# 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

## ABSTRACT

Octave equivalence serves as domain-knowledge in MIR systems, including chromagram, spiral convolutional networks, and harmonic CQT. Prior work has applied the Isomap manifold learning algorithm to unlabeled audio data to embed frequency sub-bands in 3-D space where the Euclidean distances are inversely proportional to the strength of their Pearson correlations. However, discovering octave equivalence via Isomap requires visual inspection and is not scalable. To address this problem, we define "helicality" as the goodness of fit of the 3-D Isomap embedding to a Shepherd-Risset helix. Our method is unsupervised and uses a custom Frank-Wolfe algorithm to minimize a least-squares objective inside a convex hull. Numerical experiments indicate that isolated musical notes have a higher helicality than speech, followed by drum hits.

## 1. INTRODUCTION

Octave equivalence refers to the perceived consonance of any pair of tones whose frequency ratio is a power of two. Although not universally shared across listeners, this phenomenon lies at the foundation of multiple theories of music around the globe. For example, concepts such as *Saptak* in Indian classical music or *Tonnetz* in German classical music both rely on the hypothesis that notes one octave apart can be assigned to the same pitch class.

To explain computer-generated auditory paradoxes such as the circularity of Shepard tones [1–3], music psychologists represent pitch on a 3-D helix (or, equivalently, a 2-D spiral) which makes a full turn at every octave, thereby aligning power-of-two harmonics [4].

In MIR, the earliest representation with octave equivalence is arguably the chromagram [5], i.e., a constant-$Q$ transform (CQT) followed by summation across octaves. More recently, the spiral scattering transform [6] applies a multivariable wavelet modulus operator before octave summation, thus improving discriminability while guaranteeing stability to octave transposition.

Another line of research considers octave equivalence as domain-specific knowledge in deep learning. Spiral convolutional networks [7] have receptive fields *à trous* ("with holes") spanning across octaves. Comparable ideas are found in more recent publications, such as harmonic CQT [8], folded CQT [9], Fifthnet [10], harmonic filters [11], and harmonic convolutions [12].

Despite the practical effectiveness of octave equivalence in various MIR tasks, the question of discovering octave equivalence directly from data has received less attention. One exception is [13], which applied multidimensional scaling (MDS) to visualize mutual information between time–frequency atoms learned by independent component analysis (ICA). More recently, [14] have applied the Isomap manifold learning algorithm [15] to visualize Pearson correlations between CQT activations.

One shortcoming of methods [13] and [14] is that they are purely illustrative: although they provide a scatter plot of time–frequency atoms in 3-D, the outcome is left to visual inspection, which hampers their scalability. Our paper addresses this problem by proposing a geometrical criterion as to whether a 3-D scatter plot resembles a helix.

## 2. METHODS

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:

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

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 [14] for further details.

Let $\mathbf{e}_m$ and $\lambda_m$ 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 $\chi$ denote chroma. Then, we compute chroma centroids

Companion website: github.com/sripathisridhar/sridhar2020ismir

$\widetilde{\boldsymbol{y}}[\chi] = \frac{1}{J} \sum_{j=0}^{J-1} \boldsymbol{y}[\chi + Qj]$. 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 [16] 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:

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

taken from [17, Equation 4]. In practice, we solve the problem above via a custom implementation of the Frank-Wolfe conditional gradient algorithm [18], 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:

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

We solve the problem above by linear regression.

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 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

We analyze three datasets: TinySOL (music) [19], 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.

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 $H = 0.54$.

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

## 4. CONCLUSION

We have presented a method for fitting a helix to a 3-D manifold of CQT subbands. Further research is needed to examine whether helicality matches human perception.
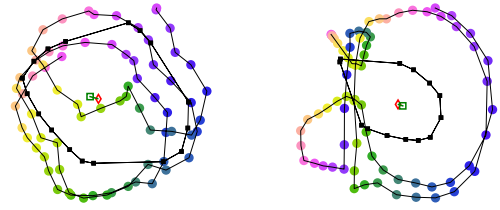


(a) Music. Helicality: 0.54.  (b) Speech. Helicality: 0.30.

**Figure 1**. Isomap embedding of music (left) and speech (right) 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 respectively denote initial and final center estimate for circle fitting.
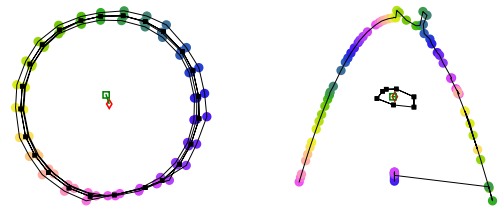


(a) Horn only. Helicality: 0.94.  (b) Drums only. Helicality: 0.28.

**Figure 2**. Isomap embedding of isolated instruments: horn (left) and drums (right). See Figure 1 for legend.
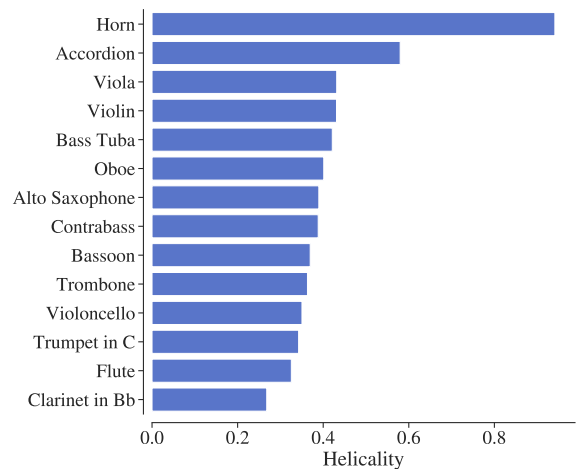


**Figure 3**. Helicality scores of TinySOL instrument classes.

## 5. ACKNOWLEDGMENT

## 6. REFERENCES

[1] D. Deutsch, "The paradox of pitch circularity," *Acoustics Today*, vol. 7, pp. 8–15, 2010.

[2] J.-C. Risset, "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds," *Journal of the Acoustical Society of America*, vol. 46, no. 1A, pp. 88–88, 1969.

[3] C. Pelofi, V. de Gardelle, P. Egré, and D. Pressnitzer, "Interindividual variability in auditory scene analysis revealed by confidence judgements," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160107, 2017.

[4] R. N. Shepard, "Circularity in judgments of relative pitch," *JASA*, vol. 36, no. 12, pp. 2346–2353, 1964.

[5] G. Wakefield, "Mathematical Representation of Joint Time–chroma Distributions," in *Proc. SPIE ASPAAI*, 1999.

[6] V. Lostanlen and S. Mallat, "Wavelet scattering on the pitch spiral," in *Proc. DAFx*, 2015.

[7] V. Lostanlen and C. E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," in *Proc. ISMIR*, 2016.

[8] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for $f_0$ estimation in polyphonic music," in *Proc. ISMIR*, 2017.

[9] J.-F. Ducher and P. Esling, "Folded CQT RCNN for Real-time Recognition of Instrument Playing Techniques," in *Proc. ISMIR*, 2019.

[10] K. O'Hanlon and M. B. Sandler, "The Fifthnet Chroma Extractor," in *Proc. IEEE ICASSP*. IEEE, 2020.

[11] M. Won, S. Chun, O. Nieto, and X. Serrá, "Data-driven harmonic filters for audio representation learning," in *Proc. IEEE ICASSP*. IEEE, 2020.

[12] Z. Zhang, Y. Wang, C. Gan, J. Wu, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Deep audio priors emerge from harmonic convolutional networks," in *Proc. ICLR*, 2020.

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

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

[15] 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.

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

[17] 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.

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

[19] 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.