



City Research Online

City, University of London Institutional Repository

Citation: Tidhar, D., Mauch, M. and Dixon, S. (2010). High precision frequency estimation for harpsichord tuning classification. Paper presented at the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, 14-03-2010 - 19-03-2010, Dallas, USA.

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <http://openaccess.city.ac.uk/4348/>

Link to published version: <http://dx.doi.org/10.1109/ICASSP.2010.5496213>

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online:

<http://openaccess.city.ac.uk/>

publications@city.ac.uk

HIGH PRECISION FREQUENCY ESTIMATION FOR HARPSICHORD TUNING CLASSIFICATION

Dan Tidhar, Matthias Mauch and Simon Dixon

Queen Mary University of London
Centre for Digital Music

ABSTRACT

We present a novel music signal processing task of classifying the tuning of a harpsichord from audio recordings of standard musical works. We report the results of a classification experiment involving six different temperaments, using real harpsichord recordings as well as synthesised audio data. We introduce the concept of conservative transcription, and show that existing high-precision pitch estimation techniques are sufficient for our task if combined with conservative transcription. In particular, using the CQIFFT algorithm with conservative transcription and removal of short duration notes, we are able to distinguish between 6 different temperaments of harpsichord recordings with 96% accuracy (100% for synthetic data).

Index Terms— Music, Pitch Estimation, Temperament

1. INTRODUCTION

Tuning of musical instruments has occupied musical and scientific minds at least since the days of Pythagoras of Samos. Musical consonance (combinations of notes that “sound good” together) is derived from the sharing of partial frequencies. As musical instruments produce harmonic tones (all partials are integer multiples of some fundamental frequency), the condition for consonance of two notes is that their fundamental frequencies are in a simple integer ratio. For example, if the ratio of frequencies f_a and f_b is $p : q$, then every p th partial of f_b is equal in frequency to each q th partial of f_a . Based on this observation, it is desirable to build musical scales out of such consonant integer ratio intervals. However, it turns out to be impossible to build a scale in which all intervals are maximally consonant. Some compromise must be made, and the way in which this compromise takes place defines the temperament, or tuning system, being used.

Attempting to build a system to classify musical recordings by temperament presents particular signal processing challenges. First, the differences between temperaments are small, of the order of a few cents, where a cent is one hundredth of a semitone, or one twelve-hundredth of an octave. For example, if A = 415 Hz is used as the reference pitch, 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 we examine in this work. 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.

The second problem is that in musical recordings, notes rarely occur in isolation. There are almost always multiple 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 the partials coincide. The third main problem is that we do not know when each note is played, so that when we detect a sinusoid, we can not be sure whether this is a fundamental frequency or a partial of another fundamental. The ability to distinguish between these cases is crucial to successful temperament classification. For example, if an A at 110 Hz is played, the signal will also have a peak at 330 Hz, which is 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 that note would be biased.

Besides posing a non-trivial research challenge, there are many ways in which temperament estimation can be useful to musicians, musicologists, and listeners. Such a classifier is necessary for music retrieval according to temperament, which would be useful for educational purposes, such as ear training, for professional users such as keyboard tuners and performers, by providing them feedback about tuning accuracy and stability, as well as helping to classify creative temperaments (i.e. those done without strictly adhering to known recipes) and determine some of their properties. Furthermore, musicologists are likely to find temperament estimators very useful when studying performance practice from recordings.

1.1. Temperament

Temperament is covered thoroughly elsewhere [1, 2], but we shall provide a brief formulation of the problem here. Consider 12 fifth steps (a fifth is 7 semitones; a pure or consonant fifth has a frequency ratio of $3/2$) and 7 octave steps (frequency ratio $2/1$). On a keyboard instrument both of these sequences of intervals lead to the same key, despite the fact that $(3/2)^{12} > 2^7$. The ratio between the two sides of this inequality is referred to as “the comma”, or more precisely, the Pythagorean comma. One way of defining particular temperaments is according to the distribution of the comma¹ along the circle of fifths. Equal temperament, for example, can be defined as such where each fifth on the circle is diminished by exactly the same amount, i.e. $1/12$ of a (Pythagorean) comma.

The six temperaments used in this work are equal temperament, Valotti, fifth-comma (Fifth), quarter-comma meantone (QCMT), sixth-comma meantone (SCMT), and just intonation. In a Valotti temperament, 6 of the fifths are diminished by a $1/6$ comma each, and the other 6 fifths are left pure. In the fifth-comma temperament we use, five of the fifths are diminished by a $1/5$ comma each, and the remaining 7 are pure. In a quarter-comma meantone temperament, 11 of the fifths are shrunk by $1/4$ of a comma, and the one

¹This research is part of the OMRAS2 project (www.omras2.org), supported by EPSRC grant EP/E017614/1.

¹For simplicity, we omit discussion of the Syntonic comma, treating it is the same as the Pythagorean.

Note	Valotti	Fifth	QCMT	SCMT	Just
C	5.9	8.2	10.3	4.9	15.6
C#	0	-1.6	-13.7	13	-13.7
D	2	2.7	3.4	1.6	-2
D#	3.9	2.3	20.6	9.8	-9.8
E	-1.9	1.9	-3.4	-1.6	2
F	7.8	6.3	13.7	6.5	13.7
F#	-1.9	-3.5	10.3	-4.9	15.6
G	3.9	5.5	6.8	3.3	17.6
G#	1.9	0.4	-17.1	11.4	-11.7
A	0	0	0	0	0
Bb	5.9	4.3	17.1	8.1	11.7
B	-3.9	-0.8	-6.8	-3.3	3.9

Table 1. Deviations (in cents) from equal temperament for the five unequal target temperaments in our experiment.

remaining fifth is $7/4$ of a comma larger than pure. In sixth-comma meantone, 11 fifths are shrunk by $1/6$ of a comma, and the one remaining fifth is $7/6$ comma larger than pure. The just-intonation temperament we used is based on the reference tone *A*, and all other tones are calculated as integer ratios corresponding to harmonics of the reference tone. The ratios we used are given by the following vector, representing twelve chromatic tones above the reference *A*:

$$(16/15, 9/8, 6/5, 5/4, 4/3, 45/32, 3/2, 8/5, 5/3, 9/5, 15/8, 2/1)$$

The deviations (in cents) from equal temperament of the five other temperaments we use, are given in Table 1.

Apart from being relatively common, this set of six temperaments represents different categories: *well* temperaments are those which enable usage of all keys (though not necessarily equally well), and *regular* temperaments are those with at least 11 fifths of equal size. Equal temperament is both well and regular, just intonation is neither well nor regular, Valotti and fifth-comma are well and irregular, and the two variants of meantone (quarter and sixth comma) are regular but not well.

1.2. Experiment outline

We set the task of distinguishing between the six different commonly-used temperaments mentioned above: equal temperament, Valotti, fifth comma, quarter-comma meantone, sixth-comma meantone, and just intonation. The framework of our classification task can be summarised as follows: given an audio recording of an unknown musical piece, we assume the instrument is tuned according to one of the temperaments mentioned above, and that we know the approximate standard tuning ($A=440$ Hz, $A=415$ Hz, etc.), but we allow for minor deviations from this nominal tuning frequency or from the temperament.

Obtaining ground-truth data for a temperament-recognition experiment proved to be non-trivial for several reasons. Most of the commercially available recordings do not specify the harpsichord temperament. 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 initial dataset ourselves. The dataset consists of real harpsichord recordings played by Dan Tidhar on a Rubio double-manual harpsichord in a small hall and of synthesised recordings rendered from MIDI using a physically-modelled harpsichord sound on Pianoteq [3]. For

each temperament and rendering alternative (real vs synthesised), we recorded 4 different tracks (i.e. a total of 48) consisting of a slow ascending chromatic scale, J. S. Bach’s first prelude in C Major from the *Well-tempered Clavier*, F. Couperin’s *La Ménéteau*, and J. S. Bach’s Variation 21 from *Goldberg Variations*. The choice of tracks encompasses various degrees of polyphony, various degrees of chromaticism, as well as various speeds. All tracks were tuned to a reference frequency of approximately $A=415$ Hz.

2. METHOD

In order to obtain accurate pitch estimates of unknown notes in the presence of multiple simultaneous tones, we developed a 2-stage approach: conservative (i.e. high precision, low recall²) transcription, which identifies the subset of notes which are easily detected, followed by an accurate frequency domain pitch estimation step for the notes determined in the first stage.

2.1. Conservative Transcription (CT)

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, so we introduce the concept of conservative transcription, which identifies only those sinusoids that we are confident correspond to fundamental frequencies and omits any unsure candidates. 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.

CT consists of three main parts: computation of framewise amplitude spectra with a standard STFT; sinusoid detection through peak-picking, which yields first frequency estimates; 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 other 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 to $f_s = 11025$ Hz, we compute the STFT using a Hamming window, a frame length of 4096 samples (370 ms), a hop size of 256 samples (23 ms, i.e. 15/16 overlap), and a zero padding factor of 2.

In order to remove harmonics of sinusoids in the amplitude spectrum $|X(n, i)|$, one has to first identify the sinusoids. We use two adaptive thresholding techniques to find peak regions. To find locally significant bins of frame n , we calculate the running weighted mean $\mu(n, i)$ and the running 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 running mean plus half a running standard deviation we consider it a locally salient bin:

$$|X(n, i)| > \mu(n, i) + 0.5 \cdot \sigma(n, i) \quad (1)$$

To eliminate noise peaks at low amplitudes we consider as globally salient only those bins which have an amplitude within 25dB below that of the global maximum frame amplitude:

$$|X(n, i)| > 10^{-2.5} \cdot \max_{j,k} \{|X(j, k)|\} \quad (2)$$

²*Precision* is the fraction of transcribed notes that are correct; *recall* is the fraction of played notes that are transcribed.

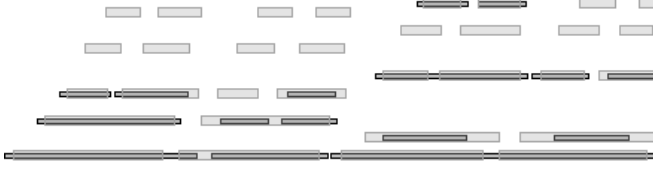


Fig. 1. The opening of Bach’s *Prelude in C Major*, shown in piano-roll notation (log frequency vs time). The performed notes are shown in light grey with the darker notes from the Conservative Transcription superimposed over them.

We will consider only those bins that are both locally and globally salient, i.e. both inequalities (1) and (2) 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 magnitude of the peak bin and its two surrounding bins [4]. This frequency estimate (denoted QIFFT) is used as the baseline pitch estimate in Section 4.

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. In addition, peaks in the same frequency bins in a neighbourhood of ± 2 frames are 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 (method SP in Table 2).

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 60 pitches ranging from MIDI note 21 to 80. 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 are given in Section 4. Remaining consecutive peaks are grouped as notes, specified by onset time, duration, and MIDI pitch number. Figure 1 depicts an extract of a conservative transcription compared to the underlying performed notes.

2.2. Pitch Estimation

Time-domain pitch estimation methods such as ACF and YIN are unsuitable due to the bias caused by the presence of multiple simultaneous tones. Thus we focus on three frequency domain techniques: the quadratic interpolated FFT (QIFFT) [4], the QIFFT with correction for the bias of the window function (CQIFFT) [5], and the instantaneous frequency calculated with the phase vocoder (PV) [6]. More advanced estimation algorithms which admit frequency and/or amplitude modulation were deemed unnecessary.

For each method, a pitch estimate is generated for each note object given by the conservative transcription. Our baseline pitch estimation, denoted QIFFT, is the frequency estimated in the CT step described above in subsection 2.1. The QIFFT estimate of the fundamental frequency is computed for each frame in the note, and the mean is returned as the pitch of the note.

Our implementation of the CQIFFT method uses different parameter values to the previous method: no downsampling, a

Blackman-Harris window with support size of 4096 samples, zero padding factor of 4, and hop size of 1024 samples. After quadratic interpolation, a bias correction is applied based on the window shape and zero padding factor [5, equations 1 and 3]. For each note, the frequency of the first 12 partials f_1, \dots, f_{12} is estimated as the median of the CQIFFT values. The final pitch estimate is described below in subsection 2.3.

The third pitch estimation method (PV) uses the same parameters as CQIFFT, but estimates frequency using the instantaneous frequency [6], which is calculated from the rate of change of phase between frames in the relevant frequency bin. The frequency estimates for each of the first 12 partials are computed with the median.

2.3. Integration of Partial

An ideal vibrating string has spectral energy at a fundamental frequency and at integer multiples of that frequency. String instruments such as the harpsichord emit approximately harmonic notes, where the inharmonicity is primarily due to the stiffness of the string and results in the frequency being slightly greater than the ideal integer multiple of the fundamental [7]:

$$f_k = k f_0 \sqrt{1 + B k^2} \quad (3)$$

where f_k is the frequency of the k th partial, and B is a constant related to the physical properties of the string. If the inharmonicity is “nearly negligible” [7, p. 343], the partial frequencies can provide independent estimates $\hat{f}(k)$ of the fundamental, by dividing the frequency by the partial number:

$$\hat{f}(k) = \frac{f_k}{k} \quad (4)$$

Välämäki et al. [8] claim that the inharmonicity of harpsichord strings is not negligible, particularly for the lower pitches, citing measured values of B between 10^{-5} and 10^{-4} .

We compare two methods of integrating the frequency estimates of the partials. First, assuming the inharmonicity is negligible, we take the median of the estimates $\hat{f}(k)$ for $k = 1, \dots, 12$ (referred to as method M in Section 4). The second method fits a line to $\hat{f}(k)$ using least squares. To remove the impact of outliers, the point furthest from this line is deleted and a line fitted to the remaining points. This outlier removal is iterated 3 times, leaving a line fitting the best 9 partial estimates. The final pitch estimate is the value of the line at $i = 1$ (denoted method L in Section 4).

3. CLASSIFICATION

We classify the 48 pieces by the temperament from which they differ least in terms of the theoretical profiles shown in Table 1. The algorithms introduced in Section 2 output a list of frequency estimates for note objects described by a MIDI note number, onset time and duration d_i . By ignoring the octave, the MIDI note number can be converted to a pitch class $p_i \in P = \{C, C\#, D, \dots, B\}$. We then convert the corresponding frequency estimates to cents deviation c_i from equal temperament. For the 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,

$$\hat{c}_k = \frac{\sum_{i:p_i=k} c_i d_i}{\sum_{i:p_i=k} d_i}, \quad k \in P \quad (5)$$

Minimum note length:		0.1		0.3		0.5		0.7	
Overtone removal:		SP	CT	SP	CT	SP	CT	SP	CT
PT	QIFFT	21	20	21	23	21	24	21	24
	CQIFFT-M	19	20	21	23	21	24	21	24
	CQIFFT-L	20	18	21	24	20	24	20	24
	PV-M	11	18	11	21	11	22	12	22
	PV-L	19	17	19	23	22	24	22	24
RH	QIFFT	17	18	17	20	18	20	17	20
	CQIFFT-M	19	20	19	22	19	23	19	23
	CQIFFT-L	17	22	17	22	17	22	16	22
	PV-M	11	16	10	16	10	17	10	18
	PV-L	18	20	20	20	20	21	20	21

Table 2. Number of pieces classified correctly (out of 24) using various algorithms and data (see text for explanation).

Given this estimate $\hat{c} = (\hat{c}_1, \dots, \hat{c}_{12})$ and a temperament profile $c^0 = (c_1^0, \dots, c_{12}^0)$, we calculate the divergence

$$d(\hat{c}, c^0) = \sum_{k \in P} w_k (\hat{c}_k - c_k^0 - r)^2 \quad (6)$$

between estimate and profile, where w_k is the squared relative duration of the k th pitch class in the note list. $r = \sum w_k (\hat{c}_i - c_i^0) / \sum w_k$ is the offset in cents which minimises the divergence and thus compensates for small deviations from 415 Hz tuning. The weight w_i favours pitch classes that have longer cumulative durations, and in particular discards pitch classes that are not in the note list. A piece is classified as having the temperament whose profile c^0 differs least from it in terms of $d(\hat{c}, c^0)$.

4. RESULTS

The classification results for the two sets of 24 pieces (4 pieces in each of 6 temperaments) are shown in Table 2. Five factors were varied: the source of data, whether synthesised data from Pianoteq (PT) or real harpsichord (RH) recordings; the pitch estimation algorithm (QIFFT, CQIFFT or PV); the method of combining frequency estimates of partials, whether median (M) or line-fitting with outlier removal (L); the minimum note length from the first-pass note identification (0.1 – 0.7 seconds); and style of first-pass note identification, whether spectral peaks (SP) or conservative transcription (CT). See Section 2 for more details.

Our observations on the results follow. Although results were better for synthetic data (all correct with several algorithms and combinations of parameters), it was also possible to classify all but one of the real recordings correctly, using CQIFFT-M. The choice of pitch estimation algorithm was inconclusive for synthetic data, with perfect results achieved by each algorithm for some parameter settings. For the RH data, CQIFFT performed best of all the methods (23 out of 24 correct). For estimating the fundamental from a set of partials, the CQIFFT approach worked better with the median of the frequency to partial number ratio (method M), while PV performed better with line-fitting (method L). Short-note deletion played an important role, giving a clear improvement in results for values up to 0.5 seconds. The improvement was greater when using conservative transcription and/or synthetic data. Finally, the use of conservative transcription improved performance of all algorithms, and was essential for obtaining accurate temperament estimates.

5. FUTURE WORK AND CONCLUSION

Various avenues for future work lie open. First, more extensive testing could be performed, using a larger set of temperaments. This might require a more sophisticated inference mechanism, which could include knowledge of relationships between temperaments, enabling partial classification (temperament family) in cases where data is insufficient to determine temperament uniquely. In order to embed the inference in the Semantic Web, we are developing an ontology of temperaments as part of the Music Ontology [9]. Second, since scores are available for most harpsichord music, an alternative approach would be to align the score to the recordings, which is likely to give a more accurate transcription of the notes, leading to more robust results. Third, more accurate pitch estimates could be obtained by estimating the inharmonicity coefficient B for each note and fitting the partials to equation 3. Fourth, statistical analysis of the data would allow us to evaluate the reliability of pitch estimates, which could then be used as weights in the classification step. Finally, we intend to explore other uses for conservative transcription, such as seeding for general polyphonic transcription, source separation and instrument identification. We believe this “precision over recall” philosophy has applicability beyond the present study.

We presented algorithms for estimating the temperament from harpsichord recordings, and reported the results of classification experiments with real and synthetic data. We showed that existing high-precision pitch estimation techniques are sufficient for the task, if combined with our conservative transcription approach. In particular, using the CQIFFT-M algorithm with conservative transcription and removal of short duration notes, we were able to distinguish between 6 different temperaments of harpsichord recordings with 96% accuracy (100% for synthetic data). We think it is unlikely that a human expert could reach this level of accuracy, but we leave the testing of human temperament estimation to future work.

6. REFERENCES

- [1] R. Rasch, “Tuning and temperament,” in *The Cambridge History of Western Music*, T. Christensen, Ed. Cambridge University Press, 2002.
- [2] C. Di Veroli, *Unequal Temperaments: Theory, History, and Practice*, Bray, Ireland, 2009.
- [3] “Pianoteq 3 true modelling,” <http://www.pianoteq.com>.
- [4] J.O. Smith, “Spectral audio signal processing: October 2008 draft,” <http://ccrma.stanford.edu/~jos/sasp/>, 2008.
- [5] M. Abe and J. Smith, “CQIFFT: Correcting bias in a sinusoidal parameter estimator based on quadratic interpolation of FFT magnitude peaks,” Tech. Rep. STAN-M-117, CCRMA, Dept of Music, Stanford University, 2004.
- [6] U. Zölzer (ed.), *DAFX: Digital Audio Effects*, Wiley, 2002.
- [7] N. Fletcher and T. Rossing, *The Physics of Musical Instruments*, Springer, 1998.
- [8] V. Välimäki, H. Penttinen, J. Knif, M. Laurson, and C. Erku, “Sound synthesis of the harpsichord using a computationally efficient physical model,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 7, pp. 934–948, 2004.
- [9] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, “The music ontology,” in *8th International Conference on Music Information Retrieval*, 2007, pp. 417–422.