Estimation of Harpsichord Inharmonicity and Temperament from Musical Recordings [a)]

Simon Dixon,[b)] Matthias Mauch, and Dan Tidhar
Centre for Digital Music,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
Mile End Road, London E1 4NS,
United Kingdom

(Dated: March 25, 2011)

The inharmonicity of vibrating strings can easily be estimated from recordings of isolated tones. Likewise, the tuning system (temperament) of a keyboard instrument can be ascertained from isolated tones by estimating the fundamental frequencies corresponding to each key of the instrument. This paper addresses a more difficult problem: the automatic estimation of the inharmonicity and temperament of a harpsichord given only a recording of an unknown musical work. An initial conservative transcription is used to generate a list of note candidates, and high-precision frequency estimation techniques and robust statistics are employed to estimate the inharmonicity and fundamental frequency of each note. These estimates are then matched to a set of known keyboard temperaments, allowing for variation in the tuning reference frequency, in order to obtain the temperament used in the recording. Results indicate that it is possible to obtain inharmonicity estimates and to classify keyboard temperament automatically from audio recordings of standard musical works, to the extent of accurately (96%) distinguishing between six different temperaments commonly used in harpsichord recordings. Although there is an interaction between inharmonicity and temperament, this is shown to be minor relative to the tuning accuracy.

PACS numbers: 43.75.Yy, 43.75.Bc, 43.75.Gh, 43.75.Xz

I. INTRODUCTION

Recent advances in music signal processing and the speed of desktop computers have facilitated the automation of many aspects of the analysis of music recordings. For example, the extraction of semantic metadata such as genre (Tzanetakis and Cook, 2002), key (Noland and Sandler, 2006), chord (Mauch and Dixon, 2010) and beat (Dixon, 2001) has been a major focus of the music informatics research community. Such work has applications in the areas of classification (e.g. organisation and navigation of music collections), recommendation (e.g. discovery and marketing of new music) and annotation (e.g. automatic transcription for education, musicological research, and music practice), and it complements traditional research methods in musicology, enabling more quantitative and larger scale analyses to be performed. For example, researchers and practictioners of early Western music debate the virtues of various tuning systems (temperaments) in terms of their theoretical properties (e.g. Di Veroli, 2009), but are unable to substantiate (or refute) claims about performance practice with empirical data, having no means of measuring temperaments from musical recordings.

In recent work on keyboard temperament estimation (Tidhar *et al.*, 2010b), we presented an automatic system for recognising temperament directly from audio recordings of unknown works. Our classifier is capable of distinguishing, with high accuracy, between six different temperaments commonly used in harpsichord recordings. The system did not take proper account of inharmonicity, and thus it was only possible to use a small number of partials in the fundamental frequency estimation step. In this paper we extend our previous work to estimate the inharmonicity of each tone, and to propose an approach for robust estimation of frequency and inharmonicity from note mixtures as they occur in standard musical works.

Building a system to estimate inharmonicity and classify musical recordings by temperament presents particular signal processing challenges. First, a high frequency resolution is required, as the differences between temperaments are small, of the order of a few cents (hundredths of a semitone). For example, if A = 415 Hz is used as the reference pitch (typical in Baroque Period music), then middle C might have a frequency of 246.76, 247.46, 247.60, 247.93, 248.23 or 248.99 Hz, based on the six representative temperaments described in subsection III.A. To resolve these frequencies in a spectrum, a window of several seconds duration would be required, but this introduces other problems, since musical notes are not stationary and generally do not last this long. Likewise, the frequency differences due to inharmonicity are even smaller for the low order partials of harpsichord tones.

The second problem is that in musical recordings, notes rarely occur in isolation. There are almost always multi-

[a)] Portions of this work were presented in “High precision frequency estimation for harpsichord tuning classification”, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (2010), pp. 61–64.
[b)] simon.dixon@eecs.qmul.ac.uk

ple notes sounding simultaneously, and this has the potential to bias any frequency estimates. To make matters worse, the intervals which are favoured in music are those where many partials coincide. In particular, it can be difficult to discern whether a spectral peak is a fundamental frequency or a partial of another fundamental. The ability to distinguish between these cases is crucial to accurate pitch estimation and thus also to successful inharmonicity determination and temperament classification.

For example, suppose the two notes A1 and E3 are played together (an interval of an octave and a perfect fifth; a very common interval). Then each partial of E3 will coincide (almost or exactly, depending on the temperament and inharmonicity) with every third partial of A1. If the tone A1 has a fundamental frequency of 110 Hz, the spectrum will also have a peak at 330 Hz, the third harmonic of the note A1, but also (if the fifth is pure) the fundamental frequency of an E. In many temperaments however, the actual note E will have a frequency different from 330 Hz (e.g. 329.6 Hz in equal temperament), so the estimation of this partial might be biased either by not being able to resolve the two partials, or if they are resolved, by assigning the partial to the wrong tone. Inharmonicity further complicates the situation, as the partials will not be precise integer multiples of the fundamental, making correct assignment of partials more difficult.

We therefore require a method for distinguishing peaks corresponding to fundamental frequencies from those which are caused by higher harmonics. This would not be the case if we assumed knowledge of the score, but this is deliberately avoided in our formulation of the problem, in order to increase the generality of our algorithms, which is important for practical applications such as the web service TempEst (Tidhar *et al.*, 2010a). To avoid the bias caused by overlapping partials, while circumventing the problem of full polyphonic transcription, which is still considered an unsolved problem (Klapuri, 2009), we introduced the concept of *conservative transcription*, which entails estimating the “safe” subset of the played notes, i.e. those whose fundamental frequencies are not harmonics of lower co-occurring frequencies. We have shown that conservative transcription is applicable in practical situations, and that it can improve temperament estimation in recordings of typical harpsichord music (Tidhar *et al.*, 2010b).

The remainder of the paper is structured as follows. In the following section we review relevant literature on temperament, inharmonicity and fundamental frequency estimation. Then, in section III, we describe the preparation of test data and the methods used in this work, consisting of signal and data processing algorithms for estimating: (1) a conservative transcription of the music; (2) the frequency of each partial of the transcribed notes; (3) the inharmonicity and fundamental frequency of each transcribed note; (4) the tuning of each pitch class relative to equal temperament; and (5) the temperament that best matches this pitch class tuning profile. In the final two sections we present and discuss the experimental results and the conclusions that can be drawn from them.

II. BACKGROUND

A. Temperament

During the past two centuries, equal temperament has been the dominant paradigm for building and describing musical scales in Western music, but since the latter part of the twentieth century there has been a revival of interest in historical performance practice of early music on period instruments, leading to increased attention to historical, unequal temperaments. Theoretical and practical aspects of temperament are covered thoroughly elsewhere (Barbour, 2004/1951; Rasch, 2002; Di Veroli, 2009), so we address them only briefly here.

The study of musical consonance has a history extending back at least as far as Pythagoras in the sixth century BC, with numerous theoretical frameworks being proposed (von Helmholtz, 1863; Lundin, 1947; Terhardt, 1977; Sethares, 1999; Palisca and Moore, 2010; McDermott *et al.*, 2010). Common to most of these frameworks is the recognition that listeners prefer sounds with harmonic spectra and without beats. For combinations of musical sounds with harmonic spectra, the sensation of consonance correlates to small integer frequency ratios between fundamental frequencies, and specifically superparticular ratios of the form $\frac{n+1}{n}$ where $n \leq 5$ (corresponding to the following pure intervals: octave, perfect fifth, perfect fourth, major third and minor third, for successive values of n). Continuous-pitch instruments and singers can dynamically adapt their intonation to form perfectly consonant intervals if required, but fixed-pitch instruments such as keyboard, some fretted, and some percussion instruments, need to commit to a tuning scheme for the duration of a piece, if not an entire concert. In the Western musical tradition at least, this gives rise to the need for temperament, because it is not possible to accommodate all pure intervals within the small set of pitch classes available.

The two most consonant intervals are the octave (frequency ratio 2:1) and perfect fifth (frequency ratio 3:2). In Western music these correspond to intervals of 12 and 7 semitones respectively. From a given starting note, either a succession of 7 octave steps or a succession of 12 perfect fifth steps will lead to the same note, despite the fact that $(\frac{3}{2})^{12} \neq 2^7$. The ratio of the two sides of this inequality (approximately 1.0136) is called the Pythagorean comma, and one way of considering temperament is according to the distribution of this comma around the cycle of fifths. For example, in equal temperament, all fifths are diminished by $\frac{1}{12}$ of a comma relative to the pure ratio 3:2.

Determining a temperament can thus be regarded as an optimisation problem, whereby keeping the octaves pure is a constraint, and various considerations lead to different compromises between pureness of fifths and pureness of major thirds. Among these considerations are the key, or set of keys, which should “work well” in the given temperament; a temperament is considered to work well for a given key if the most frequent harmonic intervals in the key (major thirds and to some extent fifths, most notably in tonic and dominant positions) are close to their pure underlying frequency ratios, and are therefore perceived as consonant. Temperaments which work well for most keys, and are bearable in all keys, are referred to as *well* temperaments. Temperaments in which all fifths but one or two are equal to each other, are called *regular* temperaments. Equal temperament is the only temperament which is both regular and well.

B. Inharmonicity

There are two main strands of research regarding inharmonicity in string instruments: investigation of the physical and acoustical properties of vibrating strings; and psychoacoustic studies relating to the perceptibility or otherwise of inharmonicity. Perceptual studies involving inharmonicity are important for understanding and developing models of pitch perception, and they bear relevance for the implementation of synthesis algorithms which aim to artificially recreate the natural sound of string instruments.

Early work on acoustics investigated the phenomenon that vibrating strings have partials at frequencies which are slightly greater than integer multiples of the fundamental frequency (Shankland and Coltman, 1939; Young, 1952). The inharmonicity of metal strings is due to two main factors: stiffness of the string and the amplitude of vibration (Shankland and Coltman, 1939); in the case of musical instruments, stiffness accounts for most of the inharmonicity. For a string with (ideal) fundamental frequency f_0 and inharmonicity constant B, the frequency f_k of the kth partial is given by (Fletcher, 1964):

$$f_k = kf_0\sqrt{1 + Bk^2} \quad (1)$$

where the constants f_0 and B are determined by the physical properties of the string. This inharmonicity contributes to the characteristic sound of the piano (Fletcher *et al.*, 1962), but is significantly less pronounced on the harpsichord (Fletcher, 1977) so that it has been described as “nearly negligible” (Fletcher and Rossing, 1998, p. 343). Välimäki *et al.* (2004) cite measured values of B ranging from 10^{-5} to 10^{-4} and show that modelling of inharmonicity is essential for realistic synthesis of harpsichord tones. Earis *et al.* (2007) report similar results for measurements of inharmonicity constants, giving a slightly higher upper limit of 1.2×10^{-4}. Rauhala *et al.* (2007) present an efficient iterative algorithm for estimating B given a recorded tone and its approximate fundamental frequency.

Various studies have investigated the effects of inharmonicity on pitch perception. Moore *et al.* (1985) found that in artificial mixtures, even a single inharmonic partial in an otherwise harmonic tone influenced the overall perception of pitch, lending support to a model in which residue pitch is a weighted average of the pitch cues provided by each partial. The experimentally determined weights of each partial varied between subjects, but in all cases were significant for only the first six partials. Anderson and Strong (2005) synthesised inharmonic complex tones based on parameters obtained from the analysis of piano tones, and asked listeners to match their pitch to synthesised harmonic tones having the same spectral and temporal envelopes. Their results showed a correlation between the inharmonicity coefficient B (obtained using a weighted least squares fit) and the perceived pitch shift relative to the fundamental. Järveläinen *et al.* (2001) investigated the threshold of audibility of inharmonicity for synthetic piano-like tones and found that the threshold increased with fundamental frequency in a linear relationship when B and f_0 are both expressed on a logarithmic scale. They concluded that the inharmonicity normally present in piano tones is audible, particularly at lower pitches, but that it nears the threshold at higher pitches.

C. Fundamental Frequency Estimation

In the vast literature on estimation of fundamental frequency and pitch, there is no single algorithm which is suitable for all signals and applications. Methods are reviewed elsewhere (de Cheveigné, 2006; Klapuri and Davy, 2006, chapters 7 and 8), but we summarise here the limitations of many current systems for music signal processing applications (Gerhard, 2003). First, assumptions are made about the signal which do not hold for most music signals, such as: that the input signal at any point in time consists of a single pitched tone (monophonicity); that the properties of the signal are stable over the duration of analysis (stationarity); and that the properties of the input signal (e.g. the instrument or class of instrument) are known or match a small set of allowed instruments. Second, fundamental frequency estimation is often equated to pitch estimation, ignoring the influences of inharmonicity and human perception, although some recent multi-pitch analysis methods do take inharmonicity into account (Klapuri, 2003; Emiya *et al.*, 2010; Benetos and Dixon, 2010). Finally, even for music analysis systems, frequency resolution is rarely much finer than one semitone, and few papers discuss the issues related to determining frequency at the resolution required for measuring temperament or inharmonicity, where differences of a few cents are decisive.

For our purposes, time-domain pitch estimation methods such as ACF and YIN (de Cheveigné and Kawahara, 2002) are unsuitable due to the bias caused by the presence of multiple simultaneous tones. Thus we choose a frequency domain technique which gives sufficient frequency resolution: the FFT with quadratic interpolation (Smith and Serra, 1987) and correction of the bias due to the window function (Abe and Smith, 2004). These techniques are described in more detail in subsections III.B and III.C respectively. In previous work we showed that this combination of methods is suitable for estimating temperament and that it outperforms instantaneous frequency estimation using phase information (Tidhar *et al.*, 2010b). More advanced estimation algorithms which admit frequency and/or amplitude modulation (Wen and

Sandler, 2009) were deemed unnecessary.

III. METHOD

A. Data

Obtaining ground-truth data for the evaluation of temperament and inharmonicity estimation algorithms presents several difficulties. Most commercially available recordings do not specify the harpsichord temperament, and even those that do might not be completely reliable because of a possible discrepancy between tuning as a practical matter and tuning as a theoretical construct. In practice, the tuner's main concern is to facilitate playing, and time limitations very often compromise precision. We therefore chose to produce our own test dataset consisting of both real and synthesised harpsichord music. The synthesised data ensures that the temperament is precise, but might not replicate the timbre of a harpsichord (e.g. coupling between strings) and typical recording conditions (e.g. reverberation and noise).

The real harpsichord recordings were played by Dan Tidhar on a Rubio double-manual harpsichord in a small hall. The synthesised recordings were performed on a digital keyboard by Dan Tidhar and rendered from MIDI files using the physical modelling synthesis software Pianoteq (Pianoteq, 2010). For each of the six temperaments (see below) and two rendering alternatives (real vs synthesised), four musical excerpts were recorded (i.e. a total of 48): a slow ascending chromatic scale, chosen as a baseline for comparison; J.S. Bach's Prelude 1 in C Major from the *Well-tempered Clavier*; F. Couperin's La Ménetou from *Pièces de Clavecin, Septième Ordre*; and J.S. Bach's Variation 21 from the *Goldberg Variations*. The choice of pieces encompasses various degrees of polyphony, various degrees of chromaticism, as well as various speeds. The tuning reference frequency for all recordings was approximately A=415Hz.

The following six temperaments were used: equal temperament (ET), Vallotti (V), fifth-comma (FC), quarter-comma meantone (QCMT), sixth-comma meantone (SCMT), and just intonation (JI). The properties of these temperaments are shown graphically in Figure 1 and described briefly below. In equal temperament, each of the fifths is diminished by $\frac{1}{12}$ of a comma, so that the frequency ratio between successive semitones is always $2^{1/12}$. In a Vallotti temperament, 6 of the fifths are diminished by $\frac{1}{6}$ of a comma each, and the other 6 fifths are left pure. In the fifth-comma temperament we used, five of the fifths are diminished by a $\frac{1}{5}$ comma each, and the remaining 7 are pure. In a quarter-comma meantone temperament, 11 of the fifths are shrunk by $\frac{1}{4}$ of a comma, and the one remaining fifth is $\frac{7}{4}$ of a comma larger than pure. In sixth-comma meantone, 11 fifths are shrunk by $\frac{1}{6}$ of a comma, and the one remaining fifth is $\frac{5}{6}$ comma larger than pure. The just-intonation tuning we used is based on the reference tone A, and all other tones are calculated as simple integer fundamental frequency ratios. The ratios are given by the following vector, representing the twelve chromatic tones above the reference A: $\{\frac{16}{15}, \frac{9}{8}, \frac{6}{5}, \frac{5}{4}, \frac{4}{3}, \frac{45}{32}, \frac{3}{2}, \frac{8}{5}, \frac{5}{3}, \frac{9}{5}, \frac{15}{8}, \frac{2}{1}\}$. The deviations (in cents) from equal temperament of the five other temperaments we use are given in Table I. Apart from being relatively common, this set of six temperaments represents different categories: Equal temperament is both well and regular, just intonation is neither well nor regular, Vallotti and fifth-comma are well and irregular, and the two variants of meantone (quarter and sixth comma) are regular but not well.

TABLE I. Deviations (in cents) from equal temperament for the five unequal temperaments (V: Vallotti; FC: Fifth Comma; QCMT: Quarter Comma Meantone; SCMT: Sixth Comma Meantone; JI: Just Intonation) used in this work.

Note	V	FC	QCMT	SCMT	JI
C	5.9	8.2	10.3	4.9	15.6
C♯/D♭	0.0	-1.6	27.4	13.0	-13.7
D	2.0	2.7	3.4	1.6	-2.0
D♯/E♭	3.9	2.3	20.5	9.8	-9.8
E	-2.0	2.0	-3.4	-1.6	2.0
F	7.8	6.3	13.7	6.5	13.7
F♯/G♭	-2.0	-3.5	-10.3	-4.9	-15.6
G	3.9	5.5	6.8	3.3	17.6
G♯/A♭	2.0	0.4	24.0	11.4	-11.7
A	0.0	0.0	0.0	0.0	0.0
A♯/B♭	5.9	4.3	17.1	8.1	11.7
B	-3.9	-0.8	-6.8	-3.3	3.9

B. Conservative Transcription

The ideal solution for estimating the fundamental frequencies of each of the notes played in a piece would require a transcription step to identify the existence and timing of each note. However, no reliable automatic transcription algorithm exists. Therefore we developed a two-stage approach in order to obtain accurate estimates of fundamental frequency and inharmonicity of unknown notes in the presence of multiple simultaneous tones. The first stage is a conservative transcription, which identifies the subset of notes which are easily detected, omitting any unsure candidates. In other words, it obtains a high *precision* (fraction of transcribed notes that are correct) at the cost of low *recall* (the fraction of played notes that are transcribed). The second stage is an accurate frequency domain f_0-estimation step for the notes determined in the first stage. We take advantage of the fact that we do not need to estimate the pitch of each and every performed note, since the tuning of the harpsichord is assumed not to change during a piece, and there are usually multiple instances of each pitch from which frequency estimates can be computed.

Conservative transcription consists of three main parts: computation of frame-wise amplitude spectra with a standard STFT; sinusoid detection through peak-picking, which yields a set of initial frequency estimates;

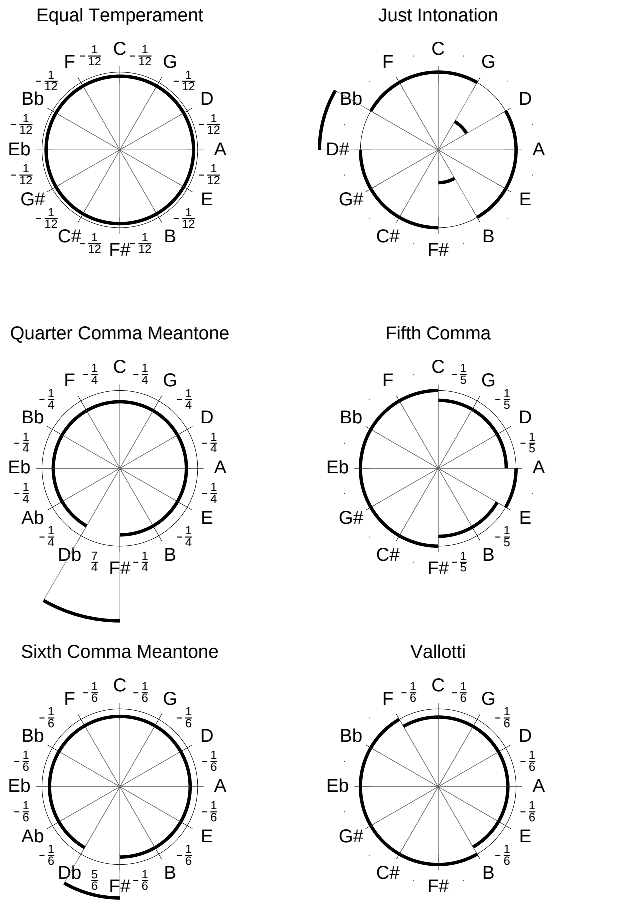

FIG. 1. Cycle of fifths representations for each of the temperaments used in this paper. The distance of the dark segments from the centre of the circle represents the deviation from pure fifths (the light circle). The fractions specify the distribution of the comma between the fifths.

and finally the deletion of sinusoids that have a low confidence, either because they are below an amplitude or duration threshold, or because they could be overtones of a different sinusoid. We describe the deletion of candidate sinusoids as "conservative" since not only overtone sinusoids, but also some of the sinusoids that correspond to fundamental frequencies could be deleted in this step.

The sinusoid detection is a simple spectrum-based method. From the time-domain signal, downsampled from $f_s = 44100$ Hz to $f'_s = 11025$ Hz, we compute the STFT $X(n,i)$, where n is the frame index and i the frequency bin index, using a Hamming window, a frame length of 4096 samples (370 ms), a hop size of 256 samples (23 ms, i.e. $\frac{15}{16}$ overlap), and a zero padding factor of 2 (i.e. FFT size $N = 8192$). In fast passages, the window will contain multiple sequential notes, but the negative effect of this is balanced by the greater frequency resolution. The use of a Hamming window rather than the Blackman-Harris window used in subsection III.C is not critical.

In order to detect each of the partials we first identify peaks in the amplitude spectrum $|X(n,i)|$ using two adaptive thresholding techniques. To find locally significant bins of frame n, we calculate the moving weighted mean $\mu(n,i)$ and the moving weighted standard deviation $\sigma(n,i)$ of $|X(n,i)|$ using a window of length 200 bins. If a spectral bin $|X(n,i)|$ exceeds the moving mean plus half a moving standard deviation we consider it a locally salient bin:

$$|X(n,i)| > \mu(n,i) + 0.5 \cdot \sigma(n,i). \tag{2}$$

To eliminate noise peaks at low amplitudes we consider as globally salient only those bins which have an amplitude not more than 25dB below that of the global maximum bin amplitude:

$$|X(n,i)| > 10^{-2.5} \cdot \max_{u,v}\{|X(u,v)|\}. \tag{3}$$

We consider only those bins that are both locally and globally salient, i.e. both inequalities (2) and (3) hold. From each region of consecutive peaks we pick the bin that has the maximum amplitude and estimate the true frequency by quadratic interpolation of the log magnitude of the peak bin and its two surrounding bins (Smith and Serra, 1987; Smith, 2010), as follows. Suppose $a_p = \log|X(n,p)|$ is a local peak in the log magnitude spectrum, that is, $a_{p-1} < a_p < a_{p+1}$ (where we drop the time index n for convenience). Then the three points $(-1, a_{p-1}), (0, a_p)$, and $(1, a_{p+1})$ uniquely define a parabola with maximum at:

$$\delta = \frac{a_{p-1} - a_{p+1}}{2(a_{p-1} - 2a_p + a_{p+1})}, \tag{4}$$

where $-0.5 \le \delta \le 0.5$ is the offset from the integer bin location p, so that the quadratically interpolated peak frequency is given by $\frac{(p+\delta)f'_s}{N}$.

The next step is the "conservative" processing, in which we delete many potential fundamental frequencies. For each peak frequency f_0, any other peak whose frequency is within 50 cents of a multiple of f_0 is deleted. Peaks in the same frequency bins as those deleted, in a neighbourhood of ±2 frames, are also deleted. For testing the efficacy of this approach, we compare it with an otherwise identical method which treats all spectral peaks as if they were fundamentals (see section IV).

In order to sort the remaining frequency estimates into semitone bins we determine the standard pitch f^{st} by taking the median difference (in cents) of those peaks that are within half a semitone of the nominal standard pitch (415 Hz). Based on the new standard pitch f^{st} each peak frequency is assigned to one of 45 pitches ranging from MIDI note 36 (C2) to 80 (G♯5). Any frequency peaks outside of this range are deleted.

In order to discard spurious data we delete any peaks which lack continuity in time, i.e. where the continuous duration of the peak is less than a threshold T. Results for various values of T were compared in previous work (Tidhar *et al.*, 2010b); in this paper we use $T = 0.3$s. Remaining consecutive peaks are grouped as notes, specified by onset time, duration, and MIDI pitch number.

C. Partial Frequency Estimation

For each note object w given by the conservative transcription, an initial estimate of the frequency f_k^w of partial k is computed using equation 1, the fundamental frequency found in the conservative transcription stage and an estimate of the inharmonicity factor B. Initially B is set to 2×10^{-5}, but after the first run of the algorithm, this is replaced by a frequency-dependent estimate of B (see subsection III.D below). We then perform STFT analysis using the following parameters: $f_s = 44100$ Hz, no downsampling, Blackman-Harris window with support size of 4096 samples, zero padding factor $z = 4$ ($N = 16384$), and hop size of 1024 samples. For each partial frequency, a spectral peak in a window of ± 30 cents around f_k^w is found and the peak location is refined using quadratic interpolation (see subsection III.B). After quadratic interpolation, a bias correction is applied based on the window shape and zero padding factor (Abe and Smith, 2004, equations 1 and 3):

$$\delta' = \delta + \xi_z \delta(\delta - 0.5)(\delta + 0.5) \tag{5}$$

where δ' is the bias-corrected offset in bin location, δ is the quadratically interpolated offset (equation 4), z is the zero-padding factor, $\xi_z = c_0 z^{-2} + c_1 z^{-4}$ is the bias correction factor and the constants $c_0 = 0.124188$ and $c_1 = 0.013752$ were determined empirically for the Blackman-Harris window by Abe and Smith (2004, table 1). If no peak is found in the search window, that partial and frame combination is ignored. For each note and frame of the note's duration (as estimated by the conservative transcription step), the frequency of the first 40 partials $f_1, ..., f_{40}$ is estimated. From these values the fundamental frequency and inharmonicity can be determined.

D. Inharmonicity Estimation

Given the frequencies f_j and f_k of any two partials j and k, equation 1 can be rearranged to obtain an estimate of B:

$$B_{j,k} = \frac{j^2 f_k^2 - k^2 f_j^2}{k^4 f_j^2 - j^4 f_k^2}. \tag{6}$$

The frequency estimates obtained above are only accurate if there is no interference from partials of other tones. Although we avoid some cases of interference using conservative transcription, it does not cover all cases, and the prevalence of musical intervals involving coincident or overlapping partials makes it impossible to avoid interference entirely. To mitigate the effects of these errors we use robust statistics to discard outliers and obtain our final estimate of the inharmonicity of each note. The median is a suitably robust measure of central tendency in the presence of measurement noise, unlike the mean which is prone to bias from outliers. B is therefore estimated as the median of all $B_{j,k}$ values, where values from successive frames of a single note are included in the median computation. We also compute a measure of the reliability of this estimate using the inter-quartile range (IQR), defined as the difference between the third and first quartiles. The IQR is chosen because it is a robust measure of the statistical dispersion of the data, as it is less susceptible to outliers than measures such as the standard deviation. Having obtained a single value of B for each note, we iterate the partial frequency and inharmonicity estimation stage using the newly estimated B to guide the search for spectral peaks (see subsection III.C above) until they converge, or if they fail to converge the iteration is terminated after ten steps. Approximately 40% of note estimates converge immediately, 20% require a further one or two steps, and 25% are terminated after ten steps.

E. Integration of Partial Frequency Estimates

Once the inharmonicity of a string has been estimated, each partial frequency provides an independent estimate $\hat{f}_0(k)$ of the theoretical fundamental, by substitution in equation 1. (The theoretical fundamental would be the fundamental frequency of the string if it had no stiffness; the first partial is sharper by a factor of $\sqrt{1+B}$.) To obtain a robust value, we take a single median over all frames and partials, and compute the inter-quartile range as an inverse measure of confidence in the estimate. The output of this stage is a list of frequency and inharmonicity estimates, together with their inter-quartile ranges, for each note identified by the conservative transcription algorithm, where the transcribed notes are described by a MIDI note number, onset time and duration d_i (where i is the index of the note). By ignoring the octave, the MIDI note number can be converted to a pitch class $p_i \in P = \{C, C\#, D, \ldots, B\}$. The corresponding frequency estimates are also converted to deviation c_i from equal temperament, measured in cents.

F. Classification

To test the fundamental frequency estimation we use the deviations c_i to classify the 48 recordings by the temperament from which they differ least in terms of the theoretical profiles shown in Table I. For each pitch class k the estimate $\hat{c}_k$ of the deviation in cents is obtained by taking the weighted mean of the deviations over all the notes belonging to that pitch class:

$$\hat{c}_k = \frac{\sum_{i:p_i=k} c_i w_i}{u_k}, \qquad k \in P, \tag{7}$$

where the note weight $w_i = \frac{d_i}{q_i}$ is the quotient of the note duration d_i and the inter-quartile range q_i of the fundamental frequency estimates for note i, and the pitch class weight u_k for pitch class k is given by $u_k = \sum_{i:p_i=k} w_i$. Given this estimate $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_{12})$ and a temperament profile $c^0 = (c_1^0, \ldots, c_{12}^0)$, we calculate the divergence

$$d(\hat{c}, c^0) = \sum_{k \in P} v_k (\hat{c}_k - c_k^0 - r)^2 \tag{8}$$

between estimate and profile, where $v_k = \left(u_k / \sum_{i \in P} u_i\right)^2$ is the squared relative weight of the kth pitch class in the note list, and $r = \sum_{i \in P} v_i(\hat{c}_i - c^0_i) / \sum_{i \in P} v_i$ is the offset in cents which minimises the divergence and thus compensates for deviations in the reference tuning frequency (pitch A4) from the 415 Hz reference assumed in previous calculations. A piece is classified as having the temperament whose profile c^0 differs least from it in terms of $d(\hat{c}, c^0)$.

The weight v_i favours pitch classes that have longer cumulative durations and lower inter-quartile ranges. In particular, any pitch classes that are not in the note list are discarded. For the four pieces selected, all pitch classes are present, but the distribution is uneven. For example, in the score of the Bach Prelude, the frequencies of occurrence of each pitch class range from 4 notes ($C\sharp$ and $A\flat$) to 113 notes (G).

IV. RESULTS

A. Inharmonicity

Figure 2 shows the inharmonicity factor B for each tone detected for a real (left) and a synthesised (right) recording of Goldberg Variation 21 (using Vallotti temperament). The two graphs show similar trends in inharmonicity, which increases with decreasing string length (B is inversely proportional to the fourth power of string length). The synthesised harpsichord exhibits a slightly steeper curve, with the difference visible on the manually fitted curves. The range of values (10^{-5} to 10^{-4}) agrees with those published elsewhere (Välimäki *et al.*, 2004). Most of the outliers are notes for which the intra-note agreement is low (represented by the size of the circles).

B. Fundamental Frequency Estimation

Using the known temperament for each recording, we compare the measured deviations from equal temperament with the corresponding expected values from Table I, and report the errors as the mean absolute differences (in cents) across various data sets in Table II. First, we show the errors across all recordings with and without conservative transcription. Using spectral peaks (SP) as fundamentals, the mean absolute error is 2.1 cents, compared to 1.5 cents using conservative transcription (CT). As in previous work, conservative transcription has a positive impact on results. The remaining results are based on the conservative transcription results only. The second section shows the effect of audio source: real and synthesised harpsichord. For the real harpsichord (RH), the mean absolute error is 2.4 cents, as compared with 0.6 cents for synthesised (PT) recordings. It is not possible to say whether this reflects inaccuracies in the tuning (either due to the limits of tuning ability or the instrument going out of tune after tuning), or the greater difficulty of signal processing due to the more complex timbres, room reverberation and noise inherent in any acoustic recording. The greater spread of deviations (see below) does not necessarily imply the latter interpretation, as each pitch class consists of a set of notes from different octaves, which could vary in their deviation values due to tuning imprecision. The third set of results, by temperament (for the real harpsichord recordings), shows some differences which suggest some variability in the accuracy of tuning across the different temperaments, with Just Intonation, Quarter Comma Meantone and Vallotti tunings being less accurate than the other tunings.

TABLE II. Comparison of the influence of four factors on the mean absolute difference between measured pitch and the corresponding temperament profiles. All values are in cents (hundredths of a semitone). The four factors are: Transcription (SP: using all spectral peaks; CT: using only peaks identified by the conservative transcription algorithm); Instrumental Source (RH: recordings of a real harpsichord; PT: recordings from the Pianoteq synthesiser); Temperament (ET: Equal Temperament; V: Vallotti; FC: Fifth Comma; QCMT: Quarter Comma Meantone; SCMT: Sixth Comma Meantone; JI: Just Intonation); and Musical Piece (Chrom: one-octave chromatic scale; Prel: J.S. Bach's Prelude 1 in C Major; Mén: F. Couperin's La Ménetou; Var21: J.S. Bach's Goldberg Variation 21).

Transcription	SP:	2.1	CT:	1.5
Instrumental Source	RH:	2.4	PT:	0.6
Temperament	ET:	1.2	FC:	1.6
	JI:	4.2	QCMT:	3.6
	SCMT:	1.2	V:	2.8
Piece	Chrom:	2.2	Prel:	2.5
	Mén:	2.8	Var21:	2.1

The final set of results (by piece, for the real harpsichord recordings) reveals some unexpected differences. The slow chromatic scale, selected as a baseline since it consists only of individual notes, yielded results slightly worse than those of the Bach Goldberg Variation 21, a polyphonic piece. Likewise the error for the Bach Prelude, where only one note at a time is played (although the notes do overlap) was higher than both of these, while the more complex and highly ornamented Couperin piece had the highest average error, as expected. Two known factors contribute to the unexpected ranking: the conservative transcription algorithm selects suitable notes for the frequencies of partials to be measured, compensating at least in part for the complexity of the piece; and the chromatic scale has only one instance of each pitch class (except the initial and final C), whereas the other pieces have many instances of each pitch class, allowing more robust estimates to be made, despite the interfering notes.

To indicate the spread of errors in pitch deviation estimation on a per-note basis, figure 3 shows the results for the Vallotti temperament recordings of the Goldberg Variation 21. The fundamental frequency difference of each measured note relative to the same note in equal temperament is shown, compared with the expected values for the Vallotti temperament in the staircase plots. Fundamental frequencies appear on the horizontal axis,

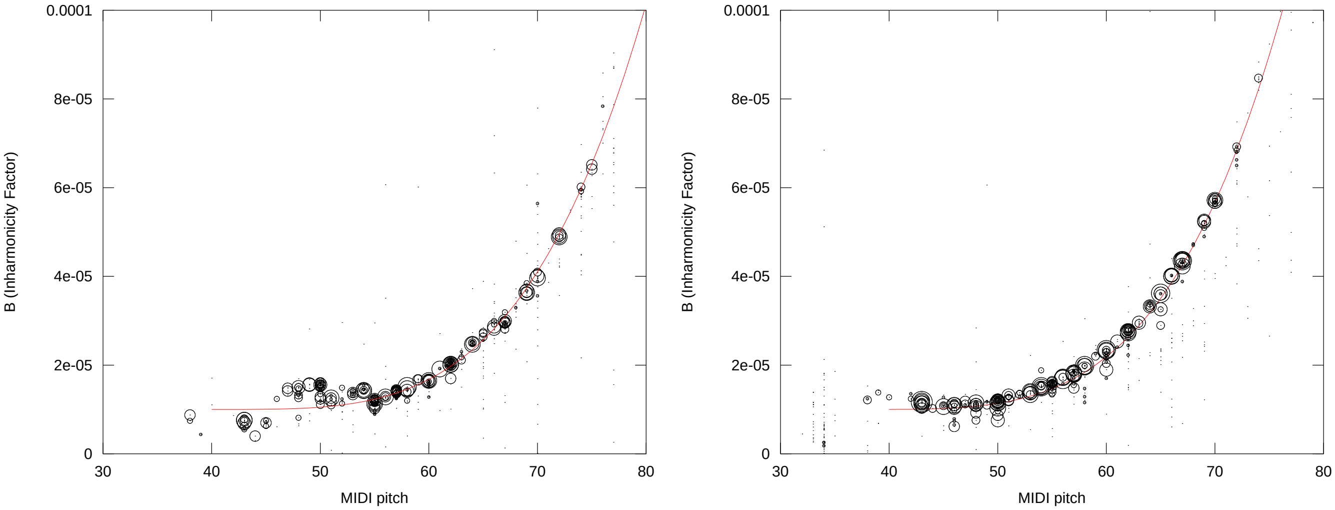

FIG. 2. Inharmonicity factor B as a function of pitch in MIDI (semitone) units for a real harpsichord (left) and the Pianoteq synthesiser (right). The data was extracted fully automatically from recordings of Bach's Goldberg Variation 21. Each circle represents a note, where the size represents the confidence, computed from the inter-quartile range (larger sizes represent smaller IQRs).

mapped to their pitch class, and the frequency difference from equal temperament is shown in cents on the vertical axis.

C. Classification Results

Table III shows classification results for various versions of the algorithm. The first row shows previous results (Tidhar *et al.*, 2010b, method M-CQIFFT), where inharmonicity was ignored, and the median of $\frac{f_k}{k}$ values for the first 12 harmonics was used as an estimate of the fundamental frequency. The second row contains results for estimation of B using equation 1 without any iteration, while up to ten iterations of the fundamental frequency and inharmonicity estimation algorithm were performed to obtain the results in the third row. The final row shows the additional effect of early deletion of any partials whose inharmonicity estimates are more than half the inter-quartile range from the median. For all cases, short note deletion was employed, so that notes with a transcribed length less than 0.3s were not used.

The classification results confirm that automatic temperament recognition can be performed with a high level of accuracy for the chosen set of temperaments. The individual differences between methods are less informative, as they involve the reclassification of a very small number of items. Figure 4 shows this more clearly, where the divergence $d(\hat{c}, c^0)$ is shown for all six temperaments for the real harpsichord (left) and synthesised (right) recordings of Couperin's La Ménetou. In the cases where misclassification occurs, the divergence of the correct temperament is very close to that of the winning temperament.

TABLE III. Percentage of recordings automatically classified with the correct temperament. The columns correspond to the use of the 24 recordings of real harpsichord (RH) and the 24 recordings synthesised with Pianoteq (PT), preprocessed with spectral peak detection (SP) or conservative transcription (CT) respectively. The rows correspond to four different approaches to inharmonicity estimation: none (ICASSP'10); using one (Single Iteration) or up to ten (Multiple Iterations) iterations of the fundamental frequency and inharmonicity algorithm; and multiple iterations with prefiltering of the data by Outlier Deletion.

Data Type:	RH		PT	
Overtone Removal:	SP	CT	SP	CT
ICASSP'10	79	92	88	96
Single Iteration	75	88	96	100
Multiple Iterations	79	92	96	100
Outlier Deletion	79	83	100	100

V. DISCUSSION AND CONCLUSION

In matching the measured frequencies to theoretical definitions of temperaments, we have used a model of temperament which ignores inharmonicity. That is, a pure fifth is defined as a frequency ratio of $\frac{3}{2}$ between fundamental frequencies, rather than the fundamental frequency ratio for which the third harmonic of the lower tone corresponds to the second harmonic of the higher tone. The latter definition would correspond better to tuning practice (where a beat-free fifth would be considered pure). Likewise, by grouping all notes within a pitch class, we assume that octaves are not stretched.

It is straightforward to compute the effect of making this assumption, given the values of B for each note, as estimated in this work. For two tones i and j with

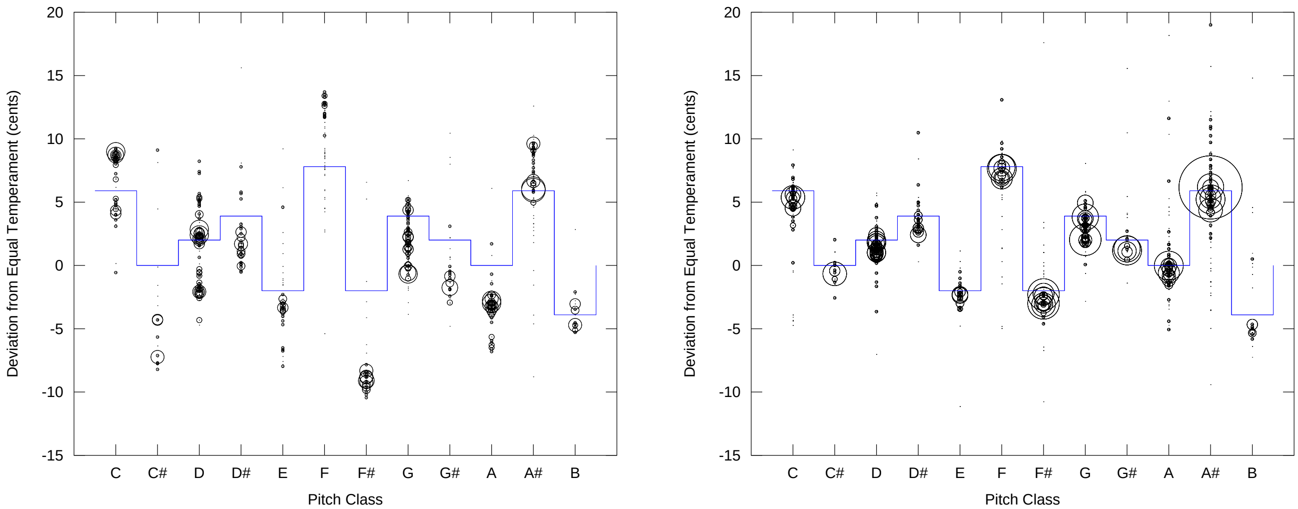

FIG. 3. Pitch difference in cents relative to equal temperament, as measured from two recordings of Variation 21 from the Goldberg Variations using Vallotti temperament (left: harpsichord; right: synthesised). Each circle represents a note, where the size represents the confidence. The staircase plot shows the expected values for the Vallotti temperament.

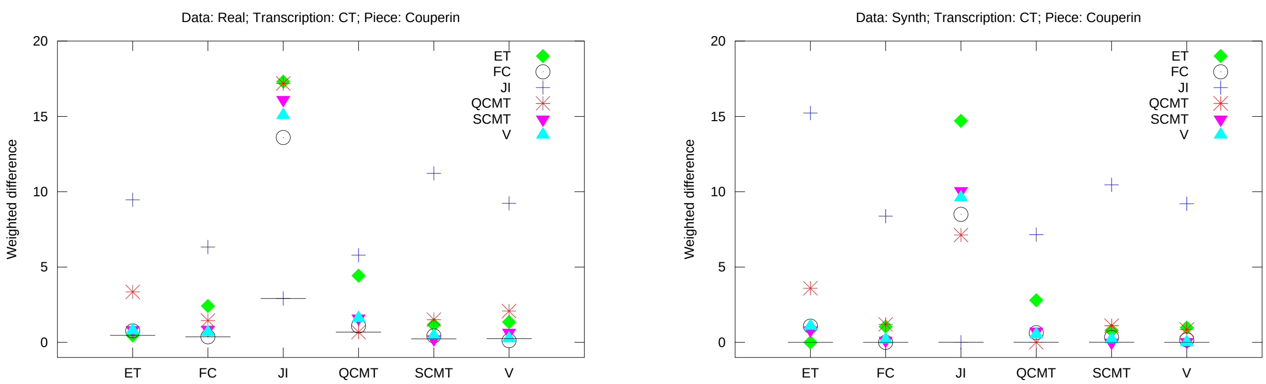

FIG. 4. Divergence $d(\hat{c}, c^0)$ between measured and theoretical temperaments for the real harpsichord (left) and synthesised (right) recordings of Couperin's La Ménetou (see equation 8). The actual temperament is shown on the horizontal axis, and the divergence on the vertical axis. Horizontal lines mark the divergence of the correct temperament in each case.

inharmonicity coefficients B_i and B_j respectively, tuned to a frequency ratio $q : p$ (so that the frequency of the pth partial of i is equal to the frequency of the qth partial of j), the deviation D in cents of the fundamental frequency ratio from the ratio $q : p$ is given by:

$$D(i, j, p, q) = 1200 \log_2 \left(\frac{\sqrt{1 + p^2 B_i}}{\sqrt{1 + q^2 B_j}} \right). \qquad (9)$$

Using our estimates of B, this deviation is less than 0.1 cent for octave intervals (ratio 1:2), 0.25 cent for fifths (ratio 2:3), and 0.5 cent for major thirds (ratio 4:5), across the whole range of the harpsichord. If the top octave is not used, the maximum deviations are smaller by a factor of 5. These deviations are small compared to our precision in frequency estimation, and thus do not adversely affect our results.

We have shown that given a standard recording of a musical work for solo harpsichord, it is possible to estimate the inharmonicity of each key and ascertain the tuning of each pitch class with a precision of 1–2 cents, which appears to be no worse than the precision of tuning the instrument. One of the difficulties in this work is the lack of ground truth data against which more extensive testing could be performed. In future work, we plan to expand our data set to include more pieces and temperaments, and to use our system in a large-scale analysis of historical harpsichord recordings. This will extend existing knowledge of inharmonicity by gathering measurements from many instruments, and inform the study of historical music performance by analysis of the temperaments employed in the recordings.

Acknowledgments

This research was part of the OMRAS2 project (`http://www.omras2.org`), supported by EPSRC grant

EP/E017614/1.

Abe, M. and Smith, J. (**2004**). "CQIFFT: Correcting bias in a sinusoidal parameter estimator based on quadratic interpolation of FFT magnitude peaks", Technical Report STAN-M-117, Center for Computer Research in Music and Acoustics, Department of Music, Stanford University.

Anderson, B. and Strong, W. (**2005**). "The effect of inharmonic partials on pitch of piano tones", J. Acoust. Soc. Am. **117**, 3268–3272.

Barbour, J. (**2004/1951**). *Tuning and Temperament, A Historical Survey*, 256 pages (Dover, Mineola, NY).

Benetos, E. and Dixon, S. (**2010**). "Multiple-F0 estimation of piano sounds exploiting spectral structure and temporal evolution", in *Int. Speech Comm. Assoc. Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA2010)*, 13–18.

de Cheveigné, A. (**2006**). "Multiple f0 estimation", in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, edited by D. Wang and G. Brown, 45–79 (IEEE Press/Wiley, Piscataway, NJ).

de Cheveigné, A. and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music", J. Acoust. Soc. Am. **111**, 1917–1930.

Di Veroli, C. (**2009**). *Unequal Temperaments: Theory, History, and Practice*, 456 pages (Bray Baroque, Bray, Ireland).

Dixon, S. (**2001**). "Automatic extraction of tempo and beat from expressive performances", J. New Mus. Res. **30**, 39–58.

Earis, A., Daly, M., Fei, S., and Thompson, R. (**2007**). "Acoustical studies of historical keyboard instruments in the Royal College of Music museum of instruments", in *Proc. 19th Int. Congr. Acoust.*, MUS–02–006.

Emiya, V., Badeau, R., and David, B. (**2010**). "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle", IEEE Trans. Audio, Speech and Lang. Proc. **18**, 1643–1654.

Fletcher, H. (**1964**). "Normal vibration frequencies of a stiff piano string", J. Acoust. Soc. Am. **36**, 203–209.

Fletcher, H., Blackham, E., and Stratton, R. (**1962**). "Quality of piano tones", J. Acoust. Soc. Am. **34**, 749–761.

Fletcher, N. (**1977**). "Analysis of the design and performance of harpsichords", Acustica **37**, 139–147.

Fletcher, N. and Rossing, T. (**1998**). *The Physics of Musical Instruments*, 756 pages (Springer, New York, NY).

Gerhard, D. (**2003**). "Pitch extraction and fundamental frequency: History and current techniques", Technical Report TR-CS 2003-06, Department of Computer Science, University of Regina, Regina, Canada.

Järveläinen, H., Välimäki, V., and Karjalainen, M. (**2001**). "Audibility of the timbral effects of inharmonicity in stringed instrument tones", Acoust. Res. Lett. Online **2**, 79–84.

Klapuri, A. (**2003**). "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness", IEEE Trans. Speech and Audio Processing **11**, 804–816.

Klapuri, A. (**2009**). "A method for visualizing the pitch content of polyphonic music signals", in *10th Int. Soc. Mus. Inf. Retrieval Conf.*, 615–620.

Klapuri, A. and Davy, M., eds. (**2006**). *Signal Processing Methods for Music Transcription*, 440 pages (Springer, New York, NY).

Lundin, R. (**1947**). "Toward a cultural theory of consonance", J. Psych. **23**, 45–49.

Mauch, M. and Dixon, S. (**2010**). "Simultaneous estimation of chords and musical context from audio", IEEE Trans. Audio, Speech and Lang. Processing **18**, 1280–1289.

McDermott, J., Lehr, A., and Oxenham, A. (**2010**). "Individual differences reveal the basis of consonance", Current Biology **20**, 1035–1041.

Moore, B., Glasberg, B., and Peters, R. (**1985**). "Relative dominance of individual partials in determining the pitch of complex tones", J. Acoust. Soc. Am. **77**, 1853–1860.

Noland, K. and Sandler, M. (**2006**). "Key estimation using a Hidden Markov Model", in *7th Int. Conf. Mus. Inf. Retrieval*, 121–126.

Palisca, C. and Moore, B. (**2010**). "Grove music online", `http://www.grovemusic.com/`, accessed 15 January 2010.

Pianoteq (**2010**). "Pianoteq 3 true modelling", `http://www.pianoteq.com`, accessed 23 November 2010.

Rasch, R. (**2002**). "Tuning and temperament", in *The Cambridge History of Western Music*, edited by T. Christensen, 193–222 (Cambridge University Press, Cambridge, UK).

Rauhala, J., Lehtonen, H.-M., and Välimäki, V. (**2007**). "Fast automatic inharmonicity estimation algorithm", J. Acoust. Soc. Am. **121**, EL184–189.

Sethares, W. (**1999**). *Tuning, Timbre, Spectrum, Scale*, 448 pages (Springer, Berlin, Germany).

Shankland, R. and Coltman, J. (**1939**). "The departure of the overtones of a vibrating wire from a true harmonic series", J. Acoust. Soc. Am. **10**, 161–166.

Smith, J. (**2010**). "Spectral audio signal processing: March 2010 draft", `http://ccrma.stanford.edu/~jos/sasp/`, accessed 23 November 2010.

Smith, J. and Serra, X. (**1987**). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation", in *Proc. Int. Comp. Mus. Conf.*, 290–297.

Terhardt, E. (**1977**). "The two-component theory of musical consonance", in *Psychophysics and Physiology of Hearing*, edited by E. Evans and J. Wilson, 381–390 (Academic Press, London).

Tidhar, D., Fazekas, G., Mauch, M., and Dixon, S. (**2010**a). "TempEst: Harpsichord temperament estimation in a semantic-web environment", J. New Mus. Res. **39**, 327–336.

Tidhar, D., Mauch, M., and Dixon, S. (**2010**b). "High precision frequency estimation for harpsichord tuning classification", in *Proc. IEEE Int. Conf. Acoust. Speech and Sig. Processing*, 61–64.

Tzanetakis, G. and Cook, P. (**2002**). "Musical genre classification of audio signals", IEEE Trans. Speech and Audio Processing **10**, 293–302.

Välimäki, V., Penttinen, H., Knif, J., Laurson, M., and Erkut, C. (**2004**). "Sound synthesis of the harpsichord using a computationally efficient physical model", EURASIP J. Appl. Sig. Processing **2004**, 934–948.

von Helmholtz, H. (**1863**). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik (On the Sensations of Tone as a Physiological Basis for the Theory of Music)*, 605 pages (Friedrich Vieweg und Sohn, Braunschweig, Germany).

Wen, X. and Sandler, M. (**2009**). "Notes on model-based non-stationary sinusoid estimation methods using derivatives", in *12th Int. Conf. Digital Audio Effects*, 113–120.

Young, R. (**1952**). "Inharmonicity of plain wire piano strings", J. Acoust. Soc. Am. **24**, 267–273.