FEATURE WEIGHTING FOR SEGMENTATION

R. Mitchell Parry *Irfan Essa*
Georgia Institute of Technology
College of Computing / GVU Center

ABSTRACT

This paper proposes the use of feature weights to reveal the hierarchical nature of music audio. Feature weighting has been exploited in machine learning, but has not been applied to music audio segmentation. We describe both a global and a local approach to automatic feature weighting. The global approach assigns a single weighting to all features in a song. The local approach uses the local separability directly. Both approaches reveal structure that is obscured by standard features, and emphasize segments of a particular size.

1. INTRODUCTION

With vast and growing digital music libraries and personal music collections, manual indexing procedures are becoming intractable. Automatic audio segmentation algorithms such as [1], [3], and [5] provide indices and reveal structure. Hierarchical schemes, such as [6], require a variable scale of analysis. We claim that a scale-specific feature weighting enables this process by revealing appropriately sized segments. The significance of this work is that we apply feature weighting to better distinguish adjacent audio segments.

Feature weighting has been exploited in the machine learning community to improve classification. For instance, features may be weighted by criterion functions such as a correlation criterion or Fisher's criterion [4], which we describe briefly. Correlation criteria correlate each input feature with its class label. One correlation criterion that we use is given by:

$$C = \frac{(x-\bar{x})\bullet(y-\bar{y})}{\|x-\bar{x}\|\|y-\bar{y}\|}, \quad (1)$$

where x is a vector of scalar samples, y is the class label (±1), and $\bar{x}$ and $\bar{y}$ are their means [4]. Fisher's criterion measures the discriminability between two classes in one dimension and is given by:

$$F = \frac{(\bar{x}_1-\bar{x}_2)^2}{\sigma_1^2+\sigma_2^2}, \quad (2)$$

where $\bar{x}_i$ and σ_i^2 is the sample mean and variance for class i, respectively [2].

2. METHODS

Although our general feature weighting method is applicable to any feature set, we focus on identifying large spectral changes as segment boundaries. Therefore, we extract principal components of the low-frequency power spectrum as suggested by Foote [3].

Our general method is to consider every potential transition (between every adjacent pair of N frames) and construct a weight vector for each. We use a criterion function to rate how well each feature discriminates between frames on either side of the boundary in a local analysis window. The size of the window indicates the scale of analysis.

2.1. Features

We partition the audio into 50 ms non-overlapping frames, apply a Hanning window to each frame and compute the logarithm of the magnitude of its Fast Fourier Transform. Continuing with Foote's methodology, the high-frequency portion is ignored (corresponding to frequencies greater than ¼ the sampling rate). Finally we use Principal Component Analysis (PCA) to find the seven primary directions of variance within the raw feature space. This preserves much of the variance in the data while reducing the dimensionality. Thus, we use a seven-element feature vector to represent each frame.

2.2. Local Weight Vectors

Let $X_{i,j}$ represent our features where i is the feature index and j is the frame number. We designate $W_k^C = \langle C_{1,k}, C_{2,k}, ..., C_{7,k} \rangle$ as the local correlation criterion weight vector for the boundary between frames k and k+1 using equation (1):

$$C_{i,k} = \frac{(x_{i,k}-\bar{x}_{i,k})\bullet(y-\bar{y})}{\|x_{i,k}-\bar{x}_{i,k}\|\|y-\bar{y}\|}, \quad (3)$$

where $x_{i,k} = \langle X_{i,t+1}, X_{i,t+2}, ..., X_{i,t+n_w} \rangle$, $t = k - n_w/2$, y is a vector of $n_w/2$ ones followed by $n_w/2$ negative ones (representing samples before and after the boundary, respectively), and n_w is the number of frames in the analysis window. Using equation (2) we designate $W_k^F = \langle F_{1,k}, F_{2,k}, ..., F_{7,k} \rangle$ as the local Fisher's criterion weight vector:

$$F_{i,k} = \frac{\left(\bar{x}_{1,k} - \bar{x}_{2,k}\right)^2}{\sigma_{1,k}^2 + \sigma_{2,k}^2}, \quad (4)$$

where $x_{1,k} = \left\langle X_{i,t+1}, X_{i,t+2}, \ldots, X_{i,t+n_w/2} \right\rangle$ represents the samples before the boundary, $x_{2,k} = \left\langle X_{i,t+n_w/2+1}, X_{i,t+n_w/2+2}, \ldots, X_{i,t+n_w} \right\rangle$ represents samples after the boundary, $\bar{x}_{i,k}$ is the mean, and $\sigma_{i,k}^2$ is the variance of $x_{i,k}$.

2.3. Global and Local Approach

We consider two main approaches to using the *N*-1 local weight vectors W_k^C and W_k^F: summarizing a *global* weight vector for application to the whole song and substituting the *local* separability as a transition rating itself.

The global approach produces one weight vector for the whole song. One advantage of this is the potential generalization if songs from the same artist or genre require similar global weight vectors. Applying a small set of likely feature weights avoids the cost of computing song-specific weight vectors. This approach smoothes local fluctuation and ensures consistency across the song. In addition, the newly weighted features may be further analyzed, such as visualizing the self-similarity matrix. Given our *N*-1 weight vectors in a song, we compute a simple average to generate the single global weighting.

The local approach uses the total separability between samples on either side of a boundary directly as the transition rating. This is analogous to using the local weight vectors to modulate features within their local analysis window for a given boundary. We use the sum of squared feature weights computed from equation (3) or sum of feature weights from equation (4) as an indication of the separability between samples. For instance, if all of the features easily discriminate between each side of a potential boundary we can rate the boundary highly. This method avoids explicit feature weighting and adapts to local changes in features, similar to [1].

As a compromise between the two approaches, we also weight each local weight vector by its separability score before summarization. Local weight vectors that provide better separation contribute more to the global weight vector.

3. RESULTS

We present the effect of feature weighting in general and results of our automatic weighting procedure. For analysis, we extract the first seven principal components of the low-frequency power spectrum normalized to zero mean as raw features. This constitutes the original weighting inherent in the data. Additionally, we construct a second set of features normalized to zero mean and unit variance. These features provide an unbiased representation on which to begin our feature weighting.

Self-similarity matrices visualize the effect of our feature weighting. Figure 1 depicts the song "Mr. Jones" by the Counting Crows using raw and normalized features on the full song and the first 40 seconds. White indicates similarity, while black indicates dissimilarity. Time proceeds down and to the right. White squares along the main diagonal reveal segments.

The raw features clearly separate large segments. At first glance the uniformly weighted features only obscure these large segments. Upon closer inspection, they reveal smaller segments. Figure 1c and 1d enlarges the first 40 seconds of each similarity matrix to show the effect of this feature weighting. The raw features provide little insight for small segments, while the normalized features, in Figure 1d, reveal this low level repeating structure. Clearly, feature weighting plays a role in what segments are revealed. The next step is to find a feature weighting automatically.

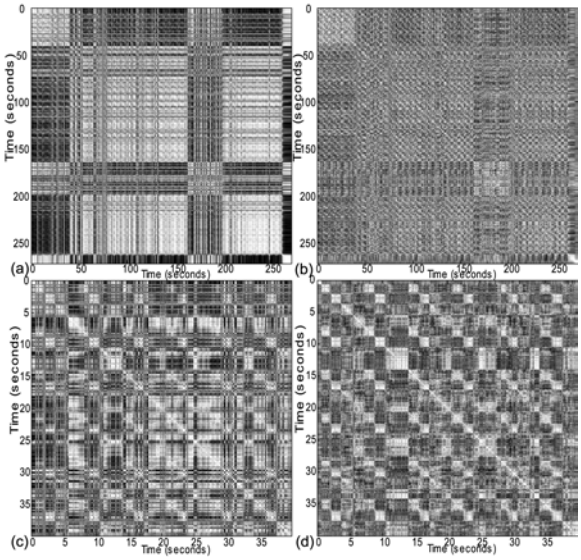

Figure 1. Similarity matrices for "Mr. Jones" by Counting Crows, using the whole song with raw features (a) and normalized features (b); and first forty seconds with raw features (c), normalized features (d).

3.1. Global Approach

We begin with the raw features shown in Figure 1c. After applying our global weighting method with a local analysis window of 3.2 seconds (1.6 seconds before and after a boundary), we would expect to generate an image similar to Figure 1d. Figure 2 compares the result of the correlation criterion (2a) and Fisher's criterion (2b) for generating a global weighting. Fisher's criterion generates feature weights that correctly discern the segments, while the correlation criterion appears muddled by the ill-suited input weights. In addition, Fisher's self-similarity matrix better reveals the segments than the normalized features from Figure 1d.

Using the normalized features for the first 40 seconds of "Porcelain" by Moby, we see that different sized analysis windows reveal different segment sizes. Figure 3 shows the similarity matrix computed with an analysis window of 3.2 s (3a) and 12.8 s (3b). The shorter analysis window reveals small measure-length segments. The larger window reveals the transition to a more complex rhythm track at 20 seconds.

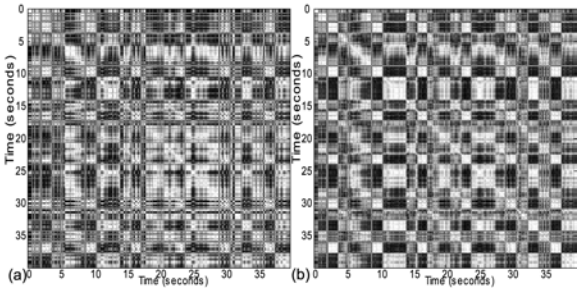

Figure 2. Self-similarity matrix for features weighted by correlation criterion (a) and Fisher's criterion (b).

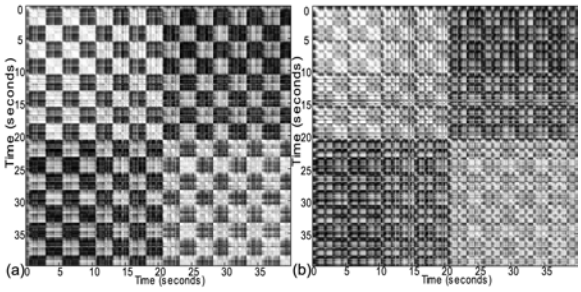

Figure 3. Fisher criterion weighted features using analysis window of 3.2 s (a) and 12.8 s (b).

We can use selective window sizes to reveal different types of segments. Figure 4 and 5 compares selected window sizes using Fisher's criterion for Schumann's "Kinderszenen (Scenes from Childhood) for piano, Op. 15 No. 10". We use a window size of 1.6 seconds to enhance approximately note-length segments. Figure 4a reveals these notes as small white squares along the main diagonal. This piano piece contains alternating soft and loud sections. Using a window size of 12.8 seconds reveals these larger sections (4b). The track begins with approximately one second of silence (the first square on the diagonal). The next relatively quite portion begins at approximately 10 seconds. The feature weights for Figure 4a and 4b are shown in 4c and 4d, respectively. The dominant feature weight in Figure 4d belongs to the first principal component, which indeed captures most of the spectral energy.

Figure 5 shows the self-similarity matrices for the whole piano piece and better depicts the alternation between quiet and loud sections in 5b. The checkerboard pattern of patches indicates alternation between two states: soft and loud. The transitions do not appear crisp because the change in loudness occurs gradually. The weights are shown in 5c and 5d. They closely resemble those computed on the first 40 seconds shown in Figure 5, indicating a consistency across the whole song. Anecdotally, this suggests a computational advantage in computing weights using a small fraction of the song and applying them to the whole song.

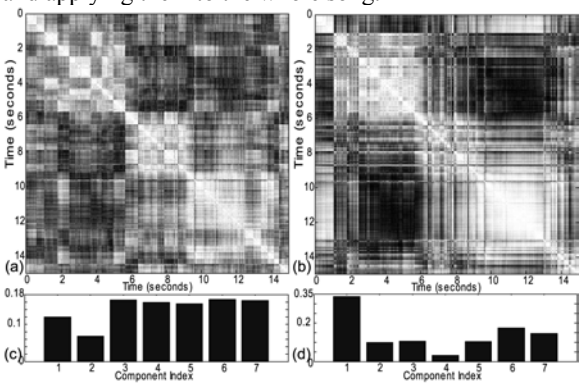

Figure 4. The first 40 seconds of "Kinderszenen" weighted using 1.6 and 12.8 second analysis window in (a) and (b), with weights in (c) and (d), respectively.

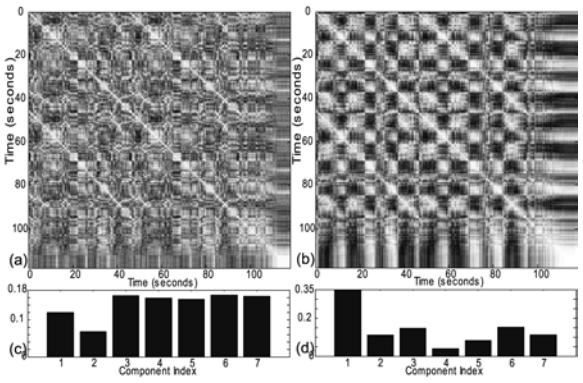

Figure 5. "Kinderszenen" weighted using 1.6 and 12.8 second analysis window in (a) and (b), with weights in (c) and (d), respectively.

3.2. Local Approach

Our second approach uses local separability directly as a transition rating. Using the raw features from Figure 1d that are not well-matched to the scale of analysis, we compare the transition rating from Foote's segmentation algorithm, the sum of the C^2 correlations and the sum of Fisher's criterion for each potential boundary. Figure 6 shows the sum-normalized plots for the raw features in Figure 1c. Fisher's criterion outperforms both of the other transition ratings by pinpointing segment boundaries with tall narrow peaks. Because Fisher's criterion is normalized by variance, feature weights do not affect its measure of discriminability.

A transition rating based on Fisher's criterion may be employed at various scales of analysis. Figure 7 shows the first 40 seconds of "Porcelain". The 1.6 second window (top) reveals each note. The 3.2 second window (middle) emphasizes two-note segments and ignores the intervening transitions. The 12.8 second window identifies the rhythm change at 21 seconds. The false peak at 3 seconds is an artifact of the analysis window overlapping with the start of the file.

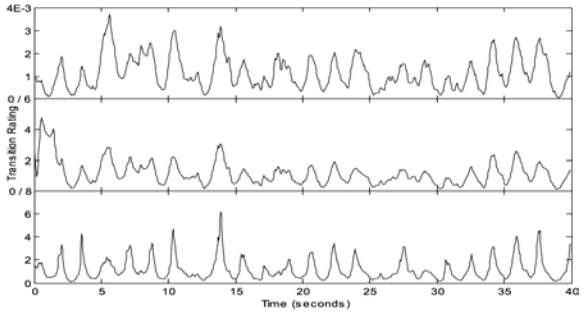

Figure 6. Transition rating computed from segmentation algorithm (top), total correlation (middle), and Fisher's criterion (bottom).

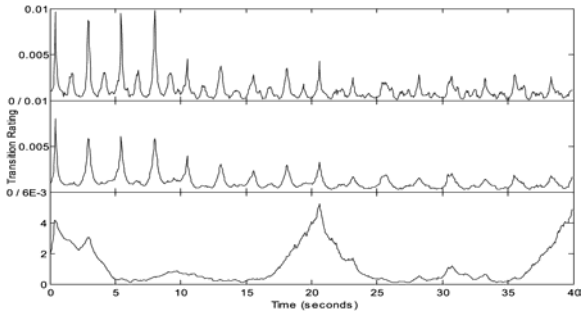

Figure 7. Transition ratings: Fisher's criterion on "Porcelain" using an analysis window of 1.6 seconds (top), 3.2 seconds (middle), and 12.8 seconds (bottom).

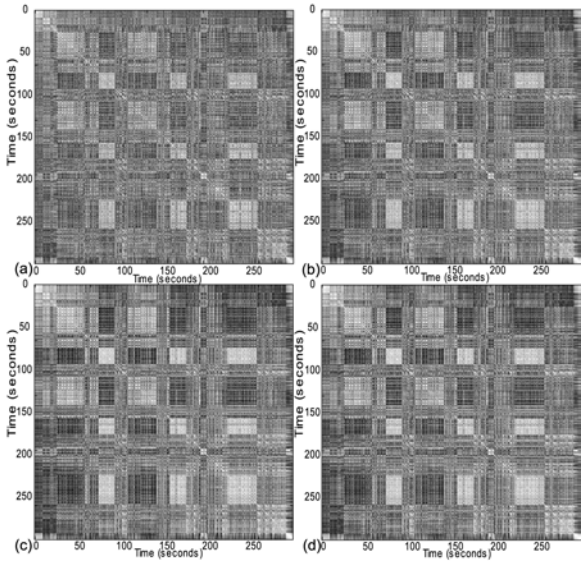

Figure 8. "Head Like a Hole" by Nine Inch Nails: (a) normalized features, (b) simple average, (c) weighted average favored, and (d) peaks only weighted average.

We also combine the two approaches by favoring local weights that provide better separation in computing the global weights. We consider weighting each local weight vector by its separability score and excluding local weights that do not produce peaks in the separability score. Figure 8 shows self-similarity matrices for the original features (a), simple average (b), weighted average (c), and weighted average with peaks only (d). The combined approaches (c and d) provide a subtle improvement in clarity.

4. CONCLUSION AND FUTURE WORK

This paper proposes the use of feature weights to reveal the hierarchical nature of music audio. Clearly, feature weighting affects the size and type of segments revealed. We describe two approaches to automatic feature weighting that emphasize features that separate a particular segment size. Regardless of the inherent weighting in the data, Fisher's criterion outperforms standard self-similarity analysis or a correlation criterion for global and local approaches. A combination of the global and local methods provides a compromise between the two.

We focus on segmenting audio at points of large spectral change. This type of segmentation is suited to locating the point of introduction or exclusion of instruments in a mix because different instruments tend to affect different portions of the frequency spectrum. If the goal is to truly distinguish different instruments in the mix, Independent Component Analysis (ICA) may be preferable to the PCA features used here [7]. Future work may explore this possibility.

5. REFERENCES

[1] Chen, S. & Gopalakrishnan, P. S. "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *DARPA Speech Recognition Workshop*, 1998.

[2] Duda, R., Hart, P., & Stork, D. *Pattern Classification*, John Wiley & Sons, New York, 2nd Edition, 2001.

[3] Foote, J. & Cooper, M. "Media Segmentation using Self-Similarity Decomposition", *Proceedings of SPIE*, 2003.

[4] Guyon, I, and Elisseeff, A. "An Introduction to Variable and Feature Selection", *JMLR*, Vol. 3, March 2003.

[5] Tzanetakis, G, and Cook, P. "Multifeature Audio Segmentation for Browsing and Annotation", *Proc. of WASPAA*, New Paltz, USA, 1999.

[6] Slaney, M. and Ponceleon, D. "Hierarchical Segmentation using Latent Semantic Indexing in Scale Space", *Proceedings of ICASSP*, Salt Lake City, USA, May 2001.

[7] Smaragdis, P. "Redundancy Reduction for Computational Audition, a Unifying Approach", Ph.D. Dissertation, Massachusetts Institute of Technology, Media Laboratory, 2001.