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

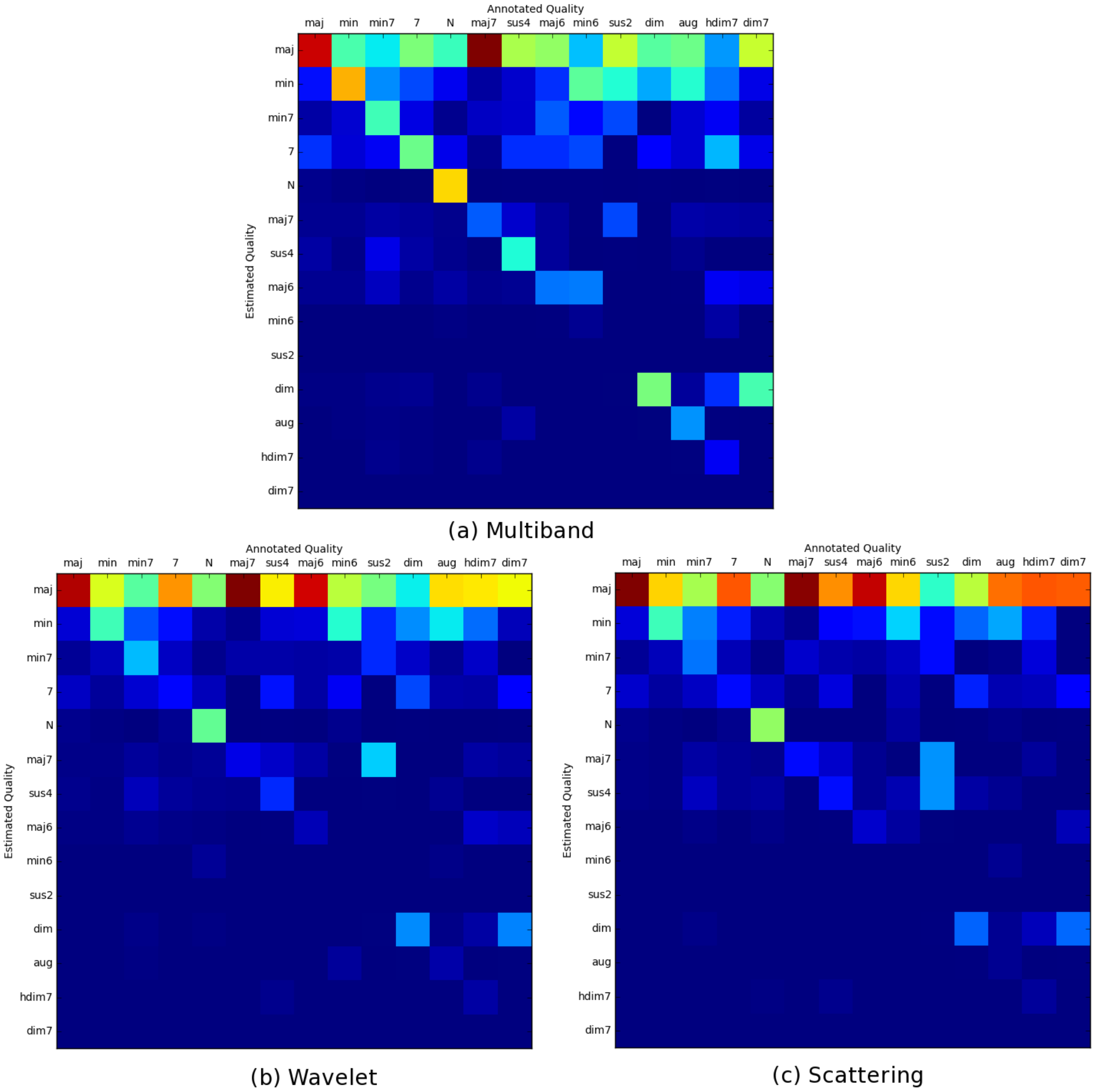
Figure 11 shows three confusion matrices – one for multiband chroma, one for the Haar wavelet representation, and one for Haar scattering, all with $K = 4$ coefficients. These confusion matrices compare the annotated "ground truth" chord for the current frame in the testing signal with the machine estimated 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



(a) Multiband

(b) Wavelet

(c) Scattering

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

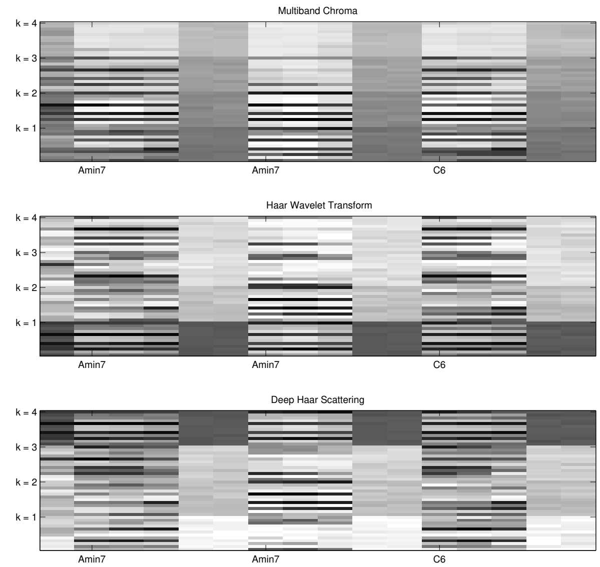


Figure 12: Normalized features from chords in Figure 8 for $K = 4$. Top: Multiband Chroma, Middle: Haar Wavelet Transform, Bottom: Deep Haar Scattering.

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

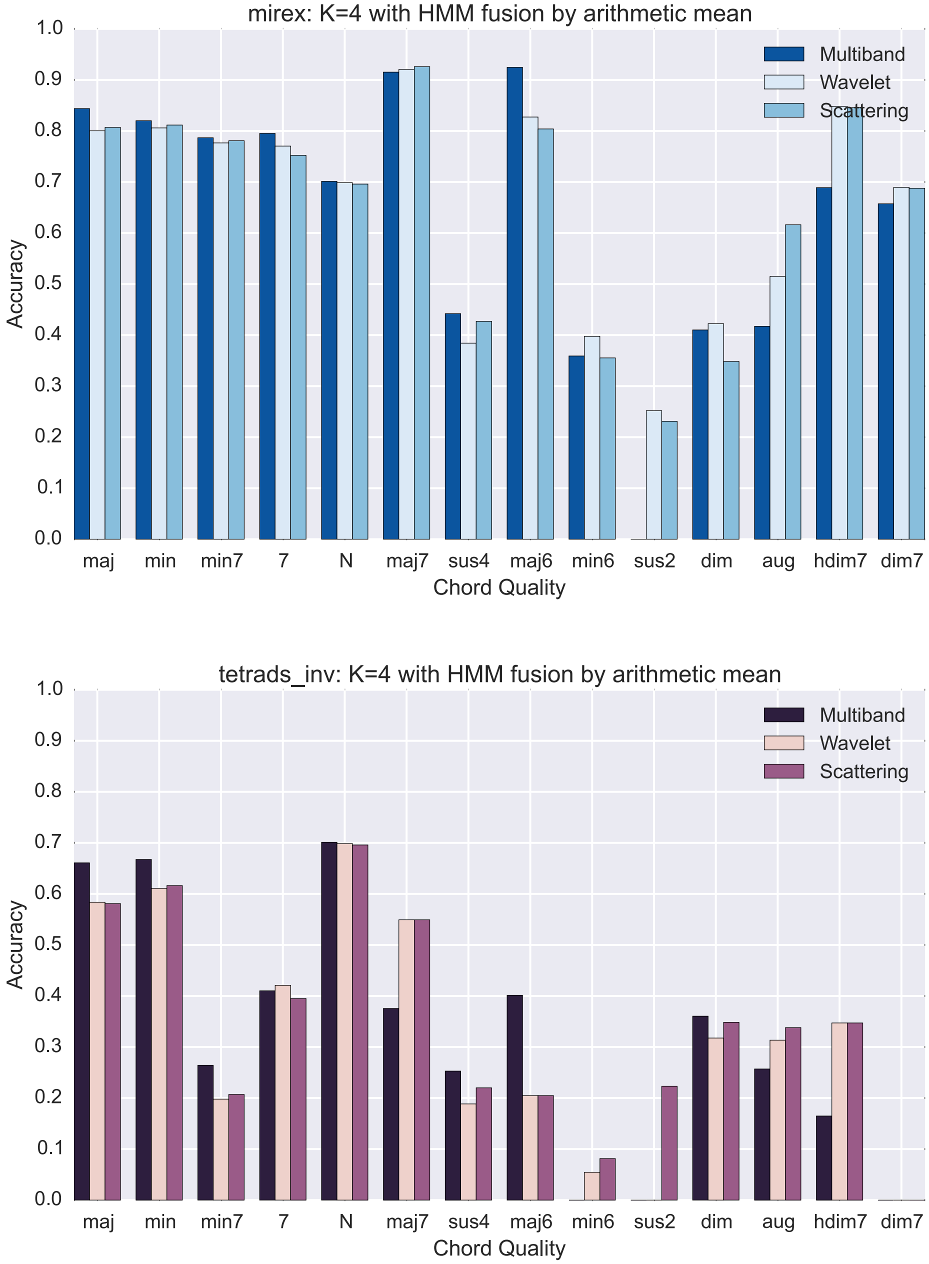


Figure 13: Accuracy for $K = 4$ for both mirex and tetrads_inv metrics, both with HMM fusion computed via arithmetic mean.

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 $k = 4$ for deep Haar scattering.