Wavelet Scattering In Chroma Space

Chris Miller

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
Steinhardt School
New York University

Advisor: Juan P. Bello
Reader: Alex Ruthmann

May 10, 2016

Abstract

State of the art automatic chord recognition systems rely on multiband chroma representations, Gaussian Mixture Model pattern matching, and Viterbi decoding. This thesis explores the use of Haar wavelet transforms and scattering in place of multiband 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 multiband chroma are discussed. Accuracy of wavelet representations used for chord recognition is analyzed over a large vocabulary of chord qualities.

Note

This thesis is based in large part on a paper submitted to the 2016 ISMIR conference and currently pending review entitled *Wavelet Scattering For Automatic Chord Estimation* by Chris Miller, Vincent Lostanlen, Stéphane Mallat, and Juan Bello.

Acknowledgements

This thesis is based on research conducted in the Fall of 2015 at École normale supérieure in Paris. As a visiting researcher in the Data Processing and Classification team, I was able to work closely with those primarily responsible for the scattering transform's many successes over its as-yet short lifespan, and contributing in any small way to this bevy of research has been a humbling and deeply gratifying experience. My many thanks to Stéphane Mallat, who so graciously invited me to his team and supported me throughout my stay. I owe much of the work in this thesis to my teammates Vincent Lostanlen and Carmine Emanuele Cella, who have guided and advised me at numerous points throughout this work. I would also like to thank Sira, Irène, Mathieu, Amos, and everyone else who so warmly welcomed me to ENS and made my stay so enjoyable and productive.

I would like to thank Juan Bello for initiating this research project, making the link with the team at ENS, and advising me throughout the initial research and multiple paper write-ups. He has been an excellent advisor, always forcing me to think deeper at each stage and discover the fundamental aspects at play in such increasingly complex problems. I would like to also thank all in the MIR team at NYU, primarily Taemin Cho for providing the state-of-the-art research in the first place, and Rachel Bittner, Eric Humphrey, and Brian McFee for their help along the way.

Agnieszka Roginska has provided endless support and guidance throughout the past two and a half years, and I am deeply grateful to her for it. It has been a pleasure being her research assistant, and I would like to thank Andrea Genovese, Areti Andreopoulou, and everyone else who has contributed to the Immersive Audio's team research. Thanks as well to all of the faculty at Steinhardt's Music Technology program and the Courant Institute who have impacted my studies.

Jordan Juras and Dave Tatasciore have proven invaluable sounding boards over these past years, putting up with many rants over even more beers.

Most importantly I would like to thank my partner Iliana for everything. Very little of this work would have been possible without her unwavering support.

Contents

List of Figures

1 Introduction

Chord sequences provide a succinct description of tonal music and are often written down on 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 [Ellis and Poliner, 2007], genre recognition [Pérez-Sancho et al., 2009], and lyrics-to-audio alignment [Mauch et al., 2012]. We refer to the review of [McVicar et al., 2014] for a recent state of the art.

All evaluation metrics for automatic chord estimation share the following basic 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 chord estimation systems rely on the chroma representation, i.e. a 12-dimensional vector derived from a log-frequency spectrum (such as the constant-Q transform) 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 less common chords, such as augmented triads and minor seventh tetrads, are conditional upon the position of the root.

Figure 1 illustrates the importance of disambiguating inversions when transcribing chords, which has previously been addressed by [Mauch and Dixon, 2010]. 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 thesis introduces two feature extraction methods that are invariant to octave transposition, yet sensitive to chord inversion.

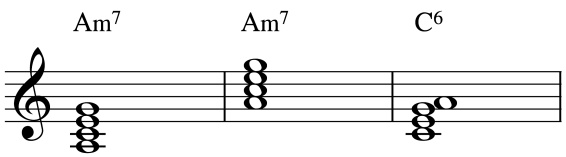

Figure 1: Three possible voicings of the pitch class set {`C,E,G,A`}, resulting either in the chord `A:min7` or `C:maj6`. See text for details.

The first consists of computing a Haar wavelet transform of the constant-Q spectrum along the octave variable and storing the absolute values of the resulting coefficients at all scales and positions. The second 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, 2013], which holds state-of-the-art performance on the McGill Billboard dataset [Burgoyne et al., 2011].

Section 2 of this thesis reviews the basic components of automatic chord estimation systems, describing in particular the multiband chroma features, as introduced by [Cho and Bello, 2013], and their integration into a multi-stream hidden Markov model. Section 3 reviews the scattering transform as introduced by Mallat in [Mallat, 2012]. Section 4 defines the Haar wavelet transform across octaves of the constant-Q spectrum and the deep Haar scattering transform, and compares these new representations with the multiband chroma. Section 5 presents and 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 provides extended analysis at all points throughout the chord recognition system. Finally, Section 8 concludes this thesis.

2 Automatic Chord Estimation

A system for automatic chord estimation typically consists of three stages: feature extraction, pattern matching, and post-filtering (decoding).

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. Chroma features are therefore typically used, and have been since Fujishima first used them in a chord estimation context [Fujishima, 1999]. In [Cho and Bello, 2013], a multi-stream (or multiband) chroma representation is introduced, and has achieved state-of-the-art results for large vocabulary chord recognition.

At the second stage, each frame in the time series is assigned a chord label among a predefined vocabulary of N chords. In this thesis we consider a large vocabulary of $N = 157$ chord labels — 13 chord qualities with roots at all 12 pitch classes, plus the no-chord label. A majority of chord recognition systems (including the one presented here) generate N chord models using Gaussian Mixture Models (GMMs), deriving the models from existing annotated music samples [Lee and Slaney, 2006]. Finally, hidden Markov models (HMMs) are applied to the estimated chord sequence using the Viterbi algorithm to filter out unlikely chord changes [Papadopoulos and Peeters, 2007].

This section presents the multi-stream approach to feature extraction, as first introduced in [Cho and Bello, 2013], followed by the pattern matching and filtering steps used in conjunction with multiband chroma for large vocabulary chord estimation.

2.1 Multiband Chroma

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 essential roles 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, resulting in a chroma representation [Cho and Bello, 2014].

To address the disambiguation of chords in an extended vocabulary, [Cho and Bello, 2013] divide the constant-Q spectrum into K bands by means of half-overlapping Gaussian windows along the log-frequency axis. The width σ of the windows is inversely proportional to the

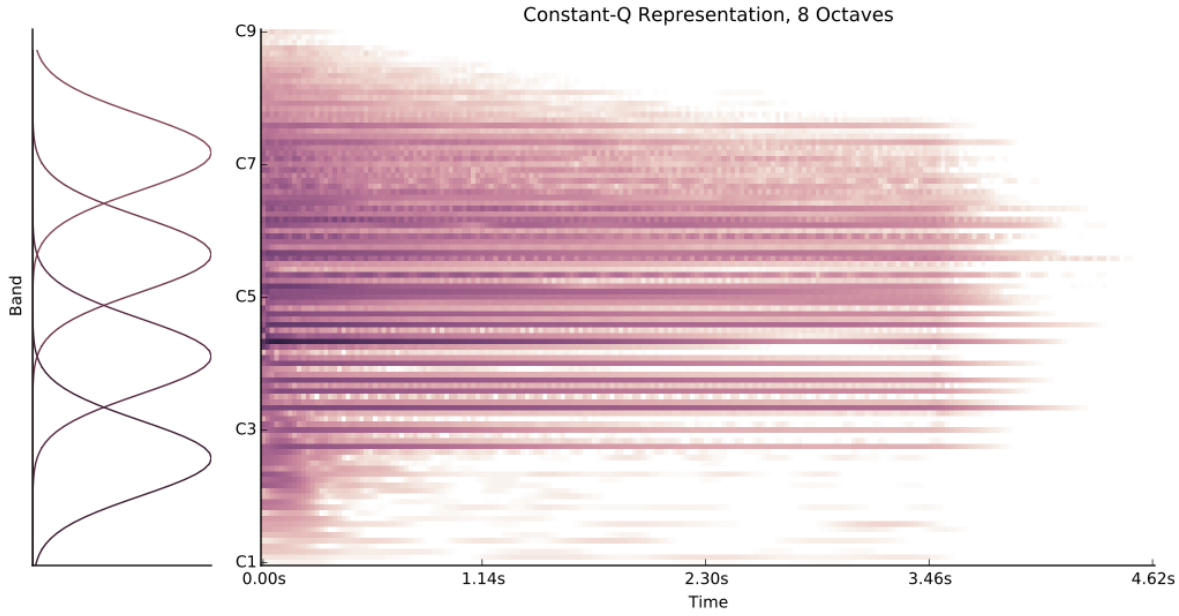

Figure 2: Multiband chroma windowing for $K = 4$ bands. Gaussian windows pictured on left covering 8 octaves in constant-Q pitch space, shown on right.

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:

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

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

2.2 Pattern Matching

2.2.1 Pre-Filtering

The constant-Q transform $\mathbf{X}[t, \gamma]$ segments the underlying signal $x(t)$ into temporal frames of size T. Per the Heisenberg uncertainty principle [Gabor, 1946], the temporal frame size T is often chosen to optimize frequency resolution for the CQT filter-bank, resulting in a frame rate that is much faster than typical rates of change of chords in musical signals. Pre-filtering of the multiband chroma representation $\mathbf{Y}[t, q, k]$ before sending to the pattern matching stage is useful to adapt the chroma features to more musically-relevant timescales. One approach uses

moving average [Fujishima, 1999] or moving median [Papadopoulos and Peeters, 2007] filters to smooth out neighboring frames in order to minimize the effect of transient behavior or noisy frames.

Another approach is to create a beat-synchronous chroma representation [Bello and Pickens, 2005], operating on the assumption that chords tend to change on beats. Frames in-between beats are summed together to generate longer-range harmonic information and smooth out noisy frames, as well as to minimize computations downstream [Cho and Bello, 2014]. The Tempogram Toolbox is used for beat extraction in all experiments conducted as part of this thesis [Grosche and Muller, 2011].

2.2.2 Likelihood Estimation

Given a dataset of audio files with annotated chord data, supervised learning techniques can be used to create chord models from a training dataset. Multivariate Gaussian Mixture Models (GMMs) are often used to model the probability distribution for a chord class based on a given chroma representation [Cho, 2013] [Papadopoulos and Peeters, 2007]. Distributions are characterized by a 12-dimensional vector of mean values μ and covariance matrix $\mathbf{\Sigma}$ estimated from the training data using the Expectation Maximization (EM) algorithm [Sheh and Ellis, 2003] [Moon, 1996].

In order to extend the limited number of annotated chords in the training dataset, each chord is first rotated and transposed to have its root at C and used to generate the chord model for the given class (major, minor, dominant 7, etc.) rooted at C. The chord is then rotated through all other 11 roots and the respective chord models are trained accordingly (see [Sheh and Ellis, 2003]).

2.3 Viterbi Decoding

A Hidden Markov Model (HMM) seeks to characterize the temporal dynamics of features as a discrete Markov process based on a set of transition probabilities between states (i.e. chords) (see [Rabiner, 1989]). With a large set of chord-annotated audio data, the matrix of transition probabilities from one chord to another can be generated automatically. Since the features $\mathbf{Y}[t, q, k]$ are beat-synchronized before chord modeling in this thesis, the symbolic chord annotations are also beat-synchronized so that the transition matrix captures the same frame-to-frame chord transition dynamics as our features.

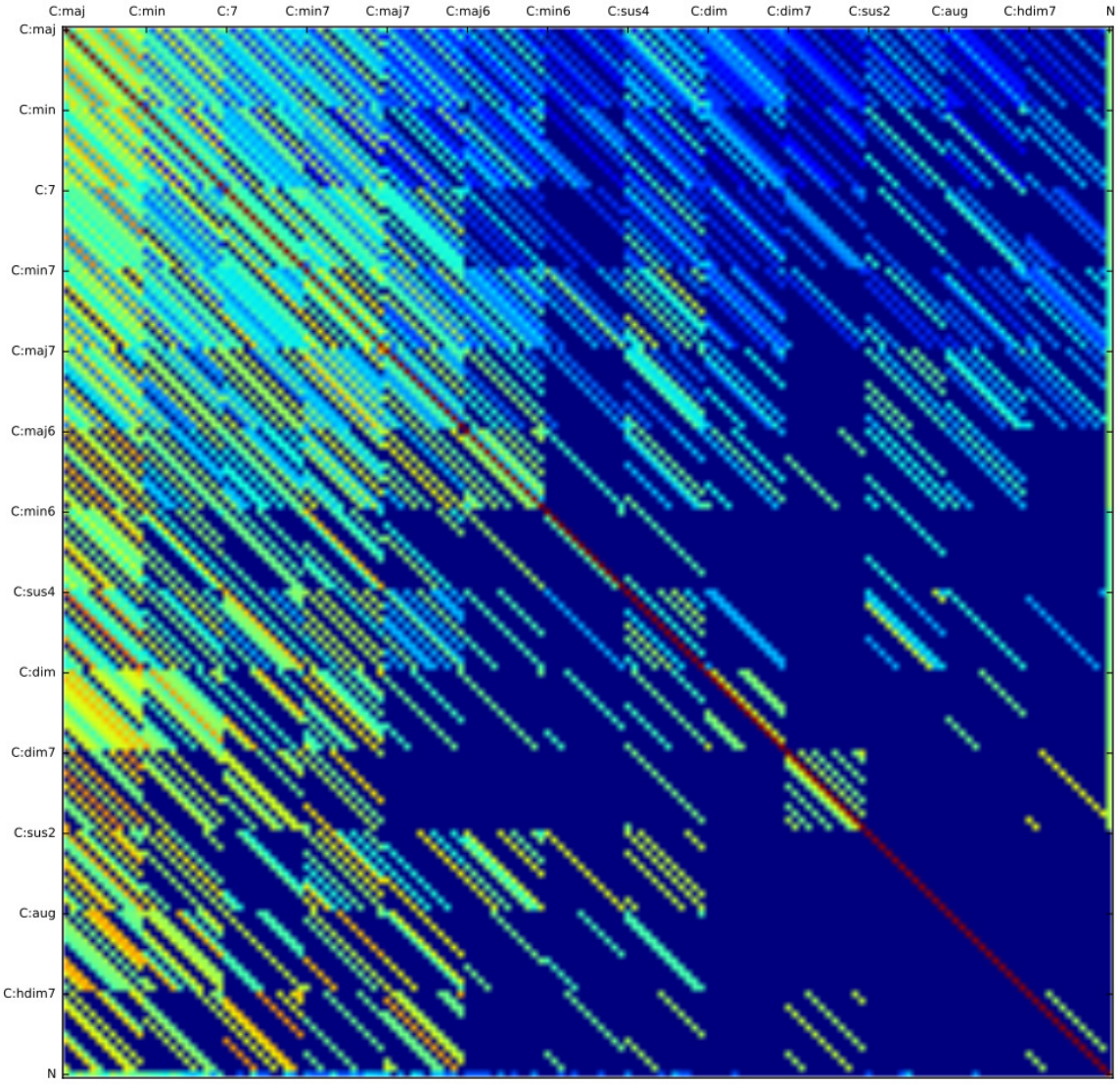

Figure 3: Transition matrix for large chord vocabulary, shown on log-probability scale.

The Viterbi algorithm is typically used to decode HMMs, in our case by finding the most likely sequence of chords (a Markov process χ_t) that results in the current kth chroma band (an observation sequence O_t). In other words, the Viterbi algorithm seeks to find the hidden state sequence χ_t that best explains the observations O_t, and does so by maximizing the conditional probability $\mathbb{P}(\chi_t|O_t)$ [1].

The generated transition matrix for our large vocabulary of chords is shown in Figure 3, displayed on a log-probability scale due to the extreme likelihood of self-transition from one chord state to itself. The deep red diagonal shows the self-transition case. The left side of the matrix

[1] For more on conditional probabilities, random variables, and Markov processes, see [Jacod and Protter, 2004]. For more on HMMs and the Viterbi algorithm, see [Rabiner, 1989].

shows the elevated probability of transition from all chords to a major quality chord, and image becomes darker moving to the right indicating the reduced likelihood of transitioning to more exotic chord qualities.

Chord models are built independently for each feature band k in the multiband chroma tensor $\mathbf{Y}[t, q, k]$ using GMMs, resulting in K end-to-end HMMs in parallel. 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 the 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.

Fed with multiband chroma features, the K-stream HMM has achieved state-of-the-art results on the McGill Billboard dataset at the MIREX evaluation campaign using $K = 4$ streams [Cho and Bello, 2013].

3 Scattering Transform

Despite their success, the multiband chroma features do not comply with the assumption of statistical independence of the K-stream HMM, owing to the overlap between Gaussian windows. Further, they remove the sensitivity to pitch inversion and invariance to octave transposition inherent in the traditional chroma representation by aggregating neighboring octaves. Scattering wavelets across octaves restores these properties. In this section, we review basic properties of the wavelet transform, detail the scattering transform, and discuss scattering for pitched and harmonic sounds.

3.1 Wavelet Transform

A wavelet transform is constructed through the dilation and translation of a mother wavelet $\psi(t) \in \mathbf{L^2}(\mathbb{R}^d)$ [Mallat, 2012] [Daubechies, 1990]. This defines a basis of wavelets $\psi_\gamma(t)$ which, in the frequency-domain, are equivalent to a bank of bandpass filters at center frequencies $\lambda \in \Lambda$. With center frequencies λ corresponding to twelve-tone pitch, i.e. $\lambda = 2^{\gamma/Q}$ for integers γ where $Q = 12$ wavelets per octave, the wavelet basis is the same as a constant-Q filter bank (see Section 2.1). Figure 4 shows the wavelet basis ψ_γ using Morlet wavelets (sinusoids localized in time by Gaussian envelopes — see [Andén and Mallat, 2014]).

A windowing function $\phi(t)$ averages the signal $x(t)$ into temporal frames of size T. ϕ is therefore a low-pass filter with frequency support of $[-2\pi/T, 2\pi/T]$, covering the lowest end of the frequency range that is missed by the bandpass filters ψ_γ. While ϕ is not derived from dilations of the mother wavelet ψ, we refer to it as a wavelet for simplicity. The wavelet transform is therefore computed by a constant-Q transform, convolving an input signal $x(t)$ with the wavelet basis $\{\phi, \psi_\gamma\}_{\gamma \in \mathbb{Z}}$:

$$\mathbf{W}\mathbf{x}[t, \gamma] = |x * \psi_\gamma| * \phi(t) \tag{2}$$

A wavelet modulus operator $\mathbf{W}\mathbf{x}[t, \gamma]$ removes the complex phase of all wavelet coefficients $|x * \psi_\gamma|$ but conserves low-frequency phase information contained in $x * \phi$. The wavelet operator $\mathbf{W}$ is contractive and, by Parseval's theorem, conserves energy, so $\mathbf{W}$ is invertible and the original signal x can be recovered completely from its wavelet representation [Andén and Mallat, 2014].

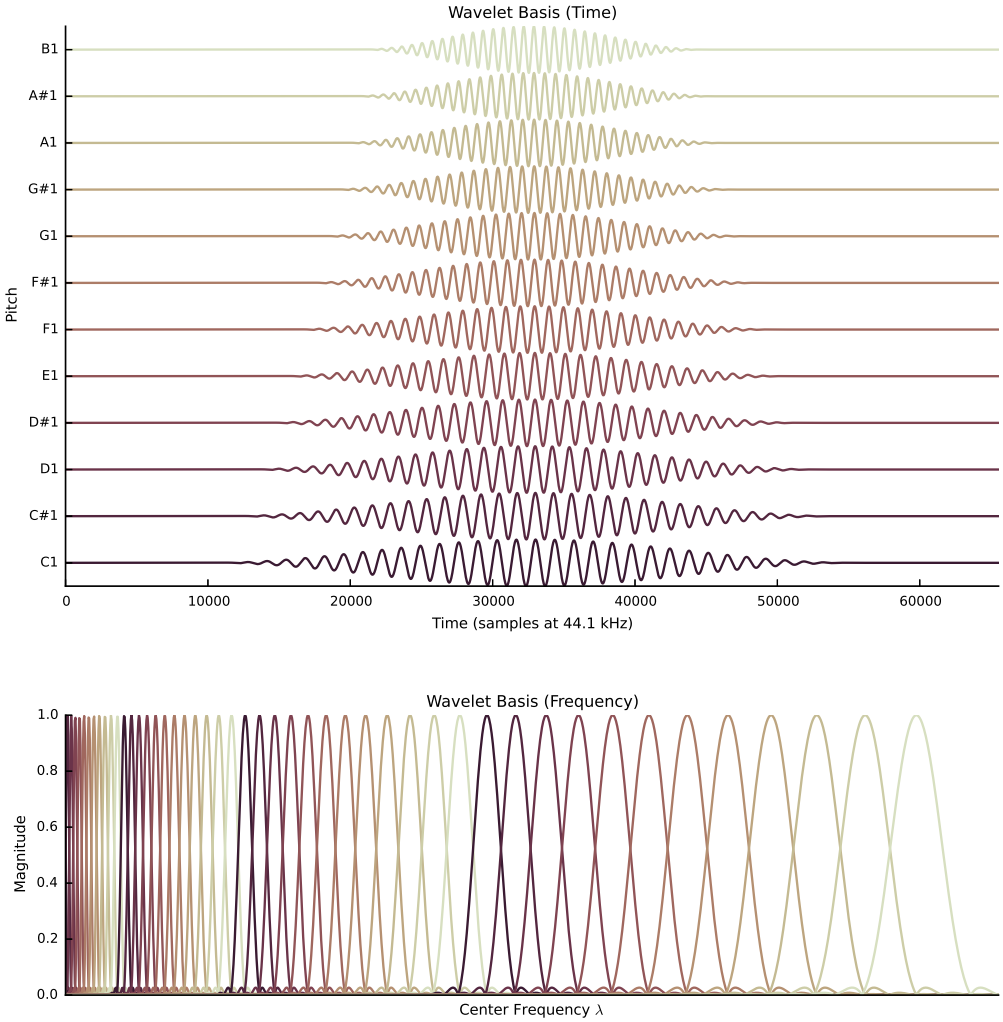

Figure 4: Morlet Wavelet basis ψ_λ. Top: Wavelets $\psi_\lambda(t)$ in time domain for one octave. Bottom: Wavelets $\hat{\psi}_\gamma(\omega)$ in frequency domain showing four octaves.

3.2 Time Scattering

The wavelet transform $|x * \psi_\gamma| * \phi(t)$ is a time-frequency representation of x in the frequency region supported by ψ_γ averaged over a temporal frame of size T. The low-pass filter ϕ removes high-frequency content, which subsequent convolution with a second wavelet bank ψ_{γ_2} recovers:

$$\mathbf{x_2}[t, \gamma_1, \gamma_2] = ||x * \psi_{\gamma_1}| * \psi_{\gamma_2}| \tag{3}$$

These coefficients encode the interferences of the signal with the wavelets ψ_{γ_1} and ψ_{γ_2}, capturing the temporal evolution of $|x * \psi_{\gamma_1}|$ in the frequency range covered by ψ_{γ_2}.

This motivates a scattering transform $\mathbf{S}_\nu$ as a cascade of wavelet transforms and modulus operations, where the order ν determines the depth of the scattering network. The scattering coefficients through a path of filters $\mathbf{p} = (\gamma_1, \gamma_2, \ldots, \gamma_\nu)$ is therefore given by

$$\mathbf{S}_\nu \mathbf{x}[t, \mathbf{p}] = ||\ldots||x * \psi_{\gamma_1}| * \psi_{\gamma_2}| * \ldots| * \psi_{\gamma_\nu}| * \phi(t) \tag{4}$$

where averaging in time by ϕ again provides invariance to translation up to the frame length T. The scattering cascade of filtering and non-linearities (modulus rectifications) can therefore be considered as a convolutional network, and as such approaches current models of the auditory cortex [Chi et al., 2005]. Second-order scattering coefficients are similar to the constant-Q modulation spectrogram as proposed by [Thompson and Atlas, 2003], and have proven effective for audio classification tasks [Andén and Mallat, 2014], texture analysis and computer vision [Bruna and Mallat, 2011], and analysis of temporal dynamics in fetal heart rates [Chudacek et al., 2014].

3.3 Joint Time-Frequency Scattering

The scattering transform decomposes each frequency band separately, and cannot therefore capture local structures across frequency. This limits the scattering transform's ability to encode deeper timbal structures such as frequency modulation or variable filters. By switching from one-dimensional wavelets in time to two-dimensional wavelets in time and log-frequency, we effectively treat the constant-Q spectrogram created by the wavelet transform $\mathbf{W}$ as an image.

The two-dimensional wavelet $\Psi(t, \gamma)$ is defined as the product of a time wavelet $\psi_\alpha(t)$ and a log-frequency wavelet $\psi_\beta(\gamma)$

$$\Psi(t, \gamma) = \psi_\alpha(t)\psi_\beta(\gamma) \tag{5}$$

The Fourier transform of $\psi_\alpha(t)$ is centered at α (a modulation frequency in Hertz), and the transform of $\psi_\beta(\gamma)$ is centered at β (in units of cycles per octave).

The second-order joint time-frequency scattering of x is therefore given as a two-dimensional convolution of the constant-Q spectrogram with subsequent temporal averaging by ϕ:

$$\mathbf{S_2}[t, \gamma_1, \gamma_2] = ||x * \psi_{\gamma_1}| * \Psi_{\gamma_2}| * \phi(t) \tag{6}$$

Joint time-frequency scattering was proposed in [Andén et al., 2015] and was inspired by the neurological auditory models proposed by Shamma and others in [Chi et al., 2005] as the "cortical transform", which decomposes the output of the cochlea (i.e. the constant-Q spectrogram) with two-dimensional Gabor wavelets.

3.4 Spiral Scattering

The joint time-frequency scattering representation, while able to retrieve deep timbral structures such as frequency modulation and transient behavior, does so ignorant of the harmonic structures of pitched sounds. A spiral scattering representation, which introduces an octave variable, has been developed in [Lostanlen and Mallat, 2015].

As before, where we wrap the CQT $\mathbf{X}[t, \gamma]$ into a chroma representation (Section 2.1), the spiral scattering representation rolls the log-frequency variable γ into a chroma spiral, making one complete rotation at each octave. γ then decomposes into an integer octave variable u and a pitch class $q \in [1, 12]$:

$$\gamma = Qu + q \tag{7}$$

where $Q = 12$ pitch classes per octave.

Another dimension is therefore added to the time-frequency wavelet $\Psi(t, \gamma)$ by splitting the log-frequency wavelet into a pitch-class and octave wavelet:

$$\Psi(t, \gamma) \rightarrow \Psi(t, q, u) = \psi_\alpha(t) \times \psi_{\beta^q}(q) \times \psi_{\beta^u}(u) \tag{8}$$

Per [Lostanlen and Mallat, 2015], the Fourier transform of the spiral wavelet $\hat{\Psi}(t, q, u)$ is centered at $(\alpha, \beta^q, \beta^u)$ and has a pitch chroma velocity of α/β^q octaves per second and a pitch height velocity of α/β^u octaves per second.

Separation of pitch into chroma height is not only driven by practices in Western music theory [Risset, 1969], but is supported by magnetic resonance imaging of the auditory cortex [Warren et al., 2003]. Given the centrality of the chroma representation to the task of chord recognition,

we use the spiral scattering representation as a jumping off point to scatter wavelets directly over the chroma representation $\mathbf{X}[t, q, u]$.

4 Haar Wavelets In Chroma Space

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.

4.1 Haar Wavelet Transform

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:

$$\forall u \in \mathbb{Z},\ \boldsymbol{\psi}[u] = \begin{cases} \frac{-1}{\sqrt{2}} & \text{if } u = 0 \\ \frac{1}{\sqrt{2}} & \text{if } u = 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

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 5 for various values of j and b.

After endowing them with the Euclidean inner product

$$\langle \boldsymbol{\psi_{j,b}} | \boldsymbol{\psi_{j',b'}} \rangle = \sum_{u=-\infty}^{+\infty} \boldsymbol{\psi_{j,b}}[u]\boldsymbol{\psi_{j',b'}}[u], \tag{10}$$

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

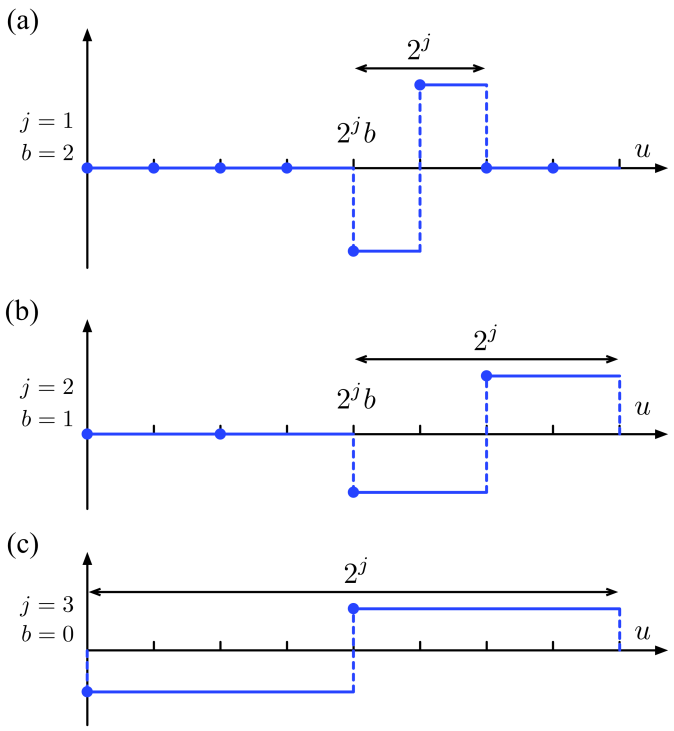

Figure 5: Three elements of the Haar wavelet basis $\{\boldsymbol{\psi_{j,b}}\}$ for various values of the scale index j and the translation index b. See text for details.

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

$$\mathbf{A_J}\boldsymbol{x} = \boldsymbol{x}[0] - \sum_{j,b} \langle \boldsymbol{x} | \boldsymbol{\psi_{j,b}} \rangle \boldsymbol{\psi_{j,b}}[0] = \sum_{u<K} \boldsymbol{x}[u] \tag{11}$$

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, 1989] develops a recursive scheme, called *multiresolution pyramid*, 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 $K + \frac{K}{2} + \cdots + 1 = \Theta(K)$.

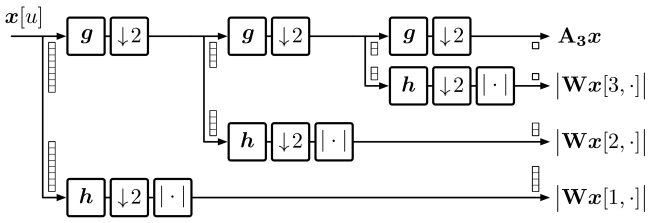

Figure 6: Discrete wavelet transform of a signal of length $K = 8$, as implemented with a multiresolution pyramid scheme. See text for details.

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

$$\mathbf{W}\boldsymbol{x}[j,b] = \left(\boldsymbol{h}_{\downarrow \mathbf{2}} \circ (\boldsymbol{g}_{\downarrow \mathbf{2}})^{(j-1)}\boldsymbol{x}\right)[b], \tag{12}$$

while the fully delocalized chroma representation rewrites as $\mathbf{A}_{\boldsymbol{J}}\boldsymbol{x} = (\boldsymbol{g}_{\downarrow \mathbf{2}})^J\boldsymbol{x}$. A flowchart of the operations involved in the wavelet transform is shown on Figure 6. We refer to chapter 7 of [Mallat, 2008] for further insight.

Since the low-pass filter $\boldsymbol{\phi}$ and the family of wavelets $\boldsymbol{\psi}_{\boldsymbol{j},\boldsymbol{b}}$'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:

$$\begin{aligned}(\boldsymbol{x} * \boldsymbol{g})[2b] &= \frac{\boldsymbol{x}[2b+1] + \boldsymbol{x}[2b]}{\sqrt{2}}, \text{ and} \\ (\boldsymbol{x} * \boldsymbol{h})[2b] &= \frac{\boldsymbol{x}[2b+1] - \boldsymbol{x}[2b]}{\sqrt{2}}.\end{aligned} \tag{13}$$

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 $(\boldsymbol{g}_{\downarrow \mathbf{2}}, \boldsymbol{h}_{\downarrow \mathbf{2}})$ 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 deferring 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.

Haar transform implementation	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.

4.2 Deep Haar Scattering

The wavelet modulus operator decomposes the variations of a signal at 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 first 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`, f_0 interferes with its perfect fifth `E` at the frequency $3f_0$. In contrast, in its third inversion labeled as `C:maj6`, 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.

In this regard, the wavelet modulus operator is neither fully invariant to octave transposition, nor does it retrieve the structure of musical chords beyond binary interactions between overlapping partials. Nonetheless, both of these desired properties can be progressively improved by cascading the wavelet modulus operator over increasing scales, until reaching the full support 2^J of the original signal $\boldsymbol{x}[u]$; a nonlinear decomposition known as the scattering transform [Mallat,

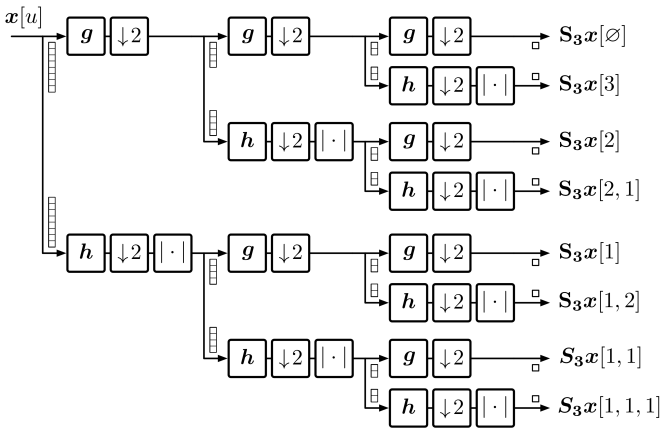

Figure 7: Deep scattering transform of a signal of length $K = 8$, as implemented with a multiresolution pyramid scheme. See text for details.

2012].

Considering that the Haar wavelet is analogous to a linear interferometer, the scattering transform is a recursive interferometric representation, whose recursion depth m varies according to the number of modulus nonlinearities encountered before reaching the scale 2^J. Whereas wavelet coefficients $\boldsymbol{W}\boldsymbol{x}[j, b]$ are indexed by a scale parameter j and a translation parameter b, scattering coefficients are indexed by sequences of scale parameters $(j_1 \dots j_m)$ called *paths*, and do not need a translation parameter since they are fully delocalized. The increasing scales in a scattering path correspond to the cumulative sum of integers j_1 to j_m. Therefore, the full sum $\sum_{n=1}^{m} j_n$ should be lower or equal to J. If it is strictly lower than J, a low-pass filtering with $\boldsymbol{\phi}$ is performed after the final wavelet modulus layer.

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 [Lostanlen and Mallat, 2015]. In addition, [Cheng et al., 2014] applies Haar scattering to the unsupervised learning of unknown graph connectivities.

Because 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, 2012] has proven that it is invariant to translation and stable to the action of small deformations. Along the octave variable u, translation corresponds to octave transposition, while small deformations correspond to variations in spectral envelope, such as those induced by a change in instrumentation or by polyphonic interference.

Haar scattering implementation	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.

Like the orthogonal wavelet transform, the scattering transform benefits from a multiresolution pyramid recursive scheme. By decomposing $\boldsymbol{x}[u]$ with subsampled quadrature mirror filters $\boldsymbol{g}_{\downarrow \mathbf{2}}[u]$ (low-pass) and $\boldsymbol{h}_{\downarrow \mathbf{2}}[u]$ (high-pass) over a full binary tree, and applying absolute value nonlinearity after each high-pass filtering, all K scattering coefficients are obtained after $\Theta(K \log K)$ operations and without allocating memory. A flowchart of the operations involved in the deep scattering transform is shown in Figure 7, and computational complexities are summarized in Table 2.

The scattering coefficient of path $(j_1, \ldots, j_m)$ is given in closed form by the following equation:

$$\begin{aligned} &\mathbf{S}_{\boldsymbol{J}} \boldsymbol{x}[j_1, \cdots, j_m] \\ &\quad = (\boldsymbol{g}_{\downarrow \mathbf{2}})^{\left(J - \sum_{n=1}^{m} j_n\right)} \bigcirc_{\sum_{n=1}^{m} j_n \leq J} \left| \boldsymbol{h}_{\downarrow \mathbf{2}} \circ (\boldsymbol{g}_{\downarrow \mathbf{2}})^{(j_n - 1)} \right| \boldsymbol{x}, \end{aligned} \tag{14}$$

where the circle symbol represents functional composition. Interestingly, the case $m = 0$ boils down to the sum across octaves $\mathbf{A}_{\boldsymbol{J}}$ already introduced in Equation 11, i.e. the chroma representation.

4.3 Representation Properties

The example chords discussed at the beginning of this paper in Figure 1 — `A:min7` (χ_1), `A:min7` up one octave (χ_2), and `C:maj6` (χ_3) — are played one after another on a piano and analyzed. Figure 8 shows all three features for this isolated chord sequence at $K = 4$ for visual simplicity.

In seeking to separate the feature profile of the `C:maj6` chord from the other two, we calculate the Euclidean distance between vectors at temporal frames in the middle of each chord activation. By maximizing the ratio $d(\chi_1, \chi_3)/d(\chi_1, \chi_2)$, the two `A:min7` chords are closer in the

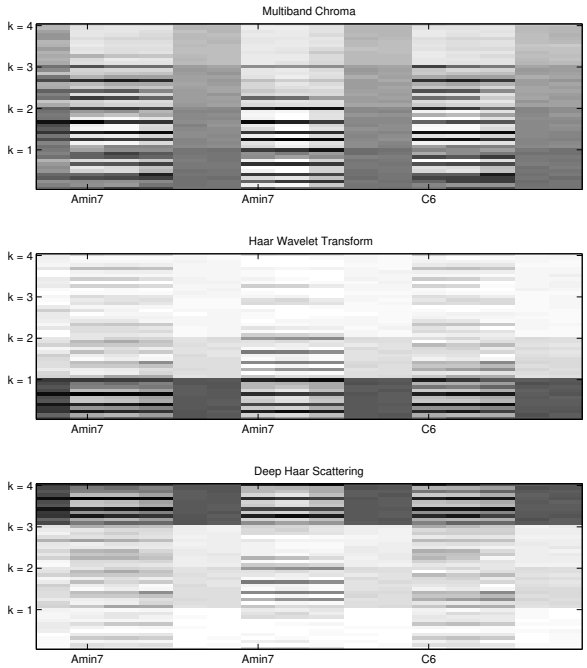

Figure 8: Features for chords in Figure 1 for $K = 4$: multiband chroma (top), Haar wavelet transform (middle), deep Haar scattering (bottom). See text for details.

K	Mode	Distance
4	Multiband	0.5653
	Wavelet	0.5920
	Scattering	0.5551
8	Multiband	0.5937
	Wavelet	0.6419
	Scattering	**0.6681**

Table 3: Ratio between d(χ_1, χ_3) and d(χ_1, χ_2), where the bigger ratio separates χ_1 and χ_3 while bringing χ_1 and χ_2 closer in the feature space for the given feature extraction method and K.

feature space while the `A:min7` and `C:maj6` are further.

Table 3 shows these distance ratios for our example chord progression. At scale $K = 4$ the wavelet transform wins, while scale $K = 8$ provides for higher disambiguation overall, with wavelet scattering separating χ_1 and χ_3 the most, further motivating the use of wavelet transforms for disambiguation of difficult chords with inversions.

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 and testing set of songs are kept constant across all experiments.

We consider a large vocabulary of chords with 13 different qualities: major, minor, minor 7^{th}, dominant 7^{th}, major 7^{th}, suspended 4^{th}, major 6^{th}, minor 6^{th}, suspended 2^{nd}, diminished triad, augmented triad, half-diminished 7^{th}, diminished 7^{th} — at all 12 roots, in addition to the null label `N`. The total number of classes in the extended vocabulary is thus equal to $12 \times 13 + 1 = 157$. For each experiment, a chord model and Viterbi transition probability matrix are generated from the training set with the band K 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).

After generating estimated chord labels for each song in the test set, Python scripts were written by the author to evaluate the results through the use of the mir_eval package [Raffel et al., 2014]. As per [Raffel et al., 2014], there is "no single right way to compare two sequences of chord labels," and mir_eval offers a broad range of metrics for automatic chord estimation. In this experiment we focus on two of these metrics: mirex, which "considers a chord correct if it shares at least three pitch classes in common" [Raffel et al., 2014], and tetrads_inv, which is much stricter and evaluates chord accuracy over the entire quality in closed voicing while taking inversions notated in the reference labeling into account.

K	Mode	mirex	tetrads inv
4	Multiband	**80.18** %	62.48 %
	Haar Wavelet	75.87 %	58.22 %
	Haar Scattering	74.38 %	56.47 %
8	Multiband	61.69 %	49.18 %
	Haar Wavelet	69.36 %	55.59 %
	Haar Scattering	68.78 %	55.44 %

Table 4: Overall accuracy for multiband chroma, Haar wavelet transforms, and deep Haar scattering at scales $K = 4$ and 8. Accuracies computed via mirex and tetrads with inversions metrics.

6 Results

Table 4 shows the accuracy of our automatic chord estimation system for all three feature extraction methods: multiband chroma, Haar wavelet transforms, and deep Haar wavelet scattering. Each method is computed for $K = 4$ and $K = 8$ streams. For multiband chroma, K refers to the number of bands in the representation, where $K = 4$ Gaussian windows cover the pitch space $\mathbf{X}[t, \gamma]$. For both wavelet transforms and wavelet scattering at scale $K = 4$, each pitch representation $\mathbf{X}[t, \gamma]$ is reduced to a 4-band multiband chroma representation, and a $J = \log_2(K) = 2$ wavelet/scattering transform is computed.

In Table 4, we see that the state-of-the-art $K = 4$ multiband results in the best accuracy under both mirex and tetrads_inv evaluation metrics. At $K = 4$, wavelet transforms and scattering suffer by roughly 5% overall for both mirex and tetrads_inv. Yet, at $K = 8$, wavelets and scattering both improve significantly on the multiband representation along both evaluation metrics. While all results for $K = 8$ are lower than their partners in $K = 4$, the Haar wavelets and Haar scattering representations certainly improve on multiband chroma when treating all octaves independently of each other. In the context of large vocabulary chord estimation, however, the vast majority of chords in our dataset are major, with minor chords more rare, and the rest of our chord qualities even rarer. This heavily skews these overall scores towards accuracy in determining major chords, and therefore a deeper analysis by chord quality is required.

Figure 9 shows accuracy by chord quality, filtering all reference labels on the given chord quality and evaluating chord estimation via mirex. Wavelet transforms and scattering improve on some rarer chord qualities for $K = 4$ (maj^6, min^6, sus^2, hdim^7) and take modest hits in the

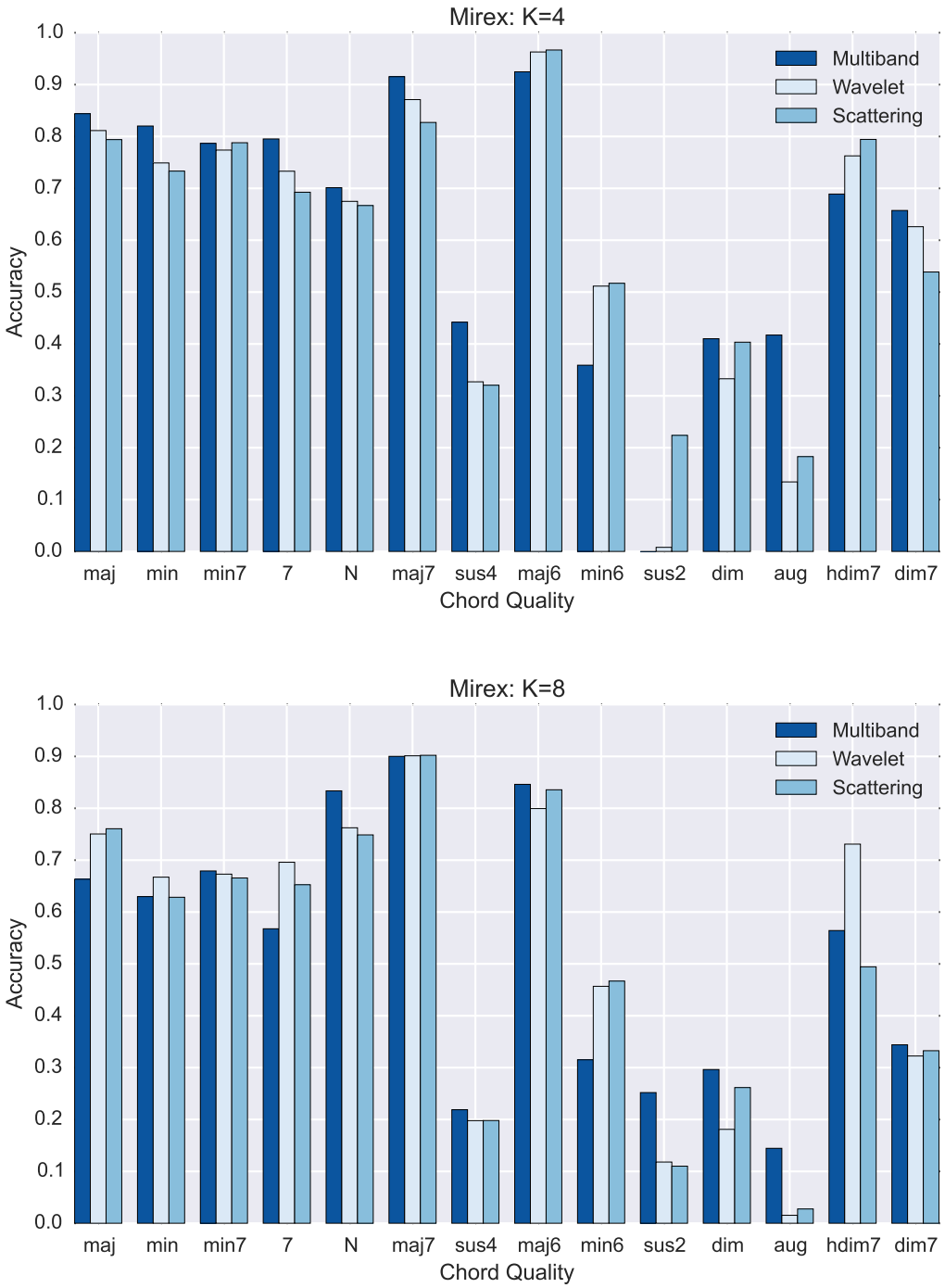

Figure 9: Multiband chroma, Haar wavelet transform, and deep Haar scattering compared for $K = 4$ (top) and $K = 8$ (bottom) streams. Chord accuracy computed via mirex.

more common chord classes. With $K = 8$, wavelet transforms and scattering actually improve on major and minor detection, as well as dominant 7^{th} and others. The mirex evaluation criteria is rather lenient for more complex chord qualities however, so we need to look at the stricter tetrads_inv metric.

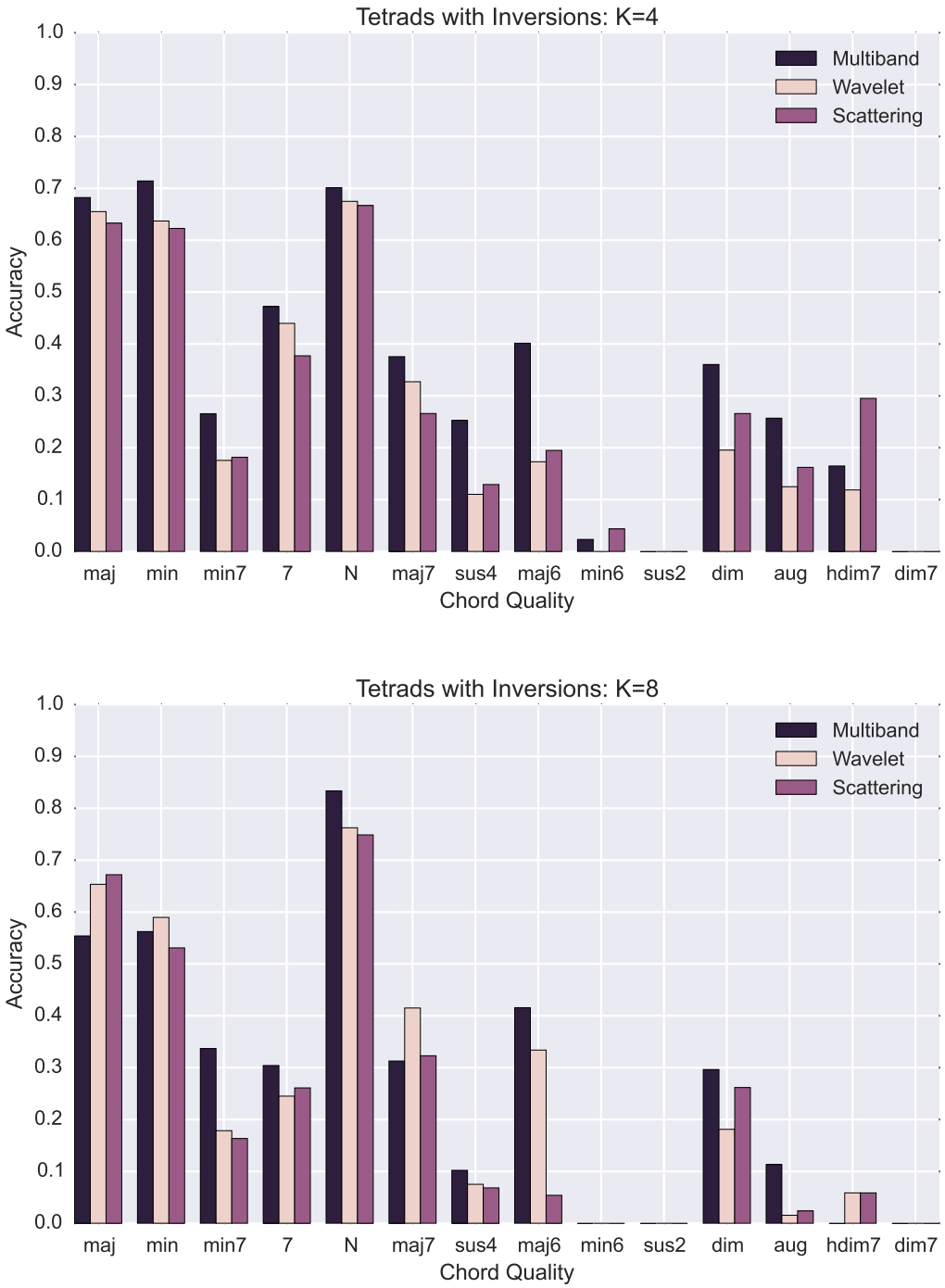

Figure 10: Multiband chroma, Haar wavelet transform, and deep Haar scattering compared for $K = 4$ (top) and $K = 8$ (bottom) streams. Chord accuracy computed via tetrads with inversions.

In Figure 10 we see accuracy by chord quality computed via tetrads_inv. For $K = 4$ streams our methods do not improve on the multiband chroma, though scattering performs slightly better for min[6] and hdim[7] qualities. Increasing scale to $K = 8$, however, we see both the wavelet

transform and scattering improve detection of major chords, while the wavelet transform provides some slight improvement to minor chords and major 7^{ths} as well.

7 Analysis

Section 6 shows that Haar wavelet representations do not improve the automatic chord estimation system overall, and, while providing some small improvements for some rarer chord qualities, provide chord estimation systems with slightly less accuracy across chord qualities. This section presents further analysis at many different points throughout the chord estimation system, not just final accuracy scores. Automatic chord estimation systems such as the one employed throughout this paper are complex and involve many moving parts — as such any alteration in one stage has ramifications all down the line.

7.1 Class Confusions

Figure 11 shows three confusion matrices – one for multiband chroma, one for the Haar wavelet representation, and one for Haar scattering, all with $K = 4$ coefficients. These confusion matrices compare the annotated "ground truth" chord for the current frame in the testing signal with the machine estimated chord quality. Python scripts written by the author are used to analyze the data and assemble the matrices, filtering out information about the chord root and inversions and simply comparing the quality. Columns are all normalized to remove the heavy skew of the dataset towards major chords.

Figure 11.a shows multiband confusion, which consists primarily of mistaking classes for major. Figures 11.b and 11.c show confusion for wavelet transform and scattering respectively, and also trend towards confusing classes for major. The multiband representation seems to lower confusion overall as it has a slightly stronger diagonal (estimated class matches annotated class) than the other two.

7.2 Inverted Annotations

The majority of chord samples in both our training and testing sets are not just major: they are also in root position. As the goal stated in the introduction to this paper was a feature extraction method that is invariant to joint octave transpositions and sensitive to chord inversions, its important to isolate the issue of chord inversions to see how our system performs on inverted data.

Any supervised learning system is only as good as the annotated data it is provided, and chord

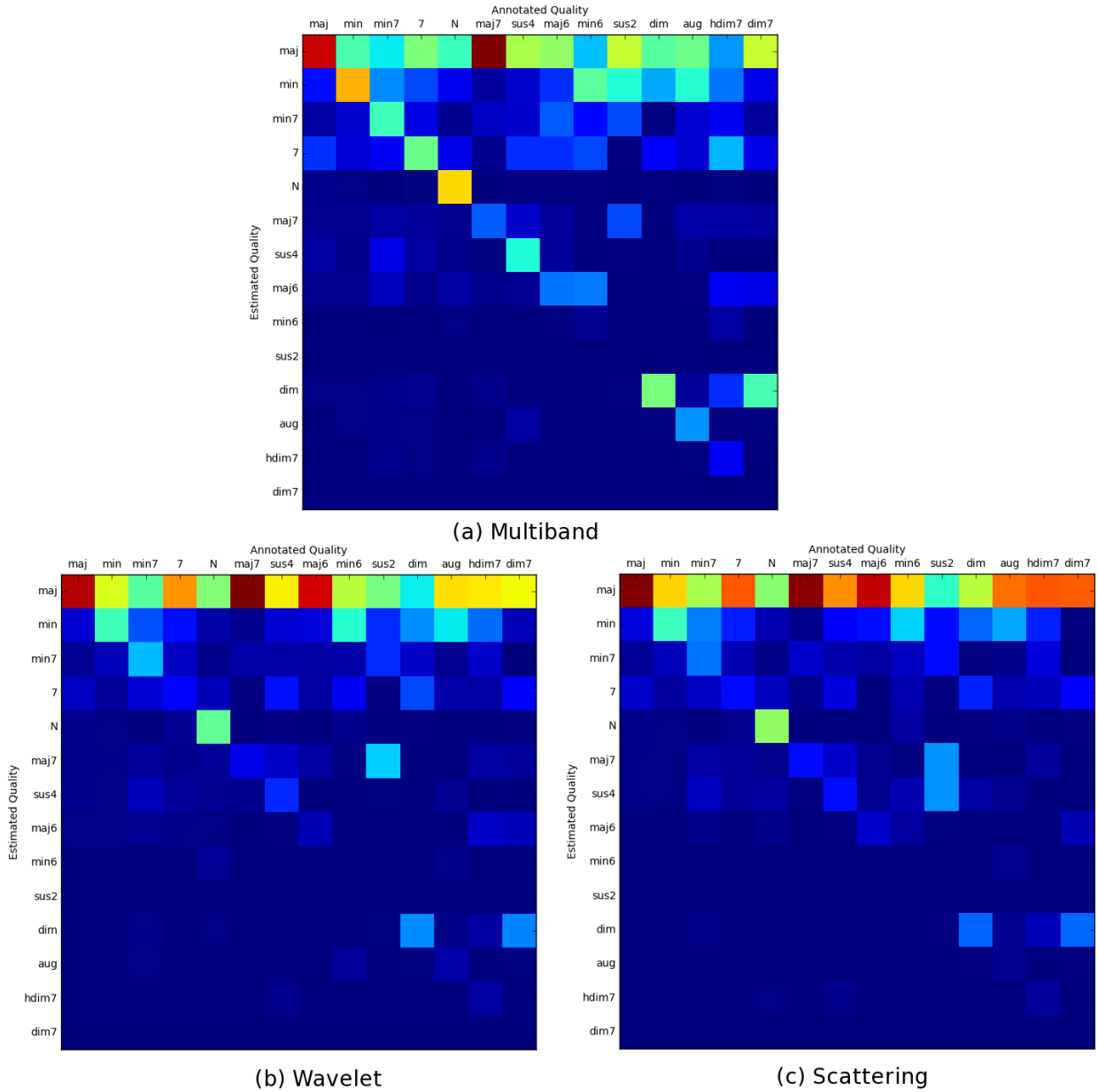

Figure 11: Chord quality confusions for (a) multiband, (b) Haar wavelet transform, and (c) Haar scattering representations. Columns are the annotated "ground truth" quality, rows are the machine estimated quality.

K	Mode	Accuracy
4	Multiband	9.14%
	Wavelet	**9.76%**
	Scattering	9.69%
8	Multiband	7.03%
	Wavelet	9.52%
	Scattering	8.73%

Table 5: Accuracy of multiband chroma, Haar wavelet transform, and deep Haar scattering on reduced testing set comprised solely of chords **not** in root position.

annotation is a notoriously ambiguous one. It is up to the human annotator to use their knowledge of and experience with music theory to disambiguate one chord label from another when confronted with the similar clusters of pitches — a task which no two experts will necessarily approach the same way. Some annotators can also be far more rigorous when it comes to labeling inversions than others.

Despite these shortcomings, Table 5 shows the accuracy of all three feature extraction methods on a subset of the testing set of frames only containing annotations with inversions. All data in our testing set is annotated along the guidelines proposed in [Harte et al.,] and thus filtering out all annotations in root position is simple. Wavelet and scattering operations increase accuracy, especially in the $K = 8$ region, though accuracy across the board is quite low. This is expected as many of these inverted annotations are for more complex chord qualities than simple major/minor triads, and are also rare (or at least left unannotated) enough to severely decrease our number of samples. That the wavelet and scattering representations perform better at both $K = 4$ and $K = 8$ indicates that they do, indeed, improve somewhat on multiband chroma for detection of the correct chord quality when chords are inverted.

7.3 Feature Normalization

As seen back in Figure 8, which shows the features for three different chords at $K = 4$, one can see that energy is far more spread out among the K chroma bands in the multiband representation as opposed to both the wavelet transform and wavelet scattering representations. Both techniques concentrate energy in coarse-scale bands and therefore affect pattern matching as bands with more sparse energy distribution vote less confidently for chord labels in the

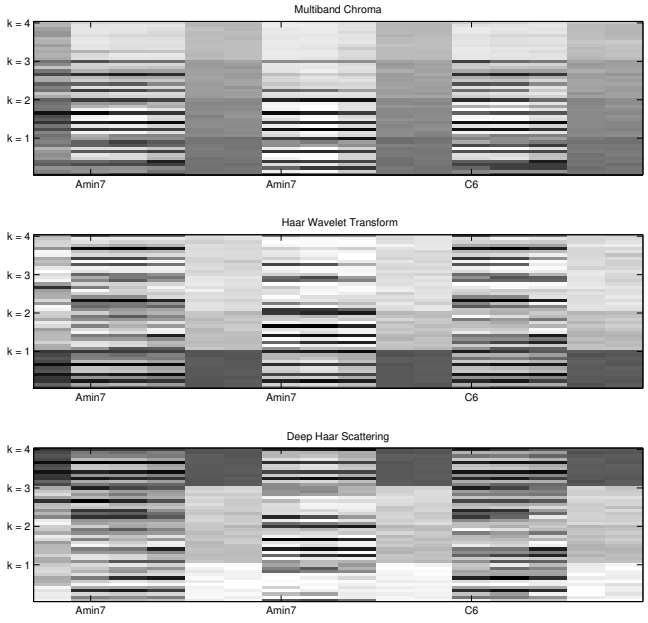

Figure 12: Normalized features from chords in Figure 8 for $K = 4$. Top: Multiband Chroma, Middle: Haar Wavelet Transform, Bottom: Deep Haar Scattering.

GMMs.

Figure 12 shows the normalized feature bands for the chords at the beginning of this paper (Figure 1), with each band normalized to unity energy. Band normalization shows an exceptionally modest, but consistent, positive effect on wavelet and scattering accuracy, ranging from gains of 0.1 to 0.3% overall accuracy along mirex and tetrads_inv evaluation.

7.4 K-Stream Hidden Markov Model Aggregation

Late fusion of the GMM probabilities plays a central role in our current chord estimation system, as each stream k is treated as independent and then combined via geometric mean after GMM likelihood estimation. Fusion by geometric mean, while sensible for multiband chroma, is more problematic for Haar wavelet based methods. Haar wavelet and scattering streams code for information at different scales and resolutions in each band rather than different regions at the same resolution as in the multiband representations.

Given a multiband chroma $\mathbf{Y}[t, q, k]$, each band along k is broken out and sent individually to the GMM for pattern matching along the chord model generated by the training stage. A likelihood matrix $\lambda[t, \chi]$ contains the likelihood of a given chord χ (‘`A:min7`’, for example)

		mirex			tetrads_inv		
K	Mode	geometric	arithmetic	max	geometric	arithmetic	max
4	Wavelet	75.87 %	76.70 %	76.67 %	**58.22** %	55.60 %	55.60 %
	Scattering	74.38 %	**76.92** %	76.91 %	56.47 %	55.35 %	55.34 %
8	Wavelet	69.36 %	70.99 %	70.41 %	55.59 %	52.72 %	52.26 %
	Scattering	68.78 %	71.10 %	71.08 %	55.44 %	52.93 %	52.84 %

Table 6: Overall accuracies for wavelet transform and deep Haar scattering coefficients at scales $K = 4$ and 8 with HMM fusion via geometric mean, arithmetic mean, and max voter "winner-takes-all". Accuracies computed via mirex and tetrads with inversions metrics.

at the frame t where each k stream is then combined via geometric mean:

$$\lambda[t, \chi] = \Big(\prod_k \mathbb{P}(\chi \mid \mathbf{Y}[t, q, k]) \Big)^{\frac{1}{K}} \tag{15}$$

$\lambda[t, \chi]$ is then sent to an implementation of the Viterbi algorithm along with the transition matrix (Figure 3) for decoding. Note that the representation $\mathbf{Y}[t, q, k]$ at this stage is any of either multiband chroma, wavelet transform bands, or deep scattering bands. Given the same maximum scale K, all representations are of identical dimensionality.

7.4.1 Stream Fusion

While geometric mean stream fusion provides good results for multiband chroma, in this section we explore two different methods for HMM fusion. In the wavelet and scattering cases, some streams have a nearly flat decision output $\mathbb{P}(\chi \mid \mathbf{Y}[t, q, k]) \; \forall \; \chi$, meaning that the band k is not confident about its vote for the most likely chord label.

One alternative fusion method is by combining the streams via arithmetic mean instead of the geometric mean:

$$\lambda[t, \chi] = \frac{1}{K} \sum_k \mathbb{P}(\chi \mid \mathbf{Y}[t, q, k]) \tag{16}$$

Another is to use a "winner-take-all" position and choose the kth stream with the most confident vote, i.e. the stream containing the maximum conditional probability $\mathbb{P}(\chi \mid \mathbf{Y}[t, q, k])$, and simply throwing out the other streams.

Table 6 compares overall accuracy for the wavelet and scattering methods with HMM fusion by geometric mean, arithmetic mean, and "winner-take-all" max voter aggregation. Evaluated along the mirex metric, arithmetic mean and max voter fusion improve results for all modes at all scales. This seems to indicate that one band is much more important than the others, as the minimization of their influence through arithmetic mean or simply throwing them out improves the mirex score.

However, evaluated along tetrads with inversions, both arithmetic mean and max voter lower accuracy. In all cases, the arithmetic mean fusion and max voter stream selection methods have very similar accuracy scores, indicating that the probability distribution in the most confident stream (the "max voter") is much more confident than all other bands, as their aggregation results in nearly the same accuracy as the most confident band taken by itself.

This is in stark contrast to the multiband approach whose accuracy plummets from an 80.18% mirex score down to 48.16% when using the max voter aggregation strategy, implying that each stream from the multiband chroma is casting informed votes and should be weighted equally.

Figure 13 shows a breakdown of estimation accuracy by chord quality for both mirex and tetrads_inv evaluation metrics. For simplicity, we show only the case where $K = 4$, and we only show HMM fusion by arithmetic mean since the max voter fusion is nearly identical. Comparing with the charts from Section 6, we see gains across the board in mirex for all chord qualities for both wavelet and scattering representations. In many cases (`min`, `7`, `sus4`) these representations catch up to the multiband representation's accuracy, while in others (`maj7`, `aug`, `dim7`) they overtake the multiband. The tetrads_inv metric, as usual, supplies lower accuracy scores across the board, but we do see some gains in some extended chord classes (`maj7`, `sus2`, `aug`, `hdim7`).

7.4.2 k-th Stream Voter Confidence

Who are the most confident voters, and how often are they the most confident? Table 7 shows the percentage of time that each band in the $K = 4$ case is the most confident voter — which is to say the band has the "peakiest" likelihood distribution across chords.

For the multiband representation, we see that the band corresponding to the lowest couple octaves in the CQT representation is the most confident voter most of the time. This is unsurprising, as the energy in lower octaves is more concentrated around the fundamental pitches of

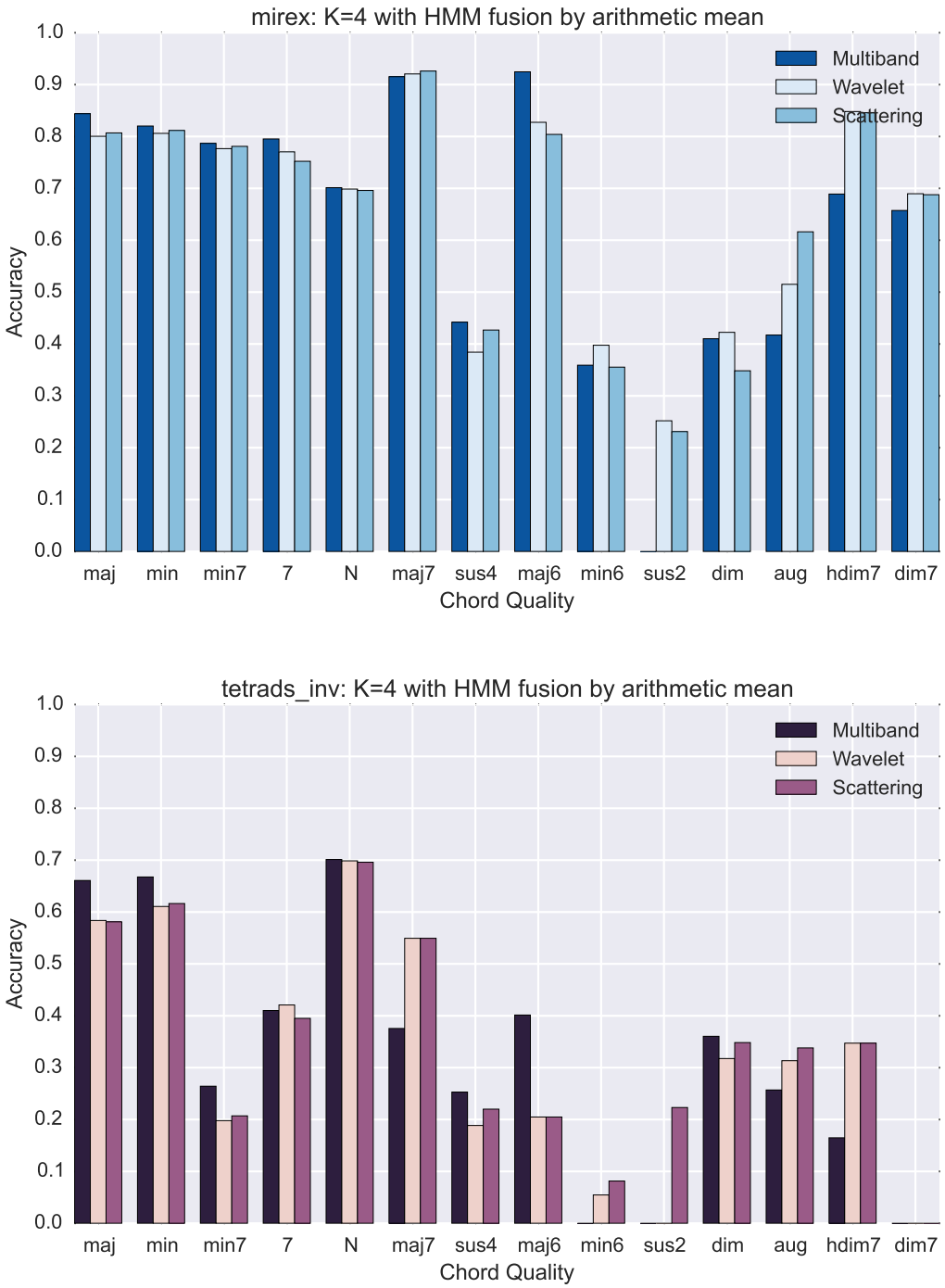

Figure 13: Accuracy for $K = 4$ for both mirex and tetrads_inv metrics, both with HMM fusion computed via arithmetic mean.

the chords, whereas their upper harmonics collide and smear out at higher octaves. As seen in Section 7.4.1, however, the lowest multiband stream is not a strong enough voter to stand on its own — fusion of all streams via geometric mean results in much improved accuracy.

In the wavelet transform and deep scattering modes, however, the max voters are overwhelm-

Mode	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Multiband	79.55%	14.77%	1.45%	4.23%
Wavelet	98.26%	0.04%	0.02%	1.67%
Scattering	0.04%	0.06%	0.18%	99.73%

Table 7: Voter stream confidence for $K = 4$ for all modes. Percent of the time each band k is the most confident voter (i.e. has the least flat probability distribution across chords).

ingly in the lowest ($k = 1$) and highest ($k = 4$) bands respectively[2]. Perhaps unsurprisingly, these are the bands of the respective methods that compute coarse-scale residual terms, i.e. the bands that overlap with the traditional chroma representation.

[2]For $K = 8$, these change accordingly to $k = 1$ for the Haar wavelet transform and $k = 8$ for deep Haar scattering.

8 Conclusions

Automatic Chord Estimation (ACE) systems comprising of beat-synchronous multiband chroma representations, GMM pattern matching, and k-stream HMM decoding have shown state of the art results for large vocabulary chord estimation [Cho and Bello, 2013]. This thesis develops a representation consisting of Haar wavelet transforms and scattering over the CQT in order to provide additional information for detecting extended and less-common chord classes. Crucially, these representations are invariant to joint octave transpositions and sensitive to chord inversions.

Our results do not yet show improved performance of the Haar wavelet transform or deep Haar wavelet scattering over the state of the art multiband chroma approaches for large vocabulary chord recognition. We do notice, however, that these two wavelet representations code for structures that multiband chroma does not when octaves are treated as independent ($K = 8$), and approach multiband results for $K = 4$ when considering alternative HMM fusion methods. Multiband and wavelet analysis differ significantly in that multiband chroma characterizes local information in K frequency regions while the wavelet representations are multiresolution approaches with iteration depth of $log_2(K)$. Perhaps the question is not whether to use one or the other but when to use both. The efficient computations for Haar analysis presented in Section 4 further motivate a complimentary approach.

As seen in Section 7.4.1, the fusion of k HMM streams plays a significant role in the outcome of the chord recognition system — a role whose effect is intricately tied to the feature extraction representation used. One as-yet unexplored avenue would forego independent HMM streams in the first place, and train a single GMM chord model on the full concatenated K band representation. For Haar wavelet and scattering representations this seems initially attractive, as each k band encodes information at different scales and resolutions. However, as the dimensionality of the representation starts to explode (a matrix of size $(12K \times T)$ where T is the number of frames in the signal), the discriminability of the GMM begins to seriously suffer.

This raises the question of whether or not the GMM is the correct classifier for pattern matching. GMMs undoubtably perform well when the multiband chroma is split up into bands of length $Q = 12$ and fused later on, but in the interest of concatenating features, perhaps a different classifier such as random forests should be used. Future work will focus on how different classifiers effect chord recognition systems fed with wavelet scattering representations.

References

[Andén et al., 2015] Andén, J., Lostanlen, V., and Mallat, S. (2015). Joint time-frequency scattering for audio classification. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE.

[Andén and Mallat, 2014] Andén, J. and Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62:4114–4128.

[Bello and Pickens, 2005] Bello, J. and Pickens, J. (2005). A robust mid-level representation for harmonic content in music signals.

[Bruna and Mallat, 2011] Bruna, J. and Mallat, S. (2011). Classification with invariant scattering representations. In *IVMSP Workshop, 2011 IEEE 10th*, pages 99–104. IEEE.

[Burgoyne et al., 2011] Burgoyne, J. A., Wild, J., and Fujinaga, I. (2011). An expert ground-truth set for audio chord recognition and music analysis. In *Proc. ISMIR*.

[Cheng et al., 2014] Cheng, X., Chen, X., and Mallat, S. (2014). Unsupervised deep Haar scattering on graphs. In *Proc. NIPS*.

[Chi et al., 2005] Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906.

[Cho, 2013] Cho, T. (2013). *Improved Techniques For Automatic Chord Recognition From Music Audio Signals*. PhD thesis, New York University.

[Cho and Bello, 2013] Cho, T. and Bello, J. P. (2013). Large vocabulary chord recognition system using multi-band features and a multi-stream hmm. In *Proc. MIREX*.

[Cho and Bello, 2014] Cho, T. and Bello, J. P. (2014). On the relative importance of individual components of chord recognition systems. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(2):477–492.

[Chudacek et al., 2014] Chudacek, V., Andén, J., Mallat, S., Abry, P., and Doret, M. (2014). Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study. *Biomedical Engineering, IEEE Transactions on*, 61(4):1100–1108.

[Daubechies, 1990] Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, 36(5):961–1005.

[Ellis and Poliner, 2007] Ellis, D. P. W. and Poliner, G. E. (2007). Identifying "cover songs" with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*.

[Fujishima, 1999] Fujishima, T. (1999). Realtime chord recognition of musical sound: A system using common lisp music. In *Proc. ICMC*.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441.

[Grosche and Muller, 2011] Grosche, P. and Muller, M. (2011). Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701.

[Harte et al.,] Harte, C., Sandler, M. B., Abdallah, S. A., and Gómez, E. Symbolic representation of musical chords: A proposed syntax for text annotations.

[Jacod and Protter, 2004] Jacod, J. and Protter, P. (2004). *Probability Essentials*. Springer-Verlag, Berlin.

[Lee and Slaney, 2006] Lee, K. and Slaney, M. (2006). Automatic chord recognition from audio using a supervised hmm trained with audio-from-symbolic data. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, pages 11–20, New York, NY, USA. ACM.

[Lostanlen and Mallat, 2015] Lostanlen, V. and Mallat, S. (2015). Wavelet scattering on the pitch spiral. In *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*.

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

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

[Mallat, 2012] Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398.

[Mauch and Dixon, 2010] Mauch, M. and Dixon, S. (2010). Approximate note transcription for the improved identification of difficult chords. In *Proc. ISMIR*.

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

[McVicar et al., 2014] McVicar, M., Santos-Rodrguez, R., Ni, Y., and Bie, T. D. (2014). 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.

[Moon, 1996] Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60.

[Papadopoulos and Peeters, 2007] Papadopoulos, H. and Peeters, G. (2007). Large-scale study of chord estimation algorithms based on chroma representation and hmm. In *Content-Based Multimedia Indexing, 2007. CBMI '07. International Workshop on*, pages 53–60.

[Pérez-Sancho et al., 2009] Pérez-Sancho, C., Rizo, D., and Inesta, J. M. (2009). Genre classification using chords and stochastic language models. *Connection science*, 21(2-3):145–159.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Raffel et al., 2014] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., and Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common mir metrics. In *ISMIR*.

[Risset, 1969] Risset, J. C. (1969). Pitch control and pitch paradoxes demonstrated with computersynthesized sounds. *The Journal of the Acoustical Society of America*, 46(1A):88–88.

[Sheh and Ellis, 2003] Sheh, A. and Ellis, D. P. (2003). Chord segmentation and recognition using em-trained hidden markov models. *ISMIR 2003*, pages 185–191.

[Thompson and Atlas, 2003] Thompson, J. K. and Atlas, L. E. (2003). A non-uniform modulation transform for audio coding with increased time resolution. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages V–397–400 vol.5.

[Warren et al., 2003] Warren, J. D., Uppenkamp, S., Patterson, R. D., and Griffiths, T. D. (2003). Separating pitch chroma and pitch height in the human brain. *Proceedings of the National Academy of Sciences*, 100(17):10038–10042.