WAVELET SCATTERING FOR AUTOMATIC CHORD ESTIMATION

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in
submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

State-of-the-art automatic chord recognition systems rely on multi-band chroma representations, Gaussian Mixture Model pattern matching, and Viterbi decoding. This paper explores the use of Haar wavelet transforms and scattering in place of multi-band chroma. Wavelets operating across octaves encode sums and differences in chroma bins at different scales. We describe both the Haar wavelet transform and deep wavelet scattering and develop an efficient algorithm for their computation. Potential benefits of wavelet representations, including stability to octave deformations, over multi-band chroma are discussed. Accuracy of wavelet representations used for chord recognition is analyzed over a large vocabulary of chord qualities.

1. INTRODUCTION

Along with lyrics and melody, chord sequences provide a succinct description of tonal music. As such, they are often written down under the form of lead sheets, for the use of accompanists and improvisers. Besides its original purpose in music education and transmission, the knowledge of harmonic content has been leveraged in music information research to address higher-level tasks, including cover song identification [4], genre recognition [11], and lyrics-to-audio alignment [9]. We refer to the review of McVicar et al. [10] for a recent state of the art.

As stressed by Humphrey and Bello [5], chord labels are not mutually exclusive categories, but instead follow a relation of partial order. For instance, the tetrad `A:min7` is contained in the triad `A:min`, which in turn is contained in the power chord `A:5`. By setting up equivalence rules for chord labels that belong to a common superset, discrepancies between conflicting annotations can be resolved up to a desired level of specificity. In spite of this variety of settings, all evaluation metrics for automatic chord estimation share the following minimal property: a chord label remains the same if all its components are jointly transposed by one octave, be it upwards or downwards.

In order to comply with this requirement, the vast majority of existing systems rely on the chroma representation, i.e. a 12-dimensional vector derived from the constant-Q spectrum by summing up all frequency bands which share the same pitch class according to the twelve-tone equal temperament. However, it should be noted that the chroma representation is not only invariant to octave transposition, but also to any permutation of the chord factors – an operation known in music theory as inversion. Although major and minor triads are unchanged by inversion, some rarer chords, such as augmented triads and minor seventh tetrads, are conditional upon the position of the root.

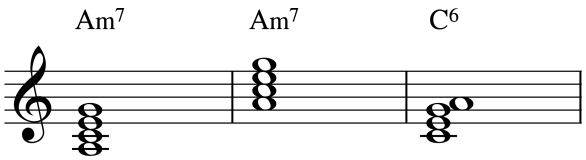

Figure 1. Three possible voicings of the pitch class set {`C,E,G,A`}, resulting either in the chord `A:min7` or `C:maj6`. See text for details.

Figure 1 illustrates the importance of disambiguating inversions when transcribing chords. The first two voicings are identical up to octave transposition of all the chord factors, and thus have the same chord label `A:min7`. In contrast, the third voicing is labeled as `C:maj6` in root position, although its third inversion would correspond to the first voicing.

With the aim of improving Automatic Chord Estimation (ACE) under fine-grained evaluation metrics for large chord vocabularies (157 chord classes), this article introduces two feature extraction methods that are invariant to octave transposition, yet sensitive to chord inversion. The former consists of computing a Haar wavelet transform of the constant-Q spectrum along the octave variable and keeping the absolute values of the resulting coefficients, at all scales and positions. The latter iterates the Haar wavelet modulus nonlinear operator over increasing scales, until reaching the full extent of the constant-Q spectrum. Both methods build upon the large chord vocabulary ACE software of Cho and Bello [3], which holds state-of-the-art performance on the McGill Billboard dataset [1].

Section 2 describes the multi-band chroma features, as introduced by Cho and Bello, and its integration into a multi-stream hidden Markov model. Section 3 defines the Haar wavelet transform across octaves of the constant-Q spectrum. Section 4 defines the deep Haar scattering transform. Section 5 discusses the experimental setup along with the evaluation metrics for chord estimation accuracy.

Section 6 presents the results of large vocabulary chord estimation comparing all three feature extraction methods and Section 7 discusses these results.

2. MULTI-BAND CHROMA FEATURES

A system for automatic chord estimation typically consists of two stages: feature extraction and acoustic modeling. At the first stage, the audio query is converted into a time series of pitch class profiles, which represent the relative salience of pitch classes according to the twelve-tone equal temperament. At the second stage, each frame in the time series is assigned a chord label among a predefined vocabulary. This section presents a multi-stream approach to acoustic modeling, as first introduced by Cho and Bello [3].

The constant-Q transform $\mathbf{X}[t, \gamma]$ is a time-frequency representation whose center frequencies $2^{\gamma/Q}$ are in a geometric progression. By setting $Q = 12$, the log-frequency variable γ is akin to a pitch in twelve-tone equal temperament. Moreover, the Euclidean division $\gamma = Q \times u + q$ reveals the octave u and pitch class q, which play an essential role in music harmony. In all of the following, we reshape the constant-Q transform accordingly, and keep the notation $\mathbf{X}[t, q, u]$ for simplicity.

To address the disambiguation of chords in an extended vocabulary, Cho and Bello have divided the constant-Q spectrum into K bands by means of half-overlapping Gaussian windows along the log-frequency axis [3]. The width σ of the windows is inversely proportional to the desired number of bands K: in particular, it is of the order of one octave for $K = 8$, and two octaves for $K = 4$. The centers of the windows are denoted by γ_k, where the band index k ranges from 0 to $K - 1$. Consequently, the multi-band chroma features are defined as the following three-way tensor:

$$\mathbf{Y}[t, q, k] = \sum_u \mathbf{X}[t, q, u] \boldsymbol{w}[Q \times u + q - \gamma_k], \quad (1)$$

where $\boldsymbol{w}[\gamma] = \exp(-\gamma^2/(2\sigma^2))$ is a Gaussian window of width σ, centered around zero.

Acoustic modeling is classically achieved with a hidden Markov model (HMM) whose states are estimated as mixtures of multivariate Gaussian probability distributions, i.e. Gaussian mixture models (GMM) in dimension $Q = 12$. In order to extend this framework to multi-band chroma features, Cho and Bello have trained K models in parallel, end-to-end, over each feature map k of the tensor $\mathbf{Y}[t, q, k]$. At test time, the emission probability distributions of each model are aggregated such that they are the predicted outputs of a single state sequence.

The computational complexity of resulting K-stream HMM grows exponentially with the number of streams K. However, by assuming synchronicity and statistical independence of the streams, the aggregation boils down to a geometric mean, thus with linear complexity in K. It must be noted that the geometric mean does not yield a true probability distribution, as it does not sum to one. Yet, it is of widespread use e.g. in speech recognition, due to its simplicity and computational tractability.

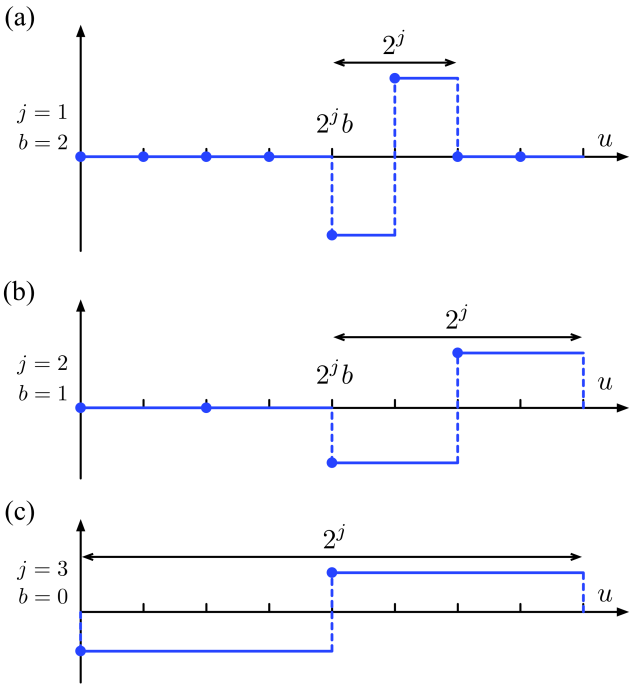

Figure 2. Three elements of the Haar wavelet basis $\{\boldsymbol{\psi_{j,b}}\}$ for various values of the scale index j and the translation index b. See text for details.

Fed with multiband chroma features, the K-stream HMM of Cho and Bello has achieved state-of-the-art results on the McGill Billboard dataset at the MIREX evaluation campaign [3].

3. HAAR WAVELET TRANSFORM

In spite of their success, the multi-band chroma features presented above are rather unsatisfying as inputs of a K-stream HMM, which relies on statistical independence. In this section, we introduce an alternative set of features for harmonic content, namely the absolute value of Haar wavelet coefficients, which satisfies statistical independence since it is derived from an orthogonal basis of $\mathbb{R}^K$. All subsequent operations apply to the octave variable u, and are vectorized in terms of time t and chroma q. To alleviate notations, we replace the three-way tensor $\mathbf{X}[t, q, u]$ by a vector $\boldsymbol{x}[u]$, thus leaving the indices t and q implicit.

Dating back to 1909, the Haar wavelet $\boldsymbol{\psi}$ is a piecewise constant, real function of compact support, consisting of two steps of equal length and opposite values. Within a discrete framework, it is defined by the following formula:

$$\forall u \in \mathbb{Z},\ \boldsymbol{\psi}[u] = \begin{cases} \frac{-1}{\sqrt{2}} & \text{if } u = 0 \\ \frac{1}{\sqrt{2}} & \text{if } u = 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The "mother" wavelet $\boldsymbol{\psi}[u]$ is translated and dilated by powers of two, so as to produce a family of discrete sequences $\boldsymbol{\psi_{j,b}}[u] = 2^{\frac{j-1}{2}} \boldsymbol{\psi}[2^{(j-1)}(u - 2b)]$ indexed by

the scale parameter $j \in \mathbb{N}^*$ and the translation parameter $b \in \mathbb{Z}$. Some Haar wavelets are shown on Figure 2 for various values of j and b. After endowing them with the Euclidean inner product

$$\langle \boldsymbol{\psi_{j,b}} | \boldsymbol{\psi_{j',b'}} \rangle = \sum_{u=-\infty}^{+\infty} \boldsymbol{\psi_{j,b}}[u] \boldsymbol{\psi_{j',b'}}[u], \qquad (3)$$

the wavelets $\{\boldsymbol{\psi_{j,b}}\}_{j,b}$ form an orthonormal basis of finite-energy real sequences. Moreover, the Haar wavelet is the shortest function of compact support such that the family $\{\boldsymbol{\psi_{j,b}}\}_{j,b}$ satisfies this orthonormality property. On the flip side, it has a poor localization in the Fourier domain, owing to its sharp discontinuities.

It must be noted that, unlike the pseudo-continuous variables of time and frequency, the octave variable is intrinsically dicrete, and has no more than 8 coefficients in the audible spectrum. Therefore, we choose to favor compact support over regularity, i.e. Haar over Daubechies or Gabor wavelets.

The wavelet transform of some finite-energy sequence $\boldsymbol{x} \in \ell^2(\mathbb{Z})$ is defined by $\mathbf{W}\boldsymbol{x}[j,b] = \langle \boldsymbol{x} | \boldsymbol{\psi_{j,b}} \rangle$. Since $\boldsymbol{x}[u]$ has a finite length $K = 2^J$, this decomposition is informative only for indices (j,b) such that $j \leq J$ and $2^j b \leq K$, i.e. $b \leq 2^{J-j}$. The number of coefficients in the Haar wavelet transform of $\boldsymbol{x}[u]$ is thus equal to $\sum_{j=1}^{J} 2^{J-j} = 2^J - 1$. For the wavelet representation to preserve energy and allow signal reconstruction, a residual term

$$\mathbf{A_J}\boldsymbol{x} = \boldsymbol{x}[0] - \sum_{j,b} \langle \boldsymbol{x} | \boldsymbol{\psi_{j,b}} \rangle \boldsymbol{\psi_{j,b}}[0] = \sum_{u<K} \boldsymbol{x}[u] \qquad (4)$$

must be appended to the wavelet coefficients. Observe that $\mathbf{A_J}\boldsymbol{x}$ computes a delocalized average of all signal coefficients, which can equivalently be formulated as an inner product with the constant function $\boldsymbol{\phi}[u] = 2^{-J/2}$ over the support $[\![0; K[\![$. Henceforth, it corresponds to the traditional chroma representation, where spectrogram bands of the same pitch class q are summed across all K octaves.

Since the wavelet representation amounts to K inner products in $\mathbb{R}^K$, its computational complexity is $\Theta(K^2)$ if implemented as a matrix-vector product. Fast Fourier Transforms (FFT) would bring the complexity to $\Theta(K(\log_2 K)^2)$. To improve this, Mallat has developed a recursive scheme, called *multiresolution pyramid* [7], which operates as a cascade of convolutions with some pair of quadrature mirror filters $(\boldsymbol{g}, \boldsymbol{h})$ and progressive subsamplings by a factor of two. Since the number of operations is halved after each subsampling, the total complexity of the multiresolution pyramid is $K + \frac{K}{2} + \cdots + 1 = \Theta(K)$.

Let us denote by $\boldsymbol{g}_{\downarrow 2}$ and $\boldsymbol{h}_{\downarrow 2}$ the corresponding operators of subsampled convolutions, and by $(\boldsymbol{g_{\downarrow 2}})^j$ the j-fold composition of operators $\boldsymbol{g_{\downarrow 2}}$. The wavelet transform rewrites as

$$\mathbf{W}\boldsymbol{x}[j,b] = \left(\boldsymbol{h_{\downarrow 2}} \circ (\boldsymbol{g_{\downarrow 2}})^j \boldsymbol{x}\right)[b], \qquad (5)$$

while the fully delocalized chroma representation rewrites as $\mathbf{A_J}\boldsymbol{x} = (\boldsymbol{g_{\downarrow 2}})^J \boldsymbol{x}$. A flowchart of the operations involved in the wavelet transform is shown on Figure 3. We refer to chapter 7 of Mallat's textbook [8] for further insight.

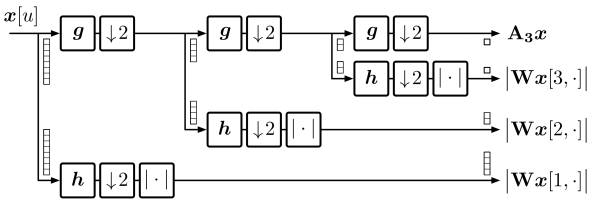

Figure 3. Discrete wavelet transform of a signal of length 8, as implemented with a multiresolution pyramid scheme. See text for details.

	operations	memory
Matrix-vector product	$\Theta(K^2)$	$\Theta(K)$
Fast Fourier transforms	$\Theta(K(\log K)^2)$	$\Theta(K)$
Multiresolution pyramid	$\Theta(K)$	$\Theta(1)$

Table 1. Computational complexity and memory usage of various implementations of the Haar wavelet transform, for a one-dimensional signal of length K. See text for details.

Since the low-pass filter $\boldsymbol{\phi}$ and the family of wavelets $\boldsymbol{\psi_{j,b}}$'s form an orthonormal basis of $\mathbb{R}^K$, any two signals $\boldsymbol{x}[u]$ and $\boldsymbol{y}[u]$ have the same Euclidean distance in the wavelet domain as in the signal domain. This isometry property implies that the wavelet representation is not invariant to translation per se. Therefore, the wavelet-based chroma features are extracted by taking the absolute value of each wavelet coefficient, hence contracting Euclidean distances in the wavelet domain. Most importantly, the distance $\|\mathbf{W}\boldsymbol{x} - \mathbf{W}\boldsymbol{y}\|$ is all the more reduced by the absolute value nonlinearity that $\boldsymbol{x}$ and $\boldsymbol{y}$ are approximate translates of each other.

In the case of Haar wavelets, the low-pass filtering $(\boldsymbol{x} * \boldsymbol{g})$ consists of the sum between adjacent coefficients, whereas the high-pass filtering $(\boldsymbol{x} * \boldsymbol{h})$ is the corresponding difference, up to a renormalization constant:

$$(\boldsymbol{x} * \boldsymbol{g})[2b] = \frac{\boldsymbol{x}[2b+1] + \boldsymbol{x}[2b]}{\sqrt{2}}, \text{ and}$$
$$(\boldsymbol{x} * \boldsymbol{h})[2b] = \frac{\boldsymbol{x}[2b+1] - \boldsymbol{x}[2b]}{\sqrt{2}}. \qquad (6)$$

Besides its small computational complexity, the multiresolution pyramid scheme has the advantage of being achievable without allocating memory. Indeed, at every scale j, the pair $(\boldsymbol{g_{\downarrow 2}}, \boldsymbol{h_{\downarrow 2}})$ has $2^{-j}K$ inputs and $2^{-j}K$ outputs, of which one half are subsequently mutated. By performing the sums and differences in place, and defering the renormalization to the end of the flowchart, the time taken by the wavelet transform procedure remains negligible in front of the time taken by the constant-Q transform.

4. DEEP HAAR SCATTERING

The wavelet modulus operator "scatters" the variations of a signal over different scales 2^j while keeping the finest

localization possible b. As such, the coefficient $|\mathbf{W}\boldsymbol{x}[j,b]|$ only bears a limited amount of invariance, which is of the order of 2^j. In this section, we iterate the scattering operator over increasing scales, until reaching some maximal scale $K = 2^J$. We interpret the scattering cascade in terms of invariance and discriminability, and provide a fast implementation with $\Theta(K \log K)$ operations and $\Theta(1)$ allocated memory.

Most of the intervallic content of chords in tonal music consists of perfect fifths, perfect fourths, major thirds and minor thirds. Quite strikingly, these intervals are also naturally present in harmonic series, as the log-frequency distances between the foremost partials. By combining the two previous propositions, we deduce that the components of a typical chord overlap at high frequencies, hence producing an interference pattern which reveals their relative positions.

In our introductory example, denoting by f_0 the root frequency of `A:min7`, the root interferes with its perfect fifth `E` at the frequency $3f_0$. In contrast, in its third inversion labeled as `C:6`, the interference between `A` and `E` only starts at $6f_0$, i.e. one octave higher. Under the same instrumentation, this inversion yields a deformation of the octave vector corresponding to `E`, which consists of the frequency bins of the form $2^u \times 3f_0$ for integer $u \in \mathbb{Z}$.

More generally, we argue that the characterization of complex interference patterns in polyphonic music is a major challenge in large-vocabulary chord estimation, as it provides a tool for disambiguating chord inversions in spite of global invariance to octave transposition.

The scattering cascade consists of iterating the process over every sequence of scale indices $(j_1 \dots j_m)$ whose sum is lower or equal to J.

Scattering has been employed as a feature extraction stage for many problems in signal classification. Initially defined as operating solely over the time dimension, it has recently been generalized to multi-variable transforms in the time-frequency domain, including log-frequency and octave [6]. In addition, Cheng et al. have applied Haar scattering to the unsupervised learning of unknown graph connectivities [2].

Since it results from the alternate composition of unitary and contractive operators, it follows immediately that the scattering transform is itself unitary and contractive. Moreover, Mallat has proven that it is invariant to translation and stable to the action of small deformations [**?**].

$$\boldsymbol{x_1}[j_1, b] = |\mathbf{W}\boldsymbol{x}[j_1, b]| = \left|\langle \boldsymbol{x} | \boldsymbol{\psi_{j_1,b}} \rangle\right|$$

$$\mathbf{S_J}\boldsymbol{x}[j_1, \cdots, j_m] = (\boldsymbol{g_{\downarrow 2}})^{\left(J - \sum_{n=1}^{m} j_n\right)} \bigcirc_{\sum_{n=1}^{m} j_n \leq J} \left|\boldsymbol{h_{\downarrow 2}} \circ (\boldsymbol{g_{\downarrow 2}})^{j_n}\right| \boldsymbol{x}, \quad (7)$$

where the circle symbol represents functional composition. Interestingly, the case $m = 0$ boils down to the sum across octaves $\mathbf{A_J}$ already introduced in Equation 4, i.e. the chroma representation. A flowchart of the operations involved in the deep scattering transform is shown on Figure 4.

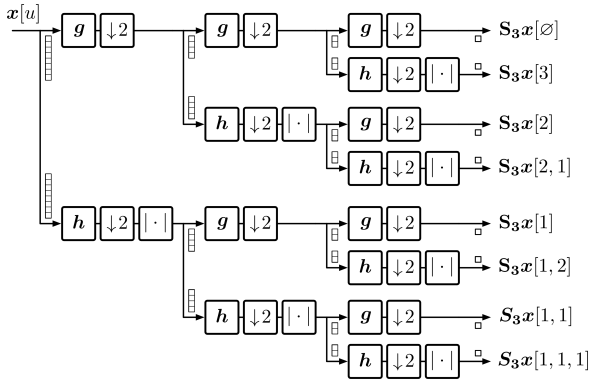

Figure 4. Deep scattering transform of a signal of length 8, as implemented with a convolutional pyramid scheme. See text for details.

	operations	memory
Matrix-vector product	$\Theta(K^3)$	$\Theta(K^2)$
Fast Fourier transforms	$\Theta(K^2(\log K)^2)$	$\Theta(K^2)$
Multiresolution pyramid	$\Theta(K \log K)$	$\Theta(1)$

Table 2. Computational complexity and memory usage of various implementations of the deep Haar scattering transform, for a one-dimensional signal of length K. See text for details.

5. EXPERIMENTAL SETUP AND EVALUATION

In all experiments, a training set consisting of 108 songs from the Beatles discography, 99 RWC pop songs, 224 songs from the Billboard dataset, and 20 Queen songs was used for a total of 451 songs. The testing dataset comprised of 65 songs from the Beatles and uspop datasets that were not part of the training set and that contained a sufficient number of examples of each chord quality. Both the training set and testing set of songs are kept constant across all experiments.

We consider a large vocabulary of chords – 13 different qualities (major, minor, minor 7th, dominant 7th, major 7th, suspended 4th, major 6th, minor 6th, suspended 2nd, diminished triad, augmented triad, half-diminished 7th, fully-diminished 7th) with their roots at all 12 pitch classes. Additionally, we consider the no-chord 'N' label, where no discernible chord is active. For each experiment, a chord model and Viterbi transition probability matrix are generated from the training set with the band K of any experiment determining the amount of chroma vectors at any given temporal window. K is therefore equivalent to the number of bands in the multiband chroma representation and the maximum wavelet scale ($K = 2^J$) in the wavelet and scattering representations (i.e. the number of wavelet coefficients). Chord recognition is then carried out on a testing set.

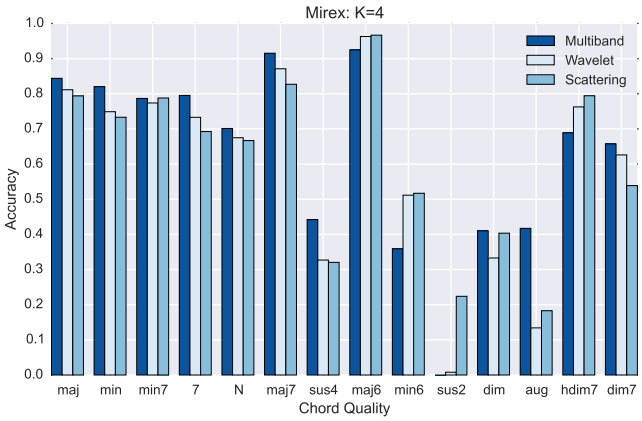

Figure 5. Multiband (chroma), Haar wavelet transform, and deep Haar scattering compared for $K = 4$ streams. Chord accuracy computed via mirex.

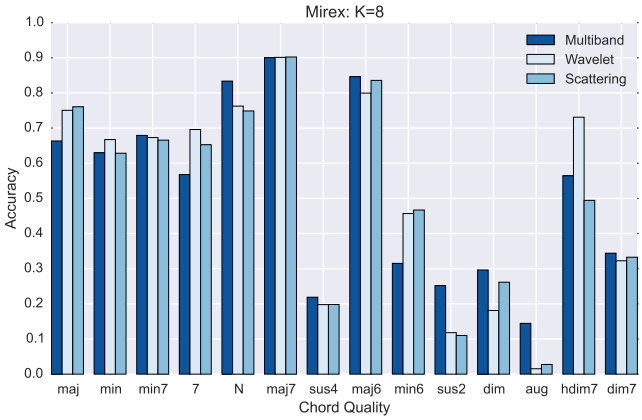

Figure 6. Multiband (chroma), Haar wavelet transform, and deep Haar scattering compared for $K = 8$ streams. Chord accuracy computed via mirex.

After generating estimated chord labels for each song in the test set, Python scripts evaluate the results through the use of the mir_eval package [12]. As per [12], there is "no single right way to compare two sequences of chord labels," and mir_eval package computes ACE accuracy based along a handful of metrics. In this experiment we consider two evaluation metrics. The **mirex** metric "considers a chord correct if it shares at least three pitch classes in common" [12]. This metric, however, is too lenient when considering inversions and rarer chord classes (such as min7, maj6). The **tetrads_inv** metric is much stricter, evaluating chord accuracy over the entire quality in closed voicing, and taking inversions notated in the reference labeling into account.

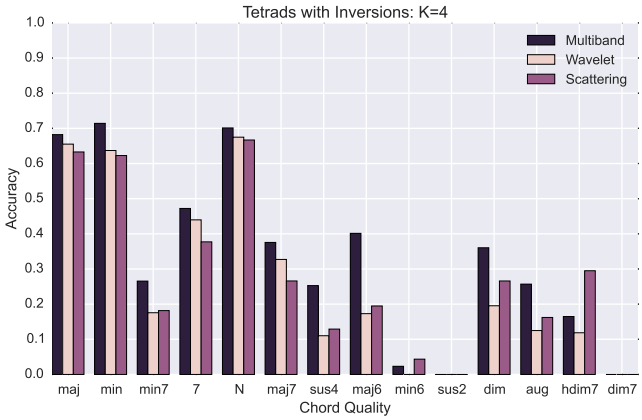

Figure 7. Multiband (chroma), Haar wavelet transform, and deep Haar scattering compared for $K = 4$ streams. Chord accuracy computed via tetrads with inversions.

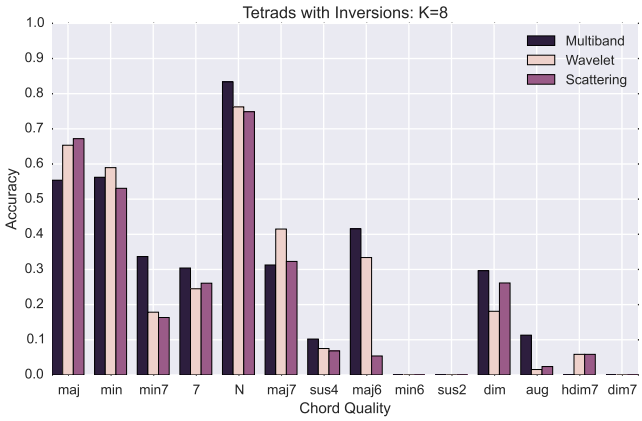

Figure 8. Multiband (chroma), Haar wavelet transform, and deep Haar scattering compared for $K = 8$ streams. Chord accuracy computed via tetrads with inversions.

Mode	k	mirex	tetrads	tetrads inv
Multiband	4	**80.18** %	64.23 %	62.48 %
	8	61.69 %	50.66 %	49.18 %
Haar Wavelet	4	75.87 %	60.03 %	58.22 %
	8	69.36 %	57.23 %	55.59 %
Haar Scattering	4	74.38 %	58.24 %	56.47 %
	8	68.78 %	57.14 %	55.44 %

Table 3. Overall accuracy for multiband chroma, Haar wavelet transforms, and deep Haar scattering at scales $k = 4$ and 8. Accuracy computed via mirex, tetrad, and tetrad with inversions metrics. State-of-the-art for large vocabulary chord recognition (multiband, $k = 4$) shown in bold.

6. RESULTS

7. DISCUSSION

8. REFERENCES

[1] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground-truth set for audio chord recognition and music analysis. In *Proc. ISMIR*, 2011.

[2] Xiuyuan Cheng, Xu Chen, and Stéphane Mallat. Unsupervised deep Haar scattering on graphs. In *Proc. NIPS*, 2014.

[3] Taemin Cho and Juan P. Bello. Large vocabulary chord recognition system using multi-band features and a multi-stream hmm. In *Proc. MIREX*, 2013.

[4] Daniel P. W. Ellis and Graham E Poliner. Identifying "cover songs" with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*, 2007.

[5] Eric J. Humphrey and Juan P. Bello. Four timely insights on automatic chord estimation. In *Proc. ISMIR*, 2015.

[6] Vincent Lostanlen and Stéphane Mallat. Wavelet scattering on the pitch spiral. In *Proc. DAF-x*, 2015.

[7] Stéphane Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, 1989.

[8] Stéphane Mallat. *A Wavelet Tour of Signal Processing, 3rd edition: The Sparse Way*. Academic press, 2008.

[9] M. Mauch, H. Fujihara, and M. Goto. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE TASLP*, 20(1):200–210, 2012.

[10] M. McVicar, R. Santos-Rodrguez, Y. Ni, and T. D. Bie. Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):556–575, 2014.

[11] Carlos Pérez-Sancho, David Rizo, and José M Inesta. Genre classification using chords and stochastic language models. *Connection science*, 21(2-3):145–159, 2009.

[12] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014.