On the Information Geometry of Audio Streams with Applications to Similarity Computing

Arshia Cont, Shlomo Dubnov, Gérard Assayag

On the Information Geometry of Audio Streams with Applications to Similarity Computing

*Arshia Cont, Shlomo Dubnov, and Gérard Assayag

***Abstract*—This paper proposes methods for information processing of audio streams using methods of information geometry. We lay the theoretical groundwork for a framework allowing the treatment of signal information as information entities, suitable for similarity and symbolic computing on audio signals. The theoretical basis of this paper is based on the information geometry of statistical structures representing audio spectrum features, and specifically through the bijection between the generic families of Bregman divergences and that of exponential distributions. The proposed framework, called *Music Information Geometry* allows online segmentation of audio streams to metric balls where each ball represents a quasi-stationary continuous chunk of audio, and discusses methods to qualify and quantify information between entities for similarity computing. We define an information geometry that approximates a similarity metric space, redefine general notions in music information retrieval such as similarity between entities, and address methods for dealing with non-stationarity of audio signals. We demonstrate the framework on two sample applications for online audio structure discovery and audio matching.**

I. Introduction

MUSIC Information Retrieval (MIR) systems deal one way or another with the information content of music signals, their transformations, or extraction of models or parameters from this information. A common question that many such systems ask at their front-end is what information is presented in the signal and to what relevancy? This question is central in almost all music information retrieval systems dealing either with temporal structures of audio data streams for search applications (query-by-humming, audio matching, music summarization etc.), or with temporal decomposition of audio (source separation, multiple-source identification, etc.).

In this paper, we seek a comprehensive framework that allows us to quantify, process and represent information contained in temporal structure of audio streams. An audio stream is a sequence of audio data presented to an algorithm one item at a time, thus capable of online processing of information. The framework introduced in this paper brings in concepts from various literatures: music signal processing, information geometry and machine learning. By this combination, we aim to investigate the natural geometric structures occupied by families of probability distributions representing audio streams that implicitly represent the ongoing information structure of the signal over time. Within this framework, music information arrives in discrete analysis windows over time and occupy statistical *points* in an information manifold. These statistical points are then analyzed within a generic mathematical framework called *Music Information Geometry*, that assures the existence of an approximate similarity metric space over data, and redefines common concepts in the MIR literature such as *similarity* and *metric balls*.

The present work inscribes itself within the more general framework of information dynamics measures for audio in relation to music cognition. Dubnov has studied information measure based on mutual information between the past and present of audio and showed its significance compared to data collected from listeners [1]. He later developed his method in [2] for non-stationary audio by separating the *data* and *model* aspects of information dynamics. One of the difficulties with this approach is determining what consists of relevant information between data and model. For instance, in the data case it is assumed that individual observations carry little information about the model, while in the model case they are represented by cluster centers, so the information between observations is ignored. The present work solves this problem by explicitly defining models on statistical points and providing mathematical tools for further processing.

Music Information Retrieval (MIR) systems mostly rely on the notion of *self-similarity measures* for music and audio [3] as a basis to compare and deduce music structures. Many MIR techniques also rely on geometric concepts in machine learning for building classifiers in supervised problems (genre classification, artist identification, query by example etc.) or clustering data in unsupervised settings (audio search engines, structure discovery etc.). Implicit in all these considerations is the fact that similarity measures, with all their variety of formulations, constitute a metric space where equivalence categories can be deduced and compared. At this stage, there is no clear boundary in the literature between *metrics* and the notion of *similarity*. Another drawback of common information processing methods in MIR is the wide use of *bag of features* models, where audio data is represented to the system *with no temporal order*. Despite their wide use, such techniques ignore the temporal dimensions of the data which is an essential criteria in many retrieval processes. The proposed *Music Information Geometry* framework aims at approximating metric spaces over a wide-variaty of signal representations and similarity measures that lie within the generic families of Bregman divergences and exponential family distributions, and by explicitly considering the temporal order of audio streams.

We emphasize that our focus here is on the mathematical properties of common distortion measures once formulated

Arshia Cont is with Institute of Research for Coordination of Acoustics and Music (IRCAM) in Paris.

using methods of information geometry, and not the difficult and open problem of characterizing which distortion measures best address the subjective quality of particular psychoacoustic characteristics of music. The major intent of this paper is to lay the theoretical groundwork for forthcoming experimental results. However, we exhibit results on two major MIR applications defined on information manifolds: *online audio structure discovery* and *audio matching*. The goal here is not to compare these results with the extensive literature within each application, but to showcase the power of information geometric formulations on complex problem sets. Details of experimental results are thus left for dedicated publications.

This paper is organized as follows: Section II introduces basic mathematical tools and theorems of information geometry over Bregman divergences and their relationship to exponential distributions. Section III, refines common tools and terms such as distortion and similarity to prepare the common ground. Section IV introduces our Music Information Geometry framework, providing tools, theorems and definitions that permit a migration from Bregman divergences to similarity metric spaces. Section V provides sample applications of the proposed framework proceeded by conclusions.

II. Preliminaries

In this section, we introduce the mathematical basis of our proposal. We start by introducing basic concepts of *information geometry* and move on to Bregman divergences and their geometric properties and introducing exponential families and their behavior in a Bregman geometry. The reader is referred to [4]–[6] for details and proofs.

A. Information Geometry of Statistical Structures

Let us consider a family of probability distributions specified by a vector parameter $p(x, \boldsymbol{\xi})$ where $\boldsymbol{\xi}$ is a vector constituting the model parameters of the probability distribution. This set can be regarded as a manifold under certain regularity conditions where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ would be its coordinate system. A manifold is an abstract mathematical space in which every point has a neighborhood which resembles a regular Euclidean space but the global structure may be more complicated. By defining probability distributions on a manifold, each *point* would then refer to a realization of a family of probability distribution. The manifold has a natural geometrical structure if the geometrical structure is (1) invariant under the coordinate system (or parameters) used to specify the distributions, and (2) invariant under rescaling of the random variable x.

Amari [4] shows that representing statistical structures within a Riemannian geometry equipped with the Fisher Information measure as inner product g, and a canonic affine connection Δ, would constitute an information geometry. This construction allows definitions for many geometrical structures such as distances, lines, volumes etc. Among such constructs, the existence of dual canonic divergences or *distance like* measure D between two points in the geometry is of extreme importance to us. Alternatively, there have been attempts to define information geometries on Riemannian manifolds with g, directly by inducing divergences instead of affine connections. Recently, Zhang [7] introduced a canonical form of affine connection that deduces many types of divergence functions which are in common use in engineering including the well-known *Bregman Divergence* family [8]. Given these findings, and within the framework introduced in [7], we can assume a geometrical structure over probability manifolds $\mathcal{S}$ using Fisher Information and Bregman Divergences.

Throughout this paper, we assume that a system under measurement generates families of probability distributions on a dual information manifold defined as $(\mathcal{S}, g, \Delta^D, \Delta^{D^*})$ where its geometric properties are induced by employing *Bregman Divergence* D. Also, the term *point* represents a family of probability distributions that belongs to a probability simplex $\mathcal{X} \in \mathbb{R}^d$. Vector mathematical constructs are notated using boldface characters in contrast to scalar constructs.

B. Elements of Bregman Geometry

Definition 1 ([6], [8]). For any two points $\boldsymbol{p}$ and $\boldsymbol{q}$ of $\mathcal{X} \in \mathbb{R}^d$, the Bregman Divergence $D_\Phi(.,.) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ of $\boldsymbol{p}$ to $\boldsymbol{q}$ associated to a strictly convex and differentiable function Φ (called *generator function*) is defined as:

$$D_\Phi(\boldsymbol{p}, \boldsymbol{q}) = \Phi(\boldsymbol{p}) - \Phi(\boldsymbol{q}) - \langle \nabla\Phi(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle \qquad (1)$$

where $\nabla\Phi = \left[\frac{\partial\Phi}{\partial x_1}, \ldots, \frac{\partial\Phi}{\partial x_d}\right]$ denotes the gradient operator and $\langle \boldsymbol{p}, \boldsymbol{q} \rangle$ the inner or dot product.

The most interesting point about Bregman family of divergence is that they can generate many of the common *distances* in the literature. Table I shows several of these canonical generations (see also [9]). Among common properties of Bregman divergences, we can easily show that they are *non-negative*, convex on the first argument, and linearly invariant on Φ. The reader is referred to [5, Appendix A] for details.

TABLE I
Bregman Divergence Generation Examples

$\mathcal{X}$	Φ	$D_\Phi(\boldsymbol{p}, \boldsymbol{q})$	Generic Name
$\mathbb{R}^+$	$x \log x - x$	$\boldsymbol{p} \log \frac{\boldsymbol{p}}{\boldsymbol{q}} - \boldsymbol{p} + \boldsymbol{q}$	Kullback-Leibler Div.
$\mathbb{R}^+_*$	$-\log x$	$\frac{\boldsymbol{p}}{\boldsymbol{q}} - \log \frac{\boldsymbol{p}}{\boldsymbol{q}} - 1$	Itakura-Saito Div.
$\mathbb{R}^d$	$\|\|\boldsymbol{x}\|\|^2$	$\|\|\boldsymbol{p} - \boldsymbol{q}\|\|^2$	Squared Euclidean

1) Dual Structure: An important property of information manifolds is the existence of a dual structure based on Legendre transformation for any geometrical structure on $\mathcal{S}$. Using statistical manifolds on Bregman divergences, this dual structure can be entirely exploited by defining the dual divergence that is generated by the Legendre transformation of Φ or $\Phi^* = \int \nabla^{-1}\Phi$. In the sequel, we denote the dual point of $\boldsymbol{x}$ as $\boldsymbol{x}' = \nabla\Phi(\boldsymbol{x})$. The following property shows the relationship between a Bregman divergence and its dual:

Property 1 ([6]). *D_{Φ^*} is also a Bregman divergence called the Legendre dual divergence of D_Φ and we have:*

$$D_\Phi(\boldsymbol{p}, \boldsymbol{q}) = \Phi(\boldsymbol{p}) + \Phi^*(\boldsymbol{q}) - < \boldsymbol{p}, \boldsymbol{q}' >= D_{\Phi^*}(\boldsymbol{q}', \boldsymbol{p}')$$

2) Bregman Balls: In analogy to Euclidean geometry, we can define a *Bregman ball*. Due to the asymmetric nature of Bregman divergences, a Bregman ball can be defined as two counterparts which are *right-type* or *left-type*. A Bregman ball of right-type centered at $\boldsymbol{\mu}_k$ with radius R_k is defined as:

$$B_r(\boldsymbol{\mu}_k, R_k) = \{\boldsymbol{x} \in \mathcal{X} : D_\Phi(\boldsymbol{x}, \boldsymbol{\mu}_k) \leq R_k\} \tag{2}$$

Similarly, the Bregman ball of left-type $B_\ell(\boldsymbol{\mu}_k, R_k)$ is defined by inverting the divergence in eq. 2 to $D_\Phi(\boldsymbol{\mu}_k, \boldsymbol{x})$.

3) Bregman Information: Let $\boldsymbol{X}$ be a random variable following a probability ν that takes values in $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^d$. Let $\boldsymbol{\mu} = E_\nu[X]$. Then the *Bregman Information* of X is:

$$I_\Phi(X) = E_\nu[D_\Phi(X, \boldsymbol{\mu})] = \sum_{i=1}^{n} \nu_i D_\Phi(\boldsymbol{x_i}, \boldsymbol{\mu}) \tag{3}$$

Well-known examples of Bregman Information are *variance* and *mutual information* (see [5]).

C. Exponential Family of Distributions

Among different distribution families, the exponential family of probability distributions are of special importance and have found their way in many pattern recognition applications. Their canonical definition is as follows:

$$p(x|\boldsymbol{\theta}) = \exp\left[< \boldsymbol{\theta}, \boldsymbol{f}(x) > - F(\boldsymbol{\theta}) + C(x)\right] \tag{4}$$

where $f(x)$ is the *sufficient statistics* and $\boldsymbol{\theta} \in \mathcal{X}$ represents the *natural parameters*. F is called the *cumulant function*, and fully characterizes the exponential family while the term $C(x)$ ensures density normalization. Many of commonly used distribution families can be generated by proper choice of *natural parameters* and *sufficient statistics* as demonstrated in table II. The expectation of X with respect to $p(\boldsymbol{x}; \boldsymbol{\theta})$ is called the *expectation parameter* or $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \int \boldsymbol{x}\, p(\boldsymbol{x}; \boldsymbol{\theta})\, d\boldsymbol{x}$.

TABLE II
EXAMPLES OF EXPONENTIAL FAMILY DISTRIBUTIONS

Distribution $p(x, \Theta)$	Natural Parameters $\boldsymbol{\theta}$	Cumulant function $F(\boldsymbol{\theta})$
$\mathcal{N}(x; \nu, \sigma^2)$ (Univ. Gaussian)	$\{\frac{\nu}{\sigma^2} \quad \frac{-1}{2\sigma^2}\}$	$-\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\nu}, \boldsymbol{\Sigma})$ (Multiv. Gaussian)	$\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} \quad -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\}$	$\frac{1}{2}\boldsymbol{\nu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$ $+\frac{1}{2}\log\det(2\pi\boldsymbol{\Sigma})$
$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$ (Multinomial)	$\left\{\log\frac{q_i}{1-\sum_{j=1}^d q_i}\right\}$	$\log(1+\sum_{i=1}^d \exp\theta_i)$

1) Duality of natural and expectation parameters: It can be shown [4] that the expectation and natural parameters of exponential families of distributions have a *one-to-one* correspondence and span spaces that exhibit a dual relationship as outlined in section II-B1. Due to the convexity of F, its dual F^* exists on Θ and the following important one-to-one mappings hold between the two spaces:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla F(\boldsymbol{\theta}) \qquad \text{and} \qquad \boldsymbol{\theta}(\boldsymbol{\mu}) = \nabla F^*(\boldsymbol{\mu}) \tag{5}$$

meaning that the expectation parameter is the image of the natural parameter under the gradient mapping and vice-versa.

2) Bijection with Bregman divergences: A natural question to ask at this point is: *What family of Bregman divergence should be chosen for a given family of exponential distributions?* The answer lies in the important property of *bijection* between exponential families and Bregman divergences as proved in [5]. This theorem implies that every regular exponential family corresponds to a *unique* Bregman divergence and vice versa, leading to a one-to-one mapping:

Theorem 1 ([5]). *Let $p(\boldsymbol{x}; \boldsymbol{\theta})$ be the probability density function of a regular exponential family of distribution with F as its associated cumulant function. Let F^* be the conjugate function of F. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the natural parameter and $\boldsymbol{\mu}$ be the corresponding expectation parameter. Then $p(\boldsymbol{x}; \boldsymbol{\theta})$ can be uniquely expressed as*

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp(-D_{F^*}(\boldsymbol{x}, \boldsymbol{\mu}))b_{F^*}(\boldsymbol{x}) \tag{6}$$

where $b_{F^*}(\boldsymbol{x})$ is a uniquely determined function.

Table III shows three examples of bijection between exponential distributions and Bregman divergences with derived *expectation parameters* corresponding to examples in table II. This information suggests that the bijected Bregman divergence for Multinomial distributions is the well-known KL divergence, whereas for a spherical Gaussian it amounts to a simple Mahalanobis distance (see [5] for more examples).

TABLE III
EXPONENTIAL DISTRIBUTIONS WITH BIJECTED BREGMAN DIVERGENCES

Distribution $p(x, \Theta)$	Expectation Parameter $\boldsymbol{\mu}$	Bijected Bregman Div. $D_{F^*}(\boldsymbol{x}, \boldsymbol{\mu})$
$\mathcal{N}(x; \nu, \sigma^2)$	ν	Squared Euclidean $\frac{1}{2\sigma^2}(x-\nu)^2$
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\nu}, \boldsymbol{\Sigma})$	$\boldsymbol{\nu}$	Mahalanobis $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma(\boldsymbol{x}-\boldsymbol{\mu})$
Multinomial	$\{Nq_j\}_{j=1}^{d-1}$	Kullback-Leibler

3) Mixture Models: In machine learning and pattern recognition literature, many stochastic sources are expressed as a mixture of k densities of the same exponential family. This yields a soft clustering where clusters correspond to the components of the mixture model, and the soft membership of a data point in each cluster is proportional to the probability of the data point being generated by the corresponding density function. Using the right side of eq. 6, the log-likelihood of an exponential mixture model with mixture weights π_i becomes:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}|\Gamma) &= \sum_{i=1}^{k} \log[\pi_i b_{F^*}(\boldsymbol{x})] D_{F^*}(\boldsymbol{x}, \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{k} \nu_i D_F(\boldsymbol{\theta}_i, \boldsymbol{x}') \end{aligned} \tag{7}$$

where property 1 and relationships in eq. 5 have been employed. Note that eq. 7 is simply a reiteration of the Bregman Information in the dual setting and up to an additive constant. This simply states that there is a duality relationship between exponential distributions and their mixtures. This is also true for more general affine connections as discussed in [4].

III. Distortions, Similarity, and Metrics

An information manifold defined on Bregman divergences or $(\mathcal{S}, g, \Delta^D, \Delta^{D^*})$ provides us with interesting information theoretic tools for qualification and quantification of parametric stream information. Once such a framework exists, it is desirable to apply it to common pattern recognition problems such as nearest neighbor search or segmentation schemes to name a few. Such measures of information have been widely referred to as *distortion measures* in the speech and audio processing literatures. The distortion between two entities represents the cost resulting when the first is reproduced by the other and is related to new information carried from one entity to other. Distortion measures have wide variety of applications in the design and comparison of systems [10]. Despite their usefulness, these measures do not guarantee *equivalence* between entities if distortion is low. This is in contrast to most information processing systems where a notion of *metric* is required to assure equivalence between classes for clustering or classifying data points or clusters. In this section, we study necessary properties for metric equivalence and discuss the behavior of Bregman distortions as metrics.

Let Ω be a nonempty set and $\mathbb{R}^+$ be the set of non-negative real numbers. A *metric* function on Ω is a function $d : \Omega \times \Omega \rightarrow \mathbb{R}^+$ if it satisfies the following properties [11]:

Property 2. $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ *iff* $\boldsymbol{x} = \boldsymbol{y}$

Property 3 (Symmetry)**.** $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$

Property 4 (Triangle Inequality)**.** $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$

We are interested in a particular type of distance, the "similarity distance". In the field of Music Information Retrieval, Jonathan Foote has been credited for promoting and using self-similarity measures for music and audio [3]. The MIR literature on database search, structure discovery, query-based retrieval and many more, rely on Foote's general notion of *similarity* as a basis to compare, retrieve, and discover music structures. Nevertheless, the metrics employed in such approaches lack one or more of the properties above, and moreover do not necessarily address any information theoretic aspect of the content. In this section, we study the notions of distortions, similarity and metric spaces with a special eye on the bijected Bregman divergences on information manifolds to pave the way for the proposal in section IV.

A. Equivalence and Similarity

Within information entities, an ideal similarity metric d and distortion measure D are inversely related. In other words, two entities have *high* similarity when the information rate between them is *low*, and vice versa. Given this intuition, we can consider two information states similar if the information carried from one to the other is minimal. Because signals can have arbitrary forms, usual choices for assessing signal difference like mean-squared error make little sense. Instead we rely on distance measures that quantify difference between the signals' probabilistic descriptions. We thus append the following definition to property 2:

Definition 2 (Similarity)**.** Two entities $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \mathcal{X}$ are assumed to be *similar* if the information gain by passing from one representation to other is zero or minimal; quantified by $D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) < \epsilon$ which depends not on the signal itself, but on the probability functions $p_X(x; \boldsymbol{\theta}_0)$ and $p_X(\boldsymbol{x}; \boldsymbol{\theta}_1)$.

Following the intimate relationship between exponential families and Bregman divergences, they would naturally fit to the above definition to detect similar entities when audio streams are modeled parametrically as exponential family of distributions. While property 2 is inherent for all D_F, Bregman divergences are *not* necessarily metrics since they are usually not symmetrical and the triangular inequality does not generally hold. We now study these two missing properties and provide the grounds to approach them in section IV.

B. Symmetrized Bregman Divergences

Bregman divergences are not necessarily symmetric and various methods exist to make them so. A common approach is to employ the J-divergence or

$$D_F^J = \frac{1}{2}\left[D_F(\boldsymbol{x}, \boldsymbol{y}) + D_F(\boldsymbol{y}, \boldsymbol{x})\right] \tag{8}$$

as in [12], [13], but this symmetrization scheme does not fit the dually flat manifold [9], and requires further considerations for use within applications that necessitate similarity computing.

C. Triangle Inequality

Among the three properties for metrics, the triangle inequality is probably the most non-trivial. The triangle inequality can be generalized for any triple $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ in $\mathcal{X}$ as follows:

$$D_F(\mathbf{x}, \mathbf{z}) + D_F(\mathbf{z}, \mathbf{y}) = D_F(\mathbf{x}, \mathbf{y}) + \langle \mathbf{x} - \mathbf{z}, \boldsymbol{y}' - \boldsymbol{z}' \rangle \tag{9}$$

If the underlying geometry is dually flat, which is the case with manifolds deduced from Bregman divergences, the generalized Pythagoras theorem states that

$$D_F(\mathbf{x}, \mathbf{y}) \geq D_F(\mathbf{x}, \mathbf{z}) + D_F(\mathbf{z}, \mathbf{y}) \tag{10}$$

where the equality hold only if for $\mathcal{X}$ being a convex simplex of $\boldsymbol{x}$ [4], $\boldsymbol{z} = \operatorname{argmin}_{\boldsymbol{q} \in \mathcal{X}} D_F(\boldsymbol{q}, \boldsymbol{y})$ which is not true for arbitrary entities $r \in \mathcal{X}$, and we are left with eq. 10 which is the *inverse* of property 4. Therefore special attention must be paid in computing similarity using regular Bregman divergences.

IV. Music Information Geometry

Using mathematical tools introduced so far, we aim at providing a framework for processing and qualifying the effectiveness of audio streams. We define an affine information geometry $(\mathcal{S}, g, \Delta^D, \Delta^{D^*})$ using Bregman divergences induced by the choice of statistical distribution over incoming data and represented hereafter as D_F, where F is deduced from theorem 1. We start by presenting the general framework, proceed to *data* and *model* information entities, and discuss information theoretic tools useful for many pattern recognition and information retrieval applications.

A. General Framework

In our framework, audio data arrives incrementally to the system as time series $\boldsymbol{X}_{t_i}$ containing sampled overlapping windows of audio where t_i is time (in seconds) of the window center. For simplicity, we drop the i index hereafter and use $\boldsymbol{X}_t$ instead where $t \in \mathbb{N}$. We assume that data underlying $\boldsymbol{X}_t$ is generated by a family of exponential distributions (or mixture there of). By this assumption, theorem 1 would provide us automatically with distortion measure D_F and underlying geometrical tools discussed so far to introduce an information processing framework. In this section, we discuss these assumptions and their consequence in the design and formulation of common problems in audio processing.

The choice of the exponential family distribution over time series $\boldsymbol{X}_t$ depends on the nature of the problem to solve and constitutes the *a priori* over modeling. Despite this limitation, generic exponential distributions (or their mixtures) are widely employed in general pattern recognition as well as audio and speech processing systems either implicitly or explicitly. For example many researchers choose a time-frequency representation over $\boldsymbol{X}_t$ as $\boldsymbol{S}_t(\omega)$ such as short-time Fourier or wavelet transforms, where each $\boldsymbol{S}_t(\omega)$ can be treated as frequency distributions or histograms of the corresponding $\boldsymbol{X}_t$. Such histogram features can be assumed, without loss of generality, to be generated by *Multinomial* distributions with the well-known KL divergence as bijected distortion measure. This choice has been empirically proved in [12] for concatenative speech synthesis. Other systems tend to use more compact representations for audio signals such as Cepstral Coefficients and/or by directly modeling through probability distributions with sparser natural parameter space. The review of such systems is out of the scope of this paper but existing literature should convince the reader of the wide use of exponential distributions and their mixtures. In summary, any design process for a given problem that involves exponential distributions (or their mixtures) as front-end has a unique information geometry defined by its bijected Bregman divergence.

Using exponential distributions over data streams, the time series X_t can be represented by their equivalent distributions $p(\boldsymbol{x}, \boldsymbol{\theta}_t)$ or by natural parameters $\boldsymbol{\theta}_t \in \Theta$. $\boldsymbol{\theta}_t$s constitute the *points* of the information manifold, referred hereafter as *data points*. Converting back and forth between the data in X_t (or relevant feature representations) and $\boldsymbol{\theta}_t$ is problem dependent but is a one-to-one mapping (see [5] for examples). In the following subsections, we employ information geometric tools for information processing of underlying audio streams.

B. From Data Information Rate to Model Information Rate

The first step in any information processing system is to introduce measures quantifying the amount of information carried through the signal. Following [2], we denote such measure as *Information Rate (IR)* within a transmission process over a noisy time-channel and defined as the relative reduction of uncertainty of the present considering the past. [2] shows that the Information Rate at time $t = T$ is equal to the mutual information carried between the past $\{X_1, \cdots, X_{T-1}\}$ (denoted in the sequel as X_1^{T-1}) and history of the signal up to present or X_1^T. It is further shown in [14] that for a stationary Gaussian process, IR can be approximated asymptotically in T using the spectral flatness measure of the time signal, or the ratio between geometrical and arithmetical means of $\{\boldsymbol{S}_t(\omega)\}$. We refer to this measure as *Data-IR*, reflecting information rates on data points. It can be proven that this *data-IR* measure is a special case of Bregman Information for Itakura-Saito (IS) divergence, widely used in speech and audio as a distortion measure on power spectra [10]. See [15, Ch. 4] for proof.

Using Bregman Information on information manifolds, the Data-IR measure can thus be extended to other representational aspects of the underlying stream. Despite this theoretical comfort, Data-IR is not useful in practice for two main reasons: (1) the underlying assumption of stationarity on X_t which is not true for real audio, and (2) extensive consideration of data-points in computation specially for long streams.

To tackle both issues, we adopt the plausible hypothesis that the signal is stationary in a finite and continuous time-interval under some *model* $\boldsymbol{\theta}_k$ and described through $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta}_k)$. Within our information manifold, this draws down to the geometric intuition that a set of continuous data points on our manifold are concentrated around a single point representative of $\boldsymbol{\theta}_k$.

Definition 3 (Models)**.** Given $(\mathcal{S}, g, \Delta^D, \Delta^{D^*})$ on a regular exponential family formed on $\boldsymbol{X}_k$, a *model* $\boldsymbol{\theta}_i$ consists of a set $\mathcal{X}_i = \{\boldsymbol{X}_k | k \in \mathcal{N}, \mathcal{N} \subset \mathbb{N}\}$ that forms a minimum enclosing *Bregman Ball* $B_r(\boldsymbol{\mu}_i, R_i)$.

A *model* defined and constructed on an audio stream refers to continuous *data points* on the information manifold that are self-contained in an information theoretic sense, and corresponds to a quasi-stationary chunk of audio in the stream. Therefore, the distinction between *data points* and *models* on an audio stream manifold is structural: data points refer to micro-structures of the audio whereas models correspond to macro-structures that give sense to the global structure of the stream. In this manner, it makes little sense to directly compute between data points where their model parameters provide enough information for intra-structural comparisons. For similarity computing, it is more useful to structurally refer to model parameters first and if needed, refer to the data points within the model. Therefore, providing a *metric space* as in section III is crucial for between-model comparisons.

The above definition requires us to formalize the following aspects: (1) How to form a minimum enclosing ball once model data-points $\mathcal{X}_i$ are given, (2) How to quantify the information rate carried from one model to another, and (3) how to incrementally achieve a segmentation of audio streams to information entities or models as defined above. The following subsections would address each question separately.

C. Centroid Computation

An important tool in both quantizer design and cluster analysis techniques is the generalized centroid of a cluster of data points. Given a cluster of points, this is the *best single representative* of the cluster and defined as the optimizer of the minimum average distance for the entire set of points in the

cluster. Banerjee et al [5] have proven the following important theorem for Bregman centroids:

Theorem 2 ([5]). *Let X be a random variable that takes values in $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ following a probability ν. Given a Bregman divergence D_F, the* right type *centroid of $\mathcal{X}$ or*

$$\boldsymbol{c}_R^F(\mathcal{X}) = \underset{\boldsymbol{c}}{\operatorname{argmin}} \sum_{i=1}^{n} \nu_i D_F(\boldsymbol{x}_i, \boldsymbol{c}) \tag{11}$$

is unique, independent of F and coincides with $\boldsymbol{\mu} = E_\nu[X]$ or for $\nu_i = 1/n$ to center of mass $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$.

It is important to notice the equivalence between eq. 11 and that of Bregman Information on uniform distributions with additional optimization. In other words, a computed centroid on an information geometric framework represents a minimum enclosing ball in terms of information content, or minimum distortion point of the given set. Given this, we can safely adopt the Bregman centroid computation for forming balls representing *models* of our information geometric framework.

In addition to the above definition and due to general asymmetry of Bregman divergences, we can define a *left-type* centroid by reversing the order of computation in eq. 11. Obviously, theorem 2 does not hold for the left-type centroid and the optimization becomes non-trivial. We can however employ the dualistic structure of our information manifold to obtain $\boldsymbol{c}_L^F$. Combining theorem 2 and property 1 we obtain:

$$c_L^F(\mathcal{X}) = (\nabla F)^{-1}(\sum_{i=1}^{n} \nabla F(x_i)) = (\nabla F)^{-1}(c_R^{F^*}(\mathcal{X}')) \tag{12}$$

stating that the left-type centroid is obtained by calculating the right-type centroid in the dual manifold using theorem 2 and converting it back to the original space.

For asymmetric Bregman divergences, a symmetrized Bregman centroid on the set $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^n \subset \mathcal{X}$ can be defined by the following optimization problem:

$$\boldsymbol{c}^F(\mathcal{P}) = \underset{\boldsymbol{c}\in\mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{D_F(\boldsymbol{c}, \boldsymbol{p}_i) + D_F(\boldsymbol{p}_i, \boldsymbol{c})}{2} \tag{13}$$

conforming to the symmetrization scheme in eq. 8. This optimization problem has been previously addressed in [16] for Kullback-Leibler divergences and by employing convex optimization techniques. It is shown in [9] that it can be extended to general Bregman divergences and simplified to a constant-size system relying on the right-type and left-type centroids by employing duality and a geodesic-walk dichotomic approximation algorithm; hence, well adapted to information manifolds of exponential distributions.

Solving eq. 13 requires an optimization framework in contrary to most literature that define (for example) symmetrized KL divergence centroids as arithmetic or normalized geometric mean of the left-type and right-type. Both approaches in [16] and [9] empirically prove this remark on image and audio processing applications. For our framework, we adopt the geodesic-walk algorithm in [9] to solve for an optimal symmetric Bregman ball. The radius of a given Bregman ball B_r with centroid c_R^F on the set $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^n \subset \mathcal{X}$ is simply the *Bregman Information* of eq. 3 on the set $\mathcal{P}$. For the symmetric construction above, it is shown that this radius is equal for the right-sided and left-sided centroids [9, Corollary 3.3].

D. Model Comparison and Data Membership Check

In our information geoemtry, an audio stream is thus represented by Bregman Balls $B_r^k(\boldsymbol{\mu}_k, R_k)$ which by themselves contain continuous data points $\boldsymbol{\theta}_t$. Here, we study the task of qualifying information rates within models and also membership of data points to models. These operations are important in many pattern recognition and information retrieval tasks that require information theoretic comparisons between entities such as similarity and nearest neighbor seach and clustering.

Following our definition of similarity, we can safely assign the information rate for passing from one model $B_r^i(\boldsymbol{\mu}_i, R_i)$ to another $B_r^j(\boldsymbol{\mu}_j, R_j)$ as the distortion between the two representative centroids of the two clusters or $D_F(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ (and similarly $D_F^J(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ for symmetrized centroids). However, checking for membership of an arbitrary data point $\boldsymbol{X}$ to a given ball $B_r^k(\boldsymbol{\mu}_k, R_k)$ containing its own set of points becomes non-trivial mostly due to the lack of triangle inequality. Note that if the triangle inequality holds, this membership check simply amounts to checking whether $D_F(\boldsymbol{X}, \boldsymbol{\mu}_k) \leq R_k$. In the absence of this property, we perform tests by projecting $\boldsymbol{X}$ onto the Bregman ball $B_r(\boldsymbol{\mu}_k, R_k)$. This projection is the unique minimizer $\boldsymbol{X}_B$ such that $\boldsymbol{X}_B = \operatorname{argmin}_{\boldsymbol{x}\in B_r^k} D_F(\boldsymbol{x}, \boldsymbol{\mu}_k)$. Once $\boldsymbol{X}_B$ is established, membership can be obtained by checking whether $D_F(\boldsymbol{X}, \boldsymbol{X}_B) \leq \epsilon$.

It can be easily shown that $\boldsymbol{X}_B$ lies on the geodesic line connecting $\boldsymbol{X}$ and $\boldsymbol{\mu}_k$ [6], [17]. Such geodesic is characterized as the set of points $\Gamma_{\boldsymbol{X}\boldsymbol{\mu}_k} = \{\nabla^{-1} F\left((1-\lambda)\boldsymbol{X}' + \lambda\boldsymbol{\mu}_k'\right) | \lambda \in [0,1]\}$. To find $\boldsymbol{X}_B$, we can first check whether $\boldsymbol{X}$ is inside the ball or not: $D_F(\boldsymbol{X}, \boldsymbol{\mu}) > R_k$. If not, then $D_F(\boldsymbol{X}, \boldsymbol{\mu}) \leq R_k$ and hence the projection is the point itself. Otherwise, $\boldsymbol{X}_B \in \Gamma_{\boldsymbol{X}\boldsymbol{\mu}_k}$ for some arbitrary $\lambda_B \in [0,1]$. Geometrically, such a point must lie on the boundary of B_r and thus combining $\Gamma_{\boldsymbol{X}\boldsymbol{\mu}_k}$ with a bisection can approximate the placement of $\boldsymbol{X}_B$ and can be found by a linear search as proposed in [18]. Note that this projection procedure assures an approximation of the Pythagoras equality in eq. 10.

E. Incremental Segmentation/Change Detection

In this section we propose a simple method for online segmentation of an audio stream to information geometric *models* and through a change detection process. Given a set of data points in $\mathcal{S}$, the problem of finding *models* according to definition 3 is a classical clustering problem in $\mathcal{S}$ where each cluster defines a minimum enclosing Bregman ball over data points. Offline Clustering over Bregman spaces has been previously addressed in [5] and as an extension to *EM* algorithms. These methods are in common use for speech and audio clustering by considering *bag of features* or *bag of frames* models and thus neglecting the importance of temporal dimensions in retrieval processes. Based on our information manifolds, we propose a simple online clustering algorithm for incremental model formation that strongly employs the temporal morphology of data over the information manifold.

Following definitions 3, we base our segmentation technique on detecting *information jumps* in an ongoing audio stream $\boldsymbol{X}_t$ for forming *models* or minimum-information Bregman balls over time. The information radius R_k defines the maximum *information gain* around a centroid $\boldsymbol{\mu}_k$ that *model* k contains through D_F^J. Employing change detection for incremental Bregman ball formation on continuous streams has the implicit assumption that the models' information gain on audio streams is a *right-continuous with left limit* function of time. This assumption is a direct consequence of our initial consideration that the signal is stationary in a finite time-frame under a model $\boldsymbol{\theta}_k$. This conforms to the intuitive nature of music information characterized by distinct events with an information onset implying a discontinuity with regards to the past.

The goal of model formation in our framework is thus to search for a proper segmentation on audio streams such that each resulting segment is quasi-stationary and homogeneous in terms of information content. The detection of a change is equivalent to accepting a hypothesis H_1 of change for time $r \leq n$ when testing against the hypothesis H_0 of no change. Algorithm 1 shows a basic online implementation of the change detection which accepts an observation sequence of length n, and initialized on $\boldsymbol{X}_0^n$ with $f = 0$.

Algorithm 1 Online *model* Segmentation/Change Detection

Require: Audio stream $\boldsymbol{X}_t$, ongoing *model* $B^k(\boldsymbol{\mu}_k, R_k)$, observation window n, first index f
Ensure: $t - f \geq n$ (minimum observation length)
1: Initialize observation vectors $\{\boldsymbol{O}_i\}_{i=0}^{t-f-1}$ to $\boldsymbol{X}_{t-f}^t$
2: Detect change point r^* in $\{\boldsymbol{O}_i\}$ with regards to B^k
3: **if** No change is detected **then**
4: set $f = t$
5: **else**
6: Initiate a new Bregman ball B^{k+1} on $\{\boldsymbol{O}_i\}_{i=r^*}^{n-1}$
7: set $f = r^*$, $k = k + 1$
8: **end if**
9: **return** next starting index f, ongoing ball $B^k(\boldsymbol{\mu}_k, R_k)$

The change detection algorithm proposed here is an adopted version of the *CuSum* algorithm [19] which has been employed in various segmentation schemes such as audio [20]. In a generic *CuSum* algorithm, the likelihood ratio of the conditional probabilities of the observations under the hypothesis H_1 and H_0 is estimated, then the maximum of the sum of the log-likelihood ratio of the sequence of observations is compared to a threshold λ to determine whether a boundary exists between two segments of the sequence. Concretely, given n observations, $c_n = \max_r \sum_{k=r}^{n} \ell_k$ where ℓ_k is the log-likelihood ratio of conditional probabilities with respect to H_0 and H_1, and compared to λ to assess the change point.

The *CuSum* algorithm assumes that the conditional probabilities of observations under both hypothesis H_1 and H_0 are known. While this is a difficulty for most applications, it does not pose any in our framework. Concretely, given an ongoing model $B^k(\boldsymbol{\mu}_k, R_k)$ and n observations, H_0 hypothesis at r assumes that $\boldsymbol{O}_0^r$ are all members of $B^k(\boldsymbol{\mu}_k, R_k)$ as explained in section IV-D, and H_1 at r assumes that $\boldsymbol{O}_r^n$ constructs a new ball. Therefore, the likelihood ratio ℓ_k becomes

$$\ell_k = D_F^J(\boldsymbol{c}^F\left[\left\{\boldsymbol{\mu}_k \cup \{\boldsymbol{O}_0^{k-1}\}\right\}\right], \boldsymbol{c}^F[\{\boldsymbol{O}_k^n\}]) \qquad (14)$$

where $\boldsymbol{c}^F$ is the symmetric Bregman centroid. The first argument of D_F^J refers to the inclusion of data points $\{\boldsymbol{O}_k^n\}$ within B^k and the second is the hypothesis of forming a new ball.

The change point threshold λ applied to ℓ_k of eq. 14 has a direct geometrical interpretation. It corresponds to the minimal discrimination information distance between two consecutive *models* on the audio stream. In other words, using the change detection algorithm above and for two consecutive models k and $k+1$, we would always have $D_F^J(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k+1}) \geq \lambda$.

Figure 1 shows the result of this segmentation on an audio excerpt corresponding to the first theme of Beethoven's 1^{st} Piano sonata performed by Friedrich Gulda. The information geometry employed for this excerpt corresponds to amplitude spectrum, assumed without lack of generality to be generated by Multinomial distributions with the bijected Bregman as Kullback-Leibler. The normalized audio waveform is superposed by ℓ_t or change point likelihoods as well as detected *model* onsets. This example consists of 583 data points (analysis frames) and leads to 44 disjoint and variable length *models*.

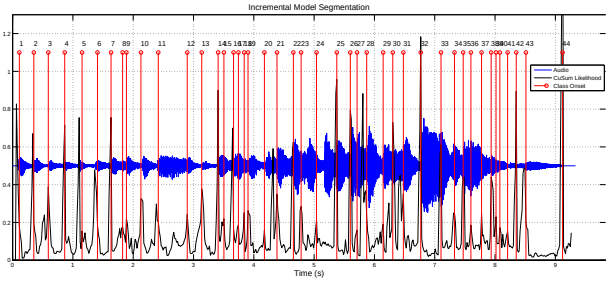


Fig. 1. Segmentation result on the first theme of Beethoven's first sonata, performed by Friedrich Gulda with $\lambda = 0.1$.

V. Sample Applications

To motivate the theoretical framework discussed above, we present two sample applications in pattern recognition and music information retrieval on audio streams. More results of these sample applications (on different types of sounds and music) can be found on our project website[1].

A. Online Audio Structure Discovery

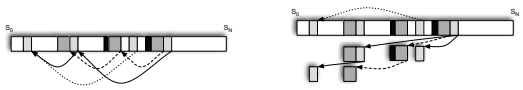
For our first sample application, we are interested in representing the repetitions and long-term regularities within an ongoing audio stream. Music structure analysis from acoustic signals has been addressed previously by various methods. A good review of existing approches can be found in [21]. We aim at obtaining an online procedure that can quickly group equivalent patterns and find the longest sequence of models in the past of the audio stream. The first problem is referred to as clustering and the second as structure discovery. We are interested in a fast method that can address both in one shot.

[1] http://imtr.ircam.fr/imtr/Music_Information_Geometry

The idea behind this application is the following: The Music Information Geometry framework provides us with information entities as minimum information Bregman balls over the time series $\boldsymbol{X}_t$, which can be compared to each other using the discussed methods as in symbolic equivalence classes but on a continuous metric and using the similarity definition on page 4. This brings out the idea of adapting common symbolic algorithms for the signal world.

Our algorithm for automatic discovery of audio structures is motivated by a technique for fast indexing of symbolic data such as text and DNA called *Factor Oracle (FO)* [22]. A time series of symbols $S = \sigma_1^n$ in a FO is learned as a state-space diagram, whose states are indexed by from 0 to n. There is always a transition called the *factor link* labelled by symbol σ_i going from state $i-1$ to state i. Navigating a FO from state 0 to n incrementally would generate the original sequence S. Depending on the structure of S, other labelled factors links as forward transitions might be created, as well as some *backward* transitions called *suffix links* with no label.

The factor and suffix links in FO have direct structural interpretations. A *factor link* going from state i to j indicates that a (variable length) history of symbols immediately before i is a common *prefix* of the sequence of symbols leading to j. A *Suffix link* from state m to an earlier state k indicates that the two states share the longest *suffix*. A suffix link goes from i to j if and only if the longest repeated suffix of s_1^i is recognized in j, connecting repeated patterns in S. The length of each repeating factor for each suffix link can be computed in linear time and denoted as $lrs(i)$ for each state i. This property of *suffix links* alone make FOs attractive on large sequences. Figure 2a shows schematically how maximum length repeated factors are interconnected by suffix links. The thickness of the chunks represents the length of the repeated factor. Following each suffix link from the head of a Factor Oracle structure to the very beginning provides a forest of disjoint tree structures whose roots are the smallest and leftmost patterns appearing in the trees, thus capturing all redundancies inside the sequence. Figure 2b shows these linked trees associated to fig. 2a.



(a) Suffix Structure Diagram (b) Expanded Suffix Trees

Fig. 2. The Suffix structure and Suffix Link forest of disjoint trees.

To extend Factor Oracles to music information geometry, *symbols* σ_i are replaced by *models* and symbolic equivalence to *similarity* as in definition 2 on bijected and symmetric D_F^J. Following figure 2, by learning audio structures we are interested in suffix links and their corresponding lengths. Figure 3 visualizes the learned structures of the Oracle on a recording of Beethoven's first piano sonata, 3^{rd} movement performed by Friedrich Gulda (recorded in 1950s). In this example, data points are constant-Q power spectrum on logarithmic musical scales as reported in [23] with an analysis window of approximately $64ms$ and an overlap factor of 2. Using these histogram features, the corresponding Bregman geometry is that of KL divergences. The information threshold for the *CuSum* algorithm is set to 0.15 and the similarity threshold ϵ for Oracle to 0.1. The three subplots show the audio waveform, the suffix structure, and the length of repeating sequence lrs associated to each state respectively. The suffix subplot is read as follows: A time t on the x-axis would send a pointer back to a time t' $(t' < t)$ indicating the longest common suffix between a factor at time t and t'. The corresponding value for t on the lrs subplot reveals the length of the detected longest sequence (as number of states) for that state. In this example, we have superposed the reference structure in terms of labelled blocks taken from explicit repetitions in the music score.

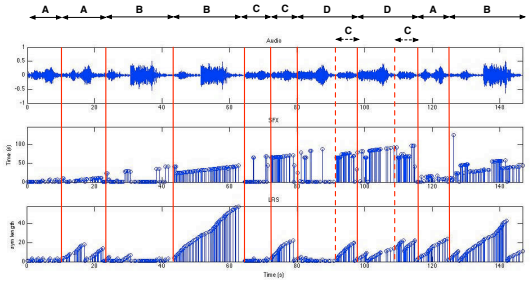


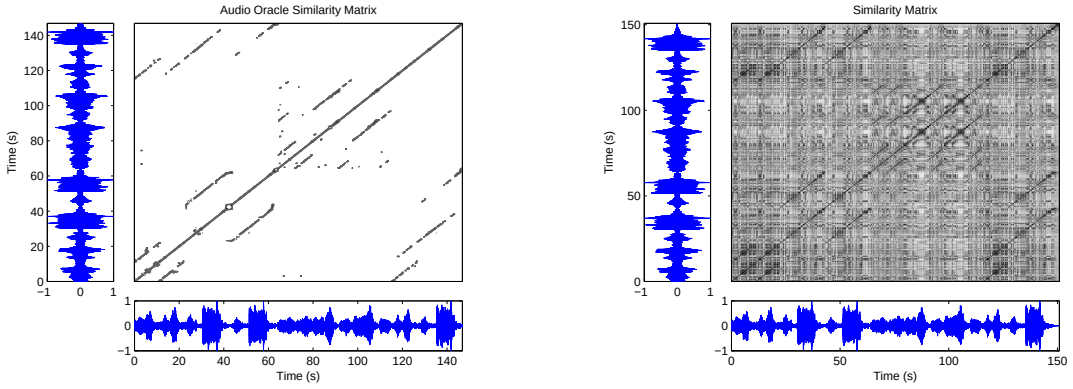
Fig. 3. Incrementally learned Oracle structure along with the segmented structure in terms of blocks from the orignal symbolic music score; Beethoven's Piano Sonata 1-movement 3, interpreted by Friedrich Gulda.

In the classical music example of figure 3, the music goes through various structural repetitions and recombinations which are mostly captured by the Oracle structure. Such repetitions in the context of a *human performance* of a piece of music are never *exact*, but nevertheless detected. This sample comprises an audio stream of 9500 analysis frames that leads to 440 learned models and states. Given this structure, we can construct a traditional similarity matrix by substituting each found suffix link by its corresponding distance or $sim(i,j) = D_F^J(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ given that a suffix link exist between i and j or zero otherwise. Figure 4a shows this similarity matrix constructed out of the Oracle in figure 3 revealing the recall and similarity structure discussed above.

To compare, figure 4b provides a classical frame-based self-similarity matrix over the same audio (and same features), in common use in the MIR literature and as proposed in [3]. Roughly, this measure is obtained by calculating the distance between all analysis frames of the entire audio against each other, using the same distortion. By segmenting the audio stream into quasi-stationary states and using symbolic equivalence rather than a distortion, figure 4a can be obtained more efficiently and online where 4b requires the entire audio. The similarity described in figure 4a also gives explicit access to equivalent entities and continuations over time through the structure of figure 3, whereas the similarity matrix of fig 4b requires further processing to deduce such relations. Furthermore, the classical similarity matrix contains exhaustive basis (9500×9500 versus 440×440 in fig 4a) for processing.

B. Similarity Queries over Information Streams

We showed in the previous section how long-term information flows can be easily captured using our information



(a) Similarity Matrix out of Oracle of figure 3

(b) Similarity Matrix based on [3]

Fig. 4. Structural Similarity matrix using structural segmentation on the music example of figure 3 using two different kernel values.

geometric framework. In this section we showcase the idea of finding the most similar stream paths on a stream database given a query. This problem is usually addressed under the topic of *audio matching*. Once again, the goal here is not to compare to all existing methods but to showcase the ease of algorithmic programming once the problem set is projected onto an information geometric framework.

Given a stream query $\boldsymbol{X}_t$ represented as a succession of models $B_X^k(\boldsymbol{\mu}, R_k)$, and a target stream in $\mathcal{T}$ as well represented by its successive Bregman balls B_T^j, the problem of finding the best match of $\boldsymbol{X}_t$ within $\mathcal{T}$ can be reduced to finding the best sequence of balls in $\mathcal{T}$ that best constructs the ball sequence in $\boldsymbol{X}_t$. The problem then can be formulated as a regular *Approximate Nearest Neighbor* search algorithm with special considerations for temporal continuity between balls. Given the compact state-space representation presented in the previous section, and given its inherent temporal and disjoint tree structures as shown in figure 2b, it would be natural to choose this data structure instead of a regular ball-tree (such as in [17]) for the search domain representation. Following an Oracle representation on the target for a sequential search has the advantage of providing results with best perceptual continuity once synthesized since they correspond to natural continuations and regularities in the original audio.

Bregman Ball sequence matching between a source (query) and target can be achieved using dynamic programming and following discussions in section IV-D. At the initialization, the program chooses the most similar balls to B_X^1 over all balls in $\mathcal{T}$, resulting into the next search domain by choosing all *factor* and *suffix* links from the found states following the Oracle of $\mathcal{T}$. This process is then repeated until either we reach the end of query B_X^T or an empty set is found during recursion. This simple dynamic programming scheme is able to trace multiple paths in a single run and provides partial matches where possible. The result is a *concatenative tree structure* on the target balls that are able to reconstruct $\boldsymbol{X}_t$ balls.

Figure 5 shows the resulting *Concatenative tree* for an audio query corresponding to Beethoven's Piano Sonata Nr.1's first musical theme (corresponding to the model sequence in Figure 1) and the entire Piano sonatas Oracle ball sequence obtained as in section V-A. This experiment was done using a similarity threshold of $\epsilon = 0.1$. Each numbered state represents a Bregman ball in the target domain $\mathcal{T}$. The tree reconstructs the query from left to right by following existing arrows. Among possible reconstruction paths, the path highlighted with gray corresponds to the *original* theme of the exact construction of the query (balls 1 to 37), which is naturally expected. Parallel to this main path, two rather continuous and alternative paths exist consisting of states 169 to 202 and 460 to 511. These paths correspond to the repetition of the main musical theme in the middle and the end of the Piano Sonata which is a main characteristic of Sonata form in classical music. Other sub-paths also correspond to reappearance of the main theme in one form or another during the *development* section of the sonata form. The explosion of states around time 11 and towards the end correspond to a specific *model* that is representing a *cadential chord* which re-appears in various places throughout the piece as an important stylistic element.

A given path of the *concatenative tree result* can be easily re-synthesized to audio by concatenating corresponding audio frames of data points within each *model* (ball) using classical concatenative synthesis techniques [24]. For more audio results we encourage the reader to check our project website.

VI. Conclusion

We proposed a preliminary framework for representation of temporal dynamics of audio streams on an information manifold. The construction of our information manifold approaches a similarity metric space where similarity is defined as control over the rate of change of information content between continuous data streams, and requires modeling the data streams as points on an information manifold generated by a family of exponential distributions. The music information geometry framework presented of section IV, provides an alternative representation of data points by incrementally gathering quasi-stationary data points within Bregman balls that represent self-contained *models* in terms of statistical information. The music information geometry in theory provides the following facilities: (1) representation of audio entities as well-behaved geometric objects with intuitive geometric properties, (2) sim-

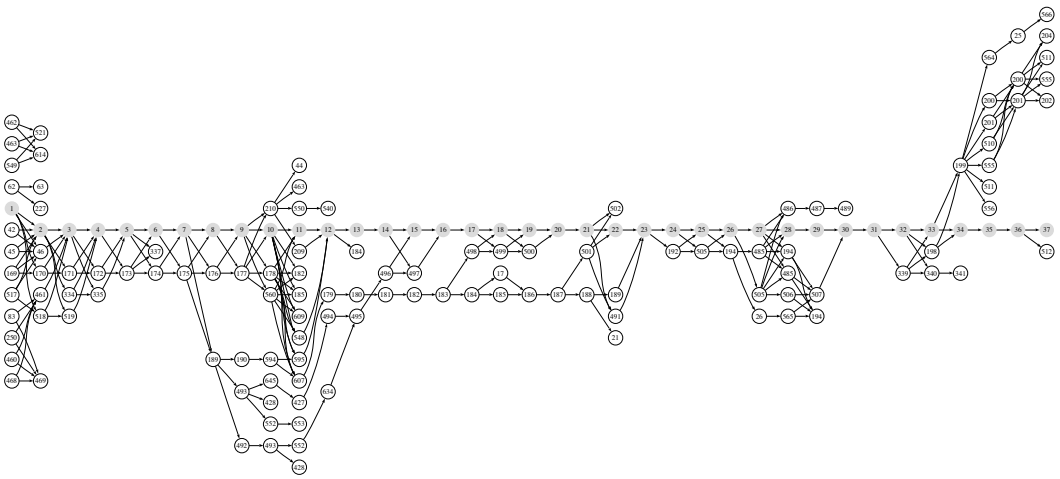


Fig. 5. Audio Matching *concatenative tree result* on Beethoven's Piano Sonata Nr.1-Mvt.1 (target) and the first musical theme as query – Showing possible audio reconstruction paths and *best path* highlighted in light-gray.

plifies optimization problems thanks to duality, (3) provides an approximate *similarity metric space*, bridging the gap between continuous and symbolic aspects of audio streams, (4) fast, sparse and incremental treatments suitable for data stream analysis, and (5) provides a generic mathematical framework extensible to more intricate models and applications.

The major intent of this paper was to lay the theoretical groundwork for forthcoming experimental results and hopefully, for other researchers interested in exploring the new possibilities offered by methods of information geometry on audio streams. We however showcased two common and complex MIR applications using the proposed framework. The promising sample results demonstrate the intuitive manner by which complex problems can be addressed within the proposed information geometry framework, and the facility to access information entities in our framework similar to symbolic processing. Further aspects and applications of music information geometry will be reported in future publications.

The study of audio streams as information geometries provide a new and challenging way to address complex problems with rather simple solutions. Besides its theoretical merit, we believe that its solutions brings out new horizons to the applications of multimedia information retrieval.

REFERENCES

[1] S. Dubnov, S. McAdams, and R. Reynolds, "Structural and affective aspects of music from statistical audio signal analysis," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1526–1536, 2006.

[2] S. Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 16, no. 2, pp. 327–337, 2008.

[3] J. Foote and M. Cooper, "Media segmentation using selfsimilarity decomposition," in *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, vol. 5021, 2003, pp. 167–175.

[4] S. Amari and H. Nagaoka, *Methods of information geometry*. Oxford University Press, 2000, vol. 191.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[6] F. Nielsen, J.-D. Boissonnat, and R. Nock, "On bregman voronoi diagrams," in *Proc. 18th ACM-SIAM Sympos. Discrete Algorithms*, 2007.

[7] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Comput.*, vol. 16, no. 1, pp. 159–195, 2004.

[8] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, pp. 200–217, 1967.

[9] F. Nielsen and R. Nock, "Sided and symmetrized bregman centroids," *IEEE Trans. Inf. Theor.*, vol. 55, no. 6, pp. 2882–2904, 2009.

[10] R. Gray, A. Buzo, J. Gray, A., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 4, pp. 367–376, Aug 1980.

[11] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[12] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2001, pp. 837–840.

[13] B. A. Carlson and M. A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 12, pp. 1255–1260, 1991.

[14] S. Dubnov, "Generalization of spectral flatness measure for non-gaussian linear processes," *IEEE Sig. Proc. Letters*, vol. 11, no. 8, pp. 698–701, Aug. 2004.

[15] A. Cont, "Modeling musical anticipation: From the time of music to the music of time," Ph.D. dissertation, University of Paris 6 and University of California in San Diego, October 2008.

[16] R. Veldhuis and E. Klabbers, "On the computation of the kullback-leibler measure for spectral distances," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 1, pp. 100–103, Jan 2003.

[17] L. Cayton, "Fast nearest neighbor retrieval for bregman divergences," in *ICML '08: Proceedings of the 25th international conference on Machine learning*. New York, NY, USA: ACM, 2008, pp. 112–119.

[18] F. Nielsen, P. Piro, and M. Barlaud, "Tailored Bregman Ball Trees for Effective Nearest Neighbors," in *Proceedings of the 25th European Workshop on Computational Geometry*, Brussels, 2009, pp. 29–32.

[19] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: theory and application*. NJ, USA: Prentice-Hall, Inc., 1993.

[20] M. K. Omar and U. Chaudhari, "Blind change detection for audio segmentation," in *In Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'05*, 2005, pp. 501–504.

[21] R. Dannenberg and M. Goto, *Handbook of Signal Processing in Acoustics*. Springer Verlag, 2009, vol. 1, ch. Music Structure Analysis from Acoustic Signals, pp. 305–331.

[22] C. Allauzen, M. Crochemore, and M. Raffinot, "Factor oracle: A new structure for pattern matching," in *Conference on Current Trends in Theory and Practice of Informatics*, 1999, pp. 295–310.

[23] H. Purwins, B. Blankertz, and K. Obermayer, "Constant q profiles for tracking modulations in audio data," in *Proceedings of the International Computer Music Conference*, Cuba, 2001, pp. 407–410.

[24] D. Schwarz, "Corpus-based concatenative synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 92–104, March 2007.