EXAMINING MUSICAL MEANING IN SIMILARITY THRESHOLDS

Katherine M. Kinnaird
Brown University
katherine_kinnaird@brown.edu

ABSTRACT

Many approaches to Music Information Retrieval tasks rely on correctly determining if two segments of a given musical recording are repeats of each other. Repetitions in recordings are rarely exact, and identifying the appropriate threshold for these pairwise decisions is crucial for tuning MIR algorithms. However, current approaches for determining and reporting this threshold parameter are devoid of contextual meaning and interpretations, which makes comparing previous results difficult and which requires access to specific datasets. This paper highlights weaknesses in current approaches to choosing similarity thresholds, provides a framework using the *proportion of orthogonal musical change* to tie thresholds back to feature spaces with the cosine dissimilarity measure, and introduces new research possibilities given a music-centered approach for selecting similarity thresholds.

1. INTRODUCTION

Since Foote introduced the self-similarity matrix as a technique for visualizing and representing audio data [7], matrix representations have been widely used to represent music-based data, such as songs or musical scores, when addressing different kinds of tasks in Music Information Retrieval [6, 11, 13, 17]. Recordings of music often contain slight variations between repeated sections either due to artistic interpretations or noise introduced by the recording environment. Addressing these MIR tasks often requires grouping time steps together using a threshold on the self-(dis)similarity matrix representation to determine which pairs of time steps are similar enough to be classified as repetitions of each other. There are two issues at play when choosing this similarity threshold: 1) selecting the best value given the task and data, and 2) using the value with the best musical interpretation.

Similarity thresholds are currently determined in ways that prioritize computational successes and ignore tangible musical interpretations. These thresholds are usually dependent on the data at hand and reported as a selection method (say a fixed percentage) instead of as a particular threshold value. These data-dependent thresholds, reported as methods, require access to common datasets in order to compare previous and current research. Furthermore, many of the processes for determining this crucial threshold do not have a mechanism for connecting that threshold back to the original feature space. For example, current methods give little understanding to what a "small-value" cosine dissimilarity measurement corresponds to in terms of musical sounds such as notes and chords.

Instead of only justifying similarity thresholds based on statistical theory or computational success, we argue musical meaning should be included in the selection and discussion of a similarity threshold. In Section 2, examples based on a jazz lead sheet offer motivation for similarity thresholds with musical context. In Section 3, we model a framework for tying a chosen threshold to a particular feature space via the concept of the maximum proportion of orthogonal musical change. In Section 4, we introduce how music-centered thresholds can enhance MIR research.

2. MOTIVATION WITH EXAMPLES

In MIR literature, there are a variety of methods for setting the similarity threshold used to decide when sections of a song are similar. Current methods have been based on statistical ideas combined with concise algorithmic explanations. In [1, 11, 15], for a given recording of a performance of a piece of music, the threshold was specified so that a fixed percentage of a matrix representation (either self-similarity matrix or self-dissimilarity matrix - SDM) would be selected. The method in [18–20] sets the meaning of "similar" for each time step by first looking for the κ nearest neighbors of a given time step and then by enforcing a mutual condition; that is that time steps i and j are determined to be similar if both time step i is a κ nearest neighbor of time step j and vice versa. In [3], the threshold was set using statistical techniques on a set of sample data. In [8–10], Goto determined a threshold using the automatic threshold selection method developed by Otsu [16] which selects a threshold using statistics of the grey-level histogram of a particular image. In the case of Goto's work [8–10], the image is a matrix representation for a song.

While the above methods are efficient and have satisfying connections to our intuition about similarity, a crucial weakness of these methods is a lack of a musical connection for the similarity threshold. For example, the fixed percentage thresholds in [1, 11, 15] are easy to set and offer clear methods for reproducing those workflows, but no musical intuition is offered for these methods. Underlying

fixed percentage thresholds is the assumption that all music has the same proportion of similarity, which certainly would not be the case in a collection with both classical and jazz recordings. The method in [18–20] to some degree addresses the flaws in this assumption, but this method leaves unanswered what it means musically to be mutual κ nearest neighbors.

The following four examples based off a jazz lead sheet use the bottom-10% paradigm for similarity threshold selection and highlight some of the issues with this method. The first example is just the chords by beat as described in the lead sheet. The second example adds absolute Gaussian noise, while the third example adds notes in a restricted manner, seeking to mimic the spontaneous composition of jazz music. The final example adds both absolute Gaussian noise and restricted "note" noise. These examples are constructed from a human coded .jazz file [1] of *Aisha* by McCoy Tyner from 1961 found in the *iRb Corpus in **jazz format* dataset [2]. Beat tracking was not used since these examples are based on a version of the piece's lead sheet. Each time step represents 8 continuous beats by concatenating adjacent 8 feature vectors (one per beat).

For each example, the distribution of dissimilarity values and the thresholded SDM are shown. All results are from single runs of the associated random processes, but similar results occur with repeated trials. For the thresholded SDM, the original SDM values are retained to further highlight contrast between examples.

Example 1 - Jazz Lead Sheet

This example is the ground truth for the true repeated structure of the lead sheet. We assume that there is neither noise nor spontaneous composition on the track.

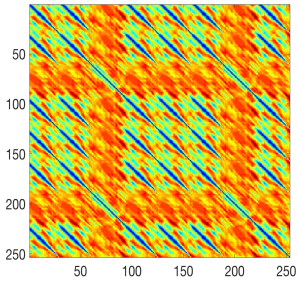
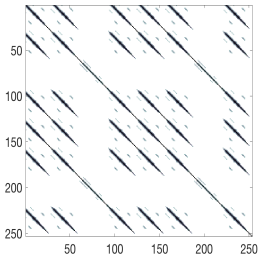


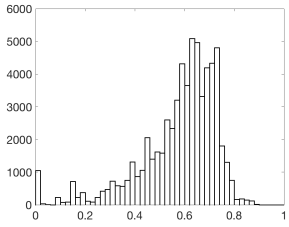
Figure 1. Complete SDM for *Aisha* lead sheet. Values near 0 are dark.

Using the bottom-10% paradigm, the threshold T is 0.375, meaning that two feature vectors with the angle between them no greater than 51.318 degrees will be deemed similar enough to be repeats of each other. This is quite a generous threshold; for example, a feature vector representing a C chord (held for 8 beats) and a feature vector representing C-minor chord (also held for 8 beats) would be deemed as repeats of each other.

[1] The .jazz file was converted to a .txt file using code by Yuri Broze [2]. Chromagrams were then extracted using a new converter file, available at `https://github.com/kmkinnaird/MusicalThresh`



(a) Thresholded SDM under bottom-10% paradigm

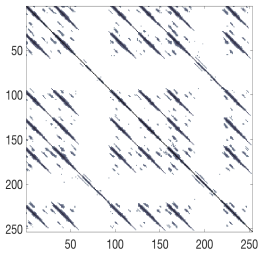


(b) Histogram of all dissimilarity values

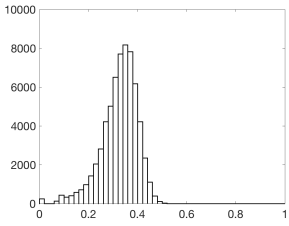
Figure 2. *Aisha* lead sheet without additions

Example 2 - Jazz Lead Sheet with Gaussian Noise

In this example, we add proxy for general noise (such as feedback in the recording environment) to the lead sheet. To each note-beat entry of the chroma matrix for the lead sheet, we add the absolute value of a random sample from the Gaussian centered at 0 with standard deviation 0.5.



(a) Thresholded SDM under bottom-10% paradigm



(b) Histogram of all dissimilarity values

Figure 3. *Aisha* lead sheet with added track noise

In addition to most of the similarity from Example 1, additional segments were classified as repeats using the bottom-10% paradigm, meaning that "similarity" is being created under this threshold selection method. However, this example's threshold value is lower, so two audio shingles must be more similar to be considered repeats than in Example 1. The threshold T is approximately 0.232, meaning that two feature vectors with the angle between them no greater than 39.818 degrees will be deemed similar enough to be repeats of each other. This shifted (and possibly contradictory) definition of similarity may be appropriate given the data but there is no musical interpretation of the threshold to support this choice. The lower threshold does reflect the compression of the distribution of dissimilarity values, shown in Figure 3(b).

Example 3 - Jazz Lead Sheet with "Note" Noise

In this third example, we add a proxy for spontaneous composition. This added "note" noise is restricted to the notes within the chord specified on the lead sheet and has its note weight randomly selected from the distribution of note values, shown in Figure 4.

The threshold T for this example is approximately 0.417, meaning that two feature vectors with the angle between them no greater than 54.357 degrees will be deemed

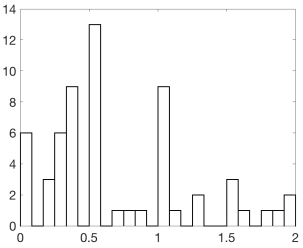
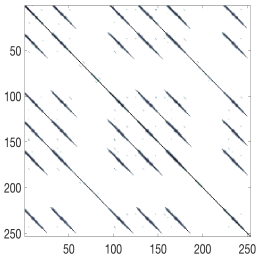
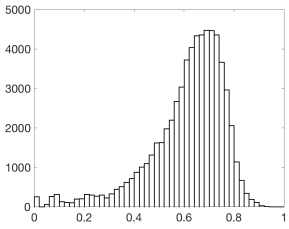


Figure 4. Histogram of note values for "note" noise

similar enough to be repeats of each other. As expected given this example's construction, this threshold is similar to the one in Example 1. However, while much of the similarity from Example 1 was found using the bottom-10% paradigm, it is clear that not all of it was. As with Example 2, this threshold may be appropriate, but there is no musical interpretation to support this choice.



(a) Thresholded SDM under bottom-10% paradigm

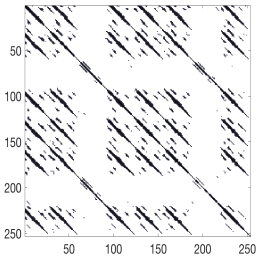


(b) Histogram of all dissimilarity values

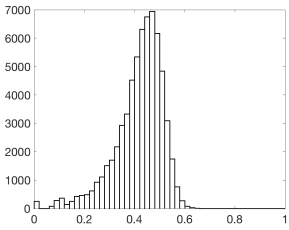
Figure 5. *Aisha* lead sheet with added "note" noise

Example 4 - Jazz Lead Sheet with "Note" Noise and with Gaussian Noise

In this example, we add proxies for both track noise (as in Example 2) and "note" noise (as in Example 3). Since we are assuming that there is both spontaneous composition and additional noise on the track, it is tempting to simply add the thresholds from Examples 2 and 3. However, we cannot, given the construction of the proxies and that cosine dissimilarity measure does not observe the triangle inequality.



(a) Thresholded SDM under bottom-10% paradigm



(b) Histogram of all dissimilarity values

Figure 6. Track and "note" noise added to *Aisha* lead sheet

Similar to Example 2, we have a possibly contradictory definition of similarity. In this example, the bottom-10% paradigm captures most of the similarity from Example 1 but also incorrectly matches additional repeated "similarity." However, the value of this threshold T is lower, at approximately 0.293, which translates to an angle no greater than 45.026 degrees between two feature vectors deemed similar enough to be repeats.

Comparing Four Examples

These four examples highlight some of the weaknesses in the commonly used fixed percentage threshold selection paradigm. First, the generous thresholds in Examples 1 and 3 allow for major and minor chords (such as C-major and C-minor) to be deemed as repeats of each other. However, the histogram from Example 3 is quite similar to Example 1, which signals that an appropriate choice of threshold for a lead sheet would also be appropriate to apply to a lead sheet with spontaneous composition.

Second, when a proxy for random track noise is introduced, as in Examples 2 and 4, major and minor chords would no longer be matched. However, a passing glance on the resulting thresholded SDMs in Examples 2 and 4 show sections of the lead sheet designated as repeats when they perhaps should not be. Additionally, the histograms for Examples 2 and 4 are much more compressed than those in Examples 1 and 3, which further signals a need in incorporate musical context into the selection of similarity thresholds.

Even though these four examples are based on a lead sheet, of which three employ random processes as proxies for track noise and spontaneous compositions, these controlled and constructed examples demonstrate the need for careful examination of the meaning and limitations of thresholds used in MIR tasks and approaches.

3. RELATING T TO MAXIMUM PROPORTION OF ORTHOGONAL MUSICAL CHANGE

In this section, we establish a framework for linking a similarity threshold T to the space of audio shingles composed of chroma feature vectors under the cosine dissimilarity measure. We define the *proportion of orthogonal musical change* (or POMC) for this feature space and prove a relationship between a given threshold T to POMC. Although we ground our discussion in one particular feature space, a similar procedure can be used to tie similarity thresholds to any feature space using the cosine dissimilarity measure.

3.1 Preliminary Definitions and Notation

We create overlapping *audio shingles* from k concatenated feature vectors, where k is a fixed integer [3–5]. For a timestep i, the chroma feature vector χ_i is the column vector of 12 non-negative entries, where each entry corresponds to one of the Western pitch classes $\{C, C\#, ..., B\}$ encoding the amount of that pitch class in the i^{th} observation [14].

For time-step i, the audio shingle of length k, incorporating local information, is the column vector α_i:

$$\alpha_i = \left[\chi_i^t, \chi_{(i+1)}^t, \chi_{(i+2)}^t, \ldots, \chi_{(i+k-1)}^t\right]^t \quad (1)$$

Each audio shingle is an element of $\mathbb{R}_{\geq 0}^{(k\times 12)}$, the non-negative closed orthant of $\mathbb{R}^{(k\times 12)}$ and can be regarded as vectors that start at the origin. Let $\theta_{\alpha_i,\alpha_j}$ be the angle between α_i and α_j. Since $\alpha_i, \alpha_j \in \mathbb{R}_{\geq 0}^{(k\times 12)}$, then $\theta_{\alpha_i,\alpha_j} \in [0, \frac{\pi}{2}]$. The pairwise cosine dissimilarity between two audio shingles α_i and α_j is defined as:

$$\mathcal{D}_{i,j} = 1 - \cos\theta_{\alpha_i,\alpha_j} \quad (2)$$

It is natural to ask: *Given the value $\mathcal{D}_{i,j}$, what are the musical differences between those two time steps?* We introduce the notion of *proportion of orthogonal musical change* (POMC), or rather the amount an audio shingle α_i must change orthogonally (before scaling) in order to become α_j. POMC encodes of how much one audio shingle can be comprised of elements perpendicular to another audio shingle before we say these two audio shingles are no longer considered to be "similar" of one another.

Consider Figure 7; the vector $\vec{\gamma}$ is orthogonal to α_i and when added to α_i will meet α_j. We can scale the vector $(\alpha_i + \vec{\gamma})$ to match α_j. Similarly $\vec{\phi}$ is orthogonal to α_j and when added to α_j meets α_i. We can scale $(\alpha_j + \vec{\phi})$ to match α_i. The length of $\vec{\gamma}$ is $||\alpha_i|| \cdot \tan\theta_{\alpha_i,\alpha_j}$, and the length of $\vec{\phi}$ is $||\alpha_j|| \cdot \tan\theta_{\alpha_i,\alpha_j}$. So $\tan\theta_{\alpha_i,\alpha_j}$ is the amount of orthogonal change for α_i to become a scalar multiple of α_j and vice versa.

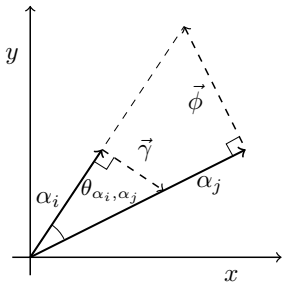


Figure 7. Visualization of orthogonal musical change of α_i onto α_j and of α_j onto α_i, represented respectively by the vectors $\vec{\gamma}$ and $\vec{\phi}$.

Definition 3.1. For a pair of audio shingles α_i and α_j, the *proportion of orthogonal musical change* (POMC) is given by $\tan\theta_{a_i,a_j}$.

3.2 Maximum POMC Given T

Suppose that we have one audio shingle, denoted $\vec{\xi}$ terminating at point ξ, and that we want to classify all audio shingles that are repetitions of $\vec{\xi}$. Let T be the threshold determining whether pairs of audio shingles are close enough to be repeats. We define θ_T as $\cos^{-1}(1-T)$.

Let Ξ be the set of audio shingles that are less than T cosine dissimilar from $\vec{\xi}$. So $\vec{v} \in \Xi$, iff $1 - \cos\theta_{\vec{v},\vec{\xi}} \leq T$, for $\theta_{\vec{v},\vec{\xi}}$. Additionally, for each vector $\vec{v} \in \Xi$, we have:

$$\cos\theta_{\vec{v},\vec{\xi}} \geq 1 - T = \cos\theta_T \quad (3)$$

Definition 3.2. Given T, the *maximum POMC*, denoted ρ, is $\tan(\theta_T)$, where $\theta_T = \cos^{-1}(1-T)$.

We begin establishing the comparison between the audio shingles in Ξ and $\vec{\xi}$ using just POMC. We first note that the set of audio shingles orthogonal to $\vec{\xi}$ is comprised of the audio shingles representing silence (i.e. those without any notes) and the audio shingles that do not have notes in common time with $\vec{\xi}$. For example, if $\vec{\xi}$ represented a C chord followed by a F chord, then an audio shingle that is orthogonal to it could be one representing a C# chord followed by an E chord. Given the importance of note and chord order in music generally, the audio shingle representing an F note followed by a C chord is orthogonal to a second representing a C chord followed by an F note. Neither of the above pairs would be mistaken as similar, and so we restrict $\theta_T \in [0, \frac{\pi}{2})$, since including $\theta_T = \frac{\pi}{2}$ would imply that orthogonal pairs of audio shingles are similar.

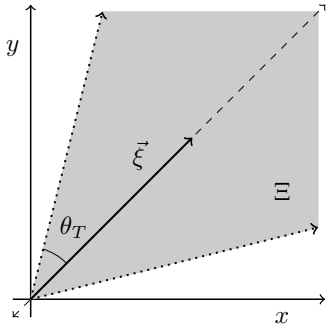


Figure 8. Visualization of Ξ, the set of audio shingles that are less than T cosine dissimilar from $\vec{\xi}$. The gray area flanked in dotted arrows is the set Ξ. The dashed line continuing from $\vec{\xi}$ is the subspace defined by $\vec{\xi}$.

More often, we want to compare pairs like a C chord followed by a second C chord with a C chord followed by a C7 chord, and determine if these two audio shingles are close enough to be deemed similar. These two audio shingles are the same save for the B♭ note in the second chord. Clearly a lone B♭ note is orthogonal to a C chord but is not orthogonal to the C7 chord. However, the C7 chord can be decomposed into the sum of a C chord and a B♭ note. In other words, the C7 chord is the C chord plus a vector orthogonal to it. Such a decomposition is at the heart of the concept of POMC.

Returning to our general case with audio shingle $\vec{\xi}$, we make the following definitions generalizing the above comparison of the C and C7 chords:

Definition 3.3. Let $\xi^\perp$ denote the hyperplane that is orthogonal to the vector $\vec{\xi}$ with the point $\xi \in \xi^\perp$.

We note that $\xi^\perp$ does not require that vectors in $\xi^\perp$ to be within Ξ. The following definition adds this restriction:

Definition 3.4. Let V_+ be the set of vectors originating at the point ξ and terminating at a point in $\xi^\perp$ such that for $\vec{v}_+ \in V_+$, we have that the cosine of the angle between $(\vec{\xi} + \vec{v}_+)$ and $\vec{\xi}$ is greater than or equal to $\cos\theta_T$.

For any vector $\vec{v}_+ \in V_+$, we have that the angle between $\vec{\xi}$ and $\vec{v}_+$ is the right angle in a right triangle with one leg along $\vec{\xi}$ with length $||\vec{\xi}||_2$ and with another leg along $\vec{v}_+$ with length $||\vec{v}_+||_2$. The tangent of the angle between $(\vec{\xi} + \vec{v}_+)$ and $\vec{\xi}$ is equal to $\frac{||\vec{v}_+||_2}{||\vec{\xi}||_2}$, which must be less than or equal to $\tan\theta_T$. So $||\vec{v}_+||_2 \leq ||\vec{\xi}||_2 \cdot \tan(\theta_T)$.

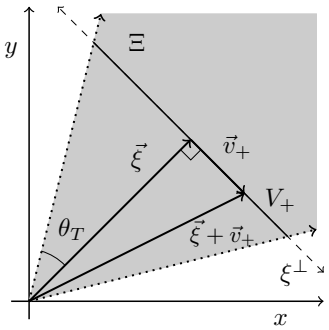


Figure 9. Visualization of the right triangle formed by $\vec{\xi}$ and $\vec{v}_+ \in V_+$ in $\mathbb{R}^2$. The solid line perpendicular to $\vec{\xi}$ represents V_+ and is $2||\vec{\xi}||_2 \cdot \tan(\theta_T)$ long. The dashed line continuing from V_+ combined with V_+ represents $\xi^\perp$.

The set V_+ represents the set of audio shingles that are created by adding to $\vec{\xi}$ an orthogonal vector of length no longer than $||\vec{\xi}||_2 \cdot \tan(\theta_T)$. For example, if $T = 0.1$ and if $\vec{\xi}$ represented one C chord, then the audio shingle representing a C7-chord would terminate in the hyperplane $\xi^\perp$, but would not be included in Ξ; however, if $T = 0.14$, then it would be included in Ξ.

3.3 Decomposition of Elements in Ξ

Thus far we have established the relationship between our audio shingle $\vec{\xi}$ and the audio shingles ending in $\xi^\perp \cap \Xi$. We are interested in understanding the relationship of T with all of Ξ. We offer the following decomposition for the elements of Ξ, which connects the definitions of the previous section to Definition 3.1 shown in Figure 7.

Proposition 3.1. *The set Ξ is equal to the set $\mathcal{S}$, given by*

$$\mathcal{S} = \left\{ w(\vec{\xi} + \vec{v}_+) | \vec{v}_+ \in V_+, w \in \mathbb{R}_{\geq 0} \right\}.$$

Proof: First, we show that $\mathcal{S} \subseteq \Xi$. We begin by letting $\vec{v}$ be an element of $\mathcal{S}$. So $\vec{v} = w(\vec{\xi} + \vec{v}_+)$ for some $\vec{v}_+ \in V_+$ and some $w \in \mathbb{R}_{\geq 0}$. Let β denote the angle between $\vec{v}$ and $\vec{\xi}$, which is also the angle between $(\vec{\xi} + \vec{v}_+)$ and $\vec{\xi}$ since $\vec{v}$ and $(\vec{\xi} + \vec{v}_+)$ are scalar multiples of each other. By the definition of V_+, we have that $\cos\beta \geq \cos\theta_T$. So $\vec{v} \in \Xi$ and thus $\mathcal{S} \subseteq \Xi$.

Next, we show that $\Xi \subseteq \mathcal{S}$. Let $\vec{v} \in \Xi$ and let $\theta_{\vec{v},\vec{\xi}}$ be the angle between $\vec{v}$ and $\vec{\xi}$. Then:

$$1 - \cos\left(\theta_{\vec{v},\vec{\xi}}\right) \leq T \tag{4}$$

Let v_+ be the point where $\vec{v}$ intersects $\xi^\perp$ (noting that it may be necessary to continue in the direction of $\vec{v}$ to intersect with $\xi^\perp$). We have a right triangle with one leg in the direction of $\vec{\xi}$ with length $||\vec{\xi}||_2$ and another leg from $\vec{\xi}$ to v_+ that is length $||v_+ - \xi||_2$. Let $\vec{v}_+$ be the vector from $\vec{\xi}$ to v_+. By definition of v_+, then $||\vec{v}_+||_2 = ||v_+ - \xi||_2$ and thus:

$$\tan\left(\theta_{\vec{v},\vec{\xi}}\right) = \frac{||v_+ - \xi||_2}{||\vec{\xi}||_2}. \tag{5}$$

Leveraging Eqn (5), we have that

$$\begin{aligned} ||\vec{v}_+||_2 &= ||v_+ - \xi||_2 = ||v_+ - \xi||_2 \cdot \frac{||\vec{\xi}||_2}{||\vec{\xi}||_2} \\ &= ||\vec{\xi}||_2 \cdot \tan\left(\theta_{v,\vec{\xi}}\right) \leq ||\vec{\xi}||_2 \cdot \tan(\theta_T). \end{aligned}$$

The last inequality is due to the angle between two vectors in Ξ is less than θ_T, that $\theta_T \in [0, \frac{\pi}{2})$, and that $\tan(x)$ is a monotonically increasing function on the interval $[0, \frac{\pi}{2})$. So $\vec{v}_+ \in V_+$. Let w be the positive scalar that we multiply $(\vec{\xi} + \vec{v}_+)$ by to get $\vec{v}$. Then $\vec{v} = w(\vec{\xi} + \vec{v}_+)$ for some $\vec{v}_+ \in V_+$ and some $w \in \mathbb{R}_{\geq 0}$. So $\vec{v} \in \mathcal{S}$ and $\Xi \subseteq \mathcal{S}$. □

This proposition gives us a decomposition for all audio shingles that are less than T cosine dissimilar from $\vec{\xi}$, regardless of how T is set. Using standard orthogonal projections, we can decompose any audio shingle into the form $w(\vec{\xi} + \vec{\alpha})$, where $\vec{\alpha}$ is audio shingle orthogonal to $\vec{\xi}$. To check if $\vec{\alpha} \in V_+$, we compute $||\vec{\alpha}||_2$ and see if $||\vec{\alpha}||_2 \leq ||\vec{\xi}||_2 \cdot \tan(\theta_T)$. Proposition 3.1 gives contextual meaning to our thresholds, that is the maximum proportion of orthogonal notes allowed between two audio shingles of at most T cosine dissimilarity.

3.4 Relating Choice of T to Audio Shingles via Maximum POMC

The above decomposition for vectors within T cosine dissimilarity measure from $\vec{\xi}$ provides us an avenue for relating our chosen thresholds directly to musical building blocks such as notes and chords, when they are represented as chroma feature vectors. This means that we can set a threshold by directly encoding acceptable musical variation for a small segment instead of setting the threshold using parameters free from musical context, such as a fixed percentage of entries from a matrix representation or a fixed-number of nearest neighbors.

We can set a threshold in one of three ways: 1) choosing T using existing methods, 2) setting the largest allowable θ_T between two audio shingles classified as similar enough, or 3) by setting ρ, the maximum POMC. Since T, θ_T, and ρ are functions of each other, fixing one inherently fixes the other two, and so we have an interpretation for that threshold in the space of audio shingles (under the cosine dissimilarity measure), returning musical context to what we mean by "similar structure."

- If we fix T, then we have $\theta_T = \cos^{-1}(1 - T)$ and

$$\rho = \frac{\sqrt{1 - (1 - T)^2}}{(1 - T)}$$

- If we fix θ_T, then $T = 1 - \cos\theta_T$ and $\rho = \tan\theta_T$.
- If we instead fix ρ, then we have $\theta_T = \tan^{-1}(\rho)$ and

$$T = 1 - \frac{1}{\sqrt{\rho^2 + 1}}$$

3.5 Returning to Motivating Examples

In Section 2, we described the thresholds for each example in terms of θ_T, which is still unsatisfactory in terms of musical intuition. Now we will interpret each T using ρ.

The thresholds in Examples 1 and 3 in Section 2 have similar interpretations, which makes sense given their constructions. In Example 1, we have $T = 0.375$. So $\rho = 1.249$, meaning that for each note in a given audio shingle $\vec{\xi}$, we can add an orthogonal vector of notes with 1.249 times the magnitude of $\vec{\xi}$ and have the result be considered similar to $\vec{\xi}$. This is quite a generous threshold. For example, an audio shingle representing a C chord whole note is considered similar to a second shingle representing a C chord whole note plus a D-minor6 chord whole note and a B♭ dotted-half note. Example 3 has a similarly generous threshold with $T = 0.417$. So $\rho = 1.395$, and we can add a few more orthogonal notes to $\vec{\xi}$ than in Example 1 and still have the result be considered similar to $\vec{\xi}$.

Examples 2 and 4 in Section 2 both include the incorporation of Gaussian noise, and their associated thresholds have similar interpretations. In Example 2, we have $T = 0.232$; so $\rho = 0.834$. In Example 4, we have $T = 0.293$; so $\rho = 1.001$. These thresholds are less generous than those in Examples 1 and 3. In Example 4, an audio shingle representing a C chord whole note is considered similar to a second shingle representing a C chord whole note plus a D-minor chord.

While the above interpretations offer a musical context for our similarity thresholds, these interpretations only regard the worst case (and less likely) scenario for comparing a given audio shingle $\vec{\xi}$ to another one; that is comparing $\vec{\xi}$ to one comprised of $\vec{\xi}$ added to an audio shingle orthogonal to $\vec{\xi}$. In addition to this interpretation, we would also advocate that when setting the similarity threshold, researchers also explore comparisons of $\vec{\xi}$ to audio shingles comprised of $\vec{\xi}$ with audio shingles that are not orthogonal to $\vec{\xi}$.

4. EXPANDING USES OF MAXIMUM POMC

Building off the examples in Section 2 and the methodology in Section 3, we propose research directions that could benefit from using a musically relevant threshold.

Within the song comparison tasks, we can use the maximum POMC to explore less well defined variants of the version detection task. For example, we can explore how much spontaneous composition is added to a jazz lead sheet, while also detecting the repeated sections in the lead sheet given a maximum POMC. In another direction, we could use the relationship between a threshold and maximum POMC to create a lower bound threshold for detecting recordings using auto-tune compared to those without.

Maximum POMC can be used beyond the song comparison tasks. We can leverage the maximum POMC to perform comparisons between genres, perhaps, by quantifying the amount of expected structure in a song from one genre, and comparing that to the expected value of another genre. Using ideas from topological data analysis, we can create diagrams quantifying the amount of structure in a given piece as we increase the maximum POMC. We could also use a dynamically set maximum POMC in generative music tasks to enforce musical style constraints given the target genre for the generated musical work.

5. CONCLUSION

Previous work in MIR determined and reported similarity thresholds as a specific method for a specific dataset preprocessed in a specific manner for a specific task, and thus it is hard to compare previous results. However we can more easily compare future work on both new and current song datasets if we choose a similarity threshold for our matrix representations that includes a tangible interpretation within the feature space.

This paper offered three contributions to the study of similarity thresholds used in MIR on the self-similarity (or dissimilarity) matrices, like those introduced in [7]. First, we demonstrated weaknesses in the current fixed percentage paradigm, using four examples based off one jazz lead sheet to show inconsistencies between the interpretations of the musical differences between sections of music that are regarded as similar.

Next we demonstrated that it is possible to link a threshold to the feature space of the original data, by providing a theoretical framework relating a given threshold to the space of audio shingles comprised of chroma vectors under the cosine dissimilarity measure. Crucial to this framework is the notion of *proportion of orthogonal musical change* (POMC), introduced here. This paper provides an avenue for interpreting and exploring the musical context of similarity thresholds (regardless of how they are determined) for self-dissimilarity matrices built from the space of audio shingles through the maximum POMC. Since the theoretical work in this paper only relied on facts of the cosine dissimilarity measure, the present framework could easily be adjusted to accommodate another feature space using the cosine dissimilarity measure.

Finally we briefly proposed new MIR research directions where contextually meaningful thresholds could provide insight. We also discussed how a contextually meaningful threshold could enhance current research directions.

Setting the similarity threshold can take into account both success on a particular task, given particular data, as well as a tangible musical interpretation of that threshold. Understanding that continuing to use current methods for determining the similarity threshold may be the best for continued computational success, this paper advocates for the inclusion of musical context into, at least, the discussion of the similarity threshold, if not the selection.

Acknowledgements

Part of this work is a portion of the author's doctoral thesis [12], which was partially funded by the GK-12 Program at Dartmouth College (NSF award #0947790). The author thanks Yuri Broze for his assistance installing his code [2] to create the .jazz files. The author also thanks Scott Pauls, Michael Casey, Dan Ellis, and Jessica Thompson for their feedback on the early versions of this work.

6. REFERENCES

[1] J. Bello. Measuring structural similarity in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2013–2025, 2011.

[2] Y. Broze and D. Shanahan. The iRb Corpus in `**jazz` format. `http://musiccog.ohio-state.edu/home/index.php/iRb_Jazz_Corpus`, 2012. [Online; accessed 28-September-2016].

[3] M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1015 – 1028, 2008.

[4] M. Casey and M. Slaney. Song intersection by approximate nearest neighbor search. *Proc. of 7th ISMIR Conference*, pages 144–149, 2006.

[5] M. Casey and M. Slaney. Fast recognition of remixed audio. *2007 IEEE International Conference on Audio, Speech and Signal Processing*, pages IV – 1425 – IV–1428, 2007.

[6] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127 –130, 2003.

[7] J. Foote. Visualizing music and audio using self-similarity. *Proc. ACM Multimedia 99*, pages 77–80, 1999.

[8] M. Goto. A chorus-section detecting method for musical audio signals. *Proc. of ICASSP*, 2003.

[9] M. Goto. SmartMusicKIOSK: Music listening station with chorus-search function. *Proc. of 16th ACM Symposium on User Interface Software and Technology (UIST 2003)*, pages 31–40, 2003.

[10] M. Goto. A chorus-section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1783–1794, 2006.

[11] P. Grosche, J. Serrà, M. Müller, and J.Ll. Arcos. Structure-based audio fingerprinting for music retrieval. *Proc. of 13th ISMIR Conference*, pages 55–60, 2012.

[12] K. M. Kinnaird. *Aligned Hierarchies for Sequential Data*. PhD thesis, Dartmouth College, 2014.

[13] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[14] M. Müller and S. Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. *Proc. of 12th ISMIR Conference*, pages 215–220, 2011.

[15] M. Müller, P. Grosche, and N. Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. *Proc. of 12th ISMIR Conference*, pages 615–620, 2011.

[16] N. Otsu. A threshold selection method from gray-level histograms. *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference*, SMC-9(1):62–66, 1979.

[17] J. Paulus and A. Klapuri. Music structure analysis using probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[18] J. Serrà, M. Müller, P. Grosche, and J.Ll. Arcos. Unsupervised detection of music boundaries by time series structure features. *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[19] J. Serrà, M. Müller, P. Grosche, and J.Ll. Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5), 2014.

[20] J. Serrà, X. Serra, and R.G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(093017), 2009.