Audio Segmentation by Singular Value Clustering

Shlomo Dubnov and Ted Apel
Department of Music, CRCA, Cal-(IT)^2, University of California, San Diego
sdubnov@ucsd.edu
tapel@ucsd.edu

Abstract

This paper presents a statistical approach to sound texture modeling based on a singular value analysis of spectral features or an eigenvector analysis of their similarity matrix. Using dimension reduction techniques we perform grouping of the signal into similar sounding audio segments that are recurrent in time. The method allows an automatic segmentation of audio signal into larger groups of similar sounding audio objects and can be used for visualization purposes, audio texture synthesis and creative audio manipulations. We present a principled approach that brings methods such as audio similarity analysis and spectral audio basis representations into one framework.

1 Introduction

Many audio texture and even musical pieces can be considered as an alternating juxtaposition of different sound types or objects. Determining segments in an audio signal that correspond to a coherent object (especially in unsupervised manner) is important for visualization purposes, audio texture synthesis and creative audio manipulations.

We are particularly interested in improving the segmentation of audio for sound texture synthesis using sound grains. Lu et al. (2002), have developed an audio texture analysis-synthesis system that uses the self-similarity of a sound over time to create new texturally similar sounds. It is hoped that by singular value clustering as presented here, synthesized sonic textures that accurately reflect the sonic object grouping of the original sound can be produced.

We present an approach that links methods of self-similarity analysis (Foote and Cooper 2001) and spectral audio basis representations (Casey and Westner 2000) into one framework. The described method has close relation to other clustering methods based on pair-wise distances, such as Normalized Cuts and graph partitioning or spectral clustering methods.

2 Analysis Method

The method presented in this paper explores the relation between Singular Value Decomposition (SVD) of spectral features and Eigenvector analysis of recurrence (also called self-similarity or pair-wise spectral distance) matrix. In the first method, SVD analysis is used to represent the data (spectral features) in terms of three matrices, which can be interpreted as audio basis vectors, their normalized expansion coefficients and corresponding variances. The second method consists of finding the eigenvectors of the recurrence matrix, which can be used for dimension reduction of the matrix. In both methods, the first k expansion coefficients or first k eigenvectors are used for clustering.

We will show that both methods are equivalent when the recurrence matrix can be written as a dot product of the spectral features. In the case when the distance measure is more complicated, one can still show that for the case of a symmetric positive definite recurrence matrix, the distances can be represented as a dot product of some (possibly non-linear) function of the features. Even if the function is unknown, clustering can be performed directly on the recurrence matrix.

2.1 Audio Feature Representation

The first step in our analysis is parameterization of the time evolving sound. The sound is windowed typically with a window size of 256 samples at a sampling rate of 18,000 Hz. A cepstral analysis is completed for each windowed segment, and the first few cepstral coefficients are used as a feature vector for that segment of sound.

2.2 Recurrence Analysis

Once we have determined our feature vectors, we create a recurrence matrix, which shows a measurement of the similarity between each pair of feature vectors. This distance measure can be calculated using different distance metrics. Here, we use the normalized dot product of the feature vectors (Foote and Cooper 2001). If d is the distance between the feature vectors X_i and X_j at frames i and j, then

$$d(i,j) = \frac{\langle X_i, X_j \rangle}{\|X_i\|\|X_j\|}. \quad (1)$$

Here $\langle X_i, X_j \rangle$ is the dot product defined as $|X_i||X_j|cos\theta$, where θ is the angle between the vectors, and $|X_i|$ is the norm of X_i. This distance measurement is large when the vectors are of high magnitude and similar, Because of the normalization, low magnitude and similar vectors also produce a large measurement. Assuming a matrix of normalized data column vectors $\mathbf{X} = [X_1 X_2 ... X_p]$, the recurrence matrix in eq. (1) can be written as $\mathbf{D} = \mathbf{X}^T\mathbf{X}$, with $\mathbf{D}_{ij} = d(i,j)$.

For each time segment, these distance magnitude values are plotted on an similarity matrix. Figure 1 shows a similarity matrix of an example sound. We will be using this similarity matrix graph as a basis for one of the methods for partitioning the sound into perceptually similar groups.

2.3 Singular Value Decomposition

Singular value decomposition takes a rectangular matrix of feature vectors (defined as $\mathbf{X}$, where $\mathbf{X}$ is a $n \times p$ matrix) in which the n rows represents the cepstral coefficients, and the p columns represents the time sequence.

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times n}\mathbf{\Lambda}_{n\times p}\mathbf{V}^T_{p\times p} \quad (2)$$

where

$$\mathbf{U}^T\mathbf{U} = I_{n\times n}, \text{ and } \mathbf{V}^T\mathbf{V} = \mathbf{I}_{p\times p}. \quad (3)$$

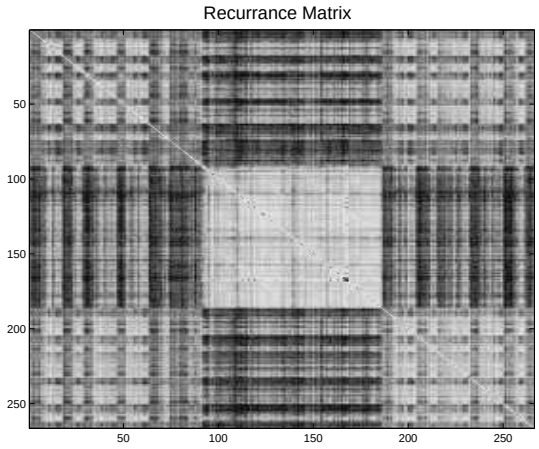

Figure 1: Similarity matrix of example sound.

Calculating the SVD consists of finding the eigenvalues and eigenvectors of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$. The eigenvectors of $\mathbf{X}^T\mathbf{X}$ make up the columns of $\mathbf{V}$, the eigenvectors of $\mathbf{X}\mathbf{X}^T$ make up the columns of $\mathbf{U}$. Also, the singular values in $\mathbf{\Lambda}$ are square roots of eigenvalues from $\mathbf{X}\mathbf{X}^T$ or $\mathbf{X}^T\mathbf{X}$. The singular values are the diagonal entries of the Λ matrix and are arranged in descending order.

For our purposes, SVD has two important properties: it factorizes the data into combinations of the column vectors $\mathbf{U}$ using expansion coefficients that are rows of $\mathbf{V}^T$. Moreover, the relative "significance" of each column of $\mathbf{U}_i$ times row of $\mathbf{V}_i^T$ combination is given by the corresponding values of $\mathbf{\Lambda}$.

2.4 Relation between the Features SVD and Similarity Matrix Eigenvectors

Using SVD of $\mathbf{X}$, and denoting by $\mathbf{S}_x = \mathbf{\Lambda}_x\mathbf{V}^T$, the expansion coefficients of $\mathbf{X}$ we write $\mathbf{X} = \mathbf{U}\mathbf{S}_x$. Due to orthonormality of the vectors $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ the recurrence matrix of $\mathbf{X}$ equals to the recurrence of the expansion coefficients $\mathbf{D} = \mathbf{S}_x^T\mathbf{S}_x$. Since most of the information is contained in the first few components, a data reduced version of $\mathbf{D}$ can be obtained by using only the first rows of $\mathbf{S}_x$. Moreover, one should note that due to orthonormality of $\mathbf{V}$, the expansion coefficients of $\mathbf{X}$ (rows of $\mathbf{S}_x$) are the transposed eigenvectors of the matrix $\mathbf{D}$.

Property I: Eigenvectors S_d of $\mathbf{D}$ are transpose of the expansion coefficients of $\mathbf{X}$, $S_d = S_x^T$.

Proof: Let us denote by $\mathbf{S_x}$ the matrix of eigenvectors S_x (and similarly for S_d). From the arguments above one can write $\mathbf{D} = \mathbf{S}_x^T\mathbf{S}_x$. $\mathbf{D}\mathbf{S}_x^T = \mathbf{S}_x^T\mathbf{S}_x\mathbf{S}_x^T$). We need to prove that the expression in brackets is diagonal matrix. Using $\mathbf{S}_x = \mathbf{\Lambda}_x\mathbf{V}^T$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ we get $\mathbf{S}_x\mathbf{S}_x^T = \mathbf{\Lambda}_x\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}_x = \mathbf{\Lambda}_x^2$.

In case when the distance measure is more complicated, one can still show that for the case of a symmetric positive definite matrix $\mathbf{D}$, the distances can be represented as a dot product of some (possibly non-linear) function $f(X_i)$. Even if the function $f(.)$ is unknown, clustering can be performed directly on $\mathbf{D}$, as shown below.

2.5 Markov Normalization

Converting the distances $\mathbf{D}$ into Markov probability matrix puts it in a generative framework. (Lu, Li, Wenyin, and Zhang 2002) This matrix represents statistics of the data in terms of probability of transition from frames i to j.

We write

$$\mathbf{P}_{ij} = P(j|i) = \frac{d(X_i, X_j)}{\sum_j d(X_i, X_j)}. \quad (4)$$

This normalizaiton takes care of the probability requirement that $\sum_j P(j|i) = 1$, i.e. that being in frame i (frame index i) we will eventually move to some other frame number j.

In the Markov case, the eigenvector analysis is done on the eigenvectors of the transition matrix $\mathbf{P}$. A Markov matrix can be written as $\mathbf{P} = \mathbf{Z}^{-1}\mathbf{D}$, with affinity matrix $\mathbf{Z} = diag(\sum_j D_{ij})$. We denote by S_p the right eigenvectors $\mathbf{P}S_p = \lambda S_p$ and show the following

Property II: The eigenvectors S_q of a symmetrically normalized matrix $\mathbf{Q} = \mathbf{Z}^{-1/2}\mathbf{D}\mathbf{Z}^{-1/2}$, are $S_q = \mathbf{Z}^{1/2}S_p$.

Proof: $\mathbf{P}S_p = \lambda S_p$ can be written as $\mathbf{Z}^{-1}\mathbf{D}S_p = \lambda S_p$. Multiplying on the left by $\mathbf{Z}^{1/2}$ gives $\mathbf{Z}^{-1/2}\mathbf{D}S_p = \lambda\mathbf{Z}^{1/2}S_p$. Using $S_p = \mathbf{Z}^{-1/2}S_q$ gives $\mathbf{Z}^{-1/2}\mathbf{D}\mathbf{Z}^{-1/2}S_q = \lambda S_q$, i.e. S_q is shown to be the eigenvector matrix of the symmetric matrix $\mathbf{Q}$.

This shows that the Markov eigenvectors S_p are scaled version of eigenvectors S_q of the symmetrically normalized matrix. Accordingly, if $\mathbf{D} = f(X)^T f(X)$, one can achieve same normalization by doing an SVD directly on a normalized data $Z^{-1/2}f(X)$.

3 Clustering Methods

We propose two alternative methods for audio segmentation:

1. In the first method, the data $\mathbf{X}$ or some possibly non-linear mapping of the vectors X is used to get a matrix $f(\mathbf{X}) = [f(X_1), f(X_2), ...]$. The choice of the mapping function is such that similarity relation between sound segments could be represented as a dot product $\mathbf{D} = f(\mathbf{X})^T f(\mathbf{X})$. SVD analysis is applied to $f(\mathbf{X})$ to find factorisation of $f(\mathbf{X})$ in terms of "basis" and "expansion" matrices $f(\mathbf{X}) = \mathbf{U}_f\mathbf{S}_f$. The values of the first k rows of S_f will be used for clustering.

2. The second method consist of finding the eigenvectors of the distance matrix $\mathbf{D}$. The first k eigenvectors of $\mathbf{D}$, $\mathbf{D}\mathbf{S}_d = \mathbf{S}_d\mathbf{\Lambda}$ are used for clustering. As was shown in previous section, both methods result in the same vectors, transposed. Accordingly, we shall denote the SVD expansion coefficients $\mathbf{S}_f$ of $f(\mathbf{X})$ and transposed eigenvectors $\mathbf{S}_d^T$ of the matrix $\mathbf{D}$ by one notation $\mathbf{S}$.

The segmentation is performed by clustering the values of the first k row vectors of $\mathbf{S}$. The choice of the first rows is in correspondence to the ordering of the eigenvalues, sorted from high to low. The clustering can be done using methods such as k-means.

For instance, partitioning a sound into two groups can be accomplished using $k = 1$, i.e. one expansion coefficient and finding a decision threshold that partitions that coefficient (row of $\mathbf{S}$) into two groups. Every time instance has a coefficient value that maybe above or below threshold and accordingly it is assigned to one or the other group. Once the assignment is done, a pointer to the original sound is used to mark a sound class to which the particular frame belongs. An example of clustering into three sound types using k-means is shown in the results section.

Going back to the sound itself, each frame at time index i is assigned to a cluster to which the appropriate coefficients of S at frame i belongs. Sequences of consecutive sound frames are joined together to compose a single macro-grain or a variant belonging to a particular sound type.

4 Results

In our experiments we analyzed several natural and musical sound that had alternating segments of different timbre. Figure 2 show the second and third eigenvectors of a Markov matrix for an example sound that contained three distinct sound types. One can see that each of the eigenvectors

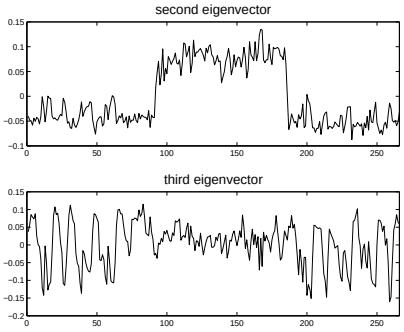

Figure 2: second and third eigenvectors.

has two areas of distinct values. Plotting the two eigenvectors against each other, i.e plotting them as coordinates in a two dimensional space, reveals two or possibly three clusters. Figure 3 shows these clusters of eigenvector values.

4.1 Comparison to Similarity Matrix Eigenvector Clustering

Clustering of $\mathbf{D}$ using eigenvectors $\mathbf{D}S_d = S_d\mathbf{\Lambda}_d$ gives sometimes different results compared to clustering using the Markov matrix eigenvectors $\mathbf{P}S_p = S_p\mathbf{\Lambda}_p$. Only in case when $\mathbf{Z}^{-1}$ is constant, i.e. all points have the same total distance from all other points, the two eigenvectors become identical. This may occur when all clusters have similar inter and intra cluster distances with equal number of points in each.

It is claimed in the literature that Markov normalization gives significantly better results for image grouping. (Shi and Malik 2000) In our experiments we did not reach any conclusive results concerning advantage of either method. One should note that in the Markov case the clustering should be done on the second largest (and possibly subsequent) eigenvectors, i.e. without taking into accound the first eigenvector. This is due to the fact that that the largest first eigenvector of Markov matrix is a unity vector, which does not allow clustering.

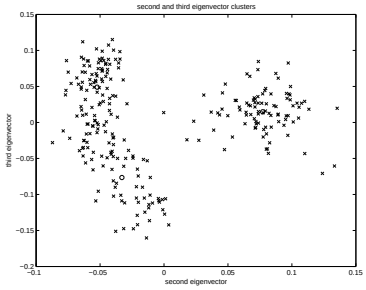

Figure 3: Second and third eigenvector showing clustering.

5 Conclusion

Our approach to sound texture modeling involves separating sound into sonic objects that are grouped by their similarity to each other. Here we have demonstrated that segmenting of sound can be equivalently achieved by a singular value decomposition of spectral features or eigenvector analysis of a similarity matrix. This analysis provides an unsupervised method of grouping similar sounds that may be applicable to sound texture synthesis systems, audio summary generation, as well as other creative audio manipulations. Questions such as automatic determining of the number of clusters (sound objects), including additional features such as pitch related distances or learning distance functions from examples are subjects for future research.

References

Casey, M. and W. Westner (2000). Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the ICMC*, pp. 154–161. ICMA.

Foote, J. and M. Cooper (2001). Visualizing musical structure and rhythm via self-similarity. In *Proceedings of the ICMC*, pp. 419–422. ICMA.

Lu, L., S. Li, L. Wenyin, and H. Zhang (2002). Audio textures. *International Conference on Acoustics, Speech, and Signal Processing*, 1761–1764.

Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22*(8), 888–905.