World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:12, 2022

On Musical Information Geometry with Applications to Sonified Image Analysis

Shannon Steinmetz, Ellen Gethner

Abstract—In this paper a theoretical foundation is developed to segment, analyze and associate patterns within audio. We explore this on imagery via sonified audio applied to our segmentation framework. The approach involves a geodesic estimator within the statistical manifold, parameterized by musical centricity. We demonstrate viability by processing a database of random imagery to produce statistically significant clusters of similar imagery content.

Keywords—Sonification, musical information geometry, image content extraction, automated quantification, audio segmentation, pattern recognition.

I. Introduction

MATHEMATICALLY speaking, audio analysis tends to involve a fairly narrow sense making apparatus. Regardless of the genre of technology, most researchers rely on some form of statistical interrogation. This includes traditional signal processing (DSP), artificial intelligence and machine learning (AI/ML), or both. AI/ML has changed the general approach; however, statistical interrogation remains central. AI/ML commonly employs convolutional neural networks (CNN) and support vector machines (SVN) [22], [12]. These still involve statistical divergence from a regression line, or hyperplane. Information geometry (IG) purports to have an advantage as observed by Cont, using *distortion measures* as the primary means of analysis. IG also offers a range of opportunities where various classes of information can be mapped to an estimator, or probability model. Unlike AI/ML, IG does not require *a posteriori* knowledge of data, but remains reliant on statistical inference and divergence. The same is true for music information retrieval (MIR), much of which was originally based around self similarity [11] and has begun to employ machine learning. MIR provides complex knowledge assessment in genre classification, timbre, "query by humming" and emotional recognition (MER) [19], [16]. Our analysis fits underneath the MIR/IG umbrella insofar as audio segmentation is concerned, but we take the statistical analysis a step further into the geometry of music.

We propose a means to analyze audio content using the foundations of information geometry, motivated by musical geometry and applied to sonified audio. This is demonstrated via pseudo-image content retrieval. We transitively associate information moving from imagery $\rightarrow$ audio $\rightarrow$ data $\rightarrow$ cluster. This seemingly non-sequitur yields a profound connection within the nature of audio geometry. Recall, fundamental frequencies aggregate musical chords and have underpinnings in the Fibonacci sequence and golden-ratio. Both are directly related to naturally occurring phenomena [2], [28]. Every waveform has an inherent geometry and the theoretical basis connecting audio to musical geometry is well established [27].

Sonification is the process of converting raw data into audio for the purposes of evaluation. Sonified audio tends to break from traditional sense making analysis such as MIR. This is because it does not benefit from assumptions involving the minimal guarantee to the listener of pleasure, composition, or sensible melody. Historically, interesting problems have been solved using sonification, but the use of the technology is not common. In 1979 Fred Scarf was one of the first scientists to use sonification on the Voyager II plasma wave research project [29]. This analysis led baffled scientists to determine micro-asteroid perturbations buried within the noise. In modern times the Laser Interferometer Gravitational-Wave Observatory (LIGO) identified *chirp* in gravitational wave signals from binary inspirals using sonification [14]. These events are often singular, unpredictable and require new and innovative methodology. By limiting the domain of inquiry we only limit innovation.

In the present paper we develop a framework using information gain and musical geometry to exploit gradient patterns within imagery. This is done through the use of sonified audio. Individually, these technologies are relatively straightforward; however, their combined use creates a complicated set of interactions. We will focus on audio segmentation using likelihood estimation within the information manifold. Structural predictors, combined with musical quantifiers map audio to correlated geometric curves. The correlated curves create image clusters having similar content.

This paper is organized as follows. Section II provides a refresher on information geometry and our audio segmentation approach. Section III discusses musical geometry as a predictor of audio segmentation and a model for calculating what is known as musical *centricity*. Section IV lays out a support architecture for sonification using the proposed information framework and defines a conic sonifier. We conclude the study in Section V with an experiment involving image content association based on the sonifier and proposed information framework.

II. Information Geometry and Exponential Predictors

Probability distributions are commonly used to predict the likelihood of occurrence of a discrete series of events. Various distributions have curvature favoring an activity based on observed behavior. Most scientists and engineers rely on two dimensional distributions, where a random variable X

Shannon Steinmetz is with University of Colorado Denver, United States (e-mail: Shannon.Steinmetz@ucdenver.edu).

is supplied in the form of a function $f(X = x) \in [0, 1]$. Information geometry, when tailored to a statistical surface, allows us to consider a space of multiple simultaneous distributions. In this space we have $f(X = x \,|\, \boldsymbol{\theta})$ with coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and manifold $\boldsymbol{M} = \{f(x \,|\, \boldsymbol{\theta})\}$. Under certain regularity conditions this surface is treated as a set of very small Euclidian neighborhoods, with every point $p \in \boldsymbol{M}$ having a local tangent plane T_p. Each tangent plane induces a product space allowing for measurement of lengths, distances, curves and other metrics. Assuming f is convex, when combined with a *metric tensor* it constitutes a Riemannian manifold (M, g) [5]. The metric tensor g is computed using partial derivatives based on Fisher Information [21], [26].

$$g_{ij} = E\left[\frac{\partial \log p(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log p(x|\boldsymbol{\theta})}{\partial \theta_j}\right]. \tag{1}$$

This metric tensor is a pseudo-projective coefficient creating relative distances within the local tangent plane. For example, using Einstein notation, one measures the invariant length of vector v

$$||v \cdot v||^2 = g_{ij}(v^i v^j) \text{ [17]}.$$

When a Riemannian manifold is equipped with an affine connection (M, g, ∇) it constitutes an *Information Geometry* (IG) [4], [9]. Affine connections are largely dependent on infinitesimal, local transitions between tangent bases. The following notation is commonly used throughout information geometry $\partial_k = \frac{\partial}{\partial \theta_k} = \frac{\partial f(x \,|\, \boldsymbol{\theta})}{\partial \theta_k}$. The derivation of intrinsic, or extrinsic products that allow us to traverse a manifolds curvature have a compact notation in the form of a Christoffel symbol.

$$\Gamma^k_{ij} = \frac{1}{2} g^{km}(\partial_j g_{im} + \partial_i g_{jm} - \partial_m g_{ij}). \tag{2}$$

This particular version is a *Christoffel symbol* of the second kind, known to be a torsion free metric compatible Levi-Civita connection [17]. Levi-Civita facilitates the *intrinsic* geometry of a surface by allowing us to view the surface from a bugs eye view, removing the need for a reference position, or origin. When this coefficient is combined with the geodesic equation

$$\frac{\partial^2 u^k}{\partial \lambda^2} = -\Gamma^k_{ij} \frac{\partial u^i}{\partial \lambda} \frac{\partial u^j}{\partial \lambda}. \tag{3}$$

We get a smooth, shortest path having no local acceleration over the information manifold.

A. Exponential Family Distributions

Even the basics of information geometry can be mathematically intense. Luckily we have extensive research on a particular type of manifold known as the *exponential family* [3], [5], [9], [17]. This distribution constitutes an information geometry [3], [6] and although we do not leverage the following fact, it is also *exponentially flat* (e-flat). A given manifold is exponentially flat when there exists

$$E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x, \boldsymbol{\theta}) \frac{\partial}{\partial \theta_k} \log p(x, \boldsymbol{\theta})\right] = 0 \tag{4}$$

for all i, j, k identically [3]. This surface forms an e-affine coordinate system in $\boldsymbol{\theta}$ and allows us to treat linearly parameterized intersecting geodesic curves as a pseudo-orthogonal coordinate system. The following is the exponential family in standard form.

$$p(x \,|\, \boldsymbol{\theta}) = h(x)\, e^{\theta^T T(x) - \psi(\theta)} \text{ [3], [7]}, \tag{5}$$

where $\boldsymbol{\theta}$ is the canonical, or natural parameters and T is a factorable function of x. $\psi(\boldsymbol{\theta})$ is known as the potential function, or cumulant. $h(x)$ is a constant, independent of the potential function. The cumulant $\psi(\theta)$ is both dually flat and convex, equivalent to the log estimator $\psi(\boldsymbol{\theta}) = -\log p(x \,|\, \boldsymbol{\theta})$ [5], [3]. The log estimate is twice differentiable at every point within regular bounds, hence it is smooth. Many well known distributions can be written in an exponential form such as the Gaussian, Multivariate and Gamma distributions, to name a few.

The exponential family is stable for models that are censored, marginal, or truncated [10]. The Gamma distribution is a member of the exponential family.

$$p(x) = \frac{\beta^\alpha x^{\alpha-1} e^{x\beta}}{\Gamma(\alpha)} \tag{6}$$

Gamma allows for transitions between Gaussian, pseudo-exponential and skewed curves, all depending on the given model $(\boldsymbol{\theta})$.

B. Geometry Based Prediction

Given audio in the form of pulse code modulation (PCM) we divide audio into geometrically consistent segments. Each segment is a series of frames, each having a duration relevant to the underlying content. We do this by predicting model parameters and treating them as segmentation rules

$$(\alpha, \beta) = (\theta_1, \frac{1}{\theta_2}) = (\text{duration}, \frac{1}{\text{frame count}}).$$

An appropriate mapping is necessary to take model parameters from the Gamma domain (θ_1, θ_2) to duration and frame count. Empirically, we found $2 \leq \theta_1 \leq 5$ and $.4 \leq \theta_2 \leq 1$ where $f_{\theta_1}(\theta_1) \mapsto [100, 1000]$ms and $f_{\theta_2}(\theta_2) \mapsto [2, 10]$ for time and frame count, respectively.

Assuming $\Omega =$ samples/millisecond, a segment is defined as a series of frames $X = \{x_1, x_2, \cdots, x_N\}$ where $x \in \mathbb{Z}$ is amplitude. This implies duration $N/\Omega = f_{\theta_1}(\theta_1)$ and frame count $\boldsymbol{X} = \{X_1, X_2, \cdots, X_M \,|\, M = f_{\theta_2}(\theta_2)\}$. The frequency spectrum of a given frame is $\boldsymbol{F} = FFT(X_i)$. In order to simplify the mathematics and ensure smoothness, the information manifold $\boldsymbol{M}$ is based on $log\, p(x \,|\, \boldsymbol{\theta})$ where p is the Gamma Distribution. A derivation of suitable metric tensor and Christoffel symbol, can be found in [8].

To quantify an audio segment, the time series is mapped to a *curve* $Q : \boldsymbol{X} \to S$ such that $S = \{s_1, s_2, \cdots\}$ where $s = \{score, \theta_1, \theta_2\}$ for each segment. Score will be discussed later, but it represents a quantified measure of underlying musical geometry.

Every segmentation interval is scored and a previous score is used to determine the next model based on our predictor

$$P(x) = e^{p(x \,|\, \boldsymbol{\theta})}.$$

Maximum likelihood is determined by searching the geodesic curve. The model to yield the highest probability given the previous score is selected as the new model. This becomes the next segmentation rule and we continue in this way until end of file is reached. Many geodesic curves can be used, but the following parameterization was found to be effective $(2, 0.997, 0.800, -0.067)$. This initial four parameter vector is a solution to the geodesic curve where we have $(\theta_1, \partial\theta_1, \theta_2, \partial\theta_2)$. The geodesic equation is a second order ODE, but is solved numerically using the Runge-Kutta method for first order ODE's [1]. The curve is stored and tagged to the original audio meta-data.

III. Centricity Quantifier

The concept of *centricity* arises from the five fundamental properties of music argued by mathematician and music theorist Dmitri Tymoczko. Tymoczko devised a set of what he calls *conjunct melodic motion*, *harmonic consistency*, *acoustic consonance* and *centricity* found in any form of audio that can be called music [27]. Tymoczko offers the caveat that a subset of these properties are largely found in western music. Formally, the full quantification of all these musical properties would necessarily require us to leverage some, or all of the existing pitch processing literature. We have state of the art analysis tools for determining pitch characteristics [23], timbre/genre [20] and instrumentation [12]. Many methods are based on machine learning and use MEL-ceptrum, or CUSUM event detection. Fortunately, centricity can be determined using a much simpler approach. Centricity is defined by the claim; *"over moderate spans of time, one note is heard as being more prominent than the others, appearing more frequently and serving as a goal of musical motion."*

Our mathematical interpretation of this concept may not be precisely identical; however, we draw inspiration from Tymoczko's work. We complicate the original definition of centricity by assuming a central frequency envelope based on prominence.

Definition 1. *Given a time series as a sequence of equally sized frames, centricity is the phenomena wherein the frequency spectra sustains a consistent centroid within the prominent envelope of the respective frame.*

In this scenario *prominence* refers to the frequency peak with the largest amplitude. This is not limited to only a single peak, which is highly dependent on the size of the selected frame.

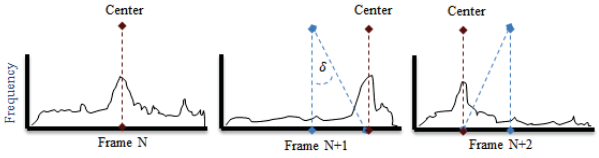

Fig. 1 Visual depiction of the Stretch Method: This shows the motion of the centric region relative to the central frequency

A. Envelope Detection and Centric Region

There are many ways to measure centricity as we have defined it, such as *spectral centroid* which works well with the short-time fourier transforms (STFT) [16]. The selection of the envelope is important to the centricity calculation, but before discussing how the envelope is chosen let us build intuition on how our centricity algorithm works.

Imagine you have a rubber band that is cut, taking it from genus 1 to genus 0. You then attach one end to a fixed point on an infinite line. Next, create a fixed axis at that same attachment position, but orthogonal to the infinite line. As you stretch the rubber band in either direction (assume we are stuck in two dimensions) the acute angular region carved out between the orthogonal vector and the rubber band, over a series of pulls, will have a maximum angle. This angle will be no greater than $\frac{\pi}{2}$, where $\pm\frac{\pi}{2}$ is somewhere at negative, or positive infinity. Fig. 1 depicts the *stretch method* algorithm. This provides the algorithm for centricity which takes samples of the time series at some interval. Each frame is transformed to the frequency domain and centered about the Cartesian origin. We use the FFT domain F_k to determine the *primary envelope* and mark the bounds. This region is where we do our centricity calculation.

The envelope is selected using the following algorithm.

Algorithm 1. *Centricity Envelope*

(1) *FFT the time frame and call it* F_k.
(2) *Set* k *equal to the frequency of the maximum peak.*
(3) *Move* $k \leftarrow k-1$ *until amplitude transitions below the standard deviation.*
(4) *Continue* $k \leftarrow k-1$ *until we again transition above the standard deviation, stop and mark position* s.
(5) *Repeat the same process to the right of* k *in the positive direction and mark the end point* e.

Algorithm 1 determines the bounds of the envelope. The region $[s, e]$ is known as the *centric region*. Once identified, we take one of two actions; *(a)* if the last centroid is not known, we mark the position $c = (s+e)/2$, otherwise *(b)* we calculate

$$\delta = sin^{-1}\left(\frac{|c_j - c_{j-1}|}{\sqrt{(A^2 + (c_j - c_{j-1})^2}}\right), \tag{7}$$

where c_j is the centroid of the current frame and $A \in \mathbb{R}$ is the max amplitude of the envelope in decibels (this smoothes angular comparisons). This has the effect of measuring the angle between two successive frames with respect to the centroid. The angles are added and averaged to compute the centricity score. As an example, Fig. 2 illustrates a detection where the algorithm is applied to a snippet of *Beethoven's minuet in G* with 6 sequential frames of 250 milliseconds each.

B. Centric Velocity

An additional quantifier can be combined with the centricity measurement algorithm. This quantifier is called *centric velocity* and works from the speed of movement between

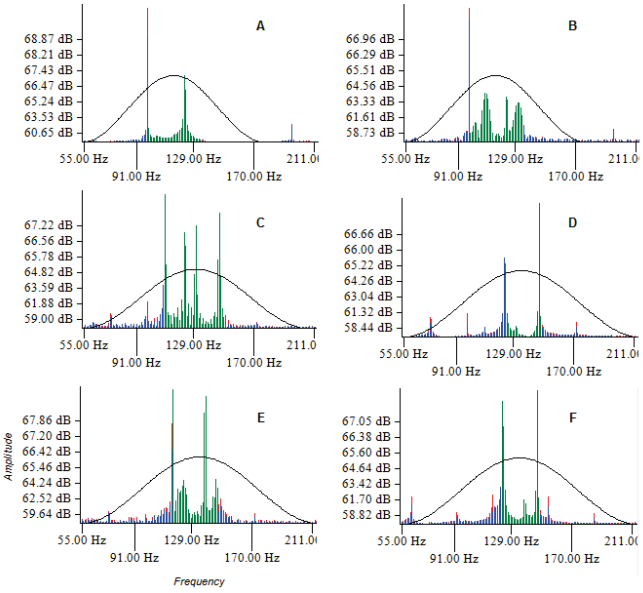

Fig. 2 Moving *centric region* on F_k at 250 ms per frame: At each frame we calculate the angle using (7)

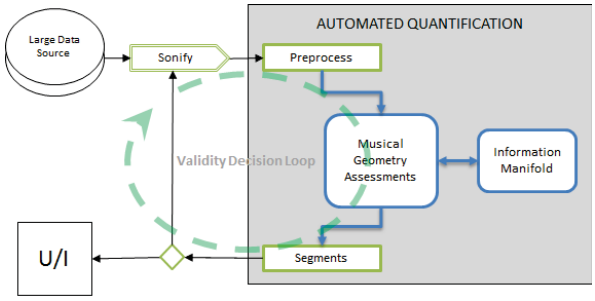

Fig. 3 Partial automation of sonified waveform analysis using musical information geometry

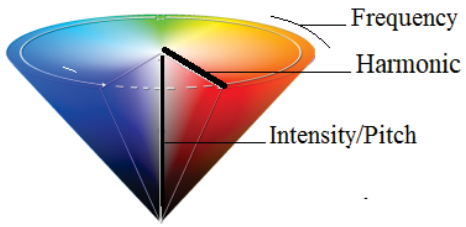

Fig. 4 HSV sonification color mapping: Frequency ≡ Hue, Harmonic ≡ Saturation, Pitch ≡ Value/Brightness

frames. The value is easily computed using the concept of physical displacement $v = d/t$ where d = distance and t = time. We measure centric velocity by determining the change in frequency displacement (in Hz) between one frame and the next. Velocity becomes $\frac{(C_j - C_{j-1})}{t}$ and yields units of Hz/ms. This measurement is combined with the centricity calculation in our experiments.

IV. HSV Sonifier

Sonification is a novel and somewhat rarely applied technique. Perhaps this is due to the voluminous nature of modern information not being suitable to a manual analysis process. In this regard, the ability to automate portions of the sonification pipeline may come in handy. Sonification and Auditory Display include sub-disciplines in Audification, PMSon (parameter mapping sonification) and Model Based Sonification [13]. Each branch offers distinct approaches to given problems, but retains shared constraints. Human perception is powerful, but limited by time and scope. Obviously, the purpose of audio transformation, in this regard, is to allow for human consumption, but how do we keep the sonification benefits (ie: finding useful information) without overwhelming the listener? The process begins by analyzing the stream of audio to find key inflections where interesting changes occur. Fig. 3 illustrates the proposed work flow which incorporates our architecture into automated sonification feedback. When raw data are received, the sonification algorithm feeds synthesized audio to our automated quantification analysis. The analysis determines which blocks of audio are relevant and responds accordingly by notifying the U/I and sonifier appropriately on how to proceed. The user is able to see and hear only interesting sections of very large datasets.

The proposed architecture ensures only paternalistically relevant information is prioritized to the listener. The full extent of this architecture is not easily identified, however we can experiment with its viability. As a use case, we developed a sonification technique designed for exploitation of conic color space. This exploit uses transitions involving color to frequency mapping, motivated by geometric parametrization of *hue*, *saturation* and *value* (HSV). When focusing on the unit conic structure, Section 2.3.1 of [24] provides a sound to frequency transformation based on *colored hearing* (a form of synesthesia). This is a road map toward interpreting conic space with respect to tone. We envision a frequency as a parameter to a complex number, not unlike the Fourier series, where a rotational coefficient accelerates a circular pattern about the origin. In real space this equates to the velocity of cycles of a sinusoid. There is an intuitive association with *hue* as illustrated in Fig. 4 wherein *hue* is rotationally bijective to color yielding a distinct map. We also know an increase in frequency is psychologically related to brighter coloring [24], thus we have a relationship with saturation and harmonic. Finally, we use *brightness (ranging from black to white)* as an additive element of *intensity* with respect to harmonic.

Let b = brightness, h = hue and s = saturation where $P = (h, s, b) \in [0, 1]^3$. We generate frequency space with the mapping $C : \mathbb{R}^3 \to \mathbb{R}$, using the *HSV Conic Sonifier* equation

$$C(P) = 2^{\lfloor (s+b)*3 \rfloor} \pi \left(16.35 + h \cdot 14.52\right) \tag{8}$$

Note: we have chosen the maximum range of harmonic to be 3. This is due to the fact that $(1 + 1) * 3 = 6$ when *saturation* and *brightness* are maximized. The human ear has a resonance frequency at approximately 3 kHz which increases sound pressure [15].

A. Audio Schematic

Sonifying an image using (8) requires a few post processing steps before frequency is synthesized. The image is scanned

from top-left to bottom-right and each pixel is mapped to an HSV Sonifier solution. From there, run length encoding is used to aggregate identical, adjacent frequency values so their total representation is transferred to relative play time. This process produces what we call an *audio schematic*. A side effect of run length includes singularities and artifacts in the distribution, hence it is not smooth. We developed an algorithm that smooths the curve based on our needs.

Assuming Cartesian coordinates and a function of the schematic curve $f(x)$, we iterate over the x-axis in the positive direction. Intuitively speaking, each vertical line segment $(x, 0)$ to (x, y) is homogeneously rotated in place in the clockwise direction until the first intersection with a subsequent line segment $(x', 0)$ to (x', y'). The following *hit test* determines intersection.

$$\text{Hit}(y) = \begin{cases} false & y < (x' - x) \\ true & \sqrt{y^2 - (x' - x)^2} \leq y' \end{cases} \tag{9}$$

Next, values below secant $(x, 0) - (x', y')$ are mapped according to the following

$$f(t) \leftarrow \begin{cases} 0 & f(t) < \frac{y'-y}{x'-x}(x' - x) \\ f(t) & otherwise, \end{cases} \tag{10}$$

assuming $t \in [x, x']$. What remains are the largest peaks and residue. Residue is zeroed and we construct the final curve by interpolating segments between non zero peaks. Fig. 5 illustrates this process.

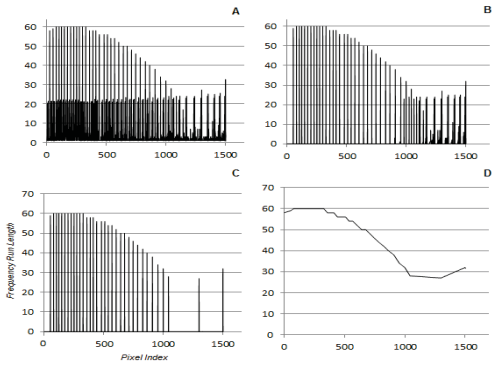

Fig. 5 Stages of audio schematic smoothing under (9) and (10)

The algorithm preserves the gradient of the image to a large extent retaining key inflections. This can be more easily understood by viewing Fig. 6 which shows anecdotal information as to the behavior of the sonifier on gradient surfaces.

V. Experimentation and Results

The theory proposed in this paper is difficult to envision from a practical standpoint. We set out to show random data that can be associated by musical geometry. As an experiment, a database was created consisting of approximately 226 different images. The images were selected based on homogenous content consisting of People, Shapes, Structures and Surfaces. Some categories were further subcategorized. For example, the category of Shapes contains Balls, Blocks, Curves and Squares. Table I provides a breakdown of the image database content.

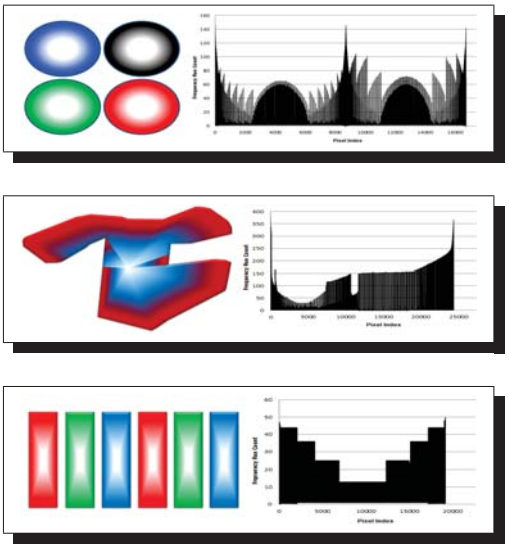

Fig. 6 Audio schematic examples of gradient imagery

TABLE I
Image Experimentation Database (Meta-Data)

Category	Total	Category	Total
People	39	Sporting Balls	31
Blocks	46	Curves	32
Squares	24	Structures	36
Surfaces	18		
			Total: 226

The experiment was done in the following way. We sonified audio for every file in the database using the algorithm from Section IV. Each audio file was classified using the framework in Section II. Geometry curves are correlated via Pearson coefficient and mapped to surface coordinates based on clustering success. Fig. 7 illustrates the homogeneity of clusters found for a given pass, where a pass consists of correlating every curve with every other at different offsets which is to say, comparing every audio file against every other audio file and transitively, every image to every other image. The x and y-axes represent the suggested offset and length to be used for correlation between curves. We must understand that each pass generates many clusters of varying consistency, the cluster with the highest homogeneity is selected as the z-value in the surface. These areas yield a set of solutions to higher performance clustering that can be employed in an application setting, over a brute force approach.

The duration of sonified audio was controlled using limits of $\{1, 2, 5, 10, 15, 25, 35, 45, 60\}$ seconds. We discovered a few substantial clusters using $1, 2$ and 3 seconds of sonified audio. The 5 second block contains the most substantial clusters, having uncontaminated groups of 3 to 16 images. Nearly every uncontaminated cluster up to 10 seconds contains the same image category, being that of Sports Balls. Table II shows a subset of the best clusters found, which is to say, near full

homogeneity of category at statistically significant sizes.

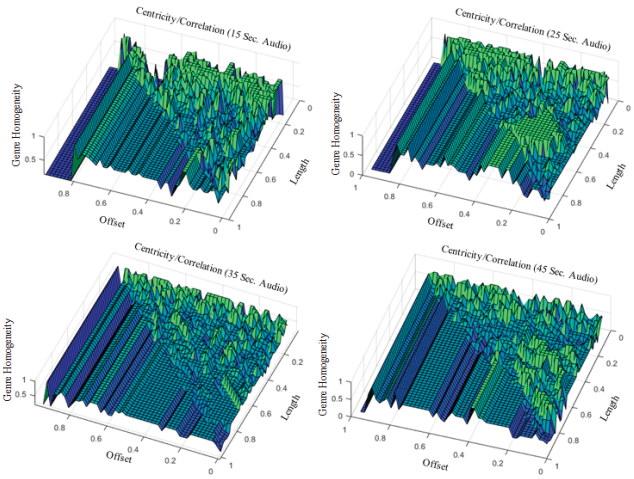

Fig. 7 Clustering success based on correlation of centricity quantified curves

We see a proliferation of image categories throughout the data as a function of audio time. Our methodology generated dozens of small to mid sized natural clusters. We have provided some examples in Figs. 8 and 9. We call them *natural* because they are driven solely by the data. When we examine the surfaces (Fig. 7) we can see several uncontaminated clusters in sequence over a specific offset, or length. Many peaks contain sequential clusters of the same category with differing content. We assume that nearby clusters with the same content can be associated in a data driven capacity. Presumably, an algorithm exists to find these and merge their content into larger clusters. In either case, these common groupings make the content readily accessible to a user interface, allowing an operator to identify and associate data quickly. This is ideal for searching large data spaces for consistent patterns.

As far as the behavior is concerned, sporting balls appear to have the most favorable outcome using this particular sonifier. Notably, the image backgrounds are quite different, containing mixed and noisy coloring for some and solid for others. It seems clear a predisposition toward the color scale in our gradient sonifier elicits the observed behaviors. The centricity algorithm links transition in gradient to frequency movement directly. The clusters of Block shapes (Fig. 9) are explained by the apparent lack of curvature, which would naturally create a contradictory equivalence class, thus sufficiently distinct and similar audio synthesis. Lastly, but perhaps most interesting, is the cluster of people, small as it may be. Fig. 10 shows a set of completely disparate photos of single and multiple humans. We conjecture this is due in part to the proliferation of curved surfaces within the images and gradient of skin tone. One final set of clusters is that of what were categorized as Structures. Structures consists of buildings, mountains and other natural phenomena. Fig. 11 shows one such cluster. There are several small clusters of this type, all containing buildings, homes or mountain views. The spurious nature of sharp edges, combined with smooth transitions, would create

TABLE II
EXAMPLE CLUSTERS FOR CENTRICITY QUANTIFIED IMAGERY

Cluster Size	Category	Homogeneity	% of Total Category
1 Second Sonified Audio			
8	Sporting Balls	100.00%	25.8%
3	Blocks	100.00%	6.5%
2 Seconds Sonified Audio			
5	Sporting Balls	100.00%	16.1%
4	Sporting Balls	100.00%	12.9%
3 Seconds Sonified Audio			
5	Sporting Balls	100.00%	16.1%
4	Sporting Balls	100.00%	12.9%
3	Sporting Balls	100.00%	9.7%
5 Seconds Sonified Audio			
16	Sporting Balls	100.00%	51.6%
15	Sporting Balls	100.00%	48.4%
12	Sporting Balls	100.00%	38.7%
11	Sporting Balls	90.91%	32.3%
7	Sporting Balls	100.00%	22.6%
5	Sporting Balls	100.00%	16.1%
4	Sporting Balls	100.00%	12.9%
3	Sporting Balls	100.00%	9.7%
10,15,20,30,35,45 Seconds Sonified Audio			
4	Structures	100.00%	11.1%
3	Structures	100.00%	8.3%
8	Blocks	100.00%	17.4%
6	Blocks	100.00%	17.4%
8	Sporting Balls	87.50%	22.6%
6	Sporting Balls	100.00%	19.4%
3	Sporting Balls	100.00%	9.7%
3	Squares	100.00%	12.5%
4	People	100.00%	10.3%
3	People	100.00%	7.7%
3	Blocks	100.00%	6.5%
4	Blocks	100.00%	8.7%
3	Curves	100.00%	9.4%
Large, Contaminated Clusters			
29	People	44.83%	33.3%
29	People	44.83%	33.3%
21	Curves	42.86%	28.1%
21	Curves	42.86%	28.1%
21	Blocks	52.38%	23.9%
21	Blocks	52.38%	23.9%

rare correlation opportunities with the sonifier which is likely why these are small and somewhat disparate. One final point to be made relates to sufficiently large contaminated clusters that retain significant numbers of similar content. This can be seen in the last subsection of Table II. For example, there is a cluster of 29 elements and although nearly half are contamination, 13 images contain single and multiple people being associated by the algorithm.

It is clear that several clustering solutions exist within the parameter variations. We have seen a consistent changes to cluster types and size as we modify the amount of synthesized audio. This is expected, but the search for new cluster solutions

Fig. 8 A natural cluster of various sports balls: There are 16 different images in this cluster from the 5 second sonification audio output

Fig. 9 Natural cluster of block shapes: There are 8 different images in the cluster from the 45 second sonification audio output

is very slow. An open problem is extended to examine performance improvement and optimal solutions for parameter selection which is to say, permutations of the information manifold, distribution range, sonification time, search pattern and correlation methodology.

There are clearly gaps in the clustering ability and some categories are completely ignored, such as the Surface category. The surface category contains various mathematical manifolds. It is apparent that the musical centricity and HSV sonifier do not express these types of images in a consistent fashion.

We surmise the likelihood of larger and better quality cluster manifestations under different parameter combinations. We also know that segmentation length and width create a distinct change in clustering (Fig. 7). An additional open challenge involves the analysis of the clustering surface. There exists a demarcation in clustering behavior above and below the point where correlation length and width are equivalent. Why the given pattern emerges and why such a distinct separation

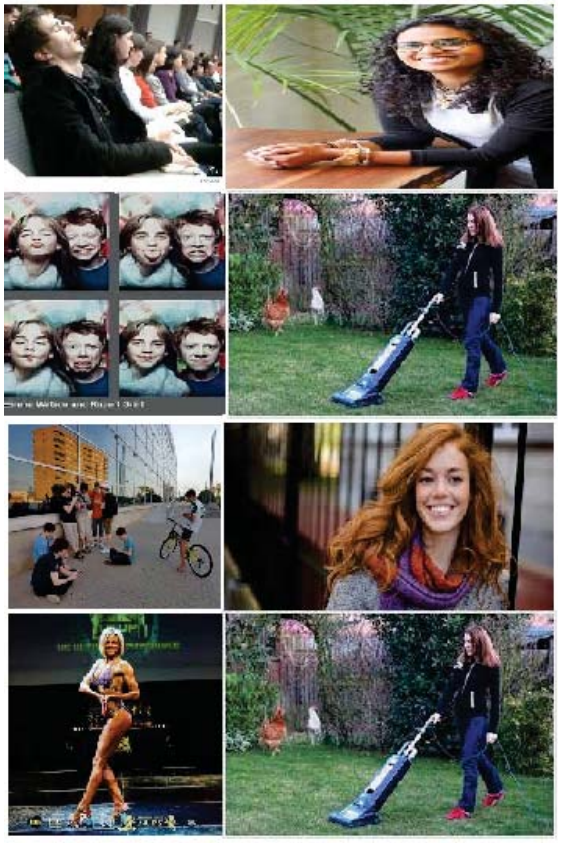

Fig. 10 Natural clusters of people: The top four images are a cluster from 10 second sonification; the bottom four are a cluster from 45 second

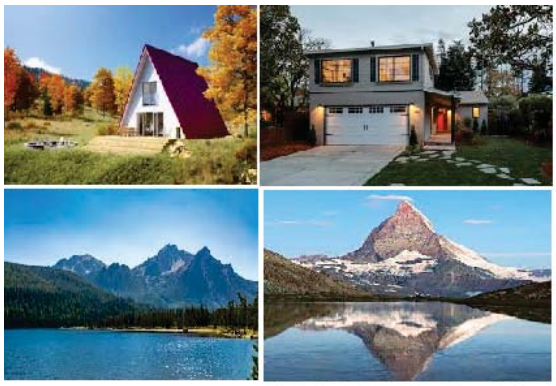

Fig. 11 Natural cluster of four structures from 20 seconds sonified audio

exists, remain unclear.

VI. Conclusions and Future Work

Some conclusions can not be drawn from the experimentation; however, we have shown that music can be used to associate the content of imagery. We have shown how it is possible to augment sonification architecture as a viable path toward automation in the user experience. We demonstrated the ability to quantify centricity and include it as a predictor of segmentation within the information manifold. We developed a sonifier and proved that it can

be used to transform imagery into audio for statistically significant content clustering. The proposed architecture exists for "plug n play" purposes, allowing one to insert different quantifiers and sonification algorithms. We surmise that different sonification techniques will induce varying outcomes for different image content. Further research is necessary to determine the role played by our unique sonifier, but we have shown musical geometry exposes a deeper level of content analysis.

A. Future Work

Our research considers several alternative implications to audio visualization and analysis; for example, the ability to provide real time feedback. This is not dissimilar to an electronic tuning device and we make no assumptions about the audio's similarity, composition, or musical relevance. There are also implications in feedback for visually based musical composition, signal detection and large data. One final ulterior motive is the improvement of synaesthetic visualization of audio (see [18], [25] for details).

REFERENCES

[1] Nassar H Abdel-All and EI Abdel-Galil. Numerical treatment of geodesic differential. In *International Mathematical Forum*, volume 8, pages 15–29, 2013.

[2] Md Akhtaruzzaman and Amir A Shafie. Geometrical substantiation of phi, the golden ratio and the baroque of nature, architecture, design and engineering. *International Journal of Arts*, 1(1):1–22, 2011.

[3] S-I Amari. Information geometry on hierarchy of probability distributions. *IEEE transactions on information theory*, 47(5):1701–1711, 2001.

[4] Shun-ichi Amari. Differential geometry of a parametric family of invertible linear systems-riemannian metric, dual affine connections, and divergence. *Mathematical systems theory*, 20(1):53–82, 1987.

[5] Shun-Ichi Amari. Information geometry and its applications: Convex function and dually flat manifold. In *LIX Fall Colloquium on Emerging Trends in Visual Computing*, pages 75–102. Springer, 2008.

[6] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[7] Ariel Caticha. The basics of information geometry. In *AIP Conference Proceedings*, volume 1641, pages 15–26. American Institute of Physics, 2015.

[8] William WS Chen and Samuel Kotz. The riemannian structure of the three-parameter gamma distribution. 2013.

[9] Arshia Cont, Shlomo Dubnov, and Gérard Assayag. On the information geometry of audio streams with applications to similarity computing. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):837–846, 2010.

[10] Arnaud Dessein and Arshia Cont. An information-geometric approach to real-time audio segmentation. *IEEE Signal Processing Letters*, 20(4):331–334, 2013.

[11] Jonathan T Foote and Matthew L Cooper. Media segmentation using self-similarity decomposition. In *Storage and Retrieval for Media Databases 2003*, volume 5021, pages 167–175. International Society for Optics and Photonics, 2003.

[12] Siddharth Gururani, Cameron Summers, and Alexander Lerch. Instrument activity detection in polyphonic music using deep neural networks. In *ISMIR*, pages 569–576, 2018.

[13] Thomas Hermann, Andy Hunt, and John G Neuhoff. *The sonification handbook*. Logos Verlag Berlin, 2011.

[14] Joachim Kopp, Ranjan Laha, Toby Opferkuch, and William Shepherd. Cuckoo's eggs in neutron stars: can ligo hear chirps from the dark sector? *Journal of High Energy Physics*, 2018(11):96, 2018.

[15] Luis Fernando Abanto León, Guillermo Kemper Vásquez, and Joel Telles. A novel fuzzy logic-based metric for audio quality assessment: Objective audio quality assessment. In *CONATEL 2011*, pages 1–10. IEEE, 2011.

[16] Tao Li and Mitsunori Ogihara. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574, 2006.

[17] Frank Nielsen. An elementary introduction to information geometry. *arXiv preprint arXiv:1808.08271*, 2018.

[18] Konstantina Orlandatou. Sound characteristics which affect attributes of the synaesthetic visual experience. *Musicae Scientiae, Vol. 19(4)*, pages 389–401, 2015.

[19] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 2020.

[20] Tae Hong Park and Sumanth Srinivasan. *The sound analysis toolbox (SATB)*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2016.

[21] CR Rao. Information and accuracy attainable in the estimation of statistical parameters. kotz s & johnson nl (eds.), breakthroughs in statistics volume i: Foundations and basic theory, 235–248, 1945.

[22] Markus Schedl, Emilia Gómez Gutiérrez, and Julián Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval. 2014 Sept 12; 8 (2-3): 127-261.*, 2014.

[23] Chetan Pratap Singh and T Kishore Kumar. Efficient pitch detection algorithms for pitched musical instrument sounds: A comparative performance evaluation. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1876–1880. IEEE, 2014.

[24] Shannon Steinmetz. *Sonic imagery: A view of music via mathematical computer science and signal processing*, pages 12–15. University of Colorado at Denver, 2016.

[25] Shannon Steinmetz. *Sonic imagery: A view of music via mathematical computer science and signal processing*. University of Colorado at Denver, 2016.

[26] Ke Sun and Stéphane Marchand-Maillet. An information geometry of statistical manifold learning. In *International Conference on Machine Learning*, pages 1–9, 2014.

[27] Dmitri Tymoczko. A geometry of music. Oxford University Press, 1 edition, 2011.

[28] Robert van Gend. The fibonacci sequence and the golden ratio in music. *Notes on Number Theory and Discrete Mathematics*, 20(1):72–77, 2014.

[29] Robert S Wolff. Sounding out images. *Computers in Physics*, 6(3):287–289, 1992.