# TONAL-ATONAL CLASSIFICATION OF MUSIC AUDIO USING DIFFUSION MAPS

**Özgür İzmirli**
Center for Arts and Technology
Computer Science Department
Connecticut College
oizm@conncoll.edu

## ABSTRACT

In this paper we look at the problem of classifying music audio as tonal or atonal by learning a low-dimensional structure representing tonal relationships among keys. We use a training set composed of tonal pieces which includes all major and minor keys. A kernel eigenmap based method is used for structure learning and discovery. Specifically, a Diffusion Maps (DM) framework is used and its parameter tuning is discussed. Since these methods do not scale well with increasing data size, it becomes infeasible to use these methods in online applications. In order to facilitate on-line classification an out-of-sample extension to the DM framework is given. The learned structure of tonal relationships is presented and a simple scheme for classification of tonal-atonal pieces is proposed. Evaluation results show that the method is able to perform at an accuracy above 90% with the current data set.

## 1. INTRODUCTION

Audio key estimation is an important aspect of MIR. It informs many other tasks including music analysis, segmentation, cover song detection, modulation tracking, local key finding and chord recognition. In order to estimate the key, most key finding models use a similarity metric between predetermined reference features and the analyzed features from the audio. All of these approaches assume that the fragment of the piece being analyzed contains tonal music and furthermore that musical content is in a single key. These models generally lack mechanisms to detect music that is not tonal and hence would make best-guess estimates regardless of the tonal quality of the input. One important question, which is the topic of this paper, is how to determine whether a piece belongs to the tonal idiom: whether there are clear and unambiguous tonal implications or not.

In this work, we explore the utility of dimensionality reduction, manifold learning and structure discovery in the context of tonal versus atonal music audio classification. We investigate the possibility of learning a low-dimensional structure representing tonal relationships among pieces. We report on experiments that utilize Diffusion Maps to perform dimensionality reduction and feature extraction from high-dimensional spectral data. We use a set of audio recordings representative of all 24 keys as the reference training set and test the model with tonal and atonal audio fragments to evaluate its performance.

The structure of the paper is as follows: The next section makes reference to related work and explains the concept of tonalness. Section 3 describes kernel methods and DM in particular. This section also discusses the tuning of the width parameter of DM. Section 4 outlines the main outcomes and describes the evaluation method. Section 5 concludes the paper.

## 2. RELATED WORK

Temperley describes a probabilistic framework on symbolic data for measuring tonal implication, tonal ambiguity and tonalness for pitch-class sets [1]. According to his definition, tonal implication is the key implied by the pitch-class set being used. Ambiguity refers to whether a pitch set implies a single key or several keys. Tonalness is the degree to which a set is characteristic of common-practice tonality. In this sense, our work relates directly to the concept of tonalness. Our assumption is that a piece that conforms to pitch distributions of common practice tonality will have certain spectral properties that distinguish it from other types of pitch distributions such as those found in twelve-tone music or polytonality. These spectral properties, or so called spectral signatures, have native representations in a high-dimensional space and therefore need to be mapped to low-dimensional features to be useful - not only for classification purposes but also for visualization and geometrical interpretations. The remainder of the paper discusses a method to classify music audio based on the degree of tonalness.

In her thesis, Gómez applied her key finding method to an atonal piece by Schoenberg [2]. She observed that the correlations of her Harmonic Pitch Class Profile (HPCP) with the major and minor profiles, that are derived from Krumhansl's work, remained low throughout the piece, indicating ambiguity.

Izmirli reported on the performance of a template based key finding algorithm using a low-dimensional representation obtained through dimensionality reduction [3]. He graphed the performance of his method as a function of the number of dimensions and noted that 2 and 3 dimensions produced acceptable accuracy for the particular model he was using.

Purwins briefly discusses poly-tone analysis and tonal ambiguity in relation to Pitch Class Profiles that he uses in his key finding algorithm [4].

## 3. DIMENSIONALITY REDUCTION, MANIFOLD LEARNING AND STRUCTURE DISCOVERY

### 3.1 Method

In general, given a set of training data we would like to infer some parameterization of it such that new data can be efficiently compared to the training data. The parameterization can then be used for many different purposes including classification. In the following we present a method that performs dimensionality reduction on a training set of tonal audio in order to find a representative structure. The resulting low dimensional representation is then used to determine whether new input data resembles the training data or not; more specifically, if it is tonal or atonal. This section describes the method of dimensionality reduction used and a scalable extension for new data.

### 3.2 Kernel Methods

In contrast to the standard linear methods such as Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) for dimensionality reduction, nonlinear methods are better suited to preserving local geometry. This is due to the fact that they attempt to approximate manifolds in the high-dimensional space by considering connectivity between neighboring points as opposed to capturing the global nature of the data. Nonlinear methods include Isometric Feature Mapping (ISOMAP), Kernel PCA and a class of kernel eigenmap methods including Laplacian Eigenmaps, Locally Linear Embedding (LLE), Hessian Eigenmaps (Hessian LLE) and Local Tangent Space Alignment (LTSA). In [5] Coifman and Lafon show that the kernel eigenmap methods are special cases of a general framework based on diffusion processes. Here, we follow a formulation for dimensionality reduction, manifold learning and data parametrization based on DM [5]. The major advantages of this approach over PCA and MDS are that it is nonlinear and preserves local structures. Kernel eigenmap methods rely on the idea that eigenvectors of a transition matrix representing the distances between points in the input space can be interpreted as coordinates on the data set.

### 3.3 Diffusion Maps

The concept of diffusion maps stems from dynamical systems and it is based on a Markov random walk on the graph of the data. The proximity of the data points is modeled as diffusion distances according to the affinity between neighboring points. DM preserves local geometry present in the high-dimensional input while performing dimensionality reduction.

Assume the data set containing k elements is given by X={$x_0$, $x_1$, $x_2$, ...,$x_{k-1}$} with $x_i$ element of R$^m$**.** A pairwise similarity matrix $L$ is calculated using a Gaussian kernel with parameter $\varepsilon$ :

$$L_{ij} = w_\varepsilon(x_i, x_j) = e^{-\left\|x_i - x_j\right\|^2 / \varepsilon^2} \tag{1}$$

Furthermore, a diagonal normalization matrix is defined to make the sum of the rows of $L$ equal 1:

$$D_{ii} = \sum_j L_{ij} \tag{2}$$

The normalized graph Laplacian is then given by the Markov matrix $M = D^{-1}L$. In order to find a mapping, $\Phi$ , from R$^m$ to R$^n$, where m > n, an eigen-decomposition of $M$ is performed. The eigenvectors and eigenvalues can be found by solving the equivalent generalized eigenvalue problem $L\phi = \lambda D\phi$ . When ε in Eq. 1 is large enough, $M$ is fully connected and has a unique eigenvalue of 1. From the remaining k-1 eigenvalues, n of the largest $1 > \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$ and their corresponding eigenvectors $\phi_1, \phi_2, ... \phi_n$ can be retained to map input samples from the high-dimensional space onto the lower dimensional feature space. The mapping is given by

$$\Phi : x_i \to [\lambda_1\phi_1(i), \lambda_2\phi_2(i), ... \lambda_n\phi_n(i)] \tag{3}$$

where index i in $\phi_n(i)$ represents the i'th element of the eigenvector.

### 3.4 Determining *ε*

The width parameter, ε, controlling the Gaussian in Eq. 1 has an effect on the locality of the structure captured. For example, a relatively small ε will capture the local structure better. However, if ε is too small then matrix $L$ will have many small elements and hence, low connectivity, which will prevent it from capturing the desired structure. An unnecessarily large value on the other hand will cause the method to overlook the local structure. Although the value of this parameter is data dependent, fortunately, its choice can be automated.

Several approaches have been proposed to determine the optimal value of ε. The average of the distances between nearest neighbors in the data set are used in [6]. Another method is to adjust the parameter until every point has a significant connection to at least one neighbor. We follow the approach used in [7]. The method consists in searching for a point on the linear segment of the log-log graph of $T(\varepsilon)$ and ε, where

$$T(\varepsilon) = \sum_i \sum_j w_\varepsilon(x_i, x_j) \tag{4}$$

The graph contains two asymptotes, $\lim_{\varepsilon\to\infty} T(\varepsilon)$ and $\lim_{\varepsilon\to 0} T(\varepsilon)$ which are connected by an approximately linear line. We choose ε corresponding to the midpoint between the asymptotes in this graph.

### 3.5 Scalability and Out-of-Sample Extensions

Kernel methods described in the previous section have been successfully applied to dimensionality reduction and manifold learning. They are, however, computationally expensive and do not scale well to large data sets. They also do not directly accommodate new data and in that sense are limited to their training set requiring a new run every time new data is to be added.

Out-of-sample extensions are approximations that utilize the original eigen-decomposition to compute the mapping of new samples that do not belong to the original data set. In [8] the authors discuss how to compute out-of-sample extensions for various kernel methods. We employ the Nyström extension to find the mapping of the new data point as follows:

$$\tilde{\phi}_j = \lambda_j^{-1} \sum_{i=0}^{k-1} w_\varepsilon(x_{new}, x_i)\, \phi_j(i) \Big/ \sum_{p=0}^{k-1} w_\varepsilon(x_{new}, x_p) \qquad (5)$$

Once $\tilde{\phi}_j$ $(j = 1..n)$ is calculated, it is substituted for the corresponding eigenvectors in Eq. 3 to obtain the position in the lower dimensional feature space.

Calculation of the Nyström extension is computationally light. The denominator of Eq. 5 can be precalculated and the numerator is just a scaled sum of k vectors.

## 4. LEARNING TONAL STRUCTURE

### 4.1 Geometric Models of Pitch and Key

Many geometric models of pitch and key space have been proposed that originate from music theory and cognitive science. These include structures such as a circle, torus, helix and double helix (See for example [9] and [10]). Furthermore, most of these geometric structures are cyclic at one if not at multiple levels. In its simplest form, we know that key arrangements of the 12 major keys moving in fifths forms a circle. Similarly minor keys follow the same pattern. Obviously, this is based on the assumption that the music is performed in an equal tempered system.

In [11] it has been demonstrated that this or another cyclic structure can be captured from the audio of musical instruments playing diatonic scales. In this 2-dimensional space, points that represent key centers are organized in such a way that if we draw lines between the closely related keys the resulting arrangement forms a closed loop visiting each key center once.

### 4.2 Learning Structure from Audio Data

In this work, we explore the utility of structure discovery in the context of tonal versus atonal music audio. We obtain a chroma representation similar to [12] from the Hanning windowed short-time Fourier Transform. A 12-element chroma vector is obtained by summing the semitone frequency ranges of the amplitude spectrum according to pitch-class equivalence. That is, the semitone frequency range around the fundamental frequency of a note, the range around its octave and its second octave etc. all map to a single bin in the chroma vector.

Initially, we employed the method outlined in Section 3 to test if it was able to learn a low-dimensional structure using only recordings of tonal music. The training data, *X*, comprised of chroma vectors calculated from initial fragments of 289 pieces containing compositions mainly from the common practice period. Each point in the data set, $x_i$, represents the average of 30 seconds of music taken from the beginning of each piece. This duration was determined experimentally and can be chosen to be shorter without significantly effecting the algorithm's output. Note that the training is unsupervised and although the key labels are known from the titles of pieces they are not part of the input. The key distribution of the data set, although not completely uniform, is such that the lowest number of pieces in the same key is 9. For a collection of this size, a completely even distribution would require 12 pieces for each of the 24 keys. Although it would have been possible to either trim all pieces to the same number or add more pieces to bring the key totals to the same level, the current distribution was kept to observe the sensitivity of the DM algorithm to the density of samples on the manifold. It should be mentioned that sampling density is a main concern for many manifold learning algorithms and may need special attention if the spatial distribution is unbalanced.

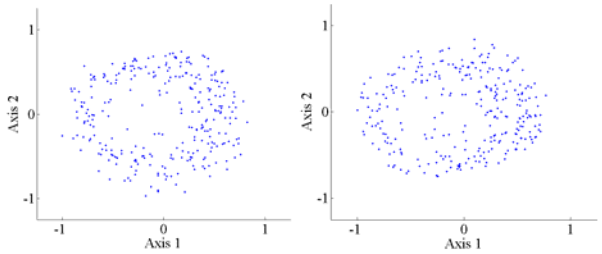

Figure 1. The input data set consisting of tonal pieces mapped to the first two dimensions. A circular structure resembling the circle-of-fifths is captured for the chroma representation (left) and for the spectral representation (right).

The left plot in Figure 1 shows the mapping Φ with $n = 2$ in response to the input data set, *X,* based on the chroma representation as described above. The out-of-sample extension is not used for this part. A circular structure is clearly visible in the figure which means it was able to capture some kind of circularity. Then again, this highly resembles the circle-of-fifths pattern. We verified the order of keys by analyzing their key labels to make sure the neighboring clusters were in a fifths relationship. There was considerable scatter within classes that belong to the same key. There was also significant overlap between classes, yet, the circle-of-fifths pattern was evident. The output for the spectral representation is shown in the right plot in Figure 1. These vectors are the same spectral vectors used to calculate the chroma representation. The reason for inclusion of the spectral vectors is to see if DM is able to obtain a mapping on par with or better than the traditional chroma representation. It should be noted that the uneven density of points does manifest itself in both plots without loss of generality of the result.

To further demonstrate the circle-of-fifths pattern we used chroma templates obtained from the audio of monophonic instrument sounds playing major scales. Each of the 12 templates consists of a single scale over multiple octaves. The details of the construction of the templates can be found in [13] and [14]. The templates were mapped using the out-of-sample method with respect to the tonal training data, X, described above. The results are shown in Figure 2. Here, each template represents an ideal key position in the feature space and the projection serves as a demonstration of the circle-of-fifths relationship among the 12 major keys. A similar order has also been observed for minor keys.

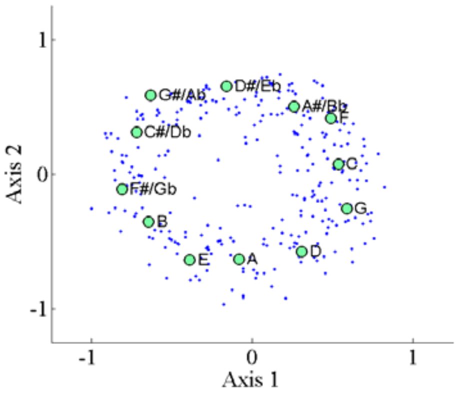

Figure 2. Mapping of audio templates to the first two dimensions. The labeled points representing the major templates are superimposed on the chroma based representation in Figure 1 (left).

### 4.3 Training and Test Data for Evaluation

Starting from the observation that a data set containing pieces in all 24 keys results in the constellations shown in Figure 1, we turn to testing the DM model with tonal and atonal data using the out-of-sample extension described above. For this part we added 25 complete atonal pieces composed by Boulez, Schoenberg and Webern. Both the tonal and atonal pieces were segmented into 10-second fragments. There are 599 atonal fragments and 925 tonal fragments in the data set. Each fragment is represented as a point, $x_i$, found by dividing the spectral or chroma vectors by their $L_2$ norm, and an associated tonal/atonal label serving as ground truth for evaluation purposes. The frequency ranges of interest for both representations are 55 - 2000 Hz. The training data set was constructed as follows: 60% of the tonal points were randomly chosen and were used to train the DM model. The remaining 40% were added to the test set accompanied by an equal number of points randomly chosen from the atonal set. After calculating the original mapping using 60% of the tonal points, the out-of-sample calculations were performed on the test set. Figure 3 shows the mapping of the test results onto the first two dimensions. These results are overlaid with the training points to show the nature of generalization the extension brings.

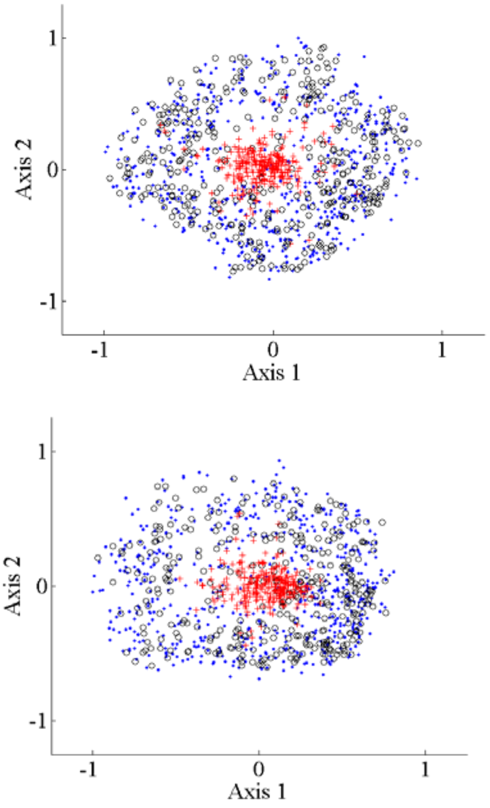

Figure 3. Training and test data mapped to the first two dimensions: chroma based inputs (top) and spectrum based inputs (bottom). Tonal training data are shown with dots (.), the tonal test data are shown with circles (○) and the atonal test data are shown with pluses (+).

### 4.4 The Tonal-Atonal Classifier

As can be easily observed from Figure 3, the tonal training points and the tonal test points tend to appear at positions closer to the outer circular pattern whereas the atonal test points tend to appear near the center. Therefore, we simply choose to use the Euclidean norm of a point in the feature space to quantify its tonalness as defined in Section 2. For the 2-dimensional case, the performance of the classifier is given by the peak classification accuracy in which a circle acts as the class boundary. It should be noted that although we treat the problem as a two-class classification task in this paper, in fact, the calculated tonalness is a continuous entity and is indeed correlated with the degree of the musical fragment's tonal implication. The distances in the feature space can be used to quantify the degree of tonalness. A study of the tonalness of transpositional type pitch class sets can be found in [15].

### 4.5 Results

An average accuracy was calculated by running the above classification 10 times. The chroma based classification resulted in an average accuracy of 91.2% and the spectrum based classification resulted in 90.4% accuracy.

As an alternative feature we ran a classification task based on the variance of the chroma and spectrum vectors ($x_i$) to see how they compared with the presented method. The intuition was that the chroma vector corresponding to tonal pieces would have more variance compared to atonal pieces because it would exhibit a strong interleaved response across bins of the vector. i.e. say, for C major, one would expect the bins corresponding to the white keys to be strong and those of the black keys to be weak. On the other hand, atonal pieces would have a more uniform spread across the bins. The chroma variance feature performed at 84.9% accuracy. The same reasoning does not really apply to the spectrum vectors because they are fairly sparse compared to the chroma vectors but nevertheless we tested the feature and obtained 64.4% accuracy; very low as expected.

## 5. CONCLUSION

In this paper we have discussed a method based on Diffusion Maps to perform tonal-atonal classification of music audio. Initially, we learn a low-dimensional structure representing pitch distributions that pertain to the tonal idiom. We then extend the learned mapping to new points and test the performance of the method. The learned cyclic structure is demonstrated through a display of the projected circular constellation of the training points and the projection of major scale templates representing ideal key locations in relation to this constellation. The use of the learned cyclic structure in quantifying tonalness is also discussed. Finally, results are presented for the tonal-atonal classification task for chroma representations as well as raw spectral representations. The results are encouraging and promising. Future work involves exploring more general mechanisms for calculating the structure similarity between training and test structures, and finding optimal training sets for faster and more efficient operation.

## 6. REFERENCES

[1] D. Temperley: "The Tonal Properties of Pitch-Class Sets: Tonal Implication, Tonal Ambiguity, and Tonalness," Eleanor Selfridge-Field and Walter Hewlett, eds. *Computing in Musicology*, *Tonal Theory for the Digital Age*, Vol. 15, 24-38, 2008.

[2] E. Gómez: “Tonal Description of Music Audio Signals,” *Ph.D. Dissertation*, Pompeu Fabra University, Barcelona, 2006.

[3] Ö. İzmirli: “Audio Key Finding Using Low-Dimensional Spaces,” *Proceedings of the International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[4] H. Purwins: "Profiles of Pitch Classes Circularity of Relative Pitch and Key – Experiments, Models, Computational Music Analysis, and Perspectives," *Ph.D. Thesis*, Berlin University of Technology, 2005.

[5] R. R. Coifman and S. Lafon: "Diffusion Maps," *Applied and Computational Harmonic Analysis*, 21, pp. 5–30, July, 2006.

[6] S. Lafon: "Diffusion Maps and Geometric Harmonics," *Ph.D. Thesis*, Yale University, New Haven, USA, 2004.

[7] A. Singer, R. Erban, I. Kevrekidis and R. Coifman: "Detecting Intrinsic Slow Variables in Stochastic Dynamical Systems by Anisotropic Diffusion Maps," *Proceedings of the National Academy of Sciences* (PNAS) 2009.

[8] Y. Bengio, J.-F. Paiement, and P. Vincent: "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering," *Advances in Neural Information Processing Systems, 16*, 2004.

[9] F. Lerdahl: *Tonal Pitch Space*. New York: Oxford University Press, 2001.

[10] H. Purwins, B. Blankertz, K. Obermayer: "Toroidal Models in Tonal Theory and Pitch-Class Analysis," Eleanor Selfridge-Field and Walter Hewlett, eds. *Computing in Musicology*, *Tonal Theory for the Digital Age*, Vol. 15, 73-98, 2008.

[11] Ö. İzmirli: "Cyclic Distance Patterns Among Spectra of Diatonic Sets: The Case of Instrument Sounds with Major and Minor Scales," Eleanor Selfridge-Field and Walter Hewlett, eds. *Computing in Musicology*, *Tonal Theory for the Digital Age*, Vol. 15, 11-23, 2008.

[12] T. Fujishima: "Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music," *Proceedings of the International Computer Music Conference* (ICMC), Beijing, China, 1999.

[13] Ö. İzmirli: "Template Based Key Finding From Audio," *Proceedings of the International Computer Music Conference* (ICMC), Barcelona, Spain, 2005.

[14] Ö. İzmirli: “An Algorithm for Audio Key Finding,” *2005 Music Information Retrieval Evaluation eXchange (MIREX) Audio Key-Finding Contest*, www.music-ir.org/evaluation/mirex-results/articles/key_audio/izmirli.pdf, 2005.

[15] Ö. İzmirli: "Estimating the Tonalness of Transpositional Type Pitch-Class Sets Using Learned Tonal Key Spaces," *Proceedings of Mathematics and Computation in Music*, New Haven, USA, 2009.