Key Estimation Using a Hidden Markov Model.

Conference Paper · January 2006

Source: DBLP

CITATIONS
33

READS
206

2 authors, including:

Mark Brian Sandler
Queen Mary, University of London
406 PUBLICATIONS **6,426** CITATIONS

SEE PROFILE

Key Estimation Using a Hidden Markov Model

Katy Noland, Mark Sandler
Centre for Digital Music,
Queen Mary, University of London,
Mile End Road,
London, E1 4NS.
katy.noland,mark.sandler@elec.qmul.ac.uk

Abstract

A novel technique to estimate the predominant key in a musical excerpt is proposed. The key space is modelled by a 24-state Hidden Markov Model (HMM), where each state represents one of the 24 major and minor keys, and each observation represents a chord transition, or pair of consecutive chords. The use of chord transitions as the observations models a greater temporal dependency between consecutive chords than would observations of single chords. The key transition and chord emission probabilities are initialised using the results of perceptual tests in order to reflect the human expectation of harmonic relationships. HMM parameters are then trained on a per-song basis using hand-annotated chord symbols, before the model for each song is decoded to give the likelihood of each key at each time frame. Examples of the algorithm as a segmentation technique are given, and its capability to estimate the overall key of a song is evaluated using a data set of 110 Beatles songs, of which 91% were correctly classified. An extension to include operation from audio data instead of chord symbols is planned, which will enable application to general music retrieval purposes.

Keywords: Key estimation, chords, harmony, HMM.

1. Introduction

The main key of a musical work, and the sequence of keys through which the music passes, are fundamental to music analysis. The home key serves as an anchor: the chord sequences in the music may lie within the home key, or may contain notes that are not part of the home key and therefore pull away from the anchor, suggesting other keys. It is the interplay between keys that gives harmonic interest to the music.

This paper describes a novel technique for estimating the key of a musical recording from chord symbols on both a frame-by-frame and a per-track basis. An estimate of the overall key of a track can be directly applied to music retrieval systems, as can a harmonic structure analysis derived from a frame-by-frame key estimate.

Our approach makes use of perceptual tests carried out by Krumhansl, described in [1], and outlined in Box 1. She uses the results in a key-finding algorithm ([1], Chapter 4), for which she used the probe tone profiles (see Box 1). The durations of each of the twelve possible pitches in the excerpt were summed to give a tone profile of the music. The correlations between the profile of the music and the twelve key profiles was then calculated, and the key with the highest correlation was taken to be correct. Performance was excellent on Bach, but the algorithm was less well able to cope with the extra chromaticisms of Shostakovich and Chopin.

Krumhansl's approach to key finding requires the music to be represented in symbolic form with note pitches and durations explicitly specified. Pauws [2] enables calculations from audio data by calculating a 12-bin chromagram for each frame, giving an energy measure for each of the 12 pitch classes. The chroma values were then used in the correlation calculations, in place of the profile of the music that Krumhansl derived from pitch durations. Gómez and Herrera [3] suggest a similar approach, but modify the probe tone profiles to emphasise pitches in the tonic, dominant and subdominant chords, and to take into account harmonics of pitches that will appear in the audio signal. Both papers suggest that analysis of the temporal structure of the music could improve accuracy.

Hidden Markov Models (HMMs) incorporate a degree of temporal dependency, and have been used to estimate tonality. Chai and Vercoe [4] estimate key changes in a musical excerpt using chromagram data as the observations for two HMMs. In the first model, each state represents a key pair (major and its relative minor); in the second, each state represents a mode (major or minor), and decoding both models gives both the root and the mode of the key. No training was carried out on the HMM parameters, and no preference was given to any key. They found that varying the probability of staying in the same key gave a trade-off between the precision of the keys extracted, and the recall and key accuracy.

Burgoyne and Saul [5] train a Dirichlet-based HMM on a pitch class profile (similar to a 12-bin chromagram) for each audio frame. Only the observation probabilities are

trained; the key transition probabilities are set according to a measure of relatedness derived from music theory. Their technique has been successful for extracting single chords, but in order to accurately extract the key a need for a more advanced harmonic model is identified.

Other related work includes that of Sheh and Ellis [6] and Bello and Pickens [7], who use HMMs to extract chord information from audio, and Temperley's Bayesian approach to symbolic key-finding, available at [8].

The algorithms described in this section have all successfully used prior musical knowledge to aid extraction of harmonic information, but all base their analysis on single chords. This paper introduces further temporal dependency into the task of key-finding by training an HMM on chord transitions rather than single chords, as well as making use of listening test results for initialisation, in order to represent expected relationships between chords and keys.

Section 2 introduces the model and explains the initialisation, training and decoding, Section 3 gives example analyses of two songs, Section 4 explains the overall key-finding method and gives the results, which are discussed in Section 5, and Section 6 concludes the paper.

2. Model of Key Space

The music-theoretic notion that a sounded chord sequence can strongly imply an underlying key, or allude to more than one key, fits well into the HMM structure, which consists of a set of underlying, unobservable states that emit observable data. Analysis of the observable data (chords) can give the most likely sequence of underlying states (keys), or the likelihood of each state at each time frame (relative importance of all keys over time). For an introduction to HMMs, together with a description of standard techniques for training and decoding, see [9] and [10]. For an introduction to music theory see [11] and [12].

So, we model tonality using a discrete HMM. Figure 1 shows a simplified diagram of the model. Only 3 keys are shown in the figure for clarity, but the 24 possible major and minor keys are included in the actual calculations. Each state represents a key, and the model is fully connected so that any key can move to any other key, or stay the same. At each time step the key generates an observable chord transition, for example C major to A minor.

It was decided that a pair of consecutive chords should be used for each observation, instead of a single chord, in order to extend the temporal dependency across a greater number of frames. The two chords that make up each chord transition can be any major, minor, augmented or diminished triad, or *no chord*, which occurs during silence or entirely percussive sections. More complex chords have been excluded from the model since chords that are not based on triads are very rare in the Western music repertoire for which this analysis is intended, and it was considered that the sequence of underlying triads, excluding extensions, is sufficient to define the key. All inversions of the same chord are treated identically, which means that the choice of bass note does not affect the estimated key.

BOX 1: PERCEPTUAL TESTS (KRUMHANSL)

More details of the following tests are given in [1], and numerical results are also given in the Appendix.

Probe tone profiles

Ten listeners were asked to judge how well each semitone fit within a given major or minor key context, on a scale of 1 (very good) to 7 (very bad). Average ratings were then calculated to represent the importance of each probe tone within a major and minor key context, giving a *probe tone profile* for each key.

Correlations between key profiles

The correlation between each pair of probe tone profiles was calculated to give a measure of how closely each pair of keys is related. The results are given in Table 2 in the Appendix.

Chord transition ratings

The chord transition rating test was used to give numerical values to how well a chord transition, or pair of chords, fits within a given key context. Listeners were asked to judge the fit on a scale of 1 to 7, and the mean rating was calculated. Ratings were given for all possible diatonic chord transitions within a major key, excepting the case where no transition is made, e.g. dominant-dominant or tonic-tonic. Table 3 in the Appendix shows the ratings.

Single chord ratings

For the single chord rating test listeners were asked to judge how well single chords fit within a key, using the same scale of 0 to 7. This test included all major, minor and diminished triads in both major and minor keys, and the average ratings are shown in Table 4 in the Appendix.

2.1. Initialisation

There are three HMM parameters that require initialisation.

2.1.1. Initial state probabilities

The initial state probabilities reflect any prior information that we may have about the most likely key, before any of the music has been heard. However, there is no reason to prefer any key above any other, so the initial probabilities for all states are set equally, to $\frac{1}{24}$.

2.1.2. State Transition Probabilities

The initial transition matrix should express how likely it is that when in a particular key, the music moves to another

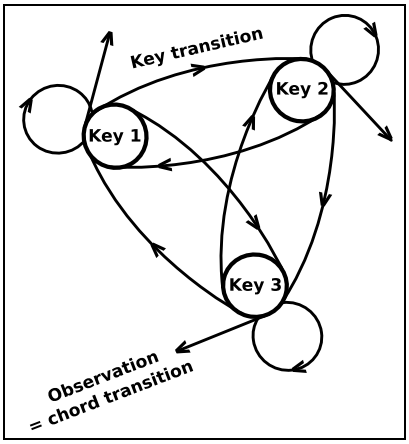

Figure 1. Simplefied diagram of the harmonic model.

key at the next time step. Intuitively it is most likely that the music will stay in the same key, and if it does change key it is most likely to move to one that is closely related and contains many of the same chords. The initial key transition matrix was created using the key profile correlations in Table 2 in the Appendix, which give numerical values to our intuitions. The values were circularly shifted to give the transition probabilities for keys other than C major and C minor; an operation that assumes G is to G major as C is to C major, etc.. The values were all made positive by adding 1, then they were normalised to sum to 1 for each key. This gave the final 24×24 transition matrix.

2.1.3. Observation Probabilities

The initial observation probabilities should reflect the human expectation of the key(s) implied by a certain chord transition. We are assuming that there is a strong correlation between the key implied by a chord transition, and the likelihood of the transition occurring in that key. This assumption is supported by Krumhansl [1] p. 195, and by finding the correlation between the chord transition ratings and the corresponding number of transitions present in our test data. Correlations of 0.39 for major keys and 0.22 for minor keys were found, both highly significant given the respective 40 and 154 degrees of freedom.

So, the chord transition B major to E major strongly implies the key of E major, since it forms a perfect cadence, so the probability of state E major emitting B-E will be very high. However, both chords are also contained in the key of B major, so the probability of state B major emitting B-E will be almost as high. Neither chord is contained in B♭ major, so the probability of state B♭ major emitting B-E will be very low.

The ratings for chord transitions within a key, given in Table 3 in the Appendix, were used to provide numerical values for the emission matrix. These only cover the diatonic chords of major keys, but the model includes all major, minor, augmented and diminished triads as well as the possibility of there being no chord, so some additional numerical values were required.

The pre-normalised probabilities for staying on the same chord, for diatonic chords, were taken from the ratings of individual chords within a key, given in Table 4 in the Appendix, and artificially boosted because repeated chords on either a frame-by-frame or beat-by-beat level are very likely. The approximate optimal increase was experimentally found to be 2. For example, the pre-normalised figure for emitting a transition from A minor to A minor in the key of C major was $3.62 + 2 = 5.62$. Pre-normalised values for transitions involving one or more non-diatonic chord are set uniformly low, to 1. For minor keys the ratings for major keys corresponding to the same scale degrees were used. The emission matrix was then normalised so that the observation probabilities summed to 1 for each key. The final emission matrix, then, had dimensions $(48 + 1)^2 \times 24 = 2401 \times 24$, since there are 48 possible chords and the possibility of *no chord* to form the chord transitions, in 24 possible keys.

2.2. Training

The expectation maximisation (E-M) algorithm, described in [9], was used to learn the HMM parameters for each individual song. If the observation probabilities, which model the relationship of each chord transition to each key, were subject to training, we could no longer be certain that the hidden states represent keys. To verify this, experiments were conducted with various combinations of HMM parameters trained.

The training data was a sequence of chord transitions, so for a chord sequence **Dm-Bdim-C** the first chord transition would be **Dm-Bdim**, and the second **Bdim-C**. For each key, each chord transition was given a numerical index from 1 to 2401. These were circularly shifted to give the values for other keys, with the exception of transitions involving a *no chord*, which has the same function in every key and so was kept at the end of the sequence.

2.3. Decoding

The Viterbi algorithm was used to find the most likely sequence of keys, and standard HMM decoding [9] was used to calculate the posterior state probabilities, giving the likelihood of being in any key at each time frame.

3. Sample Segmentation Results

The algorithm was tested on hand annotations of the start and end times of every chord in all of the first 8 Beatles albums, provided by Harte (see [13], [14]). Only simple triads were used: triad extensions were ignored, and non-triadic chords were mapped to the closest triad type according to their correlation with 4 chord templates, for major, minor, augmented and diminished chords. To simulate the kind of signal that would be obtained from audio data, with

a view to future extension of the algorithm to work from audio, the chord sequence was sampled at equal time intervals of 100 ms, such that sample times that fell between a chord start and end time were given the corresponding chord label, and any others were labelled **N**, for *no chord*.

Figures 2 and 3 show some examples of the results. The upper plots show the posterior state probabilities, and the lower plots show the most likely key sequence for the whole song. Ground truth for the key changes in the Beatles' songs is not available to our knowledge, but the figures show that the algorithm is capable of extracting meaningful structure. In *I'll Cry Instead* (see Figure 2, bottom) the two bridge passages in D major, at 42 s to 52 s and 72 s to 82 s, have been clearly separated. Similarly, in *I'm Happy Just to Dance With You* (see Figure 3, bottom) the choruses (C♯ minor) and verses (E major) have been extracted. The Beatles modified the final chorus such that the chords forming the transition back to E major are heard sooner than in previous choruses, then interrupted with C♯ minor again. This transition appears as a short E major section at about 103 to 105 s. This demonstrates one of the weaknesses of our approach, that although the chords were most closely related to E major, the key of E major was not firmly established. It is expected that this type of error would occur less frequently if the chords' position relative to the musical phrases were taken into account.

Inspection of the upper plots of Figures 2 and 3 reveals further structure that is not apparent in the hard key classification. For example, in the first E major section of *I'm Happy Just to Dance With You*, two consecutive verses are played. The hard classification of key shown in the lower plot cannot show the repetition, but the same period in the upper plot shows a pattern in the posterior state probabilities of approximately 16 s in duration that is repeated once.

Investigation of the algorithm as a musical structure extraction technique will be the subject of further research.

4. Evaluation Technique

In order to produce a quantitative evaluation of the key analysis algorithm, an experiment to test its ability to extract the overall key of a song was devised. A subjectively-assessed ground truth is available at [15], which gives a musicological analysis of the Beatles' songs. To determine the overall key, the output matrix containing the likelihood of each key at each time frame, such as those in the upper plots of Figures 2 and 3, was summed across the time domain, giving an overall likelihood for each key. The key with the largest likelihood value was taken to be the key of the song.

It should be noted that the ground truth often mentioned more than one key, in which case the first was taken to be the most important. Also several of the songs are modal, and do not directly fit this model of major and minor keys. Lydian and Mixolydian modes were treated as major, and Dorian and Aeolian as minor.

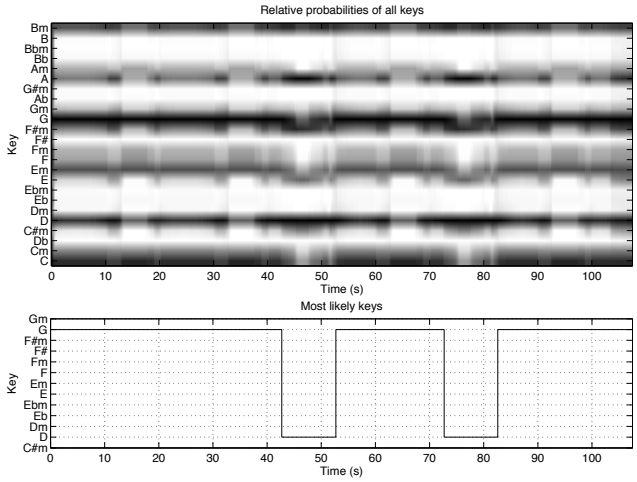

Figure 2. Probabilities of each key (top), and most likely key (bottom) for each frame for the Beatles *I'll Cry Instead*. Black indicates a high probability.

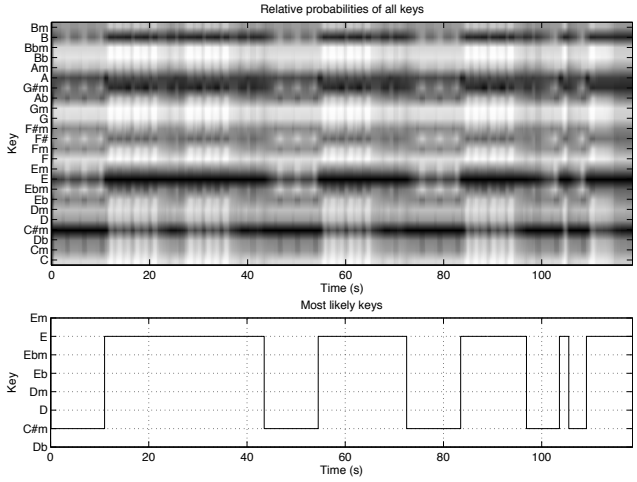

Figure 3. Probabilities of each key (top), and most likely key (bottom) for each frame for the Beatles *I'm Happy Just to Dance With You*. Black indicates a high probability.

Table 1 shows the percentage of correctly assigned overall keys for the 110 songs in the first 8 Beatles albums, with different HMM parameters trained.

Figure 4 shows the confusion matrix for the case where the transition and prior probabilities were trained, but the observation probabilities were not. Only the incorrect estimates are shown in the figure.

5. Discussion

The results in Table 1 verify the proposition that fixing the emission probabilities gives the most accurate representation of the song, since allowing adjustment alters the meaning of the hidden states. Training the prior state probabilities had little effect on the number of songs correctly classified. This is most likely due to the step where the key probabilities across the whole song were summed: the prior probabil-

Table 1. Percentage of songs correctly classified with varying training.

Probabilities subject to training			Percent
Prior	Transition	Emission	correct
yes	yes	yes	27
yes	yes	no	91
yes	no	yes	18
yes	no	no	87
no	yes	yes	28
no	yes	no	91
no	no	yes	18
no	no	no	87
Expected value for random choice of key			4

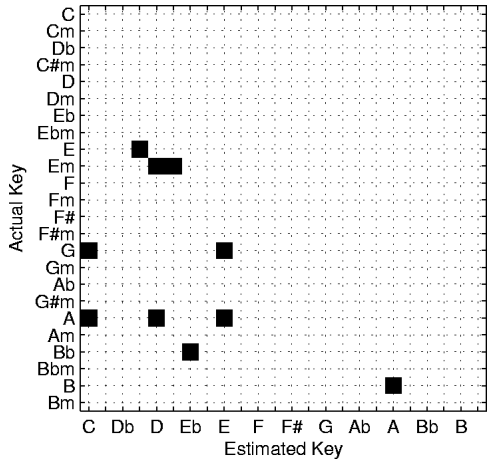

Figure 4. Confusion matrix for the case where prior and transition probabilities were trained. Only the incorrect estimates are shown. Minor keys follow their parallel major along the horizontal axis.

ities will only affect the first few frames and will therefore have limited effect on the overall key estimation. The suitability of the perception-based initialisation was confirmed by the case with no training, where 87% of songs were correctly classified. Training the transition probabilities for each song gave the optimum result of 91% of songs correctly classified.

The 91% accuracy for finding the overall key of Beatles songs is very encouraging. Closer inspection of the ground truth and confusion matrix reveal that all of the incorrect estimates can be explained, and none is unreasonably far from the ground truth.

Three of the modal songs were incorrectly classified: two Mixolydian songs were mistaken for the major key on their fourth degree, due to their flattened 7th, and one song with Dorian inflexions was mistaken for the major key on its 7th degree, due to its flattened 3rd and 7th. These errors are comparable to errors between relative major and minor keys.

One song in A major was mistaken for its dominant, E major. However, the chords that make up the song are B, E, A and D majors, which imply both keys equally when there is no context. One song in A major was mistaken for its subdominant, explained by the particular stress on the flattened 7th degree of the scale, used here to give a blues feel rather than a move to the subdominant key. These two cases would benefit from longer temporal dependencies in the model, based on phrase lengths, since it is usually the chord or cadence at the end of a phrase that defines the key.

The remaining five incorrect key estimates are for songs where more than one key is mentioned in the ground truth for the home key, and it is one of these alternative keys that has been selected by the algorithm.

6. Conclusion

An HMM initialised with results of listening tests has proved very successful for key estimation from chord symbols, and shown potential as a musical structure extraction technique. Working from chord symbols is intended as a means of testing the harmonic model without the problems associated with audio analysis. However, for most information retrieval purposes it is necessary to work with audio data to eliminate the painstaking task of hand annotation, so it is planned to extend the model to include audio-to-chord functionality. The 91% accuracy reported here will almost certainly not hold when working with audio, but we will have an understanding of how the audio-to-chord and chord-to-key errors differ. A more detailed exploration of segmentation possibilities is also planned.

References

[1] Carol L. Krumhansl, *Cognitive Foundations of Musical Pitch*, Oxford University Press, 1990.

[2] Steffen Pauws, "Musical key extraction from audio," in *Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona*, 2004.

[3] Emilia Gómez and Perfecto Herrera, "Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies," in *Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona*, 2004.

[4] Wei Chai and Barry Vercoe, "Detection of key change in classical piano music," in *Proceedings of the 6th International Conference on Music Information Retrieval, London*, 2005.

[5] J. Ashley Burgoyne and Lawrence K. Saul, "Learning harmonic relationships in digital audio with dirichlet-based hidden Markov models," in *Proceedings of the 6th International Conference on Music Information Retrieval, London*, 2005.

[6] Alexander Sheh and Daniel P. W. Ellis, "Chord segmentation and recognition using em-trained hidden Markov models," in *Proceedings of the 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA*, 2003.

[7] Juan P. Bello and Jeremy Pickens, "A robust mid-level representation for harmonic content in musical signals," in *Proceedings of the 6th International Conference on Music Information Retrieval, London*, 2005.

[8] David Temperley, "A Bayesian key-finding model," *MIREX Symbolic Key-Finding entry*, 2005, [web site], [2006 Jul 04],

Available: *http://www.music-ir.org/evaluation/mirex-results/articles/key_symbolic/temperley.pdf*.

[9] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, February 1989, vol. 77, no. 2.

[10] S. J. Cox, "Hidden Markov models for automatic speech recognition: Theory and application," *Speech and Language Processing*, 1990.

[11] Eric Taylor, *The AB Guide to Music Theory, Part I*, The Associated Board of the Royal Schools of Music (Publishing) Ltd., 1989.

[12] Eric Taylor, *The AB Guide to Music Theory, Part II*, The Associated Board of the Royal Schools of Music (Publishing) Ltd., 1991.

[13] Christopher Harte, "Chord tools & transcriptions," *Centre for Digital Music Software*, 2005, [web site], [2006 Jun 27], Available: *http://www.elec.qmul.ac.uk/digitalmusic/downloads/index.html#chordtools*.

[14] Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proceedings of the 6th International Conference on Music Information Retrieval, London*, 2005.

[15] Alan W. Pollack, "Notes on ... series," *soundscapes.info*, 2000, [web site], [2006 Jun 27], Available: *http://www.icce.rug.nl/∼soundscapes/DATABASES/AWP/awp-notes_on.shtml*.

Appendix: Perceptual test results

Table 2. Krumhansl's correlations between key profiles (see [1], p. 38).

	C Major	C Minor
C major	1.000	0.511
C♯/D♭ major	−0.500	−0.158
D major	0.040	−0.402
D♯/E♭ major	−0.105	0.651
E major	−0.185	−0.508
F major	0.591	0.241
F♯/G♭ major	−0.683	−0.369
G major	0.591	0.215
G♯/A♭ major	−0.185	0.536
A major	−0.105	−0.654
A♯/B♭ major	0.040	0.237
B major	−0.500	−0.298
C minor	0.511	1.000
C♯/D♭ minor	−0.298	−0.394
D minor	0.237	−0.160
D♯/E♭ minor	−0.654	0.055
E minor	0.536	−0.003
F minor	0.215	0.339
F♯/G♭ minor	−0.369	−0.673
G minor	0.241	0.339
G♯/A♭ minor	−0.508	−0.003
A minor	0.651	0.055
A♯/B♭ minor	−0.402	−0.160
B minor	−0.158	−0.394

Table 3. Krumhansl's chord transition ratings (see [1], p. 193).

First	Second Chord							
Chord	I	ii	iii	IV	V	vi	vii	Ave
I		5.10	4.78	5.91	5.94	5.26	4.57	5.26
ii	5.69		4.00	4.76	6.10	4.97	5.41	5.16
iii	5.38	4.47		4.63	5.03	4.60	4.47	4.76
IV	5.94	5.00	4.22		6.00	4.35	4.79	5.05
V	6.19	4.79	4.47	5.51		5.19	4.85	5.17
vi	5.04	5.44	4.72	5.07	5.56		4.50	5.06
vii	5.85	4.16	4.16	4.53	5.16	4.19		4.68
Ave	5.68	4.83	4.39	5.07	5.63	4.76	4.76	

Table 4. Krumhansl's harmonic hierarchy ratings for major, minor and diminished chords (see [1], p. 171).

	Context	
Chord	C Major	C Minor
C Major	6.66	5.30
C♯/D♭ Major	4.71	4.11
D Major	4.60	3.83
D♯/E♭ Major	4.31	4.14
E Major	4.64	3.99
F Major	5.59	4.41
F♯/G♭ Major	4.36	3.92
G Major	5.33	4.38
G♯/A♭ Major	5.01	4.45
A Major	4.64	3.69
A♯/B♭ Major	4.73	4.22
B Major	4.67	3.85
C Minor	3.75	5.90
C♯/D♭ Minor	2.59	3.08
D Minor	3.12	3.25
D♯/E♭ Minor	2.18	3.50
E Minor	2.76	3.33
F Minor	3.19	4.60
F♯/G♭ Minor	2.13	2.98
G Minor	2.68	3.48
G♯/A♭ Minor	2.61	3.53
A Minor	3.62	3.78
A♯/B♭ Minor	2.56	3.13
B Minor	2.76	3.14
C Dim	3.27	3.93
C♯/D♭ Dim	2.70	2.84
D Minor	2.59	3.43
D♯/E♭ Dim	2.79	3.42
E Dim	2.64	3.51
F Dim	2.54	3.41
F♯/G♭ Dim	3.25	3.91
G Dim	2.58	3.16
G♯/A♭ Dim	2.36	3.17
A Dim	3.35	4.10
A♯/B♭ Dim	2.38	3.10
B Dim	2.64	3.18