An information-geometric approach to real-time audio segmentation

Arnaud Dessein, Arshia Cont

An Information-Geometric Approach to Real-Time Audio Segmentation

Arnaud Dessein, and Arshia Cont

***Abstract*—We present a generic approach to real-time audio segmentation in the framework of information geometry for exponential families. The proposed system detects changes by monitoring the information rate of the signals as they arrive in time. We also address shortcomings of traditional cumulative sum approaches to change detection, which assume known parameters before change. This is done by considering exact generalized likelihood ratio test statistics, with a complete estimation of the unknown parameters in the respective hypotheses. We derive an efficient sequential scheme to compute these statistics through convex duality. We finally provide results for speech segmentation in speakers, and polyphonic music segmentation in note slices.**

***Index Terms*—Audio segmentation, real-time system, information geometry, change detection.**

I. Introduction

In this paper, we propose a generic framework for real-time segmentation on audio streams, by using methods of information geometry for exponential families. The system developed controls the information rate of the signals as they arrive in time to perform the segmentation. It also addresses shortcomings of traditional CUSUM approaches to change detection in estimating the unknown parameters before change.

A. Background

The problem of audio segmentation, also called novelty or change detection, has been widely studied, mainly for music and speech signals [1], [2]. This problem can be defined as finding time boundaries, called change points, which partition a sound signal into homogeneous and continuous temporal regions, called segments, that are inhomogeneous with the adjacent regions. This requires defining a criterion to quantify the homogeneity, and various criteria may be employed. For instance, we may want to segment a conversation in speakers, or a radio broadcast in talked parts, music and advertisement.

In many works, audio segmentation relies on automatic classification to create the segments in function of the detected classes. Such an approach has yet the drawbacks to assume the existence and knowledge of classes, to rely on a potentially fallible classification, and to require some training data. We would like to tackle this by segmenting an audio stream without any assumption on the existence of classes. Segmentation is thus distinct from classification for us.

The authors are with the MuTant Project-Team (INRIA) hosted by the Music Representations Team, UMR 9912 STMS (IRCAM, CNRS, UPMC).
Address: IRCAM, 1 place Igor Stravinsky, 75004 Paris, France.
E-mails: arnaud.dessein@ircam.fr, arshia.cont@ircam.fr.

This work was supported by a doctoral fellowship from UPMC (EDITE).

A similar approach has been considered for speaker segmentation [3]. In general, the audio frames are represented by using a cepstrum, and change points are detected by computing either a distance between successive frames, or statistics on the hypothesis of a change at different frames under normality assumptions. Other approaches without classification have also been proposed for onset detection in music signals [4]. Various signal features, such as the energy envelope or the spectrum, can be considered to build a detection function which is then used to localize the onsets by thresholding and peak-picking heuristics. In particular, distance-based and statistical methods have been used to define the detection function.

More general audio segmentation frameworks employ kernel methods to compute the distance-based segmentation in a high-dimensional feature space [5], [6], and information-theoretic methods based on entropy [7], or on test statistics such as the CUSUM algorithm [8], [9]. As discussed later, the traditional CUSUM approaches undergo approximations for parameter estimation, resulting in practical shortcomings when change points may occur rapidly such as in audio signals.

B. Motivations

A major theoretical issue of statistical approaches to sequential change detection, is to consider known parameters before change [10]. This is suitable for applications such as quality control where a normal regime is known, but limited for many real-world applications such as audio signal processing. The problem when considering unknown parameters, is that it breaks down the computational efficiency of CUSUM. Therefore, some approximations of the exact test statistics are in general made to accommodate these situations, such as learning the distribution before change on the whole window, or in a dead region where change detection is turned off.

A few specific exact statistics have yet been studied, notably for unknown mean before and after change in univariate normals with a fixed known variance [11]. Nonetheless, normality assumptions do not always model reliably the considered signals and homogeneity criteria. A more general Bayesian framework for independent observations in exponential families has been proposed to address the estimation of parameters before and after change [12]. This framework, however, relies on a geometric prior on the time between change points, which is not always appropriate to arbitrary signals. Moreover, it requires a prior knowledge on the distributions of the parameters in the respective segments, which is not always available. To overcome this, we seek to formulate a generic sequential change detection with unknown parameters before and after change, but without any a priori on the respective distributions of the change points and parameters.

C. Contributions

We formulate the problem of real-time audio segmentation in the framework of information geometry. We design a real-time modular system that can handle various types of signals and of homogeneity criteria. In general terms, the system detects changes by controlling the variation of information contained in the data stream. The information content is quantified by employing information-theoretic measures on statistical descriptions of the signal. As a by-product, the quantified units can be characterized with representative probabilistic models that are suitable for further processing.

We notably consider densities from exponential families. Information geometry provides a rich theoretical framework to study their properties from both geometrical and statistical perspectives. In this framework, we devise a computationally efficient scheme for change detection, which finds statistical grounds in the use of sequential generalized likelihood ratio tests, and geometrical interpretations in the dually flat geometry. It thus provides a unifying view of change detection for many statistical models and corresponding distance functions.

Last but not least, we address the problem of CUSUM schemes for parameter estimation, by using exact generalized likelihood ratio test statistics, where the unknown parameters are estimated separately in each hypothesis. This breaks the inherent simplicity of CUSUM algorithms, but we are nonetheless able to obtain an efficient scheme with sequential computation of the test statistics by using the convex duality of exponential families. This scheme is applied to different segmentation tasks in speech and music signals for validation and demonstration of the generality of the approach.

II. Proposed framework

We first describe the general architecture of the proposed segmentation system. We then present the framework of exponential families and dually flat information geometry. We finally formulate the proposed approach to change detection.

A. General architecture of the system

We consider an audio stream that arrives incrementally to the system as successive time frames. These frames are represented with a short-time sound feature to provide a time series of observations $x_1, x_2, \ldots$, and these observations are modeled with probability distributions from a given family. The segmentation paradigm then tries to detect when a distributional change occurs. As a by-product, each segment can be characterized by a statistical prototype corresponding to a description of the distribution in that segment. The segmentation scheme can be summed up as follows.

We start with an empty window $\vec{x} \leftarrow (\,)$. For each time increment n, we accumulate the incoming observation in the growing window $\vec{x} \leftarrow \vec{x}\,|\,x_n$, and attempt to detect a change at any time i of the window. When a change point is detected, we discard the observations before the change point and start again the procedure with an initial window $\vec{x} \leftarrow (x_{i+1}, \ldots, x_n)$. The sequential change detection problem can thus be reduced to finding one change point in a given window $\overline{x} = (x_1, \ldots, x_n)$.

We notice, however, that this reduction is disputable. Indeed, it supposes that if no change point has been detected yet, then adding one extra observation may introduce only one change point. This is a reasonable assumption in general, but it does not account for the possibility that a change point has been missed. It may occur when the distributions before and after change are very similar and not enough observations are available to discriminate between them, or when a small drift in the distribution occurs. Nevertheless, we focus on the widespread framework of abrupt change detection where considerations on smooth changes such as drifts are left aside.

Concerning the sound features (e.g., MFCCs, DFTs) and statistical families (e.g., normal, multinomial), their choice is left to the user depending on the types of signals and criteria for homogeneity considered. The framework of information geometry allows a modular change detection paradigm independent of this choice for a wide range of common families of distributions, namely exponential families.

B. Information geometry of exponential families

A minimal regular exponential family is a parametric statistical model whose densities can be expressed as follows:

$$p_\theta(x) = \exp\left(\theta^\top T(x) - F(\theta) + C(x)\right) \quad , \tag{1}$$

where the natural parameters θ belong to a convex open subset Θ of $\mathbb{R}^d$, the log-normalizer $F\colon \Theta \to \mathbb{R}$ is smooth and strictly convex, the carrier measure $C\colon \mathcal{X} \to \mathbb{R}$ and the sufficient statistic $T\colon \mathcal{X} \to \mathbb{R}^d$ are measurable [13], [14]. These families encompass most distributions commonly employed in statistical learning (e.g., Bernoulli, Dirichlet, Gaussian, Laplace, Pareto, Poisson, Rayleigh, Von Mises-Fisher, Weibull, Wishart, log-normal, exponential, beta, gamma, geometric, binomial, negative binomial, categorical, multinomial). Moreover, the class of exponential families is stable under various statistical constructs such as truncated and censored models, marginals, conditionals through linear projections, joint distributions of independent variables and in particular i.i.d. samples.

Exponential families can be reparametrized by the expectation parameters $\eta(\theta) = \nabla F(\theta)$, where the gradient ∇F is actually one-to-one. Considering the framework of convex duality, F is of Legendre type with Legendre-Fenchel conjugate $F^\star$. The conjugate $F^\star$ is also of Legendre type and we have $\nabla F^\star = (\nabla F)^{-1}$, so that $\theta(\eta) = \nabla F^\star(\eta)$. Under mild assumptions, the maximum likelihood estimate $\widehat{\theta}$ of k i.i.d. samples $x_1, \ldots, x_k$, possesses a closed-form expression as the average of the sufficient statistics in expectation parameters:

$$\widehat{\eta} = \frac{1}{k} \sum_{j=1}^{k} T(x_k) \ . \tag{2}$$

This relation actually holds iff the average lies in the expectation parameter space, which happens with probability increasing up to one as the sample size grows to infinity.

These notions can be studied within the framework of information geometry [15]. In particular, an exponential family endowed with the well-known Fisher information metric g is a Riemannian manifold, and can be enhanced with a family of dual affine α-connections $\nabla^{(\alpha)}$. In addition, this

statistical manifold is a dually flat space for the dually flat affine connections $\nabla^{(1)}$ and $\nabla^{(-1)}$, where θ and η form dual affine coordinate systems. The dually flat geometry generalizes the self-dual Euclidean geometry, with two dual Bregman divergences B_F and $B_{F^\star}$ instead of the self-dual Euclidean distance, where the Bregman divergence B_G generated by a smooth strictly convex function $G\colon \Xi \to \mathbb{R}$ on a convex open set Ξ is defined as follows:

$$B_G(\xi\|\xi') = G(\xi) - G(\xi') - (\xi - \xi')^\top \nabla G(\xi') \ . \quad (3)$$

Finally, the two dual Bregman divergences on the parameters are linked with the Kullback-Leibler divergence on the corresponding densities through the following relation:

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta'}) = B_F(\theta'\|\theta) = B_{F^\star}(\eta(\theta)\|\eta(\theta')) \ . \quad (4)$$

C. Formulation of change detection

We now formulate an information-geometric framework for change detection. As discussed previously, we seek to detect one change point in a given window $\overline{x} = (x_1, \ldots, x_n)$. We assume that all samples x_j are drawn independently according to distributions from a given statistical model $\{p_\xi \colon \xi \in \Xi\}$.

The usual approach under these assumptions is first to suppose that the parameters ξ_{bef}, ξ_{aft}, before and after change are known, and to test statistical hypotheses H_0 of no change and H_1^i of a change at time i. To choose between these hypotheses, a CUSUM test can be employed by computing the likelihood ratio Λ^i at time i, expressed as follows:

$$\frac{1}{2}\,\Lambda^i = \log \frac{\prod_{j=i+1}^n p_{\xi_{\mathrm{aft}}}(x_j)}{\prod_{j=i+1}^n p_{\xi_{\mathrm{bef}}}(x_j)} = \sum_{j=i+1}^{n} \log \frac{p_{\xi_{\mathrm{aft}}}(x_j)}{p_{\xi_{\mathrm{bef}}}(x_j)} \ . \quad (5)$$

The maximum of Λ^i across potential change points i is compared to a chosen threshold λ, and a change is detected at the corresponding time if the maximum exceeds this threshold.

When the parameter after change is unknown, a CUSUM test can still be employed by computing the generalized likelihood ratio, where we replace ξ_{aft} with its maximum likelihood estimate $\widehat{\xi}_1^i$ in H_1^i. However, when the parameter before change is unknown, the test cannot be written in its cumulative sum form anymore when also replacing ξ_{bef} with the maximum likelihood estimates $\widehat{\xi}_0$ and $\widehat{\xi}_0^i$, in H_0 and H_1^i. Hence, some approximations have been proposed to keep the simplicity and tractability of the statistics, by assuming the equality of the estimates before change $\widehat{\xi}_0 = \widehat{\xi}_0^i$ in all hypotheses, and by computing the estimation either on the whole window or in a dead region at the beginning of the window where change detection is turned off. We argue here that we can still construct computationally efficient test statistics for any exponential family, without using such approximations.

We now suppose that the statistical model is an exponential family, and that both the parameters before and after change are unknown. We thus consider the following hypotheses:

$$H_0 : \ x_1, \ldots, x_n \sim p_{\theta_0} \, ; \quad (6)$$

$$H_1^i : \ x_1, \ldots, x_i \sim p_{\theta_0^i} \, , \quad x_{i+1}, \ldots, x_n \sim p_{\theta_1^i} \ . \quad (7)$$

Considering the maximum likelihood estimates $\widehat{\theta}_0$, $\widehat{\theta}_0^i$, $\widehat{\theta}_1^i$, the generalized likelihood ratio is expressed as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i = \log \frac{\prod_{j=1}^i p_{\widehat{\theta}_0^i}(x_j) \prod_{j=i+1}^n p_{\widehat{\theta}_1^i}(x_j)}{\prod_{j=1}^i p_{\widehat{\theta}_0}(x_j) \prod_{j=i+1}^n p_{\widehat{\theta}_0}(x_j)} \ . \quad (8)$$

Replacing the densities with their exponential canonical form, the carrier measures and the logarithms simplify, leading to:

$$\begin{aligned}\frac{1}{2}\,\widehat{\Lambda}^i = &\sum_{j=1}^{i} \Big((\widehat{\theta}_0^i - \widehat{\theta}_0)^\top T(x_j) - F(\widehat{\theta}_0^i) + F(\widehat{\theta}_0) \Big) \\ &+ \sum_{j=i+1}^{n} \Big((\widehat{\theta}_1^i - \widehat{\theta}_0)^\top T(x_j) - F(\widehat{\theta}_1^i) + F(\widehat{\theta}_0) \Big) \ . \quad (9)\end{aligned}$$

Introducing the maximum likelihood estimates and using duality between natural and expectation parameters, we have:

$$\begin{aligned}\frac{1}{2}\,\widehat{\Lambda}^i = &\, i \Big(F(\widehat{\theta}_0) - F(\widehat{\theta}_0^i) - (\widehat{\theta}_0 - \widehat{\theta}_0^i)^\top \nabla F(\widehat{\theta}_0^i) \Big) \\ &+ (n-i) \Big(F(\widehat{\theta}_0) - F(\widehat{\theta}_1^i) - (\widehat{\theta}_0 - \widehat{\theta}_1^i)^\top \nabla F(\widehat{\theta}_1^i) \Big) \ . \quad (10)\end{aligned}$$

This makes two Bregman divergences B_F on the natural parameters appear, which can be rewritten with Kullback-Leibler divergences D_{KL} on the densities as follows:

$$\frac{1}{2}\,\widehat{\Lambda}^i = i D_{\mathrm{KL}}\Big(p_{\widehat{\theta}_0^i} \Big\| p_{\widehat{\theta}_0} \Big) + (n-i) D_{\mathrm{KL}}\Big(p_{\widehat{\theta}_1^i} \Big\| p_{\widehat{\theta}_0} \Big) \ . \quad (11)$$

This statistical test can geometrically be interpreted as computing the weighted divergences between the maximum likelihood estimates before change, resp. after change, and the maximum likelihood estimate with no change. Moreover, the test is invariant under reparametrizations, choice of a dominating measure, and sufficiency. Therefore, the test is intrinsic to the underlying information-geometric manifold.

From a computational viewpoint, the statistics can be calculated sequentially quite inexpensively through convex duality. Indeed, rewriting the statistics with dual Bregman divergences $B_{F^\star}$ on the expectation parameters and developing, we have:

$$\begin{aligned}\frac{1}{2}\,\widehat{\Lambda}^i = &\, i \Big(F^\star(\widehat{\eta}_0^i) - F^\star(\widehat{\eta}_0) - (\widehat{\eta}_0^i - \widehat{\eta}_0)^\top \nabla F^\star(\widehat{\eta}_0) \Big) \\ &+ (n-i) \Big(F^\star(\widehat{\eta}_1^i) - F^\star(\widehat{\eta}_0) - (\widehat{\eta}_1^i - \widehat{\eta}_0)^\top \nabla F^\star(\widehat{\eta}_0) \Big) \ . \quad (12)\end{aligned}$$

Because of the barycentric relation $n\widehat{\eta}_0 = i\widehat{\eta}_0^i + (n-i)\widehat{\eta}_1^i$, between the maximum likelihood estimates, we finally obtain:

$$\frac{1}{2}\,\widehat{\Lambda}^i = i F^\star(\widehat{\eta}_0^i) + (n-i) F^\star(\widehat{\eta}_1^i) - n F^\star(\widehat{\eta}_0) \ . \quad (13)$$

Since the maximum likelihood estimates between successive windows are related by simple time shifts or barycentric updates in expectation parameters, it provides a computationally efficient scheme for calculating the statistics sequentially.

III. Experimental results

The generic GLR segmentation scheme presented above is capable of controlling information rate changes in real time given that the audio representation is modeled with a member of the ubiquitous exponential families. Below, we showcase this scheme on two famous problems discussed in the literature, by adapting the proposed framework to the signals and homogeneity criteria considered at hand.

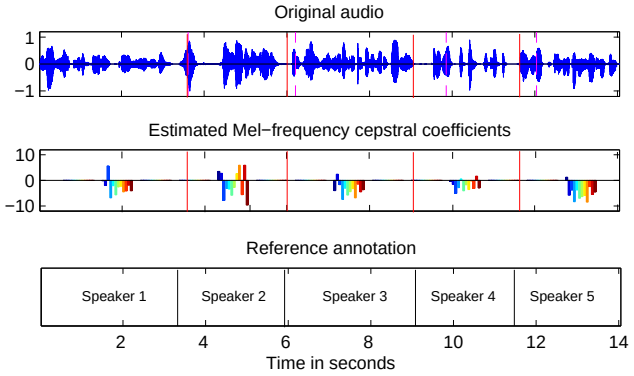

Fig. 1. Segmentation of a speech fragment based on different speakers.

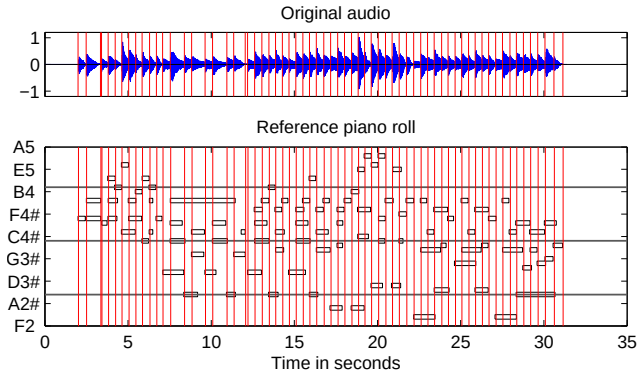

Fig. 2. Segmentation of polyphonic music based on different note slices.

A. Speech segmentation in speakers

In a first experiment, we applied the system to speech segmentation into speakers. Figure 1 shows the results on a speech fragment constructed by concatenating utterances from 5 different speakers. We considered a timbral homogeneity criterion through cepstra that provide information on the spectral envelope of the source. We computed 12 MFCCs with half-overlapping windows of 512 samples at a sampling rate of 11025 Hz, and modeled them with spherical normal distributions. We used a threshold $\lambda = 100$ set empirically on the example, and the algorithm ran about 10 times faster than real time. We compared the results with a CUSUM scheme.

The top plot shows the detected segments on the waveform, with solid lines for GLR and dashed lines for CUSUM, while the bottom plot shows the ground-truth segmentation. This proves that both GLR and CUSUM have detected the speaker turns given the usual tolerance of 1 s in this context, although GLR is more precise than CUSUM in the estimation of the changes. The middle plot depicts the estimated MFCCs in the different segments detected with GLR. These prototypes can be used in further applications such as speaker recognition.

B. Polyphonic music segmentation in note slices

In a second experiment, we considered polyphonic music with the goal of slicing the audio into stationary polyphonic chunks. Figure 2 shows the results on a piano excerpt. We used a spectral homogeneity criterion through magnitude spectra for information on the frequency content. We computed 257-bin DFTs with the same analysis parameters as above, modeled as frequency histograms with multinomial distributions. We used a threshold $\lambda = 10$ set empirically on the example, and the algorithm ran about 30 times faster than real time.

The top plot shows the detected segments on the waveform, while the bottom plot explores their relevancy compared to the hand-labeled reference pitches presented as a piano roll. This is to confirm that the system has successfully detected change points that actually correspond to note onsets and offsets.

We finally assessed quantitative results on a well-known dataset with standard evaluation methods [16]. The results show that GLR has slightly outperformed CUSUM, with respective F-measures of 66.3 % and 65.3 %, confirming that both approaches have performed relatively well, and that GLR has improved the estimation of changes compared to CUSUM.

IV. Conclusion

We formulated an information-geometric framework for real-time audio segmentation. The proposed scheme is generic for all exponential families, and addresses shortcomings of CUSUM in estimating the unknown parameters. We showed the relevancy of the approach on speech and music signals.

In future work, we want to perform thorough evaluations of the system on various segmentation tasks. We also think the approach may be useful in other domains of signal processing.

References

[1] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation," in *WASPAA*, 1999, pp. 103–106.
[2] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *ICME*, 2000, pp. 452–455.
[3] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.
[4] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Audio, Speech, Language Process.*, vol. 13, no. 5, pp. 1035–1047, 2005.
[5] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2961–2974, 2005.
[6] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe, "A regularized kernel-based approach to unsupervised audio segmentation," in *ICASSP*, 2009, pp. 1665–1668.
[7] M. Liuni, A. Robel, M. Romito, and X. Rodet, "Rényi information measures for spectral change detection," in *ICASSP*, 2011, pp. 3824–3827.
[8] M. K. Omar, U. Chaudhari, and G. Ramaswamy, "Blind change detection for audio segmentation," in *ICASSP*, vol. 1, 2005, pp. 501–504.
[9] A. Cont, S. Dubnov, and G. Assayag, "On the information geometry of audio streams with applications to similarity computing," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 837–846, 2011.
[10] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., 1993.
[11] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *The Annals of Statistics*, vol. 23, no. 1, pp. 255–271, 1995.
[12] T. L. Lai and H. Xing, "Sequential change-point detection when the pre- and post-change parameters are unknown," *Sequential Analysis: Design Methods and Applications*, vol. 29, no. 2, pp. 162–175, 2010.
[13] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, ser. Probability and Mathematical Statistics. Wiley, 1978.
[14] L. D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, ser. Lecture Notes-Monograph. Institute of Mathematical Statistics, 1986, vol. 9.
[15] S.-i. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical Monographs. American Mathematical Society, 2000, vol. 191.
[16] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *ISMIR*, 2004, pp. 72–75.