Dan Tidhar, Simon Dixon, Emmanouil Benetos and Tillman Weyde

The temperament police

Arguably among the most important determinants of musical phenomena, tuning and temperament have received much attention from music theorists and musicians throughout the entire recorded history of Western music. Keyboard temperaments in particular have been explored in historical treatises dating from as early as the 16th century. More recently, as part of the interest in historically informed performance practice in the 20th and 21st centuries, keyboard temperaments have been the subject of numerous scholarly articles, monographs and textbooks (including works by J. Murray Barbour,[1] Owen Jorgensen,[2] Mark Lindley,[3] Bradley Lehman,[4] Ross Duffin[5] and Claudio Di Veroli,[6] to name just a few). The historical sources, together with more recent scholarly work, form a comprehensive corpus of studies, which are by and large theory driven, and are characterized by a prescriptive attitude, i.e. one in which appropriate temperaments are determined and characterized. Based on historical sources or theoretical considerations, the appropriateness of various temperaments can be determined according to a set of parameters such as musical style, national style, time of composition, and musical characteristics such as key and harmonic structure. Such a temperament is, in turn, a well-defined entity in terms of its theoretical frequency ratios, besides often being described in the form of a practical tuning recipe.

Without questioning the importance and value of the vast existing corpus of temperament-related studies, the present work aims to offer a so-far unique point of view by providing a glimpse into what is actually done in practice by harpsichordists and tuners, as documented by harpsichord solo recordings over many decades. This novel viewpoint is facilitated by automatic temperament estimation methods that are based on recent advances in digital signal processing and music information retrieval. Our 'police' title metaphor should, of course, be taken with a substantial grain of salt. While we do compare notated to measured temperament, our aim is not to promote a prescriptive agenda by pointing out inaccuracies in tuning. Rather, the approach remains purely descriptive, and although we obviously cannot apply our analysis directly to 18th-century sound recordings, we aim to learn as much as possible from harpsichord temperaments in modern recordings, thus providing a practical context to the current temperament-related discourse, with some possible inferred relevance to the historical sources as well. The reasons for deviating from notated temperaments, or indeed from the implicit imperative to adhere to a single temperament per piece, may vary. Without further research, one can merely speculate about phenomena such as inaccuracies in the tuning itself or inaccuracies in temperament labelling (for example due to lack of time, attention or vocabulary). It seems plausible, and certainly compatible with personal experience, that a certain degree of tuning creativity, sometimes even ad hoc, is exercised by tuners or by players who tune. On an even more speculative note, it seems plausible to assume that discrepancies between tuning theory and tuning practice may have been at least as common in the 17th and 18th centuries as they are today, thus licensing deviations from well-known temperaments as potentially historically appropriate.

Our previous work has demonstrated the feasibility of temperament estimation from recordings, and several further developments, including the estimation of string inharmonicity as it is apparent in recordings, and the application of temperament estimation technology to a set of CDs that specify the temperament on their packaging. This study

Early Music, Vol. xlii, No. 4 © The Authors 2014. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1093/em/cau101, available online at www.em.oxfordjournals.org
Advance Access publication October 11, 2014

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

applies the same approach to a dataset of over 2,000 tracks, from over 90 CDs. About 20 per cent of these specify temperament information (notated temperament). Our measured temperaments are compared with the specified ones, where these are available. For the entire dataset, including CDs that do not provide temperament information, measured temperaments are analysed in terms of within-piece and within-CD consistency. Additionally, we report preliminary results identifying trends over time. This is an ongoing project, with a constantly growing dataset. Beyond specific dataset-related results, the main take-home message of this article is that automatic temperament estimation from harpsichord solo recordings is indeed possible, and can be used to promote a descriptive approach to tuning in practice, to complement and contextualize the abundant theoretical discourse.

The discussion below outlines the technological challenges and provides an overview of relevant previous work, describing the methods developed and applied, including automatic transcription, high-precision pitch estimation and temperament classification. We provide some detail about the current dataset, discuss the temperament estimation results obtained, and conclude by presenting some future directions that this work is expected to take.

Background

Given the current ubiquity of electronic tuning devices and applications, readers may wonder about the degree of technological novelty in temperament estimation from recordings. The technological challenges that are involved in fulfilling this task are, however, rather substantial. For the processing of real-world recordings one cannot assume that notes appear in isolation, as this is normally not the case. Further, one cannot assume knowledge of the score, which in some cases will simply not be available, and in others will be obfuscated by ornamentation or improvisation. Not knowing in advance which notes are played when, and not normally having them appear in isolation, places serious challenges to automatic transcription and to precise pitch measurement, as described below.

In a 2010 article, the idea of automatic analysis of harpsichord tuning from standard audio recordings of musical works was introduced.[7] We built a software system that distinguished between six different temperaments with 96 per cent accuracy (100 per cent for synthesized recordings). The concept of conservative (high-precision, low-recall) transcription was introduced and shown to be beneficial for the task. Several signal processing algorithms were compared for pitch estimation, with the discrete Fourier transform combined with quadratic interpolation and bias correction performing best. Based on this work, we developed a Semantic Web ontology for representing and reasoning about temperament, and released a web service for analysing temperament.[8] Two years later, a further article revised the previous temperament estimation approach by additionally modelling the inharmonicity of harpsichord strings, which allowed a greater number of partials to be analysed, and provided a more robust method of estimating the temperament profile.[9]

In 'The temperament police: the truth, the ground truth and nothing but the truth' we further improved on the automatic temperament estimation system and used it to compare measured temperaments with those annotated in CD sleeve notes for over 500 harpsichord tracks (this project was the beginning of the 'Temperament police').[10] A new harpsichord-specific conservative transcription system was developed as the initial processing stage. The temperament classifier was extended to recognize 15 different temperaments, as well as rotations of these temperaments. In most cases the estimated temperament matched the sleeve-note descriptions, although there were several CDs for which a large discrepancy was observed, leading to interesting questions on the nature of human annotations and their use as 'ground truth' for training and evaluating computational methods.

Digital Music Lab

Our empirical work on temperament has recently been continued in the context of the project Digital Music Lab—Analysing Big Music Data (DML).[11] This project aims at developing new technologies for the automated analysis of large music collections, in collaboration with the British Library. It is engaged in work that combines audio features with metadata to facilitate and perform large-scale musicological research. The temperament estimation project, including the collection, curation and analysis of the current dataset, is

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

one of the research avenues into analysis techniques in the DML. Although some aspects of the audio analysis presented here are harpsichord-specific, most are of wider applicability. The integration into the project's software framework enables larger-scale temperament and tuning analysis that will become more relevant as the dataset increases. It thus provides an example of techniques that can and will be applied to larger datasets in the DML.

Automatic transcription

The first step in estimating the temperament of a recording is to detect the existence and timings of notes, a process which is called automatic music transcription.[12] The problem of automatic transcription is fundamental in the field of music information retrieval (due to numerous applications in computational musicology, interactive music systems and in automatic organization and annotation of music collections) and is considered open for the case of multiple-instrument music, as well as for recordings with a large number of simultaneous notes.

In this article we propose and employ a harpsichord-specific transcription system suitable for large-scale audio analysis; this work is also motivated by the DML project. In our 2011 article[13] (for which the data collection was significantly smaller), a transcription system was proposed using non-negative matrix factorization, which is a machine-learning technique suitable for analysing audio spectrograms. This system was able to detect notes on a semitone scale, which was suitable for a rough transcription.

Here, the proposed transcription system is based on probabilistic latent component analysis (PLCA), which is able to efficiently analyse audio recordings and estimate pitches at a higher frequency resolution (in this case, 20-cent resolution). The proposed system is based on the model of Emmanouil Benetos, Srikanth Cherla and Tillman Weyde,[14] specifically adapted for harpsichord recordings. For training the system, we created a dictionary of harpsichord spectral templates taken from isolated note recordings from the RWC (Real World Computing) database.[15] Three different harpsichord models are included in the training dictionary, performing at 8′ pitch, using both single and double manuals. The complete note range of each harpsichord model was used (F2–F7 for the first two models and G2–D6 for the third).

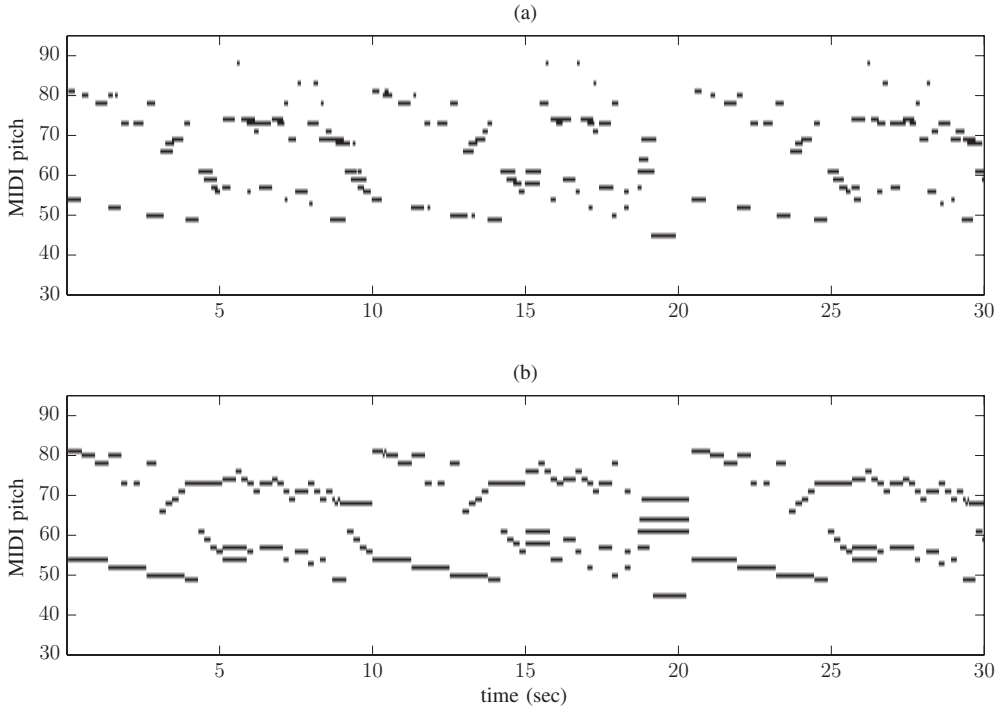
The transcription model takes as its input a harpsichord recording, computes a time-frequency representation (in this case, a log-frequency spectrogram), and, using the PLCA method, computes a binary pitch activation matrix (of pitches on a semitone scale over time frames), along with a pitch salience, indicating the energy of each potential note over time. For post-processing of the pitch activation matrix, detected notes with small durations (less than 60 milliseconds) are removed. As an example of the performance of the transcription system, the transcription output of J. S. Bach's Menuet in G minor, BWV Anh.11, is shown in illus.1a–b, along with the pitch ground truth. In most cases, notes are identified correctly in terms of their pitch and onset time, although shorter durations are estimated (due to the fast decay of harpsichord notes).

The performance of the proposed transcription system was evaluated using a validation set of seven recordings from the RWC database,[16] for which aligned ground truth (in MIDI format) is also available. Since for the present temperament-estimation system a 'conservative' transcription is favourable (meaning that relatively few notes are detected, for which there is high confidence), the transcription system was set to have a low 'false alarm' rate (i.e. the rate of false notes introduced by the process without having actually been played—in this case, 5.3 per cent), which gave a missed detection rate of 44.2 per cent (see Poliner[17] for metric definitions).

Precise frequency estimation

The second step in estimating the tuning of a harpsichord from a musical recording is to obtain precise estimates of the fundamental frequency of each note. Although there is a vast literature on frequency and pitch detection (see reviews by Alain de Cheveigné[18] and Anssi Klapuri and Manuel Davy[19]), many approaches are not suitable for analysing our harpsichord data set, as they are based on assumptions that do not hold in our case. The three most common assumptions are that the music is monophonic, the signal is stationary (which implies that notes do not decay) and that each tone is harmonic (the frequencies of partials are exact integer multiples of the fundamental frequency), none of which hold for our harpsichord data set. Few approaches address high-precision frequency estimation to a resolution

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014



1 Automatic transcription of J. S. Bach's Menuet in G minor, BWV Anh.11, from the RWC database: (a) transcription system output, in a piano-roll representation; (b) pitch ground truth

of a cent (hundredth of a semitone), which is what is required for this study.[20] In previous work,[21] we compared several methods and found that the highest precision is obtained using the discrete Fourier transform with quadratic interpolation and correction of the bias due to the window function.[22]

We first need to establish the tuning-reference frequency, which is expressed as the fundamental frequency of the note A4. For early music, the tuning reference frequency usually lies in the range 392–440Hz, nearly always lower than the modern standard of 440Hz. This raises a problem for automatic analysis: if the system is not supplied with the score (or at least the key) of a piece of music, it is not possible to distinguish an A4 with fundamental frequency 415Hz (where this is also the reference frequency) from a G♯4 with fundamental frequency 415Hz (where the reference frequency is 440Hz). To overcome this semitone-level ambiguity, the dataset was annotated with initial tuning frequency estimates ('392', '415', or '440') based on listening. To estimate the tuning reference, a selection of 40 frames, equally spaced throughout the piece, are analysed for note content using the method described below. We then compute the frequency ratios of detected notes and their nominal frequencies (obtained using an initial estimate of the tuning frequency) to give a set of deviations from the estimated tuning frequency. A weighted average of these deviations is employed to update the tuning-frequency estimate, and this process is repeated five times (or until the update is less than one cent, if sooner).

For each note identified in the conservative transcription, spectral peaks corresponding to the partials of the note are searched for. Then, for each local peak a parabola is fitted to the peak point and the two surrounding points in the log magnitude spectrum, to determine the position of the maximum of the parabola, which gives a much more accurate estimate of the true frequency of the partial than the centre frequency of the peak bin. This estimate is further refined by correcting for the bias due to the window shape and zero padding factor.[23]

Then the fundamental frequency and inharmonicity of each note are estimated jointly. For a string

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

with ideal fundamental frequency f_0 and inharmonicity constant B, the frequency f_k of the kth partial is given by Harvey Fletcher (1964):[24]

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

From measurements of the frequency and the partial numbers of any two partials of a note, it is possible to solve for $\boldsymbol{f_0}$ and $\boldsymbol{B}$. This estimation can be repeated for all pairs of partials of a note, over the duration of the note and over each instance of the note within a recording, to give a large number of frequency (and inharmonicity) estimates for each note. Here we omit any partials from simultaneous notes that overlap in frequency, as these lead to unreliable estimates. We take advantage of the large number of independent measurements (typically thousands) and use a robust statistic, the median, to ameliorate noise in the estimation process, as described below. Also, we use the inter-quartile range as an inverse measure of confidence in the estimates.[25]

Temperament estimation

A temperament profile for each recording is computed as follows. The previously determined fundamental frequency estimates are expressed as pitch differences in cents from their nominal frequencies on an equal-tempered scale tuned to the estimated tuning reference. Assuming the tuning has pure octaves, all such pitch differences can then be combined into a single value for each pitch class, using a median (or weighted median, where the weights are given by the pitch salience computed during the transcription step). This gives a 12-dimensional vector, called the 'temperament profile', which can be compared with the profiles of known theoretical temperaments. For simplicity the pitch class is represented by an integer from 0 (C) to 11 (B), corresponding to the MIDI pitch number modulo 12.

The set of temperaments currently recognized by our classifier includes the six temperaments used in our initial 2010 experiment,[26] and then all the specified temperaments on sleeves of CDs in our dataset. The set, which may further grow in the future, currently includes the following temperaments:

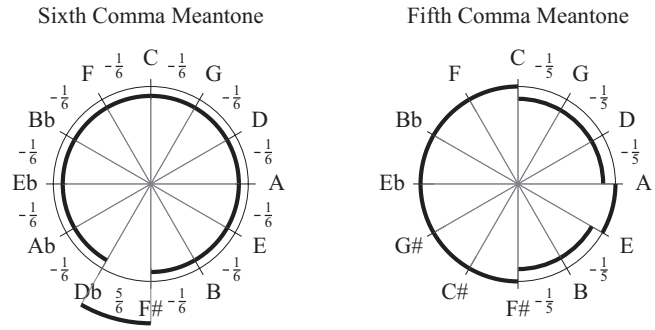
equal (standard 12-tone equal temperament)
fifth comma (a practical variation on Kellner, with a pure A–E interval; see illus.2)
Vallotti (hereafter Vall)
quarter-comma meantone (QCMT)
fifth-comma meantone (FCMT)
sixth-comma meantone (SCMT)
Kellner
Werckmeister 3
Lehman 'Bach'
Neidhardt (versions 1, 2 and 3, all from 1724)
Kirnberger (2 and 3)
just intonation[27]

The definitions used are specified in full detail via this article's weblink.[28] We also recognize rotations (different starting pitches) of these temperaments, although this is not a typical tuning practice for all temperaments, as illustrated by the example of the Young II temperament, a rotation of the Vallotti temperament, which is considered a different temperament in its own right. Rotations are specified via the wolf interval where applicable (for example, SCMT-FD has wolf interval F♯–D♭, as in illus.2), otherwise by the number of semitones rotated (for example, Vall+7). Allowing rotations of temperaments is an elegant way to deal with the tuning ambiguity discussed above.

Given an observed temperament profile, we can calculate its divergence from any theoretical profile via a weighted average of the squared differences between corresponding profile elements, adjusted for any offset in tuning frequency.[29] The weights are based on the number of frequency estimates available for computing each profile element. While this favours the most reliable estimates, it does not solve the problem of missing observations, where not all 12 pitch classes are played (or detected) in a recording. Observed profiles can be classified by finding the nearest theoretical profile, and the space of temperaments can be explored by clustering observed profiles in order to investigate tuning practice, some of which are described in the following sections.

In the first 'temperament police' article[30] we validated the reliability of our approach using four pieces recorded with six different temperaments using the physical modelling synthesiser Pianoteq.[31] Each of the 24 recordings was classified correctly from the set of 180 possible temperaments (15 temperaments by 12 rotations). For CD recordings, we can gauge confidence in classification results by considering the divergence between the observed and the selected temperament profiles (low values

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014



2 Circle-of-5ths representations for two temperaments used by our classifier. The deviation of each 5th from a pure 5th (the lighter circle) is represented by the positions of the darker segments. The fractions specify the distribution of the comma between the 5ths (if omitted, the 5th is pure).

implying high confidence) and the consistency of results between related recordings (for example, movements of the same piece).

The dataset

The dataset we use in this project is a constantly growing collection of harpsichord solo recordings. The analysis presented here is applied to a snapshot of the dataset at the time of writing; it currently includes 92 CDs, comprising 2,021 tracks. (Full details are provided via the article weblink.[32]) The current dataset consists of commercially released CDs by established performers. It covers many decades of recordings of both 17th- and 18th-century music in major national styles, including Italian, German, French and English repertory. It is currently biased towards Johann Sebastian Bach due to the special interest in studying the *Goldberg Variations* and *The Well-Tempered Clavier* (see below). We do not claim that this evolving dataset is perfectly balanced or representative at any given time; one of the challenges involved is to accumulate sufficient material to compensate statistically for any bias in the dataset. Table 1 shows the distribution of the current dataset over time (in decades).

Table 1 Distribution of recording dates in the current dataset (in decades)

Period	*Number of CDs in dataset*
1930–1939	1
1940–1949	0
1950–1959	0
1960–1969	1
1970–1979	4
1980–1989	8
1990–1999	23
2000–2009	31
2010–2014	24

Results and analysis

The analysis reported here is for two separate though interrelated strands: for the entire dataset, we look at temperament consistency within a CD and within a piece; for the smaller set of CDs which do provide temperament information, we determine the measured temperament and compare it to the notated one. In both cases, the analysis is done on a per-track basis. Whereas the former can be obtained by directly analysing the frequency estimates per pitch class (i.e. without necessarily labelling the temperament), the latter requires the additional step of temperament estimation (i.e. subjecting the pitch-class frequency estimations to a classifier, as described above).

The first result is related to the entire dataset, and is obtained with only a very minimal amount of metadata. In fact, this could be considered an agnostic approach relative to pieces, keys and even particular temperaments. For this particular exercise we only compare the measured frequencies per pitch class across different tracks, without employing the temperament classifier. The only information used, apart from the frequency estimates themselves, is the order of tracks within each CD and, for each track, which CD it belongs to.

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

As a bird's-eye indication of the congruence of temperament within CDs, we compare the average distance—in terms of temperament—between tracks belonging to the same CD to the overall average distance between tracks. The average distance between successive tracks provides an approximation of the congruence within-pieces without involving further metadata. It is indeed an approximation because not all pairs of successive tracks belong to the same piece, though most pairs do. In the next step, we introduce additional metadata about the pieces, and specifically which tracks belong to which piece, which enables us to directly measure the within-piece agreement of temperament. It is worth keeping in mind that for pieces by J. S. Bach our metadata follows BWV notation, and therefore a large piece like the *Goldberg Variations* is considered one piece for the purpose of the within-piece agreement analysis, whereas in the case of *The Well-Tempered Clavier*, each pair of preludes and fugues has a separate BWV number and are therefore considered a unit on their own. Including the entire *Well-Tempered Clavier* as one monolithic piece would have worsened the within-piece agreement reported here, as discussed in more detail below.

For each of the 2,021 tracks in the current dataset we calculate a temperament profile as explained above. We then apply the Euclidean distance metric[33] to the profiles and calculate average pairwise distances, as summarized in Table 2. The distances are pairwise significantly different from each other according to a Mann-Whitney test (with the exception of successive tracks vs track pairs within piece, where the former can be seen as an approximation of the latter) which suggests that the following claims hold for this dataset:

Table 2 Euclidean distances (in cents) between temperament profiles of pairs of tracks across the dataset

Set of track pairs	*Average distance*
All track pairs	22.7
Track pairs within CD	14.6
Successive tracks	11.5
Track pairs within piece	10.8

Tracks within the same CD tend to resemble each other in terms of temperament more than they resemble tracks from other CDs.

Tracks within the same instance of a piece tend to resemble each other in terms of temperament more than they resemble tracks from other instances or other pieces.

In other words, temperament congruence on the dataset as a whole is, on average, just as one would expect it to be. Next, we look at the specific subset of CDs which provide temperament information in their accompanying booklets.

First, we consider results that have previously been reported in the first 'temperament police' article.[34] These relate to about 25 per cent of the current dataset, consisting of CDs that specify the temperament(s) used for the recordings. The results are summarized in Table 3, and full details are provided on that article's website.[35] Column 1 is our internal CD index, where letters are used to distinguish groups of tracks with different temperament metadata. Column 2 shows the annotated reference tuning, while the mean and standard deviation (over tracks) of the estimated reference tuning are given in columns 3 and 4 respectively. Columns 5 to 8 give the annotated temperament, the average distance of tracks from this temperament, the most frequent classification result from the temperaments listed in the temperament classification section above, and the average difference in distance between the annotated temperament and the classified temperament. The table also includes the ground truth data from the 'High precision frequency estimation' study,[36] which appears in the bottom two rows as 'RH' for real harpsichord recordings and 'PT' for recordings synthesized with Pianoteq, as discussed below.

The results for tuning (pitch level) show agreement with the notated values where they were available, with the exception of CD 19, which had only two tracks at A440. The CDs generally show tuning consistency across all tracks, with high standard deviations (> 2Hz) being due to a bimodal distribution of tuning frequency (CD 18) and five outlier tracks (CDs 2, 7, 19). Summarizing by CD assumes that the same tuning (pitch level) is used for all tracks on a CD, which is clearly not always the case.

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

Table 3 Summary of results from Dixon, Tidhar and Benetos, 'The temperament police' (2011), with columns for CD number, notated reference tuning, estimated reference tuning, standard deviation across tracks of CD, notated temperament, highest-ranked temperament, and average difference in distance between notated and highest-ranked temperaments. The last two rows refer to the data from Tidhar, Mauch and Dixon, 'High precision frequency estimation for harpsichord tuning classification' (2010)

CD	*Pitch in Hz*			*Temperament*			
	Notated	*Estimated*	*StD*	*Notated*	*Div.*	*Estimated*	*ΔDiv.*
1		417.6	0.2	Ordinaire		Neid2	
2	405	405.7	3.2	FCMT	21.8	Various	16.4
3a		416.8	0.2	SCMT-BG	3.3	FCMT-BG	2.5
3b		413.9	0.2	Kellner*	8.5	Various	1.2
3c		414.2	0.2	Kellner	3.3	Kellner	0.0
4b		416.9	0.3	FCMT-FD	1.1	FCMT-FD	0.0
5	415	417.1	0.9	QCMT	1.4	QCMT-GE	0.0
6		413.8	0.7	Late17		Vall+7	
7		432.6	4.8	FCMT	7.6	Various	4.1
8b		416.8	0.4	QCMT	1.2	QCMT-GE	0.0
9	415	415.3	0.3	Neid	1.1	Neid 1/2	0.0
10	415	416.5	0.4	Werck3	3.4	Various	1.7
11	415	416.6	0.6	Werck3	3.0	Various	0.9
12	415	415.3	0.2	Kirn3	11.1	Neid1	9.4
13	415	415.1	0.3	Kirn3	7.3	Neid1	5.9
14a	(415)	412.7	0.3	QCMT	10.0	Various	7.0
14c	(415)	435.2	0.2	QCMT	2.7	QCMT-GE	0.0
15		415.7	1.3	Werck3	3.4	Werck3	0.5
16		416.1	1.1	Werck3	0.0	Werck3	0.9
17		413.9	1.2	QCMT	6.0	FCMT	2.2
18		440.5	2.4	QCMT	5.0	QCMT-GE	2.7
19	440	447.6	5.6	QCMT	19.5	FCMT	15.2
20		412.9	0.6	Werk3	2.6	Various	0.8
21		414.5	1.6	FCMT	1.0	FCMT-GE	0.0
22		408.7	0.3	Lehman	1.1	Lehman	0.1
RH	415	415.5	0.8	Various	7.1	Various	0.3
PT	415	415.6	0.7	Various	0.1	All correct	0.0

The temperament results vary from close agreement to the metadata (CDs 4, 5, 8, 9, 16, 21, 22), to moderate agreement (for example, CDs 15, 18), to disagreement (for example, CDs 12, 13, 17). For a number of tracks it was not possible to find a single 'best fit', as some temperaments are only distinguished by the tuning of a pitch class (a chromatic note) that does not appear (or is not detected) within the piece. The large divergences of CDs 2 and 19 may be explained by the tuning frequency being at the half-way point between two semitones relative to the A440 reference assumed by the transcription algorithm, making the transcriptions less reliable.

'Temperament ordinaire', specified by CD 1, is an ambiguous title, which may not have been a proper name so much as a description meaning 'the usual temperament', the nature of which would depend on the context. Some usage examples of 'Temperament ordinaire' as a proper name exist, usually referring to QCMT in the 17th century and to a 'well' temperament in the 18th century. The measured temperament, Neid2, agrees to a certain extent with the second interpretation.

On CD 17 and some other tracks specifying QCMT, the temperament was often closer to FCMT. This is an interesting tendency, as the two are fairly

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

similar, with FCMT being milder (slightly larger major 3rds and a smaller wolf interval). It seems plausible that QCMT was intended but then tempered to bring it (inadvertently) closer to the less extreme FCMT. However, the opposite tendency appears on CD 3a. Werckmeister 3 is specified on five CDs, but only fulfils the claim on two. The reason may be that Werckmeister 3 is popular as a starting point for tuners while they experiment and develop their own temperaments, or that it is very close to other temperaments such as Kellner (note the low value of ΔDiv in each case).

Since we suggest that CD sleeve notes are a questionable source of 'ground truth', an independent means of ascertaining the reliability of our system is needed. The bottom row of Table 3 shows the results for four pieces recorded with six different temperaments using the physical-modelling synthesizer Pianoteq.[37] Using the current approach, these tracks were all classified correctly from the set of 180 possible temperaments (15 temperaments by 12 rotations). Confidence in classification results can also be gained by considering the divergence value and consistency of results (i.e. if a number of related tracks are classified with the same label, and low divergence from the given temperament).

Fashions and trends

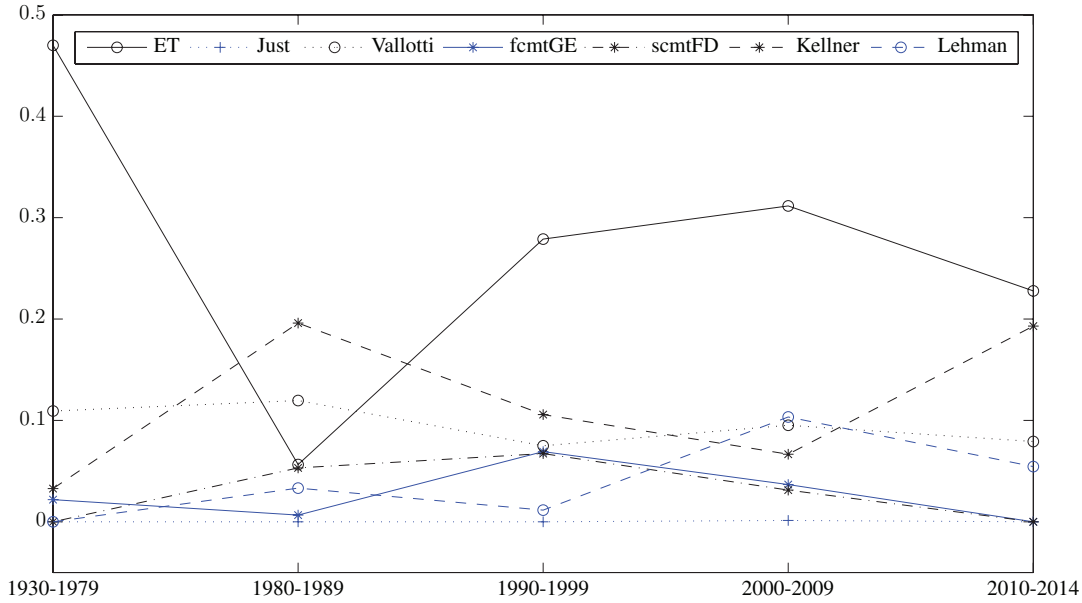
Similarly to other cultural phenomena, tuning and temperament are subject to fashions and trends. Historical temperaments being rediscovered, new ones being invented, developments in knowledge, understanding and ideologies relating to performance practice, as well as individual preferences of prominent performers, are all factors that can dramatically affect the popularity of certain temperaments at certain times. Identifying and characterizing such trends is one of the aims of our project. Despite the generally balanced appearance of our current dataset, it is very far from being a complete record of available harpsichord recordings, and is by no means a rigorously chosen representative sample. Our results relating to the dataset as a whole are, therefore, of a somewhat tentative nature, and are likely to require adjustments as the dataset grows. Illus.3 shows the current measurements of the popularity of selected temperaments and how it changes over time. Keeping this in mind, the graph can be interpreted as demonstrating a few trends in our current datasets.

Equal temperament appears to have been very much in use in the first period graphed (which conflates five decades due to data sparsity). It then significantly loses popularity in the 1980s, presumably as a result of the Historically Informed Performance (HIP) movement, and regains some of its popularity in the decades to follow, perhaps (and this is highly speculative) due to loosening up of some of the HIP criteria. The two nearly horizontal lines (just intonation near the bottom and Vallotti further up along the vertical axis) demonstrate that just intonation is not actually ever used for keyboard tuning (strictly speaking, it is not even a temperament), and that Vallotti's popularity is more or less constant. Kellner and fifth-comma meantone tend to vary more, with the latter being significantly less popular, perhaps because it is used for distinctively earlier repertory that is not well represented in the current dataset. Lehman's version of his 'Bach' temperament was first published in 2005, which explains why it only gains popularity relatively late. The fact that it appears to have a small but positive value for the 1980s could be explained by a measurement error, or perhaps ad hoc temperaments which happened to be close enough to be classified as Lehman.[38]

Specific repertory

In general, the current dataset displays a significant degree of within-piece and within-CD temperament agreement, as statistically indicated above. Certain pieces are of particular interest due to, for example, the requirement placed on the temperament by key changes between movements. Such is the case with J. S. Bach's *Goldberg Variations* and *The Well-Tempered Clavier*, both of which are quite well represented in the dataset (Tables 4–5).[39] The former is in G major apart from three variations in G minor, which are of a highly chromatic character. The tuning challenge involved arises from the fact that the variations are (unless interrupted by Count Kaiserling, falling asleep) deemed to be played as one piece and the temperament chosen needs to accommodate both G major and minor. It is therefore particularly revealing to check the within-piece temperament agreement for this particular case, as well as to estimate the different temperaments chosen for recordings of the piece.

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014



3 Trends in popularity of some selected temperaments as reflected in relative number of occurrences in the dataset across decades. These include equal temperament (ET), just intonation (Just), Vallotti, fifth-comma meantone with a G♯–E♭ wolf (fcmtGE), sixth-comma meantone with an F♯–D♭ wolf (scmtFD), Kellner and Lehman's 'Bach' temperament.

Table 4 Recordings of Bach's *Goldberg Variations* in the current dataset

Wanda Landowska	1933 (1999 remaster)
Gustav Leonhardt	1965
Gustav Leonhardt	1978
Kenneth Gilbert	1987
Ton Koopman	1988
Lars Ulrik Mortensen	1989
Richard Egarr	2006
Matthew Halls	2007
Yoshiko Ieki	2007
Barbara Dobozy	2009
Aapo Hakkinen	2010
David Shemer	2011

Table 5 Recordings of Bach's *Well-Tempered Clavier*, Book 1, in the current dataset

Gustav Leonhardt	1973
Christiane Jaccottet	1989
Bob van Asperen	1999
Gary Cooper	2000
Ottavio Dantone	2001
Peter Watchorn	2006

The case of *The Well-Tempered Clavier* is slightly more complicated, for several reasons. First, it covers all 24 major and minor keys (twice). Second, the title indicates a strong relevance of temperament, a fact that has given rise to different interpretations over the years. Third, the question whether it is a piece rather than a collection of pieces, i.e. whether it is intended to be played in one go, is under debate (at least regarding Book 1). The current datatset includes six recordings of the first book of *The Well-Tempered Clavier*, dating from between 1973 (Gustav Leonhardt) and 2006 (Peter Watchorn), none of which seems to be in one single temperament throughout. The maximal degree of homogeneity is observed in the latest recording (Watchorn 2006), in which two-thirds of the tracks were classified as being in Lehman, and the remaining one-third in two variants of Neidhardt. The other examples typically include five or six different temperaments in each version, mixing fifth-comma-based temperaments with sixth-comma, and occasionally quarter-comma-based ones (such as Kirnberger 3).

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

Wanda Landowska's 1933 recording of the *Goldberg Variations* is homogeneously classified as being in equal temperament, which is hardly any surprise. More surprising is the clear tendency to use different temperaments within the piece, which is observed in quite a few of the other 11, more recent, *Goldberg* recordings in the current dataset. The more homogeneous recordings include (in chronological order) Lars Ulrik Mortensen's 1989 recording, which is mostly in Neidhardt 2; Ton Koopman's 1997 CD, mostly in Neidhardt; Kenneth Gilbert's 2001 disc, which is mostly in Neidhardt 3; Richard Egarr's 2006 recording, in which most tracks were classified as Lehman; and Matthew Halls's 2007 performance, which was mostly classified as Neidhardt 2. The result for Egarr's recording is in accordance with the fact that he has been an outspoken proponent of Lehman's 'Bach' temperament. The other relatively homogeneous renditions of the piece, which span nearly two decades, all seem to concentrate around different Neidhardt temperaments, which is an interesting observation. The remaining six versions included in the current dataset are characterized by greater divergence and each seems to contain five or six different temperaments, similarly to the non-homogeneous recordings of *The Well-Tempered Clavier*.

Future research

The corpus of harpsichord solo recordings is ever growing and we are currently working on extending this study to larger datasets. This is an ongoing endeavour, which includes collecting and curating audio recordings and textual metadata, maintaining and extending the set of temperaments known to the system, and of course performing the analysis and interpreting its results. The analysis of particular pieces will benefit from a larger dataset as well as additional observations on the track level (for example, particular variations, as in the case of *Goldberg Variations*). Other directions include further developing the Semantic Web work published in 2010,[40] as well as the availability of temperament estimation as a web service and as a VAMP plugin for use with Sonic Visualiser.[41] In the longer term, we envisage extending the applicability of temperament estimation technology to other keyboard instruments including the organ and the piano, and to non-solo recordings as well.

Dan Tidhar is a Research Fellow at the Music Informatics Research Group of City University London, and a member of the Centre for Music and Science and Wolfson College at Cambridge. He is also active as harpsichordist and tuner specializing in historical keyboard instruments. His research interests cover a broad spectrum of topics in computational musicology, historical musicology and music cognition. dut20@cam.ac.uk

Simon Dixon is a Reader in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He is President of the International Society for Music Information Retrieval and his research interests include high-level music signal analysis and the representation of musical knowledge.

Emmanouil Benetos is a Research Fellow at the Music Informatics Research Group of City University London. He holds a BSc and MSc in Informatics and a PhD in automatic music transcription from Queen Mary University of London. His research interests include audio-signal processing, music-information retrieval and machine learning.

Tillman Weyde is a Senior Lecturer in Computer Science at City University London, where he leads the Music Informatics Research Group. His research interests include machine learning in music and other areas, large-scale data analysis, and the combination of signal processing with symbolic models.

This work is supported by the AHRC 'Digital Music Lab—Analysing Big Music Data' project, grant no. AH/L01016X/1. *Emmanouil Benetos is supported by a City University London Research Fellowship. The authors are very grateful to all those who helped collect the set of recordings used for this work.*

1 J. M. Barbour, *Tuning and temperament: a historical survey* (East Lansing, MI, 1951, R/2005).

2 O. Jorgensen, *Tuning the historical temperaments by ear: a manual of eighty-nine methods for tuning fifty-one scales on the harpsichord, piano, and other keyboard instruments* (Marquette, MI, 1977).

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014

3 M. Lindley, 'Instructions for the clavier diversely tempered', *Early Music*, v/1 (1977), pp.18–23.

4 B. Lehman, 'Bach's extraordinary temperament: our Rosetta Stone', *Early Music*, xxxiii/1 (2005), pp.3–23, and xxxiii/2 (2005), pp.211–31.

5 R. E. Duffin, *How equal temperament ruined harmony (and why you should care)* (New York, 2007).

6 C. Di Veroli, *Unequal temperaments: theory, history, and practice* (Bray, 2009, 3/2013).

7 D. Tidhar, M. Mauch and S. Dixon, 'High precision frequency estimation for harpsichord tuning classification', in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Piscataway, NJ, 2010), pp.61–4.

8 D. Tidhar, G. Fazekas, M. Mauch and S. Dixon, 'TempEst: harpsichord temperament estimation in a semantic-web environment', *Journal of New Music Research*, xxxix/4 (2010), pp.327–36.

9 S. Dixon, M. Mauch and D. Tidhar, 'Estimation of harpsichord inharmonicity and temperament from musical recordings', *Journal of the Acoustical Society of America*, cxxxi/1 (2012), pp.878–87.

10 S. Dixon, D. Tidhar and E. Benetos, 'The temperament police: the truth, the ground truth and nothing but the truth', in *12th International Society for Music Information Retrieval Conference* (2011), pp.281–6.

11 Digital Music Lab—Analysing Big Music Data, http://dml.city.ac.uk/.

12 E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff and A. Klapuri, 'Automatic music transcription: challenges and future directions', *Journal of Intelligent Information Systems*, xli/3 (2013), pp.407–34.

13 Dixon, Tidhar and Benetos, 'The temperament police'.

14 E. Benetos, S. Cherla and T. Weyde, 'An efficient shift-invariant model for polyphonic music transcription', 6th International Workshop on Machine Learning and Music, Prague, September 2013.

15 M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, 'RWC music database: music genre database and musical instrument sound database', in *4th International Conference on Music Information Retrieval* (2003), pp.229–30.

16 See n.15.

17 G. Poliner and D. Ellis, 'A discriminative model for polyphonic piano transcription', *EURASIP Journal on Advances in Signal Processing*, viii (January 2007), pp.154–62.

18 A. de Cheveigné, 'Multiple F_0 estimation', in *Computational auditory scene analysis: principles, algorithms and applications*, ed. D. L. Wang and G. J. Brown (Piscataway, NJ, 2006), pp.45–79.

19 A. Klapuri and M. Davy (eds.), *Signal processing methods for music transcription* (New York, 2006).

20 We cannot, of course, be certain about tuning precision at the time historical temperaments were developed, but the mathematical definitions are precise and so require this level of precision.

21 Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

22 M. Abe and J. Smith, '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, Stanford University (2004).

23 See n.22.

24 H. Fletcher, 'Normal vibration frequencies of a stiff piano string', *Journal of the Acoustical Society of America*, xxxvi/1 (1964), pp.203–9.

25 For technical details of the approach, see Dixon, Tidhar and Benetos, 'The temperament police'.

26 Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

27 Just intonation is not typical for keyboard instruments, and is included in the classification scope for legacy reasons as it was part of the experiments in Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

28 See http://dml.city.ac.uk/temperament/.

29 Dixon, Tidhar and Benetos, 'The temperament police'.

30 See n.28.

31 Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

32 See n.28

33 The Euclidean distance between profiles

$$t = (t_1, t_2, ..., t_{12}) \text{ and } u = (u_1, u_2, ..., u_{12})$$
$$\text{is } d(t,u) = \sqrt{\sum_{i=1}^{12} (t_i - u_i)^2}$$

34 Dixon, Tidhar and Benetos, 'The temperament police'.

35 www.eecs.qmul.ac.uk/~simond/ismir11.

36 Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

37 www.pianoteq.com, and Tidhar, Mauch and Dixon, 'High precision frequency estimation'.

38 As explained in the section dealing with temperament classification, the classifier can only choose between the temperaments that are known to the system. Additional temperaments can be easily introduced to the system by providing their temperament profiles, but at the time of writing this classification is limited to the set of temperaments specified above. Note that this does not affect the other forms of analysis presented here, as they are based on statistical analysis applied directly to the pitch measurements, rather than on classification results. In the particular case of an early recording classified as Lehman, one can speculate about a tuner seeking a milder version of another temperament and coming close enough to fool our classifier, or perhaps even close enough to be said to have serendipitously tuned to the temperament itself.

39 For detailed work discographies, see for example www.bach-cantatas.com/NVD/BWV988.htm and www.bach-cantatas.com/NVD/BWV846-869.htm.

40 Tidhar, Fazekas, Mauch and Dixon, 'TempEst: harpsichord temperament estimation'.

41 C. Cannam, C. Landone and M. Sandler, 'Sonic Visualiser: an open source application for viewing, analysing, and annotating music audio files', in *MM'10, Proceedings of the International Conference on Multimedia* (2010), pp.1467–8.

Downloaded from http://em.oxfordjournals.org/ by guest on November 13, 2014