Calculating Musical Rhythm Similarity

Eric Battenberg

***Abstract*—We present a method for comparing the rhythm of two songs. We use self-similarity features called beat spectra that are used to train a Gaussian mixture model which then describes the rhythm of a song. The mixture model parameters of two songs are compared using a perceptually-motivated approximation of the KL-divergence between mixture models.**

***Index Terms*—rhythm similarity, music information retrieval, machine learning**

I. Introduction

AUTOMATIC music information retrieval is an important topic in today's world due to the increasing prevalence of digital media in our daily lives. More and more people are amassing countless gigabytes of compressed digital audio on their hard drives and portable players. Most listeners are limited to sorting their music by artist, album, genre and other types of metadata tagged to their music files. This type of organization makes it very difficult to find songs that actually sound similar or have a common feel. Music information retrieval attempts to solve this problem by automatically providing supplemental information about a song that can be used to compare it to others. There are a number of techniques that compare songs by the timbre, or spectral qualities, of the audio; however, rhythmic similarity, though very important perceptually, hasnt seen as much action. Some authors have pursued methods to calculate individual characteristics of rhythm, such as tempo , meter, and swing, but systems which characterize the patterns and overall rhythmic feel of a song are lacking. We attempt to make progress in this area by modeling and comparing the self-similarity in songs.

II. Methods

A. Feature Extraction and Clustering

The first step in calculating self-similarity within a song is to compute features which describe the spectral content within a small window of time. The most basic feature set with this property is the short-time Fourier transform. A method which more efficiently describes the spectral envelope with respect to the human auditory system involves computing mel-frequency cepstral coefficients (MFCCs), a feature set which is widely used within the speech recognition community. MFCCs are created by summing the energy in sub-bands distributed according the mel scale, a perceptual auditory scale. The log of this vector of sub-band energies is taken and then transformed using a discrete cosine transform (DCT). Only a relatively small number of (usually the first 13-20) DCT coefficients need to be retained in order to adequately describe the spectral envelope of a 23ms window of time.

After MFCCs are computed for each half-overlapping 23ms frame of a song, self-similarity values can be computed between pairs of frames. Frames are compared using the squared Euclidean distance between their MFCCs, i.e. if $\vec{x_i}$ and $\vec{x_j}$ are vectors containing the MFCCs of two frames, we compute:

$$d(i,j) = (\vec{x_i} - \vec{x_j})^T(\vec{x_i} - \vec{x_j}) \quad (1)$$

We then store the results in a matrix where the (i,j)th entry contains $d(i,j)$. This type of similarity matrix was introduced by Foote in [1]. An example similarity matrix is shown in Figure 1.

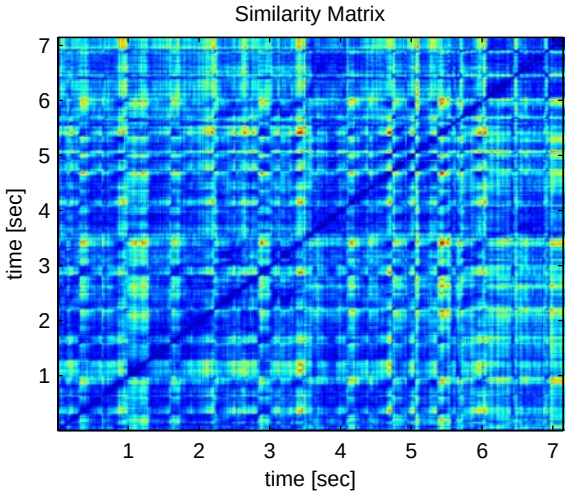

Fig. 1. Similarity matrix from 7 seconds of AC/DC's *Back in Black*. The axis labels are scaled to show time rather than frame number.

To get a final representation of the rhythmic periodicities of an audio signal, sums can be taken over a certain range, R, along each of L diagonals of the similarity matrix:

$$B_m(l) = -\sum_{i=mR}^{(m+1)R-1} d(i, i+l), \quad \text{for } l = 0, 1, \ldots L-1. \quad (2)$$

This results in a signal which Foote calls the "beat spectrum" [1]. The beat spectrum is basically the strength of the similarity at each lag time l.

Beat spectra, $B_m(l)$, are extracted for values of m that cover the entire song (about 100 beat spectra for a four minute song). Then the first 116ms of each 4.75 sec beat spectrum is thrown out as suggested in [2], leaving 400 samples. Next the DCT is taken and only the first 50 coefficients are retained, sine faster variations in the beat spectrum are unimportant to rhythmic description.

Finally, the beat spectra of a song are clustered using expectation-maximization on a Gaussian mixture model

Eric Battenberg, Ph.D. Candidate, Department of EECS, University of California, Berkeley, eric@cnmat.berkeley.edu

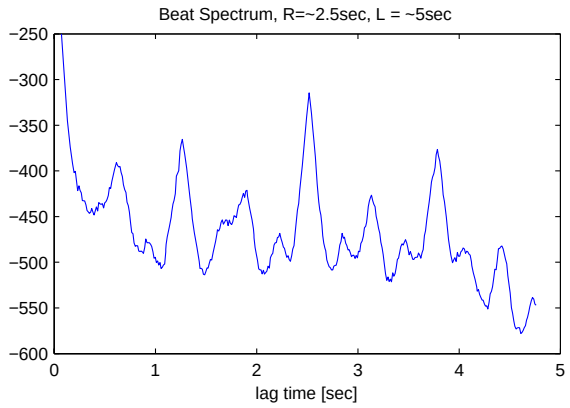

Fig. 2. Example beat spectrum from AC/DC's *Back in Black*. Strong similarity is apparent at lags of 1.25, 2.5, and 3.75 seconds.

(GMM) (initialized using k-means on random initial means). A diagonal covariance matrix was assumed since the DCT hopefully decorrelates the components. Five clusters were adequate to describe rhythmic variations in typical pop songs. For each cluster, c, we arrive at a mean vector, $\vec{\mu_c}$, and covariance matrix, Σ_c. These are the final parameters used to describe the rhythm in a song. A block diagram summarizing the feature extraction and clustering process for a single song is shown in Figure 3.

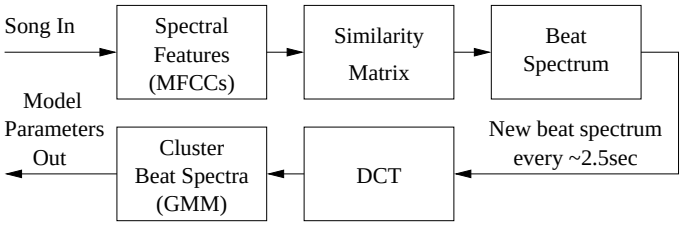

Fig. 3. Block diagram of feature extraction and model training.

B. Model Comparison

The goal of this project was not only to extract a rhythmic description of a song, but to make a meaningful rhythmic comparison between songs. To achieve this we need to compare the model parameters of two songs somehow. An obvious choice to compare two probability distributions would be the Kullback-Leibler divergence; however, there is no analytic formula for the KL divergence between two GMMs.

The most accurate method of approximation is Monte Carlo sampling [3]; however, its accuracy comes with computational cost. Another popular method is the earth mover's distance [4] proposed for image retrieval. This, however, requires a complex dynamic programming algorithm.

A relatively simple method that performs fairly well is the Goldberger approximation [5]. This method involves matching the most similar clusters to each other and then computing a weighted sum of the KL divergence between each of the matched Gaussian clusters. If f_a and g_b are the distributions of the ath and bth clusters of mixtures f and g and π_a and ω_b are the priors of each cluster, the best matching clusters are determined as:

$$m(a) = \arg\min_b D_{\mathrm{KL}}(f_a||g_b) - \log(\omega_b) \tag{3}$$

Goldberger's approximation is then:

$$D_{\mathrm{KL}}(f||g) = \sum_a \pi_a \left(D_{\mathrm{KL}}(f_a||g_{m(a)} + \log \frac{\pi_a}{\omega_{m(a)}} \right) \tag{4}$$

Although this approximation does not possess as many desirable theoretical properties as other approximations, it still performs well empirically [3].

C. Tempo-Similarity Distance

The KL divergence between clusters can be calculated from the following expression:

$$D_{\mathrm{KL}}(N0||N1) = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) - N \right. \tag{5}$$

$$\left. +(\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) \right)$$

The second line is interesting because it is basically the Euclidean distance between mean vectors in which the individual differences are weighted by the inverse covariance matrix. In our case, with a diagonal covariance matrix, the differences are simply divided by the variance of each component. This Euclidean distance weighted comparison is similar to the beat spectrum comparison proposed by Foote in [2]:

$$D(\vec{B_1},\vec{B_2}) = (\vec{B_1}-\vec{B_2})^T(\vec{B_1}-\vec{B_2}) \tag{6}$$

The problem with this comparison is that it leaves out the fact that peaks at nearby lag times would sound almost indistinguishable to a person. As an extreme example, consider two beat spectra with very narrow peaks at 1.16 sec and 1.22 sec respectively. The Euclidean distance between beat spectra would be very large since the peaks occur where the opposite beat spectrum takes a value close to zero. To a listener though, these periodicities represent tempos of 52 and 49 beats per minute, an undetectable perceptual distance. To remedy this problem, we propose the *tempo-similarity distance* (TSD):

$$D_{\mathrm{TS}}(\vec{B_1},\vec{B_2}) = (\vec{B_1}-\vec{B_2})^T S_T(\vec{B_1}-\vec{B_2}) \tag{7}$$

where S_T is a matrix which holds the perceived similarity between beat spectrum components. We implement S_T by stacking normalized triangular kernels centered at each lag time with widths proportional to the "just noticeable difference" (JND) at the corresponding tempo. According to [6], the JND is approximately 8% of the reference tempo. The TSD is compared to Euclidean distance for a basic rock beat at various tempos in Figure 4.

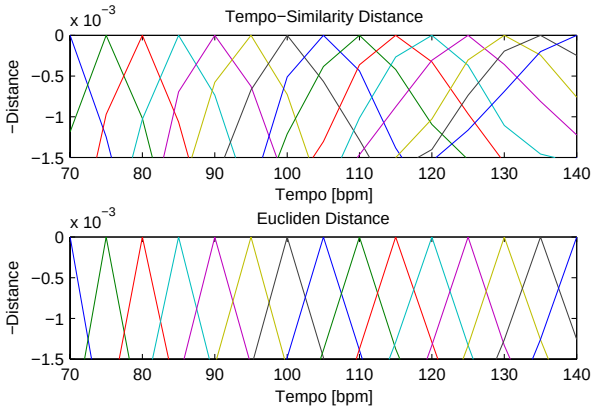

Fig. 4. The distance between a basic rock beat at various tempos. For the tempo-similarity distance, notice how faster tempos have a more gradual slope as tempo is varied linearly.

Now that we've introduced the tempo-similarity distance, we must factor it into our cluster KL divergence calculations. The diagonal covariance matrix can be factored and S_T placed between, yielding the modified distance-related term from the second line of eqn. 5:

$$(\mu_1 - \mu_0)^T \Sigma_1^{-\frac{1}{2}} S_T \Sigma_1^{-\frac{1}{2}} (\mu_1 - \mu_0) \tag{8}$$

However, the beat spectrum components have been transformed by the DCT, so the tempo-similarity matrix needs to be calculated using the center frequency of each DCT component, rather than the lag time of each beat spectrum component as in eqn. 7.

This modified distance term in the KL divergence bears strong resemblance to the tempo-similarity distance in eqn. 7. It is basically the TSD between the mean beat spectra of two clusters with the differences divided by the standard deviation of the corresponding component. We can make the following simple transformation to a cluster's covariance matrix so that tempo similarity properties will be inherent to the original KL divergence:

$$\Sigma_c' = \Sigma_c^{\frac{1}{2}} S_T^{-1} \Sigma_c^{\frac{1}{2}} \tag{9}$$

The first term on the first line of eqn. 5 remains unchanged from this transformation due to properties of the determinant. The second term, involving the trace, does change and must be computed using covariance matrices transformed as in eqn. 9.

III. Results

The rhythm similarity system described in the previous section was tested on a set of six songs. Two rhythmically similar songs were chosen from each of three artist: Bob Marley, Santana, and AC/DC (see Figure III). The genre of each artist can be described as reggae, Latin rock, and hard rock, respectively. All songs had relatively slow tempos in order to demonstrate the discriminative ability of the system. The rhythmic distance between songs, calculated using transformed covariance matrices in the Goldberger approximation of the KL divergence between models, is displayed in Figure III. For comparison, we also include the distance computed when using the modulation spectrum of note onset energies to train the mixture models. This feature set is another valid description of rhythmic periodicities. Onset energies are frequently used in tempo estimation as in [7]. Though effective for pure tempo estimation, the onset energies do not perform nearly as well as the beat spectrum in our implementation.

	Artist	Song
1.	Bob Marley	Is This Love
2.	Bob Marley	Buffalo Soldier
3.	Santana	Black Magic Woman
4.	Santana	Evil Ways
5.	AC/DC	Back in Black
6.	AC/DC	Deep in the Hole

Fig. 5. Songs used in distance calculations

The results achieved using the beat spectrum features seem to match better with perceived differences. The AC/DC songs both have a slow, heavy, driving beat and our system shows them to be strongly similar, though very different from the other artists. The songs by both Bob Marley and Santana are similar to each other with Evil Ways having a slightly faster tempo, and, therefore, a larger distance from the other songs.

The results achieved using the onset modulation spectrum are not completely undesirable, since similarities are still evident within two of the artists. As with any real-world implementation, the results are highly dependent on the variable parameters chosen in the particular implementation. Though the beat spectrum performs better in this instance, both feature sets may prove useful for describing rhythm in this type of system.

IV. Conclusion

The main contribution of this project seems to be the perceptually-motivated tempo-similarity distance used to compare two rhythm models. This technique more accurately models how humans judge tempo differences and it performs well in our very limited trials. Along with the TSD, the beat spectrum was presented as a valid feature to be used in rhythmic comparison. Its discriminative capabilities are very good when used to train a mixture model to represent the rhythm of a song.

The system presented here has much room for improvement. There are many parameters used in the implementation that can be optimized to yield better results and more efficient computation. A first step in future work would be to test this system using a much larger catalog of songs to gauge its true potential.

References

[1] J. Foote and S. Uchihashi, "The beat spectrum: a new approach to rhythm analysis," *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 881–884, 2001.

[2] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," *Proceedings of the International Conference on Music Information Retrieval*, vol. 3, pp. 265–266, 2002.

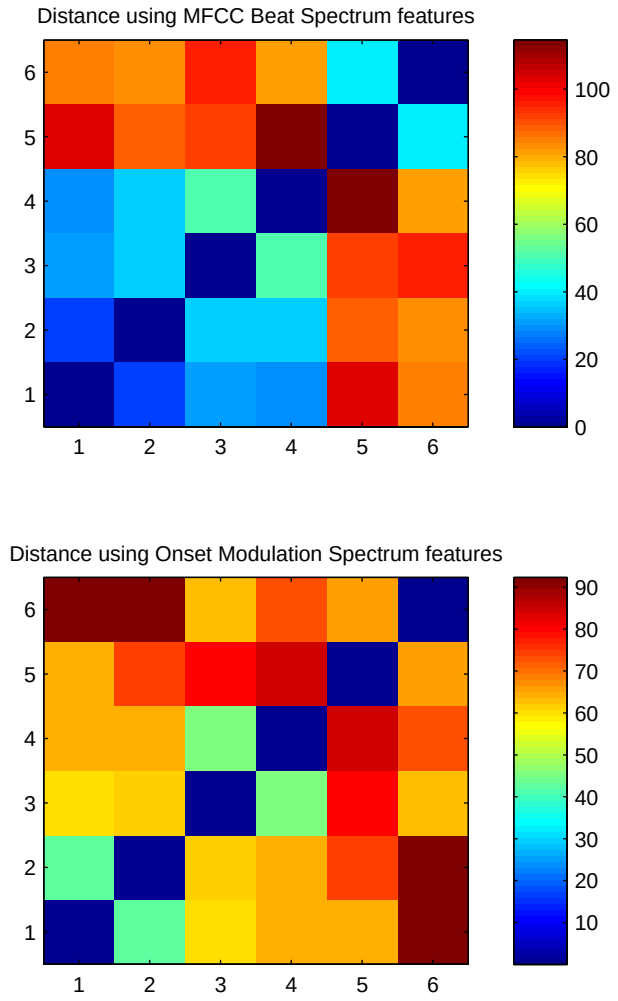

Fig. 6. Distances between the mixture models describing each song. The top matrix uses beat spectrum features, while the bottom matrix uses onset modulation spectrum features.

[3] J. Hershey and P. Olsen, " Approximating the Kullback Leibler divergence between gaussian mixture models," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. 317–320, 2007.

[4] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[5] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 487–493, 2003.

[6] K. Thomas, "Just noticeable difference and tempo change," *Journal of Scientific Psychology*, pp. 14–20, May 2007.

[7] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, vol. 14, no. 1, pp. 342–355, 2006.