Wavelet Scattering In Chroma Space

Christopher 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: TheNameOfYour2ndReader

March 28, 2016

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.

Acknowledgements

This is where you acknowledge those who have contributed to your work in some way.

I would like to thank my cat and my hamster for sticking with me through the sleepless nights.

Iis igitur est difficilius satis facere, qui se Latina scripta dicunt contemnere. in quibus hoc primum est in quo admirer, cur in gravissimis rebus non delectet eos sermo patrius, cum idem fabellas Latinas ad verbum e Graecis expressas non inviti legant. quis enim tam inimicus paene nomini Romano est, qui Ennii Medeam aut Antiopam Pacuvii spernat aut reiciat, quod se isdem Euripidis fabulis delectari dicat, Latinas litteras oderit?

Contents

List of Figures

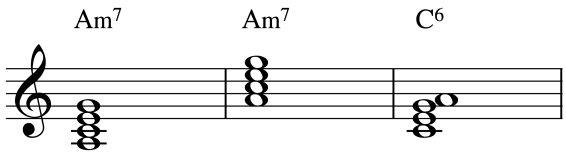

Figure 1: Three possible voicings of the pitch class set {C, E, G, A}, resulting either in the chord A:min7 or C:maj6. See text for details.

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 [**?**], genre recognition [**?**], and lyrics-to-audio alignment [**?**]. We refer to the review of [**?**] for a recent state of the art.

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 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 rarer 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 [**?**]. 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 first 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 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 [**?**], which holds state-of-the-art performance on the McGill Billboard dataset [**?**].

Section 2 describes the multi-band chroma features, as introduced by [?], and their 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 shows an example for the three representations while Section 6 presents and discusses the experimental setup along with the evaluation metrics for chord estimation accuracy. Section 7 presents the results of large vocabulary chord estimation comparing all three feature extraction methods and Section 8 summarizes our findings and presents ideas for future work.

2 Scattering Transform

3 Chroma Representations

3.1 Multiband Chroma

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 in [?].

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, [?] 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 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.

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, [?] train K end-to-end models in parallel 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 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 [?].

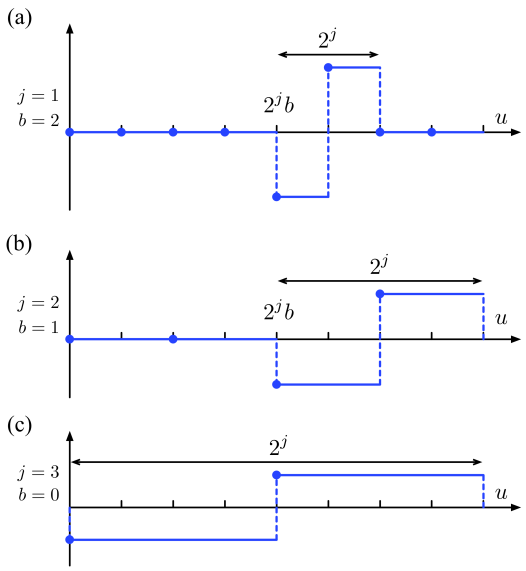

Figure 2: Three elements of the Haar wavelet basis $\{\boldsymbol{\psi_{j,b}}\}$ for various values of the scale index j and the translation index b. See text for details.

4 Haar Wavelets In Chroma Space

4.1 Haar Wavelet Transform

In spite of their success, the multi-band chroma features presented above do not comply with the assumption of statistical independence of the K-stream HMM, owing to the overlap between Gaussian windows. 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.

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:

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

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

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

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

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

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, [**?**] 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 $K + \frac{K}{2} + \cdots + 1 = \Theta(K)$.

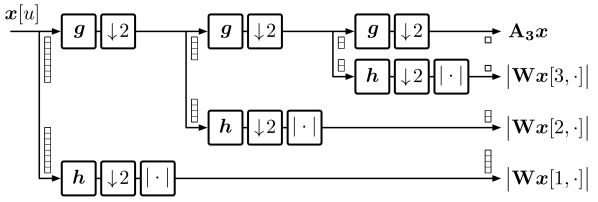

Figure 3: Discrete wavelet transform of a signal of length $K = 8$, as implemented with a multiresolution pyramid scheme. See text for details.

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

$$\mathbf{W}\boldsymbol{x}[j, b] = \left(\boldsymbol{h}_{\downarrow \mathbf{2}} \circ (\boldsymbol{g}_{\downarrow \mathbf{2}})^{\backslash} j - 1)\boldsymbol{x}\right)[b], \tag{5}$$

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 3. We refer to chapter 7 of [**?**] 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:

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

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

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 [**?**].

Considering that the Haar wavelet is analogous to a linear interferometer, the scattering transform is a recursive interferometric representation, whose recursion

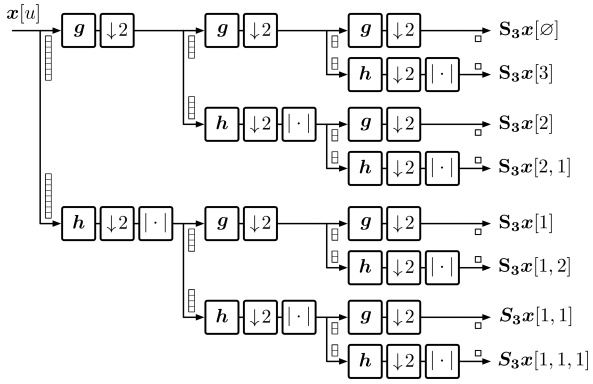

Figure 4: Deep scattering transform of a signal of length $K = 8$, as implemented with a multiresolution pyramid scheme. See text for details.

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 [**?**]. In addition, [**?**] 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, [?] 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.

Like the orthogonal wavelet transform, the scattering transform benefits from a multiresolution pyramid recursive scheme. By decomposing $\boldsymbol{x}[u]$ with subsam-

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

Table 1: 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.

pled 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 on Figure 4.

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

$$\begin{aligned}&\mathbf{S}_{\boldsymbol{J}}\boldsymbol{x}[j_1, \cdots, j_m] \\ &\quad = (\boldsymbol{g}_{\downarrow \mathbf{2}})^{\left(J - \sum_{n=1}^{m} j_n\right)} \bigcirc_{\sum_{n=1}^{m} j_n \leq J} \left| \boldsymbol{h}_{\downarrow \mathbf{2}} \circ (\boldsymbol{g}_{\downarrow \mathbf{2}})^{(j_n - 1)} \right| \boldsymbol{x}, \end{aligned} \tag{7}$$

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 4, 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 5 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

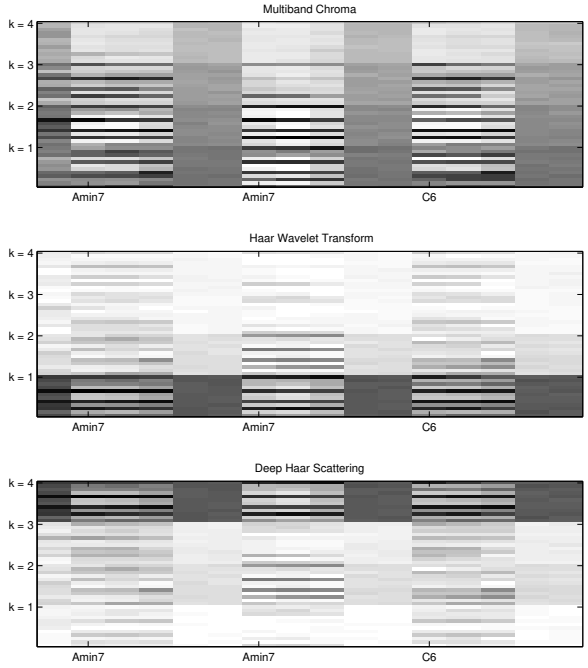

Figure 5: Features for chords in Figure 1 for $K = 4$: multiband chroma (top), Haar wavelet transform (middle), deep Haar scattering (bottom). See text for details.

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

Table 1 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 Automatic Chord Estimation

5.1 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 evaluate the results through the use of the mir_eval package [**?**]. As per [?], there is "no single right way to compare two sequences of chord labels," and mir_eval computes 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" [?], 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.

5.2 Results

Table 2 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 $K = 4$ windows the pitch space $\mathbf{X}[t, \gamma]$ with Gaussians covering approximately two octaves. 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

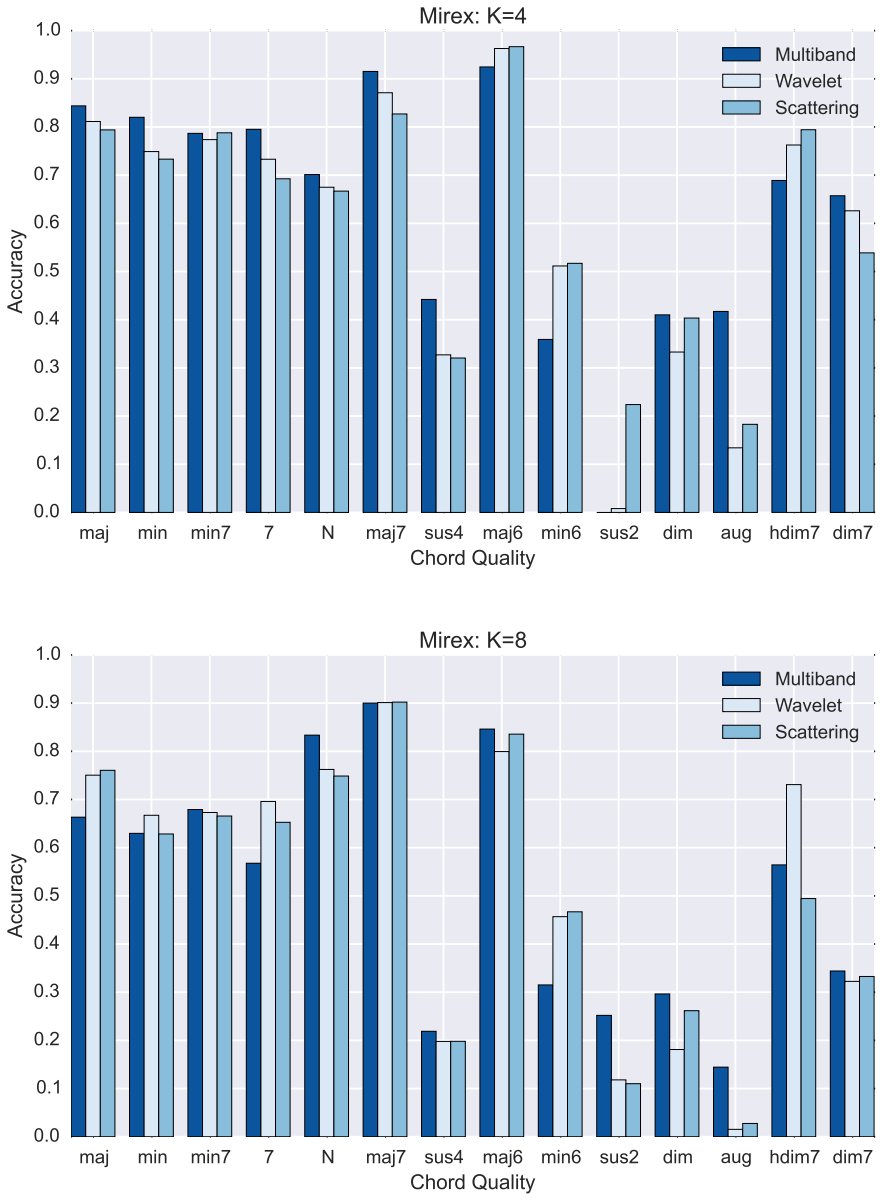

Figure 6: Multiband chroma, Haar wavelet transform, and deep Haar scattering compared for $K = 4$ (left) and $K = 8$ (right) streams. Chord accuracy computed via mirex.

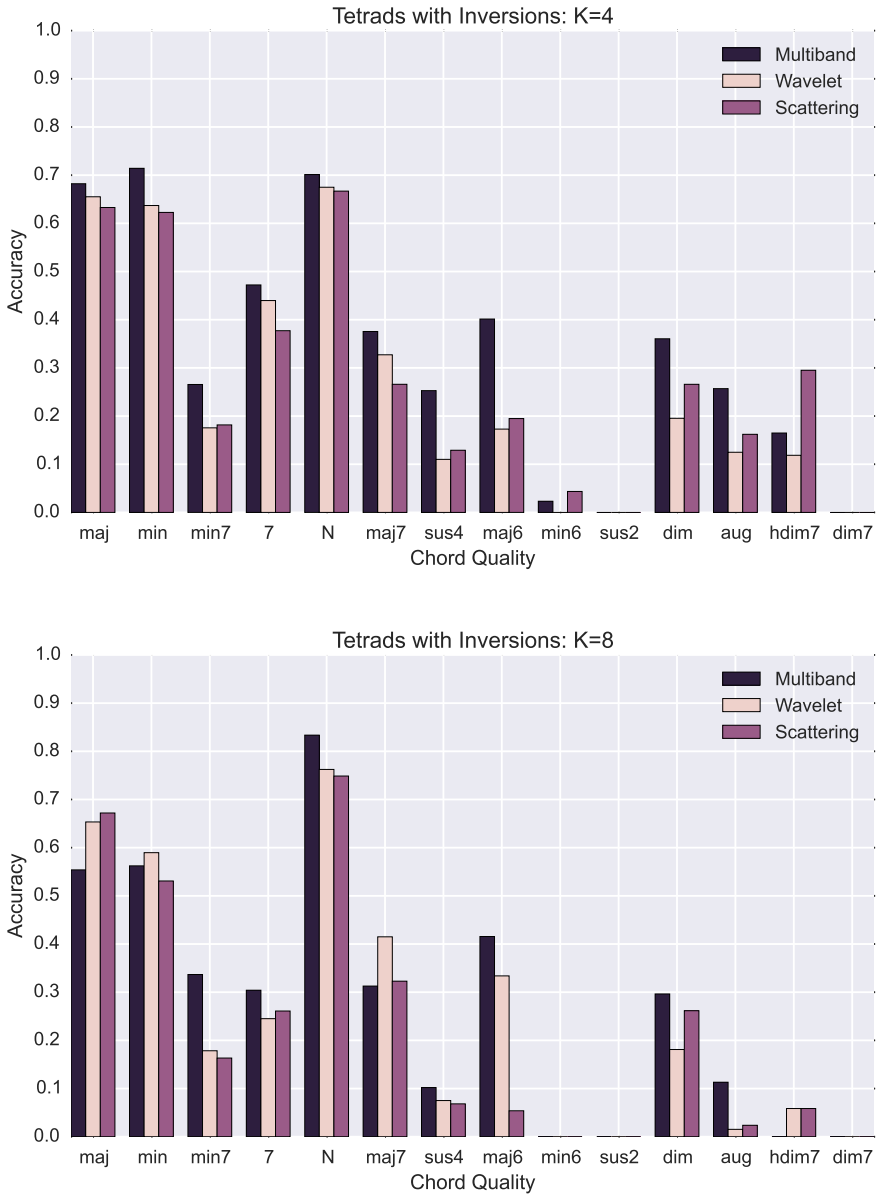

Figure 7: Multiband chroma, Haar wavelet transform, and deep Haar scattering compared for $K = 4$ (left) and $K = 8$ (right) streams. Chord accuracy computed via tetrads with inversions.

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 2: 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.

representation, and a $J = \log_2(K) = 2$ wavelet/scattering transform is computed.

In Table 2, we see that the state-of-the-art $K = 4$ multiband results in the best accuracy under both mirexand tetrads_invevaluation metrics. At $K = 4$, wavelet transforms and scattering suffer by roughly 5% overall for both mirexand 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 6 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 more common chord classes. With $K = 8$, wavelet transforms and scattering actually improve on major and minor detection, as well as dominant 7$^{\text{th}}$ and others. The mirexevaluation criteria is rather lenient for more complex chord qualities however, so we need to look at the stricter tetrads_invmetric.

In Figure 7 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.

6 Conclusions

Summarize your thesis here and add some relevant discussion.

6.1 This is a subsection

Contra quos omnis dicendum breviter existimo. Quamquam philosophiae quidem vituperatoribus satis responsum est eo libro, quo a nobis philosophia defensa et collaudata est, cum esset accusata et vituperata ab Hortensio. qui liber cum et tibi probatus videretur et iis, quos ego posse iudicare arbitrarer, plura suscepi veritus ne movere hominum studia viderer, retinere non posse.

6.1.1 This is a subsubsection

Sive enim ad sapientiam perveniri potest, non paranda nobis solum ea, sed fruenda etiam [sapientia] est; sive hoc difficile est, tamen nec modus est ullus investigandi veri, nisi inveneris, et quaerendi defatigatio turpis est, cum id, quod quaeritur, sit pulcherrimum. etenim si delectamur, cum scribimus, quis est tam invidus, qui ab eo nos abducat?

A Here is an Appendix

Here you can add any appendices you see fit.

B Here is another Appendix

Synephebos ego, inquit, potius Caecilii aut Andriam Terentii quam utramque Menandri legam?