly see the structure of an audio file. Regions of high audio similarity, such as silence or long sustained notes, appear as bright squares on the diagonal. Repeated figures, such as themes, phrases, or choruses, will be visible as bright off-diagonal rectangles. If the music has a high degree of repetition, this will be visible as diagonal stripes or checkerboards, offset from the main diagonal by the repetition time. Below are some examples; the time scales are seconds. For reasons of resolution and space most images are from small excerpts of longer works.

3.4 "Drum Solo" Example

Figure 1 is a sampled "drum solo" taken from an audio test CD. The different drums are visually distinct. The solo starts with a snare drum roll, followed by a syncopated alternation of kick and snare hits and cymbal accents. Figure 3 zooms in to the first ten seconds. With the higher time resolution, the individual snare hits in the beginning roll are visible. 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

[1] As feature vectors come from discrete windows, we use discrete time indexes throughout this discussion.

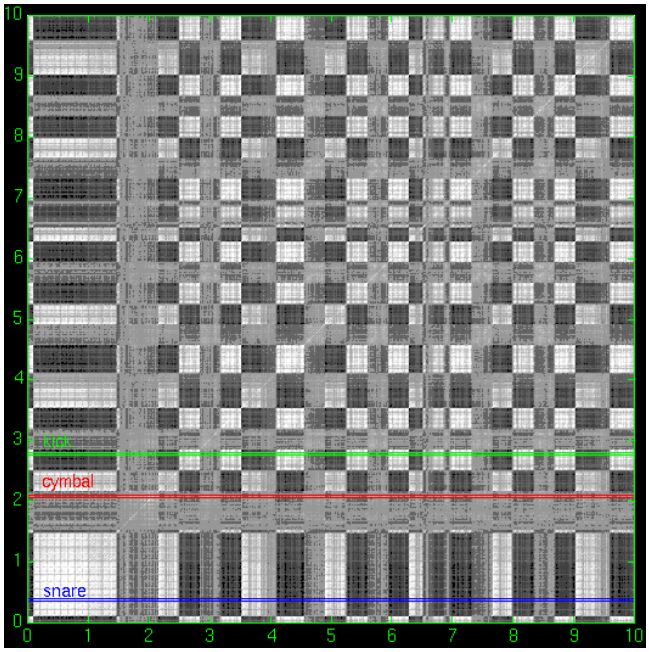

Figure 3. Self-similarity visualization of drum pattern

is a snare drum hit followed by a kick drum hit. This sequence is reversed (kick, then snare) between seconds 3 and 4. To clarify the visualization, stripes marked "snare," "cymbal," and "kick" have been indicated on the Figure. These rows indicate the time similarity of the audio to the indicated instruments because they are the autocorrelations with reference windows containing the respective instruments. For example, the stripe marked "snare" starts brightly, because the audio starts with a snare roll. The different instruments can be clearly distinguished. Of course it helps that they are spectrally very different; it is generally more difficult to differentiate between instruments of similar range and timbre, for example a flute and a clarinet.

Figure 4 shows the autocorrelation stripes as a more conventional plot. Looking at Figure 4, it is clear that a simple maximum would do a very good job at both

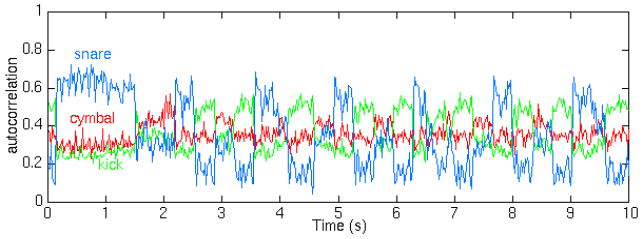

Figure 4. Graph of autocorrelation vs. time

segmenting the audio and classifying the different drum hits 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. This plot highlights features that are not so apparent in Figure 3; for example the kick-drum syncopation clearly visible at 7 seconds. Note particularly the way the high-hat (cymbal) accents are visible at 4 and 7 seconds.

4. More Examples

This section presents additional visualizations across a variety of musical genres. The Electronic Version of this paper includes the playable source audio as well as full-color versions of the annotated visualizations.

4.1 Bach Prelude

Figure 6 shows the first seconds of Bach's *Prelude No. 1 in C Major*, from *The Well-Tempered Clavier*, BVW 846. This lovely 1924 piano performance is by Ferruccio Busoni. The image is fuzzy due to the extremely poor audio quality of the 1924 recording. (Indeed, conventional audio analysis techniques would make little headway due to the poor bandwidth and extremely high noise level of this audio). The striations at the very beginning are clicks and pops due to surface noise from the 78 RPM recording. The visualization makes both the structure of the piece and details of performance visible. For an example of the

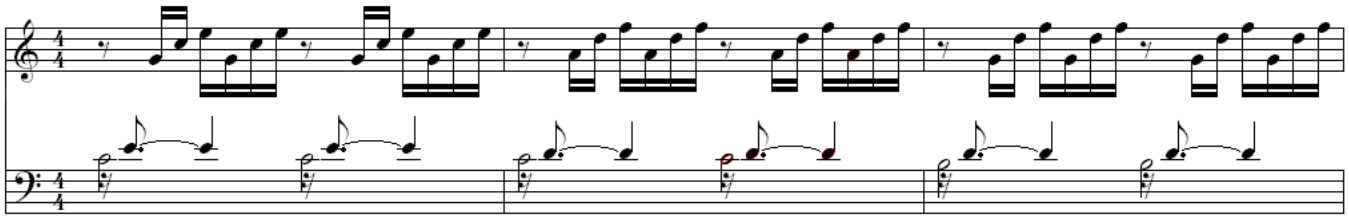

Figure 5. First bars of Bach's *Prelude No. 1 in C Major*, BVW 846, from *The Well-Tempered Clavier*

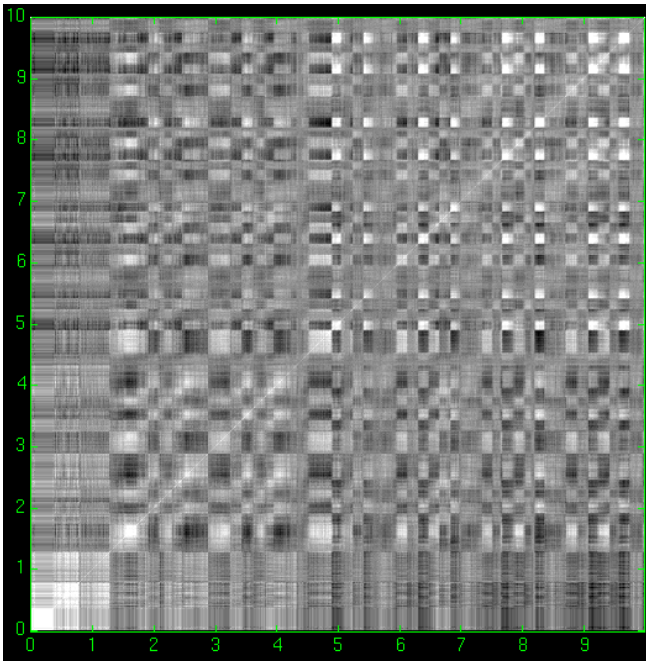

Figure 6. Visualization of Bach's *Prelude No. 1*

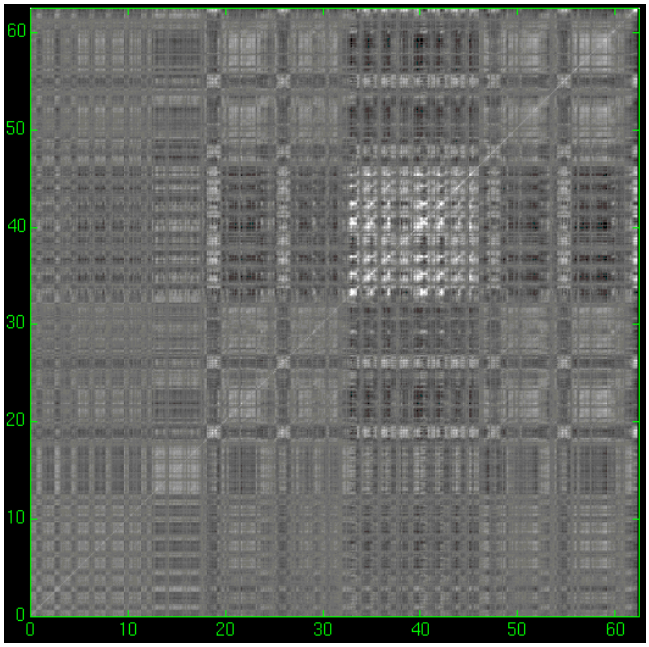

Figure 7. *Take 5* by the Dave Brubeck Quartet

latter, note the slow first notes and the gradual *accelerando* (speedup) as the checkerboard patterns get closer together. The musical structure is clear from the repetitive motifs; multiples of the repetition time can be seen in the off-diagonal stripes parallel to the main diagonal. Figure 6 shows the first few bars of the score: the repetitive nature of the piece should be clear even to those unfamiliar with musical notation.

4.2 Brubeck's *Take 5*

Figure 7 shows the beginning of the Dave Brubeck composition *Take 5* as performed by the Dave Brubeck Quartet. The eponymous 5/4 time signature is visible as a 3-2 subdivision, particularly in the lower left corner. The especially bright regions are due to high notes from the alto saxophone.

5. Mozart's *Horn Concerto*

Figure 8 shows the start of the *Rondo* movement from W. A. *Mozart's Horn Concerto No. 4*. The statement of the theme by the horn and *tutti* restatement by the ensemble are visible in the lower left. While the two statements are melodically identical, they appear dissimilar because of the different timbres. The sustained high horn note causes the bright quartet near the 20-second mark.

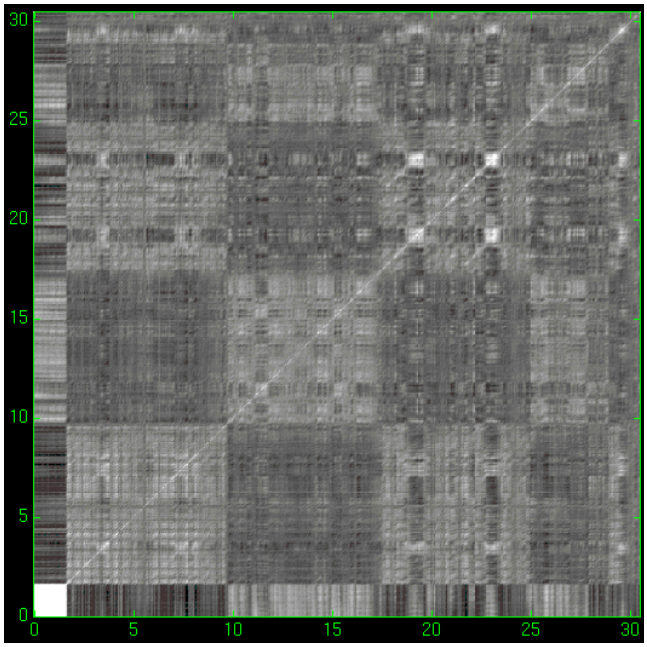

Figure 8. *Rondo* from Mozart's Horn Concerto No. 4

5.1 *Day Tripper* by the Beatles.

Figure 9 shows the entire song *Day Tripper* by the Beatles. The image has been annotated to show the canonical pop song structure, which is: intro verse, chorus, second verse, chorus, bridge, third verse and chorus, coda, and "outro." Vocals in the first verse start

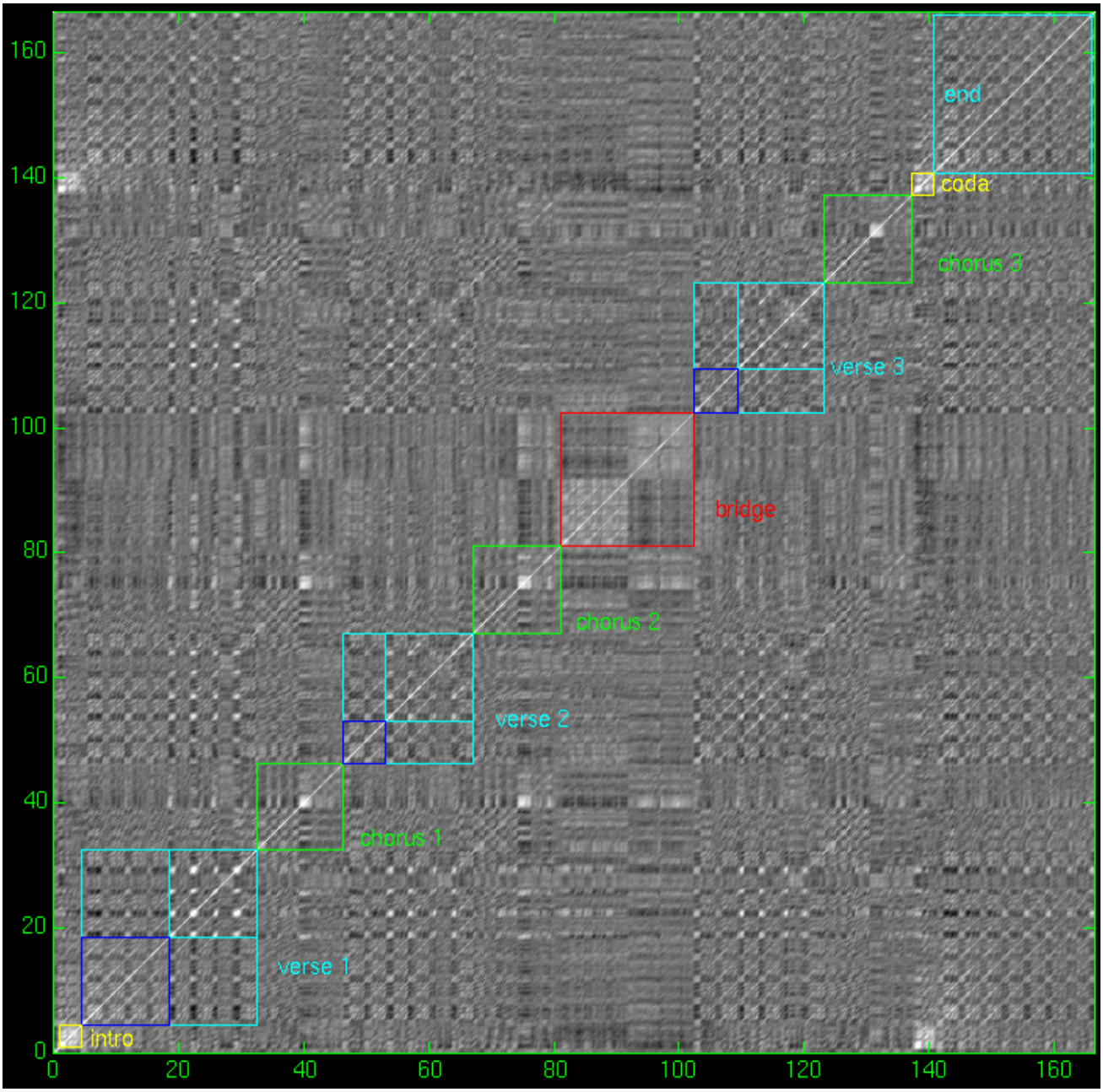

Figure 9. ***Day Tripper*** **by Lennon/McCartney, performed by the Beatles**

at about 18 seconds; the 4 vocal phrases (“Got a good reason/For taking the easy way out.”) can be seen echoed in the second verse (“She's a big teaser...”) about 20 seconds later. The chorus (“She was a day tripper”) 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; it is sung in falsetto, approximately an octave plus a minor third higher than the first two. The first half of the bridge is instrumental while the second contains background vocals (“ah”), the last half can be seen to be similar to first and second “so’s” from the chorus. The repetitive 11-note guitar/bass 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.

6. Applications of Visualization Techniques

This technique should aid musicological analysis. Having visual representations of, for example, two different performances of the same symphonic movement would allow comparisons of tempo and emphasis in the two realizations. The grayscale visualizations could be colored to add another variable dimension; for example, using a colormap to indicate relative volume, so that, for example, fortissimo passages are colored reddish while softer passages vary down the spectrum to a blue pianissimo. Differences in both dynamics and tempo would then be clearly visible.

6.1 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 frames of the source commercial and the TV broadcast, resulting in a rectangular similarity matrix. Commercial 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 [8]. 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 symphonic movement should have a similar structural visualization regardless of how or when they were performed or recorded, or indeed the instruments used.

6.2 Structure/Tempo extraction

This last point emphasizes a particularly promising application of the similarity measure. Because self-similarity is being determined rather than any particular audio characteristic, important information can be automatically derived from a the similarity measure. This can be particularly useful; as discussed for Figure 4, it would be possible to generate a MIDI representation from an audio source, even in the absence of pitch information. A very attractive possibility is the automatic determination of tempo. Given the audio of a particular performance and a MIDI file representation of the same piece, 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. This might be particularly useful for archival performances such as the Bach piece of Section 4.1.

7. ACKNOWLEDGEMENTS

Thanks to S. S. Ghosh for the Boccioni recording of the Bach prelude (suggested by Steven Smoliar). Thanks to the staff of the Institute for Systems Science (now Kent Ridge Development Laboratory) in Singapore, where this work was undertaken. This work was funded by a William J. Fulbright Fellowship administered by the Committee for the International Exchange of Scholars.

8. 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] Moritz, William, "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] 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.

[7] 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

[8] 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.