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

Influences of Signal Processing, Tone Profiles, and Chord Progressions on a Model for Estimating the Musical Key from Audio

Tonality analysis is an important part of studying music, and in recent years, automatic key estimation from audio input has become an important part of music information retrieval. The primary key is often given in the title of classical music compositions, together with some information about the structure or style, such as *Sonata in F* or *Scherzo in G*, which suggests that it is considered by composers to be an important means of classification. This article investigates the effects of three important aspects of many key-estimation algorithms: low-level digital signal processing parameters, tone-profile values, and harmonic progression through time, using a key estimation algorithm based on a Hidden Markov Model (HMM).

The terms "key" and "tonality" are often used interchangeably. In this article, we use "key" to refer to a single, discrete tonal center and its associated scale, with the acknowledgment that there may be simultaneous keys present in a given piece of music. We use "tonality" to refer to the more abstract concept of the music's relationship to all possible keys, which can be modeled as a position within some kind of psychologically informed geometrical tonal space, such as Chew's Spiral Array (Chew 2001).

We begin by explaining the parameters under investigation, and then we describe the technique using HMMs on which we base our investigations. We also explain how the model has been altered to test the importance of DSP parameters, tone profile values, and harmonic progression through time. We provide details of our experiments, which test the algorithm for global key estimation on a set of Beatles songs and a set of preludes and fugues from J. S. Bach's *Well-Tempered Clavier*. We then present the results, which strongly support the use of musically informed tone profiles, and which also suggest that limiting the frequency range is beneficial. Large variations in the results show that investigation into the low-level parameters is worthwhile. Finally, we summarize our findings and suggest some directions for further research.

Computer Music Journal, 33:1, pp. 42–56, Spring 2009

Some Important Issues in Tonality Estimation

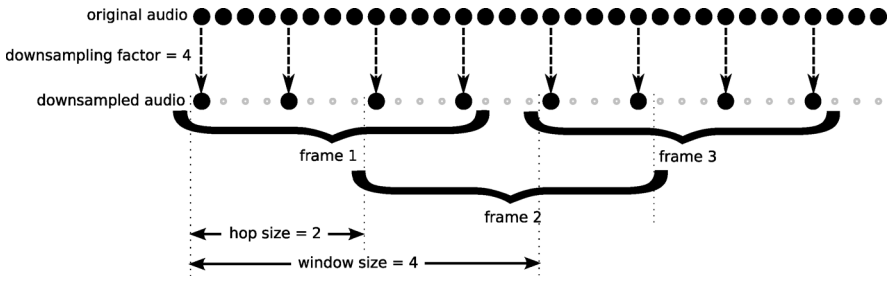
Several aspects of automatic tonality estimation are common to many published algorithms. We investigate three of these aspects in the context of one particular tonality-estimation method that is based on our previous work (Noland and Sandler 2006, 2007).

Low-Level Digital Signal Processing

To work with music in digital-audio form rather than in a symbolic representation such as MIDI, some low-level digital signal processing (DSP) is required to transform the raw audio into a form that is useful for tonality estimation. Most audio tonality-estimation systems use a logarithmically spaced frequency representation of the signal as their foundation, because these map elegantly to the notes of the equal-tempered chromatic scale.

Calculation of these features often starts with downsampling to reduce the amount of data and therefore the running time. However, it is well known that frequencies above half the sampling rate cannot be represented, so they must be removed before downsampling, which causes any high-frequency information to be lost. With real filters it is not possible to completely suppress the high frequencies—only to reduce their level—so

Figure 1. Illustration of the downsampling factor, window size, and hop size.



after downsampling, there will always be small components derived from them that appear at a lower and usually inharmonic pitch. This is known as aliasing. Downsampling also results in fewer samples per second, which means that the time resolution of the signal is reduced by the same factor as the sampling rate. For a more detailed introduction to DSP, see for example McClellan, Schafer, and Yoder (1998).

It is then necessary to divide the audio into frames, and decisions must be made regarding frame length (also called the window size) and hop size (which determines the overlap between frames). We must also decide whether to use a data-driven technique for calculating optimal frame sizes such as beat detection (Davies and Plumbley 2004) or tonal-change detection (Harte, Sandler, and Gasser 2006). Figure 1 illustrates the downsampling factor, window size, and hop size.

Such parameters are often chosen in a somewhat ad hoc fashion, with efforts concentrated on higher-level parameters, such as tone-profile values. However, we believe that the choice of DSP parameters can significantly affect the results, so we present an investigation into their effects.

Tone Profiles

A central idea to many tonality-estimation algorithms is the tone profile. This is a twelve-element vector in which each element represents one of the twelve semitones of the equal-tempered chromatic scale. The value of each element is proportional in some way to the importance of the pitch within a given key. It is generally assumed (with strong support from music theory and practice) that the tone profiles have rotational symmetry, such that the C element has the same value in C major as the G element in G major. This leads to two distinct profiles: one for major keys and one for minor keys.

There are different ways of arriving at the tone profiles, which can be divided into three broad categories: those based on music theory (Chai and Vercoe 2005; Gómez Gutiérrez 2006, p. 19), those based on cognitive studies (Krumhansl 1990; Aarden 2003), and those based on collecting statistics of real music (Gómez Gutiérrez 2006, p. 26; Noland and Sandler 2007). Temperley (2004) presents a comparison of selected tone profiles and their application to key finding from a symbolic representation of music.

The profiles based on music theory tend to be simple, such as the flat profiles used by Gómez Gutiérrez (2006, p. 19), where an element's value is set to 1 (or just over 14 in our normalized case) if the note is contained in the scale, and 0 otherwise, as shown in Figure 2.

The profiles derived from cognitive studies depend on having reliable listening test subjects and suitably generic test stimuli, and they have been criticized for measuring relationships that subjects have learned through music-theory training rather than relationships derived from subjects' unbiased judgments (Purwins 2005). They do, however, provide more information about the relative importance of pitches in a given key than do the flat profiles, and we show that they are more effective for key recognition. Figure 3 shows the profiles derived by Krumhansl using listener rating tests, in which listeners judged how well each tone fit with a key context, using a scale of 1–7 (Krumhansl 1990, p. 30).

The third approach to creating profiles is dependent on the corpus from which the statistics are

Figure 2. Flat tone profiles (normalized to add to 100).

Figure 3. Krumhansl's probe tone profiles (normalized to add to 100).

Figure 4. Statistical profiles derived from Bach (normalized to add to 100).

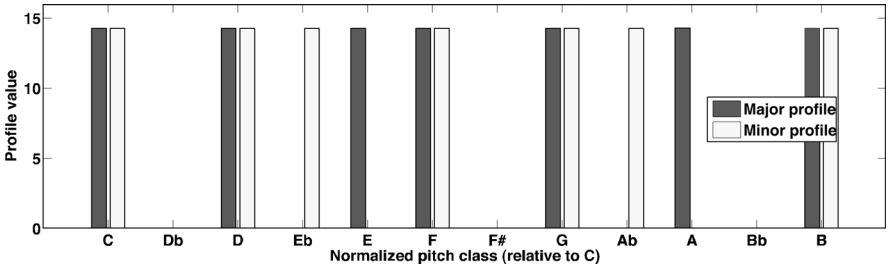


Figure 2

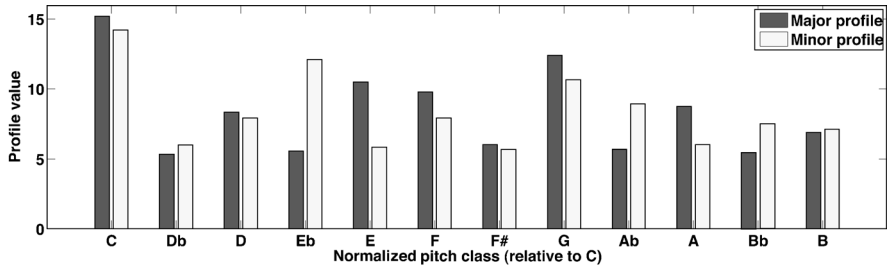


Figure 3

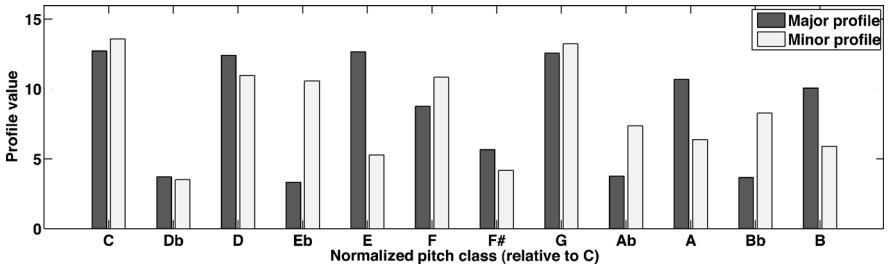


Figure 4

derived, and so a large and varied corpus is required for a general profile, or a style-specific profile can be produced from a more limited database. Figure 4 shows the profiles used by Noland and Sandler (2007) generated from Bach recordings. These do show some obvious similarities with the probe-tone profiles but exhibit greater differences between the major and minor profiles for the sixth, seventh, raised-sixth, and raised-seventh degrees of the scale, as well a much higher relative value for the second scale degree.

There is also some disagreement regarding the meaning of the profiles. Those derived from finding the pitch distribution in real music represent the likelihood of each pitch occurring within a given key, but those profiles derived from listening tests do not have such a clear meaning. Krumhansl shows that there is a strong correlation between her probe-tone profiles and the statistical distribution of pitches in a database of tonal music, and she later uses them in a key-finding algorithm based on the pitch distribution. However, Aarden (2003) shows that the probe-tone profiles are more closely related to the expectancy of pitches at the ends of phrases owing to the nature of the probe-tone test, and he derives an alternative profile that he argues represents the expectancy of pitches within a continuing phrase.

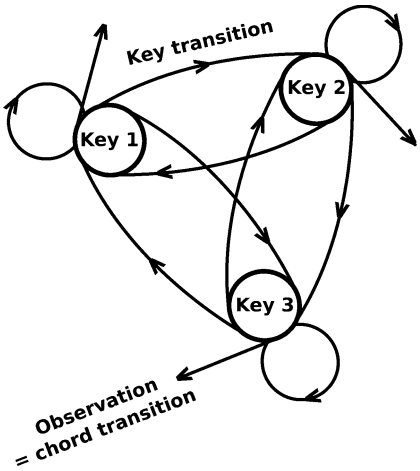
Tone profiles have been used in various audio key-finding algorithms. Pauws (2004) enables calculations from audio by calculating a twelve-bin chroma vector for each frame, thereby giving an energy measure for each of the twelve pitch classes. The correlation between the chroma values and Krumhansl's tone profiles was calculated in all twelve possible rotations, the highest correlation pointing to the most likely key. Gómez and Herrera (2004) suggest a similar approach, but they modify the tone profiles to emphasize diatonic pitches and to take account of upper partials in the signal. İzmirli (2005) uses a key-confidence measure to improve results.

There are also approaches that map the twelve-dimensional chroma vector onto a lower-dimensional space that is in some way musicologically relevant, such that harmonically close keys are located close together. These methods bear much similarity to the probe-tone techniques, in effect differing only in the way the distance from the key profile is calculated. Examples of these approaches include Chuan's extension to audio of Chew's center of effect generator (Chuan and Chew 2005) and the six-dimensional tonal model of Harte, Sandler, and Gasser (2006).

Harmonic Progressions Through Time

Good results for key finding have been obtained using note distributions without regard for their order or position within musical phrases (Gómez and Herrera 2004; Pauws 2004; Chuan and Chew 2005). However, we are interested in the notion of chord progressions, because we believe that understanding and modeling the sequence of chords through time can help define the tonal center. Previous approaches that embrace analysis of the progression through time of various musical features include Chai's and Vercoe's HMM-based technique (Chai and Vercoe 2005), which estimates key changes in a musical excerpt using individual frames of chromagram data as the observations for two first-order HMMs—one for the key note and one for the mode. Noll and Garbers (2004) have built a software tool, the HarmoRubette, which allows the adjustment of parameters of various similar HMM-based key-estimation techniques on a particular excerpt of music. Shmulevich et al. (2001) take a similar approach to estimate key based on tone relations in a monophonic melody. The latter two approaches, however, work with symbolic input and do not function on audio directly.

Figure 5. Simplified diagram of the harmonic model.



The Basic Model

Although the model used for the experiments in this article operates on audio input, it is based on our earlier model (Noland and Sandler 2006), which operates on a symbolic representation of musical chords. In this section, we review that model as a background to the subsequent section, which shows the model's extension to handle audio input.

In the basic model (and also in its extension to audio), tonal progressions are modeled by a discrete HMM, which consists of a set of underlying, unobservable states that emit observable data. (An introduction to HMMs is given by Rabiner 1989.) The model follows the work of Bello and Pickens (2005), who use a similar approach to extract chord-level harmonic features.

Figure 5 shows a simplified diagram of the tonal model. Each state represents a key, and each observation is a pair of consecutive chords or a chord transition, for example G major to E minor. Only three keys are shown in the figure for clarity, but the 24 possible major and minor keys are included in the

actual calculations. The model is fully connected so that any key can move to any other key or stay the same. The two chords that make up each chord transition can be any major, minor, augmented, or diminished triads, or "no chord," which occurs during silence or entirely percussive sections. All inversions of the same chord are treated identically, which means that the choice of bass note does not affect the estimated key.

Initialization of the Three Sets of HMM Parameters

All HMMs include three sets of probabilities that must be initialized, and these can later be adapted to better fit the observation data by a training procedure. These sets are the initial state probabilities, state-transition probabilities, and observation probabilities.

Initial State Probabilities

For each model state, the initial state probabilities define a probability value that represents the likelihood of the model's starting in that state at the first time frame. In the case of our model, they reflect any prior information that we may have about the most likely key before any of the music has been heard. Because we do not have any such information, the initial probabilities for all states are set to be equal.

State-Transition Probabilities

State-transition probabilities represent the likelihood of changing from any one state to any other; these are stored in the transition matrix. For our model, the initial transition matrix should express how likely it is that, when in a particular key, the music moves to another 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. The initial key-transition matrix was created using the key-profile correlations given by Krumhansl (1990, p. 38), which give numerical values to our intuitions.

Observation Probabilities

The observation probabilities (also called the emission probabilities) represent the likelihood, given that the model is in one state, of the model's emitting a particular observation. They are stored in the emission matrix. The initial observation probabilities for our model should reflect the human expectation of the key(s) implied by a certain chord transition. We assume that there is a strong correlation between the key implied by a chord transition and the likelihood of the transition's occurring in that key. This assumption is supported by Krumhansl (1990, p. 195), and by a z-test to measure the statistical significance of the correlation between the chord transition ratings and the corresponding number of transitions present in hand-annotated Beatles chords. (For an explanation of z-tests, see for example Downie and Heath 1974, pp. 224–226.)

The ratings given by Krumhansl (1990, p. 193) for chord transitions within a key 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 no chord, so some additional numerical values were required.

The probabilities of staying on the same chord were based on the chord ratings given by Krumhansl (1990, p. 171) but adjusted to be higher than the ratings for changing the chord, because it is most likely that from one frame to the next the chord will not change. (The typical length of a chord is several frames.) Values for transitions involving one or two non-diatonic chords were set uniformly low, corresponding to a value of 1 on Krumhansl's seven-point rating scale, or "fits poorly" within the key context. Krumhansl does not give ratings for chord transitions in minor keys, so the major-key ratings for the corresponding harmonic minor scale degrees were used.

Training

The expectation maximization (EM) algorithm, described by Rabiner (1989), was used to learn the

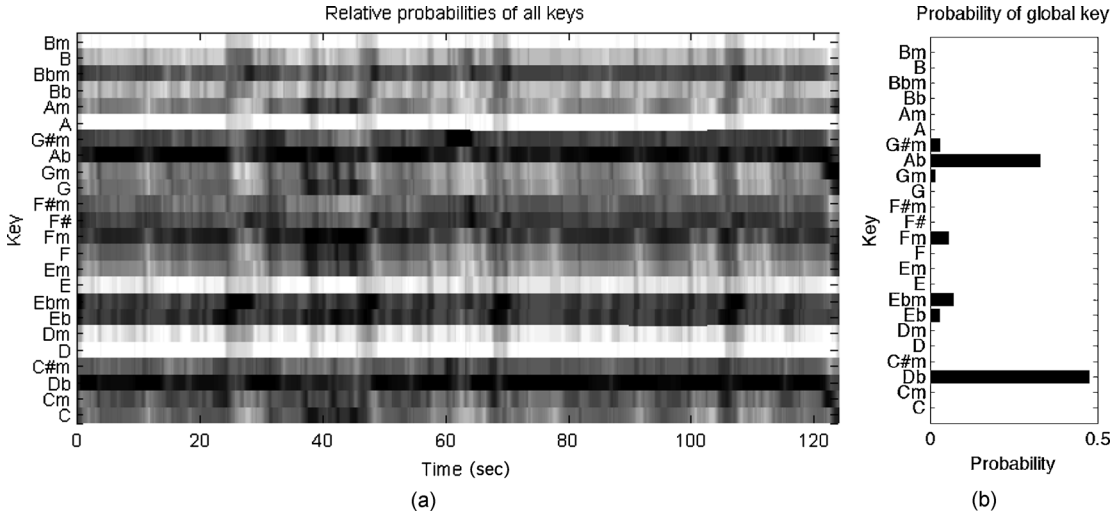


Figure 6. (a) Example of the posterior state probabilities, which give the relative key probabilities at each time frame (black indicates a strong key); (b) sum of the posterior state probabilities over time, which gives an estimate of the global key. The track is in D-flat major.

HMM parameters for each track, but the observation probabilities were fixed to ensure that the hidden states would always represent keys. The training data were chord transitions, which for the original experiments (Noland and Sandler 2006) were taken from hand annotations, but which are derived from audio in the present work (as described subsequently in "Extensions to the Basic Model").

Decoding

Standard HMM decoding (Rabiner 1989) was used to calculate the posterior-state probabilities, giving the likelihood of being in any key at each time frame. This gives a detailed view of how different keys interact through the track. For the purposes of global key estimation, the posterior-state probabilities were added across the time domain, and the key with the largest likelihood value was taken to be the key of the song. Figure 6 shows an example of the posterior-state probabilities together with their sum over time, for a track in D-flat major.

Extensions to the Basic Model

Results from the original model (Noland and Sandler 2006) were promising, with 91 percent correct key classification for the songs of eight Beatles albums. In this section, we describe the extension of that model to make it applicable to audio, and we further investigate the effects of DSP parameters, tone-profile values, and modeling harmonic progression through time.

Extension to Audio

For the model to work directly on audio, we start with a chord-recognition step. We chose the method described by Harte and Sandler (2005), which begins with a short-time constant-Q transform. This gives a representation of the signal in the frequency domain such that the frequency bin centers are logarithmically spaced and can therefore be made to correspond directly to the twelve semitones of the equal-tempered scale. We use 36 bins per octave to give a resolution of one-third of a semitone.

Brown (1991) gives a mathematical derivation of the constant-Q transform, which is defined as

$$X_{cq}(k_{cq}) = \frac{1}{N(k_{cq})} \sum_{n=0}^{N(k_{cq})-1} W(n, k_{cq}) x(n) e^{-2\pi j Qn/N(k_{cq})}$$

where X_{cq} is the constant-Q frequency representation of $x(n)$, k_{cq} is the frequency index, N is the window length (as a function of the frequency bin),

Table 1. Chord Templates Used by the Chord-Recognition Algorithm

Chord Type	*Template*
major	100010010000
minor	100100010000
augmented	100010001000
diminished	100100100000

The first element of each template corresponds to the root of the chord.

Q is the quality factor, and W is a variable-length Hamming window.

We investigate the effects of changing the minimum and maximum analysis frequencies (the range of k_{cq}), which define the window length. In addition, the efficient implementation used (Brown and Puckette 1992) converts the calculation into the frequency domain and allows a threshold to be set on the transform kernels below which all values are set to zero, so the number of necessary multiplications is reduced. We also investigate the effects of changing the kernel threshold value.

Chromagram and Chord Recognition

All octaves of the constant-Q spectrogram are added together to give a chromagram, which shows the energy at each one-third pitch class, regardless of the octave at which it was sounded. The chromagram is then tuned using the method described by Harte and Sandler (2005), to take into account recordings where A4 was not tuned to the standard 440 Hz. After tuning, the three bins corresponding to each pitch class are added to give a twelve-bin chromagram.

The tuned chroma vector for each frame is multiplied by all possible rotations of four binary chord templates that correspond to major, minor, augmented, and diminished chords, as shown in Table 1. The template and rotation giving the highest value are used to indicate the correct chord.

Harte and Sandler (2005) tested this audio-based chord-recognition algorithm on two of the Beatles albums, *Please Please Me* and *Beatles for Sale*, and a recognition accuracy of 62.4 percent was reported. The chord-recognition accuracy will directly affect the key estimation, so we expect that key estimates will be less accurate when working from audio than when working with symbolic input. However, the mechanics of the HMM mean that performance should not be degraded as much as for a simpler key-estimation method, because the emission probabilities allow unusual chords to be played occasionally within a given key. In addition, a single incorrect chord estimate that is harmonically close to the correct chord is likely to still produce a valid chord sequence within the key of the song.

Initialization Parameters

To investigate the importance of the initialization parameters, we tested the model using four different sets: those described by Noland and Sandler (2006) that are derived from the Krumhansl probe-tone experiments *(probe tone 1)*, a variation on probe tone 1 that differentiates chord transitions with one non-diatonic chord and those with two non-diatonic chords *(probe tone 2)*, a flat version, and a random version. For each type of initialization (probe tone 1 or 2, flat, or random), we change all three initialization parameters (initial state, transition, emission) together to give a complete model based on one profile type.

The emission matrix of probe tone 2 differs from probe tone 1 in its treatment of non-diatonic chord transitions. Transitions with only one non-diatonic chord are given a probability twice as high as those with two non-diatonic chords, equivalent to a value of 2 in Krumhansl's rating scale. Otherwise, it is identical to probe tone 1. For the flat emission matrix, chord transitions are initialized with probability 0.99 when both chords are within the key and 0.01 otherwise, then normalized to add to 1 for the key. Similarly, for the flat transition matrix, the probability of staying in the same key was initialized to 0.99, and that of changing key to 0.01; then, the values are normalized. The elements of the random transition matrix were taken from a uniform probability density function and normalized.

Table 2. Parameters for the Reference Key Estimation Model and Alternative Parameters Tested

Parameter	*Reference Value*	*Alternative Values Tested*
Downsampling factor	4	8, 16
Hop size (frames)	1/8	1/2, 1
Max. frequency (Hz)	1760	880, 3520
Min. frequency (Hz)	110	55, 27.5
Sparse kernel threshold	0.0055	0.5, 0.055, 0.06, 0.005, 0.00055
Observation type	Chord transitions	Single chords
Profile type	Probe tone 1	Probe tone 2, flat, random

Harmonic Progressions Through Time

We also wish to investigate the importance of harmonic progression through time. We do this by comparing two variations of the model. The first is the model as previously described, and the second uses only single chords as the observations.

For the single-chord model, the same key-transition probabilities were used as for the chord transition model for all initialization types. The chord emission probabilities for the Krumhansl initialization were based on the ratings of single chords in a given key (Krumhansl 1990, p. 171). For the flat initialization, they were set to 0.99 for chords within the key and 0.01 otherwise, then normalized. For the random initialization, they were taken from a uniform probability density function and then normalized.

Experiments

Investigating all possible combinations of the extensions mentioned would lead to a prohibitively large set of experiments, so we select one set of parameters to use as a reference to compare others against based on the settings suggested by Harte and Sandler (2005) and Noland and Sandler (2006). These settings are given in Table 2, together with the alternative parameter values tested.

We evaluate the algorithm according to the percentage of tracks labeled with the correct global key. Identifying the global key is not the only goal of an ideal tonal analysis; it is also desirable to follow key changes within the piece at various different levels of detail. However, this is much more difficult to evaluate and will thus be left as a topic for further research.

We used two different test sets. The first consisted of the songs from the first eight Beatles albums, for which key information is given in a musicological study that is available online (Pollack 2000). This gave a total of 110 songs. The second test set consisted of recordings of all preludes and fugues from J. S. Bach's *Well-Tempered Clavier*, Book 1, played on piano by Glenn Gould during 1963–1965 (Gould 1993). This gave 48 tracks with the keys given in the titles.

Results

The percentage of tracks for which the main key is correctly identified, for each test set, is shown in Table 3. The most obvious difference in performance is between the two test sets. The algorithm performed far better on the Bach than the Beatles data in almost all variations. This is most likely due to the content of the recordings: the Beatles songs contain a drum kit and many other percussive and distorted sounds, which will add noise to the chromagram and so affect the chord recognition, whereas the clean piano recordings have a much more harmonic spectrum, leading to more accurate chord recognition. Figure 7 shows the spectra up to 5 kHz for a typical C chord taken from the Beatles and one from Bach. The harmonic peaks are much clearer in the Bach spectrum. There are also several Beatles songs for which the global key

Table 3. Percentage of Global Key Estimates that Are Correct for the Beatles and Bach Test Sets

Changed Parameter	*Parameter Value*	*Total Correct (%)*	
		Beatles	*Bach*
Downsampling factor	**4**	**72**	**94**
	8	71	96
	16	74	92
Hop size (frames)	**1/8**	**72**	**94**
	1/2	73	75
	1	69	75
Maximum frequency (Hz)	880	75	98
	1760	**72**	**94**
	3520	72	94
Minimum frequency (Hz) [frame length (sec)]	**110 [0.74]**	**72**	**94**
	55 [1.48]	70	92
	27.5 [2.96]	68	77
Kernel threshold	0.5	71	90
	0.055	73	98
	0.006	73	92
	0.0055	**72**	**94**
	0.005	72	94
	0.00055	75	96
Profile type, with chord transitions model	**probe tone 1**	**72**	**94**
	probe tone 2	72	94
	flat	60	79
	random	4	2
Profile type, with single chords model	probe tone 1	73	65
	flat	55	48
	random	5	4

The reference model is shown in **bold**.

is not obvious: either there is a key change partway through, or the song is neither major nor minor, but modal. All of the Bach compositions, however, clearly start and end in the title key.

Figure 8 shows how the incorrect estimates relate to the correct key for the case with a hop size of one frame, or 0.74 sec. Most of the confusion is with closely related keys for both test sets, but the proportion of errors in the "other" category is much higher for the Beatles data. This is partly due to the few Beatles songs that do jump between two distant keys during the song, but it could also be influenced by the higher proportion of non-tonal sounds in the Beatles corpus, which would lead to errors that are not likely to be related to the harmony.

It is also interesting that far more Bach tracks were mistaken for their relative minor key than Beatles tracks. This is most likely a result of the harmonic content of the music. The majority of the Beatles songs stay in the same key throughout the song, even though some do use complex harmonies. However, even in the short preludes and fugues (with durations of no more than a few minutes), Bach takes the music through various tonal centers, and many of the tracks do include large sections that lean toward the relative minor. This type of error could be reduced by either mapping the combined key vector onto a musicologically relevant tonal space and finding the "center of effect" after Chew (2001), or more simply by adding the posterior state

Figure 7. Power spectra up to 5 kHz for a typical C chord from (a) the Beatles and (b) Bach test sets.

Figure 8. Error distribution relative to ground-truth values for key estimation, using the model with a hop size of one frame and all other parameter values as in the reference set.

Figure 9. Error distribution for the chord algorithm from Harte and Sandler (2005).

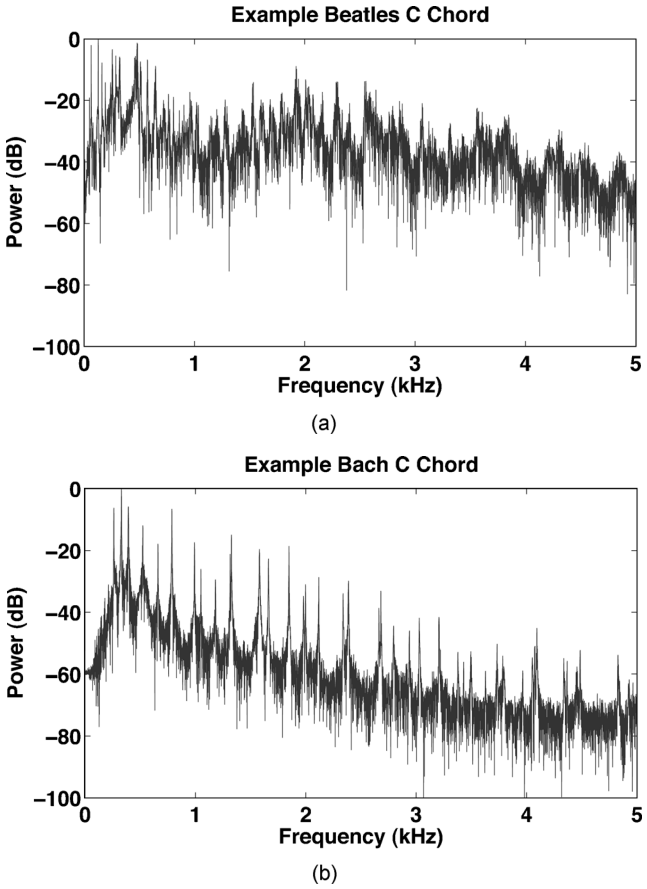


Figure 7

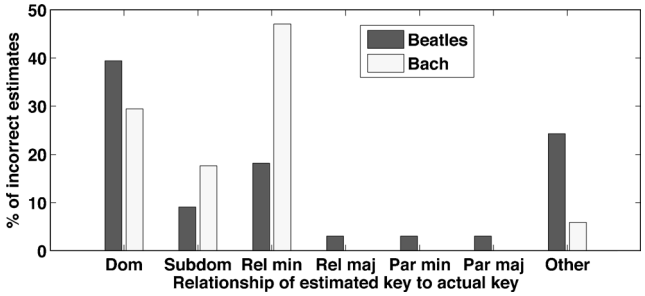


Figure 8

probabilities only over the start and end of the track, after Mardirossian and Chew (2005).

Figure 9 shows the error distribution for the chord-recognition algorithm on two of the Beatles albums, as reported by Harte and Sandler (2005). Note the higher proportion of harmonically distant errors for the chord recognition than for the keys, which shows that our probabilistic framework is capable of disregarding occasional unlikely chord estimates. The harmonically related chord errors follow a similar distribution to the Beatles key errors. This conforms to our expectations, because emission probabilities for closely related keys are similar, so a harmonically related chord error may be just enough to cause a key error in the same harmonic direction.

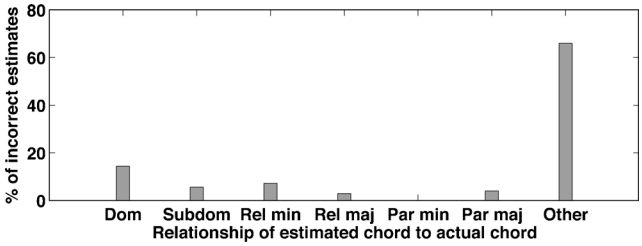


Figure 9

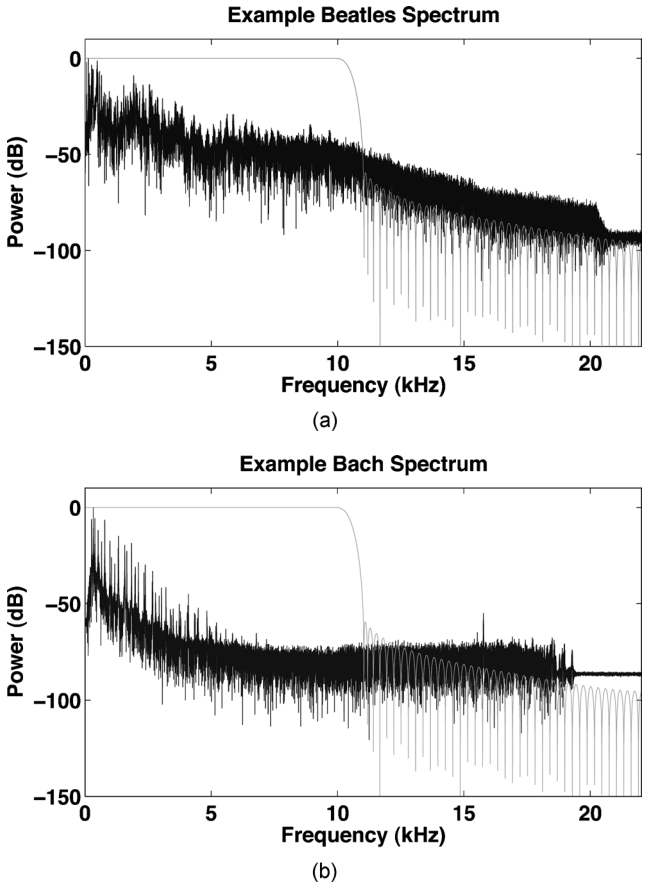
Effects of Low-Level Digital Signal Processing

Downsampling

The downsampling factor did not make a significant difference to the performance of the algorithm. The downsampling filter gave a stopband attenuation of 60 dB. Typical spectra for both test sets are shown in Figure 10. They show a dynamic range of less than 60 dB, so any aliased components will be well below the noise floor and will not affect the results.

The maximum downsampling factor of 16 gives an audio bandwidth of 1.4 kHz, which implies that the important information for key estimation is contained in the frequencies up to 1.4 kHz. This has implications for reducing computation time for calculating the chroma features, because downsampling means there are fewer samples to process, and if the higher frequencies are to be disregarded, then the constant-Q transform need not encompass such a wide frequency range. The loss in time resolution is not significant in the context of chord changes, because the sampling period at

Figure 10. Typical full-audio-bandwidth power spectra (black lines) from (a) the Beatles and (b) Bach test sets, with the antialiasing filter response for a downsampling factor of two (gray lines).



2.8 kHz is 0.36 msec, much shorter than any chord duration.

Hop Size

With the minimum frequency set to 110 Hz, the frame size was 8,192 samples, or 0.74 sec. The performance was clearly best with a hop size of one-eighth of a frame for the Bach (an improvement of 19 percent over the longer hop sizes), but there was a much less dramatic effect on the Beatles. The hop size will affect the model probabilities, because for a short hop size the probability of staying on the same chord and in the same key will be much higher than for a longer hop size. However, it is likely that the chords change much more frequently in Bach's music than in the that of the Beatles, which would mean that the loss of temporal resolution would affect the Bach more severely. Further experiments showed that for similar degradation on the Beatles test set, hop sizes had to be increased to approximately three times the reference frame length.

In general, finer resolution produced better results. We conducted an additional experiment on both data sets at a hop size of one-sixteenth of a frame to test whether even higher accuracy could be achieved. Results were 70 percent for the Beatles and 88 percent for the Bach; both of these figures are lower than the experiment in which a hop size of one-eighth of a frame was used. These facts led to a recommendation of hop sizes around 0.1 sec (approximately 0.74/8) for the given HMM parameter initialization.

Maximum Frequency

The maximum constant-Q frequency did not significantly affect the results, although the highest accuracy was achieved with the lowest cutoff frequency. This is consistent with the finding that downsampling by up to 16 times, which limits the bandwidth to about 1.4 kHz, still gives acceptable accuracy, and it even suggests that the higher frequencies are detrimental to key estimation. The lowest cutoff frequency of 880 Hz is A5 (taking middle C as C4), which roughly corresponds to the highest fundamental frequency. (The highest note in the Bach preludes and fugues is C6.) The chord-recognition algorithm assumes fundamentals have the greatest power, and it does not take harmonics into account, so it is reasonable that including the harmonics might decrease the recognition rate.

Minimum Frequency

The key-estimation algorithm was least effective when the lowest frequencies were included (down to 27.5 Hz), particularly for the Bach data. This implies that the pitches between 27.5 Hz (A0) and 55 Hz (A1) are detrimental to key recognition. However, halving the minimum frequency also results in a doubling of the frame length, which means that when the lower frequencies are included, temporal resolution is reduced. We have already noted in

our analysis of hop size that the Bach test set is particularly sensitive to loss of temporal resolution.

To test whether it is the loss of resolution or the addition of lower frequencies that causes the loss of accuracy, we ran an experiment on the Bach with a minimum frequency of 27.5 Hz and a hop size of 1/32 of a frame, and another with a minimum frequency of 55 Hz and a hop size of 1/16 of a frame. Both cases give a hop size of the same duration (approximately 0.1 sec), and both achieved 96 percent accuracy. To improve computation time, this was carried out after downsampling by a factor of 16 so the results should be compared to line 3 of Table 3, 92 percent accuracy. The higher value of 96 percent implies that the temporal resolution is of the most importance, and inclusion of frequencies down to 55 Hz is beneficial, but including frequencies down to 27.5 Hz offers no additional improvement. A frequency of 55 Hz (A1) roughly corresponds to the lowest fundamental frequency (for the Bach data the lowest note is B-sharp1 and for the Beatles, although we do not have a transcription, the bass line is almost entirely played by bass guitar, which in standard tuning cannot play below E1), so we conclude that only low frequencies corresponding to notes played are required for our analysis. This finding is as we might expect, because lower frequencies must come from unpitched percussion or extraneous noise sources, which are unlikely to be related to the harmony and so will not benefit key recognition.

Kernel Threshold

Variations in the spectral kernel threshold made only small differences, with only a slight drop in performance at the highest threshold value of 0.5; the highest accuracy for the Bach was achieved at the second-highest threshold of 0.055. The maximum kernel value is approximately 0.54, so a threshold of 0.5 removes all but the highest peak, leading to small inaccuracies in the constant-Q representation. Our results show that these inaccuracies do not significantly reduce performance of the key-estimation algorithm, so a high threshold can be used to improve computation time, because the multiplication operation for kernel values of zero need not be performed.

Effects of Profile Type

Initialization with the Krumhansl probe-tone profiles proved far superior to initialization with the flat or random profiles in both the chord-transition and single-chord models. There was no difference between the two probe-tone versions, indicating the transitions between two diatonic chords are of the greatest importance. Using the flat profiles resulted in a loss of between 12 and 18 percent accuracy, and the random initialization was unsuccessful. For a random choice of key, we would expect to be correct for 1 in 24 tracks (4 percent), and the performance here was no higher than 5 percent. In fact, for the random initialization, there is no reason why the hidden states should correspond to keys. Within one song, examples of every possible key change from which to learn transition probabilities are not present, so it is much more likely that the hidden states correspond to a low-level harmonic atom, such as a small group of chords.

The EM training algorithm is well known to be dependent on good initialization, which is supported by our results. It would be an interesting study to tie some of the parameters during the training phase; for example, the key transition C major–G major should be updated whenever the G major–D major transition is updated. This would give a better general model, but may not be so useful for per-song training.

Effects of Observation Type

The choice between using pairs of chords and single chords as observations for the HMM made a significant difference to the Bach test set, but not to the Beatles set. This suggests that the idea of harmonic progression through time is more important in Bach's music than the Beatles'. Certainly, Bach uses a more traditional harmonic grammar, whereas the Beatles, although they do make excursions to various keys, often do so in a way that does not reflect standard modulation techniques. Because the single-chord model has lower complexity, it would be preferred for the Beatles, but the more complex chord transition

Table 4. Parameters Giving the Best Results for Each Test Set

	"Best" Value	
Parameter	*Beatles*	*Bach*
Downsampling factor	16	8
Hop size (frames)	1/2	18
Maximum frequency (Hz)	880	880
Minimum frequency (Hz)	110	110
Sparse kernel threshold	0.00055	0.055
Observation type	single chords	chord transitions
Profile type	probe tones (1 or 2)	probe tones (1 or 2)

model is recommended for Bach. It is likely that such recommendations could be extended to a wider range of popular and classical music after further investigation.

Best Parameters

The parameters that gave the highest scores for each test set are shown in Table 4. We tested a version of the model with all the "best" parameters, which gave scores of 71 percent for the Beatles and 98 percent for Bach. In addition, we tested the all the "best" parameters for the Beatles but with the chord-transition model, which gave 75 percent correct global key assignment. The "best" chord-transition models gave accuracy equal to the highest scores in Table 3, which suggests that for a given model structure, the signal-processing parameters can be individually optimized and then combined. However, the "best" Beatles parameters with the single-chord model reduced the accuracy, so the previous DSP parameter optimization did not hold for a different model structure.

The highest scores achieved by our model are encouraging. Gómez Gutiérrez (2006, p. 117) compares simple tone-profile approaches using a test set of Beatles songs and Bach fugue subjects. The test sets used by Gómez Gutiérrez are not identical to ours, but we would expect them to give similar results, because there is a significant overlap and they are from the same band and composer. However, it should be noted that Gómez Gutiérrez's Bach test collection included only the fugue subjects, which are much shorter than the whole fugues and much less likely to include modulations. Her highest scores are 66 percent correct for the Beatles test set and 88 percent for the Bach, which means that our high scores of 75 percent and 98 percent accuracy compare favorably. This shows that our use of an HMM to model frame-by-frame key progressions and apply a statistical approach to reducing the influence of unlikely chords was successful. The notion of harmonic progression through time is of interest over longer a time span than a single frame or single chord transition, so in the future we will investigate higher-order models and hierarchies of HMMs.

Conclusions

We have investigated the effects of low-level DSP parameters, tone-profile values, and observation type for an HMM-based key-estimation algorithm on two different test sets: one of Beatles songs, and one of Bach's music played on a piano. The algorithm performed better on the Bach test set, which consists of very "clean" recordings of a harmonic instrument, whereas the Beatles recordings contain more inharmonic sounds from the drum kit and other percussive sounds. Further studies will investigate the effects of including preprocessing stages such as transient removal (Duxbury, Davies, and Sandler 2001) to isolate the harmonic parts of the signal, along with beat detection (Davies and Plumbley 2004) to give beat-length analysis frames.

We have found that, although higher accuracy can be achieved with the Bach test set, it was much more sensitive to temporal resolution than the Beatles test set. The Bach set was also more sensitive with respect to simplifying the model to use only single chords as observations. At this stage, it would be unwise to assume these results apply to a broader range of music, but further investigation could lead to more general recommendations about the model complexity required for analysis of different musical styles.

Results suggest that the most important frequencies for harmonic analysis are between 55 and 880 Hz (between A1 and A5), which is approximately the range of the fundamental frequencies of the music we tested. For our algorithm, which only refers to the fundamental pitches, the inclusion of higher frequencies leads to only small differences in performance, which means that the algorithm can be made more efficient by downsampling and by reducing the number of constant-Q bins. It is not clear, however, whether another algorithm that makes use of or takes account of upper partials would be improved by including higher frequencies. We found that further computational efficiency can be obtained by applying a high threshold to the spectral kernels for the constant-Q transform without affecting the key estimation.

The results show that the initialization parameters are of great importance. The probe-tone profiles gave the best initialization, and results showed that the transitions between diatonic chords are the most significant. This means that a simpler model that only takes diatonic transitions into account may give equivalent results. Using flat profiles, which contain information about which chords are within a key but no details on the hierarchy of the chords' importance, reduced the performance by 12–18 percent. Using random initialization, which contains no musical information, performed little better than a random choice of key.

The large variation in results shows that investigation into low-level parameters is worthwhile. The same parameters will not necessarily be best for every algorithm, as supported by our previous work (Noland and Sandler 2007), and the various parameters are interrelated, which means that finding the best combination of parameters is a complex process. However, for a constant model structure, it was possible to individually optimize the lower-level signal processing parameters and then combine them. The optimal parameters will also vary for different types of musical recordings, so some generalization is necessary.

The highest scores show that accurate global key estimation can be obtained using our HMM structure. Our future research will include improving the performance on noisy recordings of percussive music and investigating methods of evaluating a more detailed tonal analysis.

References

Aarden, B. J. 2003. "Dynamic Melodic Expectancy." PhD dissertation, Ohio State University.

Bello, J. P., and J. Pickens. 2005. "A Robust Mid-Level Representation for Harmonic Content in Musical Signals." *Proceedings of the Sixth International Symposium on Music Information Retrieval*. London: Queen Mary, University of London, pp. 304–311.

Brown, J. C. 1991. "Calculation of a Constant Q Spectral Transform." *Journal of the Acoustical Society of America* 89(1):425–434.

Brown, J. C., and M. S. Puckette. 1992. "An Efficient Algorithm for the Calculation of a Constant Q Transform." *Journal of the Acoustical Society of America 92*(5):2698–2701.

Chai, W., and B. Vercoe. 2005. "Detection of Key Change in Classical Piano Music." *Proceedings of the Sixth International Symposium on Music Information Retrieval*. London: Queen Mary, University of London, pp. 468–473.

Chew, E. 2001. "Modelling Tonality: Applications to Music Cognition." *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey: Lawrence Erlbaum, pp. 206–211.

Chuan, C.-H., and E. Chew. 2005. "Polyphonic Audio Key Finding Using the Spiral Array CEG Algorithm." *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*. Amsterdam. New York: Institute of Electrical and Electronics Engineers. Available online at imsc.usc.edu/papers/techreports/05001.html.

Davies, M. E. P., and M. D. Plumbley. 2004. "Causal Tempo Tracking of Audio." *Proceedings of the Fifth International Symposium on Music Information Retrieval*. Barcelona: Universitat Pompeu Fabra, pp. 164–169.

Downie, N. M., and R. Heath. W. 1974. *Basic Statistical Methods*, 4th ed. New York: Harper and Row.
Duxbury, C., M. Davies, and M. Sandler. 2001. "Separation of Transient Information in Musical Audio Using Multiresolution Analysis Techniques." Paper presented at the COST G-6 Conference on Digital Audio Effects (DAFX-01). Limerick, Ireland, 6 December.
Gómez, E. and P. Herrera. 2004. "Estimating the Tonality of Polyphonic Audio Files: Cognitive Versus Machine Learning Modelling Strategies." Paper presented at the Fifth International Symposium on Music Information Retrieval, Barcelona, 11 October.
Gómez Gutiérrez, E. 2006. "Tonal Description of Music Audio Signals." PhD dissertation, Barcelona: Universitat Pompeu Fabra.
Gould, G. 1993. *Bach: The Well-Tempered Clavier I. The Glenn Gould Edition*. Audio compact disc. New York: Sony Classical SM2K 52 600.
Harte, C., and M. Sandler. 2005. "Automatic Chord Identification Using a Quantised Chromagram." Paper presented at the 118th Convention of the Audio Engineering Society, Barcelona, 29 May.
Harte, C., M. Sandler, and M. Gasser. 2006. "Detecting Harmonic Change in Musical Audio." Paper presented at the Audio and Musical Computing for Multimedia Workshop 2006, Santa Barbara, 27 October.
İzmirli, O. 2005. "An Algorithm for Audio Key Finding." 2005 Music Information Retrieval Evaluation eXchange (MIREX) Audio Key-Finding Contest. Available online at www.music-ir.org/evaluation/mirex-results/articles/key_audio/izmirli.pdf.
Krumhansl, C. L. 1990. *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.
Mardirossian, A., and E. Chew. 2005. "SKeFiS—A Symbolic (MIDI) Key Finding System." 2005 Music Information Retrieval Evaluation eXchange (MIREX) Audio Key-Finding Contest. Available online at www-scf.usc.edu/~mardiros/papers/MIREX2005.pdf.
McClellan, J. H., R. W. Schafer, and M. A. Yoder. 1998. *DSP First: A Multimedia Approach*. Upper Saddle River, New Jersey: Prentice Hall.
Noland, K., and M. Sandler. 2006. "Key Estimation Using a Hidden Markov Model." Paper presented at the Seventh International Symposium on Music Information Retrieval, Victoria, British Columbia, 10 October.
Noland, K., and M. Sandler. 2007. "Signal Processing Parameters for Tonality Estimation." Paper presented at the 122nd Convention of the Audio Engineering Society, Vienna, 8 May.
Noll, T., and J. Garbers. 2004. "Harmonic Path Analysis." In Guerino Mazzola, Thomas Noll, and Emilio Lluis-Puebla, eds. *Perspectives of Mathematical and Computational Music Theory*. Osnabrück: epOs-Music, pp. 395–427.
Pauws, S. 2004. "Musical Key Extraction from Audio." Paper presented at the Fifth International Symposium on Music Information Retrieval, Barcelona, 11 October.
Pollack, A. W. 2000. "Notes on . . . Series." Available online at www.icce.rug.nl/~soundscapes/DATABASES/AWP/awp-notes_on.shtml.
Purwins, H. 2005. "Profiles of Pitch Classes Circularity of Relative Pitch and Key - Experiments, Models, Computational Music Analysis, and Perspectives." PhD dissertation, Technischen Universität Berlin.
Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77(2):257–286.
Shmulevich, I., et al. 2001. "Perceptual Issues in Music Pattern Recognition: Complexity of Rhythm and Key Finding." *Computers and the Humanities* 35(1):23–35.
Temperley, D. 2004. "Bayesian Models of Musical Structure and Cognition." *Musicae Scientiae* 8(2):175–205.