Helicality: An Isomap-based Measure of Octave Equivalence in Audio Data

Sripathi Sridhar
New York University
New York, NY, USA

Vincent Lostanlen
Cornell Lab of Ornithology
Ithaca, NY, USA

1. INTRODUCTION

Octave equivalence refers to the perceived consonance of any pair of tones whose frequency ratio is a power of two. Despite the wealth of evidence on the practical effectiveness of octave equivalence in MIR, the question of discovering octave equivalence directly from data has received less attention. One exception is [1], which applied multidimensional scaling (MDS) to visualize mutual information between time–frequency atoms learned by independent component analysis (ICA). More recently, [2] have applied the Isomap manifold learning algorithm [3] to visualize Pearson correlations between CQT activations.

While both methods [1] and [2] are unsupervised, the latter operates on atoms whose center frequency is defined a priori by the CQT, thus making their pitch assignment readily available. Yet, an important shortcoming of these methods is that they are purely illustrative: although they provide a scatter plot of time–frequency atoms in 3-D, the subsequent task of recognizing a helix in this scatter plot is left to visual inspection, which hampers their scalability.

In this paper, we present a new algorithm in computational geometry whose goal is to fit a helix to a cloud of points as represented in cylindrical coordinates. Specifically, we combine the Quickhull algorithm for convex hull estimation and the Frank-Wolfe algorithm for constrained minimization of total least squares.

2. METHODS

2.1 Isomap embedding of CQT log-magnitudes

Given an unlabeled audio dataset of N audio files, we use librosa v0.8.0 to compute a CQT representation of every file with $Q = 24$ bins per octave. We restrict the dataset in the time domain to the loudest CQT frame in each audio file; and in the frequency domain, to the $J = 3$ octaves of greatest variance. This results in a matrix $\mathbf{X}$ with $P = 24 \times 3 = 72$ rows and N columns.

We extract squared Pearson correlations $\boldsymbol{\rho}^2[u, v]$ across all pairs of features, and apply the following formula:

$$\mathbf{D}_{\boldsymbol{\rho}^2}[p, q] = \sqrt{-\frac{1}{2} \log \boldsymbol{\rho}^2[p, q]} \tag{1}$$

to convert them into pseudo-Euclidean distances. Following the methodology of Isomap, we use $\mathbf{D}_{\boldsymbol{\rho}^2}$ to compute a nearest-neighbor graph with $k = 3$ neighbors per vertex. With scikit-learn v0.20.0 , we apply classical multidimensional scaling (MDS) to build an embedding space in which Euclidean distances approximate geodesic distances on the graph. We refer to [2] for further details.

2.2 Circle fitting with Frank-Wolfe algorithm

Let $\mathbf{e}_m$ and λ_m be the eigenvectors and eigenvalues resulting from MDS. We rank eigenvalues in decreasing order without loss of generality. We represent every subband p by a point $\boldsymbol{y}[p] = (\mathbf{e}_1[p], \mathbf{e}_2[p])$ on the plane. Let χ denote chroma. Then, we compute chroma centroids $\widetilde{\boldsymbol{y}}[\chi] = \frac{1}{J} \sum_{j=0}^{J-1} \boldsymbol{y}[\chi + Qj]$. Our postulate is that if $\mathbf{X}$ has a property of octave equivalence, then the set of points $\mathcal{Y} = \{\widetilde{\boldsymbol{y}}[1] \dots \widetilde{\boldsymbol{y}}[\chi]\}$ should form a circle.

We apply the Quickhull algorithm [4] to extract the convex hull of $\mathcal{Y}$, denoted by $\mathcal{H}$. We denote by $\boldsymbol{c}_0$ the barycenter of vertices in $\mathcal{H}$. Then, we fit a circle to $\mathcal{Y}$ by seeking a point c inside $\mathcal{H}$ which minimizes the following objective:

$$V_{\mathrm{circle}}(\boldsymbol{c}) = \sum_{\chi=1}^{Q} \left\| \boldsymbol{c} - \widetilde{\boldsymbol{y}}[\chi] \right\|_2^2 - \frac{1}{Q} \left(\sum_{\chi=1}^{Q} \left\| \boldsymbol{c} - \widetilde{\boldsymbol{y}}[\chi] \right\|_2 \right)^2, \tag{2}$$

taken from [5, Equation 4]. In practice, we solve the problem above via a custom implementation of the Frank-Wolfe conditional gradient algorithm [6], initialized at $\boldsymbol{c}_0$.

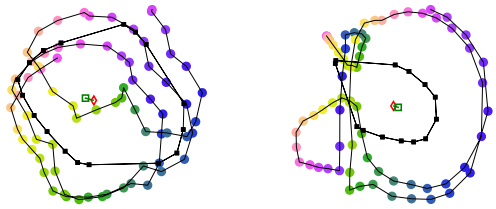
Likewise, we seek two parameters a and b such that the affine function $p \mapsto (a \times p + b)$ approximates the sequence $\boldsymbol{z} = \mathbf{e}_3$ by minimizing the following objective:

$$V_{\mathrm{line}}(a, b) = \sum_{p=1}^{P} \left\| a \times p + b - \boldsymbol{z}[p] \right\|_2^2 \tag{3}$$

We solve the problem above by linear regression.

2.3 Helicality as projection Euclidean distance

On the point cloud $\boldsymbol{\psi}[p] = (\boldsymbol{e}_1[p], \boldsymbol{e}_2[p], \boldsymbol{e}_3[p])$, we fit a helix based on the circle and line estimates, denoted by $\boldsymbol{\psi}'[p]$. Then, we define helicality as the inverse of the mean squared Euclidean distance between the embedding points



(a) Music. Helicality: 0.54. (b) Speech. Helicality: 0.30.

Figure 1. Isomap embedding of music and speech data. The hue of the colored dots and the grey line denote pitch chroma and pitch height. The black squares and the solid black line represent the convex hull. The red diamond and green square denote the inital circle center estimate and the final center after gradient descent.

and the projected points in Equation 4. This measures the deviation of the embedding from an ideal helix, denoting the extent of octave equivalence in the frequency content of the audio dataset.

$$\boldsymbol{H} = \frac{1}{\frac{1}{P}\sum_{p=1}^{P} \left\| \boldsymbol{\psi}[p] - \boldsymbol{\psi}'[p] \right\|_2^2}, \tag{4}$$

3. RESULTS

3.1 Dataset

We analyze three datasets: TinySOL (music) [7], which contains 2913 recordings from 14 instruments; ENST-drums (drums), from which we select the subset of 107 isolated drum hits from 3 drummers; and North Texas Vowel Dataset (speech), containing 3190 recordings of voiced vowels by 50 American English speakers.

3.2 Cross-dataset comparison

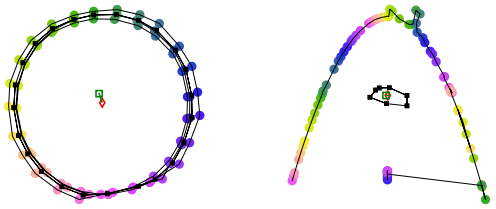
On visual inspection, music data has a more helical embedding topology than speech data in Fig. 1. Accordingly, speech data has a lower helicality score H = 0.30, while music data has a higher score H = 0.54.

3.3 Cross-instrument comparison

Horn (H = 0.94) produces the most helical embedding topology, seen in Fig. 2, followed by *Accordion* (H = 0.58). Surprisingly, *Trumpet* (H = 0.34) has a low helicality score, despite its characteristic harmonic structure. We refer to Fig. 3 for a comparison of different instrument classes. Isolated drum hits receive a helicality score of H = 0.28.

4. CONCLUSION

We have introduced an algorithm to quantify the "helicality" between frequency subbands in a given audio dataset. Further research is needed to examine whether helicality matches human perception.



(a) Horn only. Helicality: 0.94. (b) Isolated drum hits only. Helicality: 0.28.

Figure 2. Isomap embedding of horn and isolated drum hits.

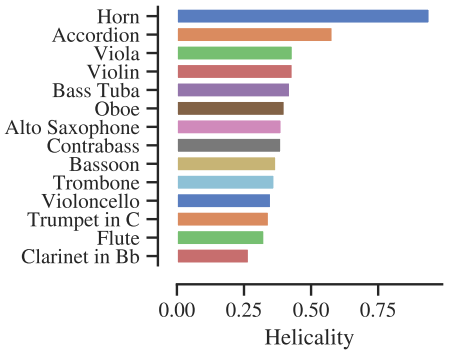


Figure 3. Helicality scores of TinySOL instrument classes

5. REFERENCES

[1] S. A. Abdallah and M. D. Plumbley, "Geometric ICA Using Nonlinear Correlation and MDS," in *Proc. ICA/BSS*, 2003.

[2] V. Lostanlen, S. Sridhar, B. McFee, A. Farnsworth, and J. P. Bello, "Learning the helix topology of musical pitch," in *Proc. IEEE ICASSP*, 2020.

[3] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[4] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *TOMS*, vol. 22, no. 4, pp. 469–483, 1996.

[5] I. D. Coope, "Circle fitting by linear and nonlinear least squares," *Journal of Optimization Theory and Applications*, vol. 76, no. 2, pp. 381–388, 1993.

[6] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. ICML*, 2013.

[7] C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "OrchideaSOL: a dataset of extended instrumental techniques for computer-aided orchestration," in *Proc. ICMC*, 2020.