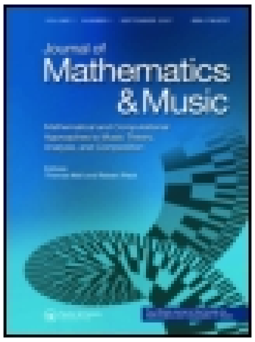


Journal of Mathematics and Music

Mathematical and Computational Approaches to Music Theory, Analysis, Composition and Performance

ISSN: 1745-9737 (Print) 1745-9745 (Online) Journal homepage: http://www.tandfonline.com/loi/tmam20

The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation

Luwei Yang, Khalid Z. Rajab & Elaine Chew

To cite this article: Luwei Yang, Khalid Z. Rajab & Elaine Chew (2017): The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation, Journal of Mathematics and Music, DOI: 10.1080/17459737.2016.1263897

To link to this article: http://dx.doi.org/10.1080/17459737.2016.1263897

Published online: 14 Mar 2017.

Journal of Mathematics and Music, 2017
http://dx.doi.org/10.1080/17459737.2016.1263897

The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation

Luwei Yang[a*], Khalid Z. Rajab[b], and Elaine Chew[a]

[a] *Centre for Digital Music, Queen Mary University of London, London, UK;*
[b] *Antennas & Electromagnetics Group, Queen Mary University of London, London, UK*

(*Received 7 December 2015; final version received 19 November 2016*)

We present a novel approach to frame-wise vibrato detection and estimation in music signals using the Filter Diagonalisation Method (FDM). In contrast to conventional fast Fourier transform-based methods, the FDM's output remains robust over short time frames, allowing frame sizes to be set at values small enough for accurately identifying local vibrato characteristics and pinpointing vibrato boundaries. FDM decomposes the local fundamental frequency into sinusoids and returns their frequencies and amplitudes, which the system uses to determine vibrato presence and vibrato parameter values. We test two decision mechanisms – the decision tree and Bayes' Rule – for vibrato detection. The systems are tested against state-of-the-art techniques on monophonic datasets consisting of string, woodwind, brass, and voice excerpts. In addition to using existing datasets, we have created a new monophonic dataset consisting of performances of an entire music piece on erhu and violin, with annotations of vibrato presence and parameters. We show that FDM-based techniques consistently yield the best results in both frame-level and note-level evaluations. Furthermore, FDM with Bayes' Rule leads to better F-measure results – 0.84 (frame-level), 0.41 (note-level) – than FDM with decision tree – 0.80 (frame-level), 0.31 (note-level). FDM's accuracy for determining vibrato rates is above 92.5%, and for vibrato extents is about 85%.

Keywords: vibrato; vibrato detection; filter diagonalisation method; expressive music performance analysis; music information retrieval

2010 Mathematics Subject Classification: 00A65
2012 Computing Classification Scheme: applied computing

1. Introduction

This paper proposes a novel solution to the problem of vibrato detection and estimation. Vibrato constitutes an important expressive device in music performance whereby the musician modulates the fundamental frequency of a pitch in a periodic fashion at a rate typically between 4 and 8 Hz. Vibrato is frequently used to enhance and make selected notes more prominent in music performance (Palmer and Hutchins 2006). Vibrato use is central to music making on strings, woodwind instruments, and the human voice. It varies significantly across cultures, musical periods, and individual musicians' styles (Nwe and Li 2007; Regnier and Peeters 2009; Mitchell and Kenny 2010; Özaslan, Serra, and Arcos 2012; Yang, Chew, and Rajab 2013;

*Corresponding author. Email: l.yang@qmul.ac.uk

Yang, Tian, and Chew 2015). The computational study of musical expressivity is a growing field (Liebman, Ornoy, and Chor 2012; Fabian, Timmers, and Schubert 2014). Precise characterisation and measurements of vibrato features via computational means can reveal differences between performance styles and performers' skills, and has direct impact on ethnomusicological studies of the use of vibrato in world musics, the tracing of musical influences in musicological phylogenetic studies, and expressive performance pedagogy, analysis, and synthesis.

Automatic detection and estimation of vibrato, the focus of this paper, would greatly speed vibrato analysis, systematic expressive performance research, music expression synthesis, and automatic music transcription. In the long term, our goal is to create an automatic vibrato analysis system that can be widely used for comparing different musical styles and performers across instruments, genres, and cultures, and in computer-assisted education. One such application can be found in the AVA software, an interactive vibrato and portamento analysis system, described in Yang, Rajab, and Chew (2016). As an illustration, Figure 1 demonstrates the vibrato detection process. The top graph shows the spectrogram and the middle one plots the f_0 (the estimate of the fundamental frequency time series). The bottom graph shows the vibratos detected by the Filter Diagonalisation Method (FDM) with Decision Tree (DT), and the Filter Diagonalisation Method (FDM) with Bayes' Rule (BR). The FDM method and the decision mechanisms will be described in Section 3.

Prior efforts in automatic vibrato detection have focused primarily on applying the Fourier transform to the f_0 of the audio, to determine whether the spectral peak resides in the expected vibrato frequency range (Herrera and Bonada 1998; Ventura, Sousa, and Ferreira 2012). Due to the uncertainty principle for Fourier transforms, choosing the best window size for computing the spectrogram presents a challenge when applying the Fourier transform to f_0. The Fourier transform decomposes f_0 into a number of sinusoids. In frame-wise vibrato detection, spectral peaks (sinusoids with the largest amplitudes) will be blurred if the frame size is too large, containing both vibrato and non-vibrato segments; the precise location of the boundary would also be hard to identify. If the window is too small, the resolution in the frequency domain will be too low to show if the spectral peak resides in the vibrato rate range. This paper offers a new solution

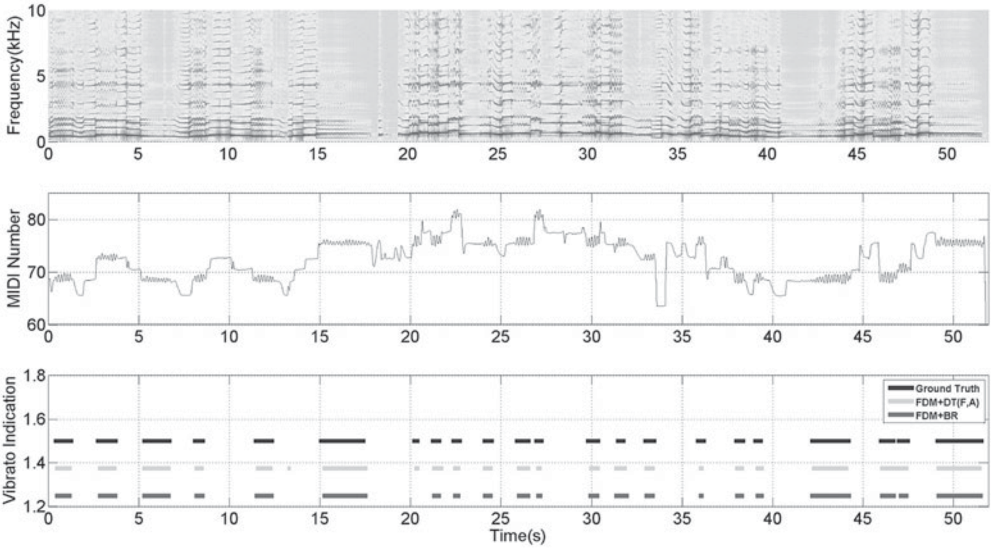


Figure 1. Vibrato detection demonstration of a real vibrato passage. Upper plot: spectrogram; middle plot: f_0; lower plot: vibrato detection results. FDM + DT(F,A) and FDM + BR are the proposed methods.

to this problem via a computational technique that can be applied to short-term signals with high frequency resolution.

We propose that FDM presents a highly competitive alternative technique for frame-wise vibrato detection. FDM is an harmonic inversion method. It is realised by constructing a filtered local frequency-domain signal matrix and then diagonalising it. FDM is especially well suited to extraction of spectral features that occur over a very small time span. The greater precision afforded by FDM allows for high resolution extraction of vibrato boundaries and characteristics, above and beyond current techniques. The method is also amenable to real-time implementation due to its frame-wise processing and small window size. Like other frame-wise methods, FDM bypasses the segmentation process of note-wise vibrato detection methods (Rossignol et al. 1999; Pang and Yoon 2005; Özaslan and Arcos 2011; Weninger et al. 2012), and thus presents a first step towards a fully automatic vibrato detection system. We will show that the performance of the FDM-based system exceeds that of state-of-the-art vibrato detection methods and obtains high accuracy values for vibrato parameter estimation. To our knowledge, our study represents the first application of FDM to the music domain.

FDM offers two advantages over Fourier transform-based methods. It can obtain the frequencies and amplitudes of a given number of sinusoids in a selected frequency range for minute time frames. Its parametric fitting technique is capable of extracting the sinusoids directly from the f_0 time series without deriving them from spectral information, bypassing the error-prone peak-picking (Keiler and Marchand 2002) intermediate step in the frequency domain. We will show that these two advantages significantly improve vibrato detection beyond the state-of-the-art for short time frames.

The structure of the paper is as follows: Section 2 presents related work in automatic vibrato detection; Section 3 introduces the proposed method; evaluations are shown and discussed in Section 4; and Section 5 presents the conclusions.

2. Related work

In this section, we focus on prior work related to vibrato detection. In the literature, there exist two classes of vibrato detection methods: note-wise and frame-wise methods. Note-wise methods require a note segmentation pre-processing step, usually carried out via manual annotation, before determining if the note contains vibrato. See, for example, Rossignol et al. (1999), Pang and Yoon (2005), Özaslan and Arcos (2011), and Weninger et al. (2012). The requisite pre-processing step in these methods precludes real-time detection. In contrast, frame-wise methods can be applied in real time, by dividing the audio stream, or the extracted f_0 information, into a number of uniform frames. Potential vibrato behavior in the frame is then detected. The focus of the present paper is frame-wise vibrato detection. The leading frame-wise methods are described below.

Figure 2 shows a flow chart describing a general frame-wise method. The time-varying signal is first subdivided into overlapping frames. The fundamental frequency, f_0, or together with amplitude[1] $\mathcal{A}$, of the audio signal is extracted from each frame to form a time series. The time series extracted serves as input to the feature extraction module where salient features of the vibratos are determined. Finally, vibrato existence is determined via a decision-making mechanism. We next describe the three existing frame-wise vibrato detection methods. Table 1 summarises the key components of the above three frame-wise methods.

Herrera and Bonada (1998) applied a Short-Time Fourier Transform (STFT) to f_0 to obtain a spectral representation of the time series. Because vibrato rates tend to be in the 4–8 Hz range,

[1] Usually obtained from the root mean square of the audio intensity.

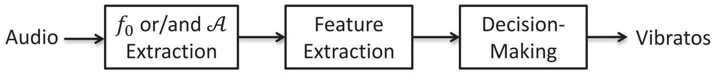


Figure 2. Basic framework of frame-wise vibrato detection methods.

Table 1. Comparison of existing vibrato detection methods.

Method	$f_0/\mathcal{A}$	Feature extraction	Decision-making
Herrera–Bonada	f_0	STFT(f_0) + Parabolic Interpolation	DT(F)
Ventura–Sousa–Ferreira	f_0	STFT(f_0) + RecSine Peak Estimation	DT(F)
von Coler–Röbel	f_0 and $\mathcal{A}$	Cross correlation of STFT(f_0_mod) and STFT($\mathcal{A}$_mod)	DT(corr)

Note: f_0 = fundamental frequency; $\mathcal{A}$ = amplitude of audio signal; DT(F) = decision tree using frequency; DT(corr) = decision tree using cross correlation.

parabolic interpolation was employed to determine if the spectrogram peaked around 5–6 Hz. Ventura, Sousa, and Ferreira (2012) proposed a similar method, but their method used a combined rectangular-sine window frequency interpolation method described in Sousa and Ferreira (2010) to improve the spectral peak-picking.

von Coler and Röbel (2011) presented a cross-correlation-based frame-wise method for vibrato detection. Their method assumes that frequency modulations in physical instruments cause amplitude modulations. They extracted the frequency and amplitude modulation time series. A cross correlation of the STFT of both the frequency and amplitude modulation time series was then computed. The resulting curve showed positive peaks in parts with vibrato; however, the method cannot be used for pure frequency modulation vibratos.

3. Methodology

The vibrato detection and estimation system follows the basic framework outlined in Figure 2. Section 3.1 describes the FDM method for extracting spectral features from f_0. Section 3.2 presents the two decision-making methods paired with FDM for vibrato detection.

3.1. *The filter diagonalisation method*

FDM was developed as a tool for the efficient extraction of high resolution spectral information from short time signals, and has been used for a range of applications ranging from nuclear magnetic resonance to quantum dynamical systems (Neuhauser 1990; Wall and Neuhauser 1995; Mandelshtam and Taylor 1997; Mandelshtam 2001; Martini et al. 2013). As FDM is new (to the best of our knowledge) to musical analysis, we will briefly describe its formulation and application to harmonic inversion.

While FDM can, in general, be used to determine all the fundamental frequencies and harmonics of a waveform audio signal $x(t)$ over an arbitrary frequency band, we are concerned in particular with the characterisation of vibrato. As vibrato is an oscillating pitch, it can be characterised by properties of the oscillations over very short time periods. One may simply apply the Fourier transform to the time-varying fundamental frequency but, as will be discussed, the STFT is ill-suited to this task even if some peak-picking methods have been employed. We will, however, show that the FDM algorithm outputs a good representation of the vibrato signal.

First, we define the fundamental frequency time series, which describes the variation of the fundamental frequency with time as $f_0(t) = f_0(n\tau)$, where $n = 0, \ldots, N$ and τ is the sampling period for the fundamental frequency. The time series of the fundamental frequency can be extracted from a musical audio signal waveform, $x(t) = x(n'\tau')$, where $n' = 0, \ldots, N'$ and τ' is the sampling period for the audio signal waveform. A frame $x(nN_s\tau' + 1 : (n+1)N_s\tau')$ is defined over a short time segment of N_s samples, and then one of a number of fundamental frequency extraction methods (Boersma 1993; de Cheveigné and Kawahara 2002; Mauch and Dixon 2014; Childers, Skinner, and Kemerait 1977) can be applied to determine the fundamental frequency of that particular frame. By iterating the frames across the entire time segment signal, a time-dependent fundamental frequency function results:

$$f_0(t) = f_0(n\tau) = f_0(nN_s\tau') = \mathcal{T}\{x(nN_s\tau' + 1 : (n+1)N_s\tau')\}, \tag{1}$$

where $\mathcal{T}$ stands for the fundamental frequency extraction transform, $N = N'/N_s$. If the final Δ samples of a frame are used as the first Δ samples of the next frame, then the total number of frames is given by $N = (N' - N_s)/(N_s - \Delta)$. Further analysis will be applied to the signal $f_0(t)$ in order to characterise the spectra resulting from vibrato oscillations.

3.1.1. *Outline of FDM*

Determination of the vibrational spectrum of a dynamical system is typically performed using one of two classes of techniques: through calculation of the Fourier transform of a signal; or by diagonalisation or inversion of a matrix representing a short time segment of a signal. It is well understood that the straightforward application of the Fourier transform, while effective at extracting large numbers of frequencies at arbitrary spectral ranges, is restricted by the uncertainty principle. For discretely sampled data this implies the need for a long time signal $T = 1/\Delta f$, which makes computation prohibitively expensive. Furthermore in many dynamical systems, including music, the harmonic profile may be changing rapidly enough throughout the analysis window that there is insufficient time to capture the data necessary for the inversion, resulting in low resolution results.

The second class of techniques is typically more useful in these applications as they rely on the determination of relevant information (harmonic frequencies, decay rates, amplitude, and phase) simultaneously through analysis of a short time segment of signal. Essentially, the purpose of these algorithms (including Prony's method, MUSIC (Schmidt 1986), ESPRIT (Paulraj, Roy, and Kailath 1985), etc.) is to fit the relevant parameters to represent the signal as a sum of exponentially decaying sinusoids:

$$f_0(t) = f_0(n\tau) = \sum_{k=1}^{K} d_k e^{-in\tau\omega_k}, \quad \text{for } n = 0, 1, \ldots, N, \tag{2}$$

where K is the number of sinusoids required to represent the signal to within some tolerance; ω_k and d_k are fitting parameters which are defined as the complex frequency and complex weight of the kth sinusoid, respectively. In general, the real part of the complex frequency represents the sinusoidal frequency while the imaginary part represents the decay rate (damping factor). The complex weighting parameter d_k represents the relative amplitude (real part) and phase (imaginary part) of each sinusoidal component. The aim is to solve for a total of $2K$ unknowns, representing all ω_k and d_k.

While these techniques use a variety of methods to achieve this approximation, a common feature necessary for computational efficiency is the need to convert a nonlinear fitting problem to a linear algebraic one. Typically this may lead to large or ill-conditioned problems when there are "too-many" frequencies (Mandelshtam and Taylor 1997). The method of Wall and Neuhauser was introduced for high resolution spectral analysis of a time signal defined over a short time segment, and was shown to be exceptionally efficient (Wall and Neuhauser 1995) compared to linear prediction algorithms (MUSIC, ESPRIT, etc.), not least because all parameters ω_k and d_k are given through a single diagonalisation, rather than requiring multiple procedures (Hu et al. 1998).

While the technique was originally applied to continuous segments, it was later extended to discrete signals in Mandelshtam and Taylor (1997). The key novelty in this procedure was the association of the time signal, $f_0(t)$, with an autocorrelation function, such that

$$f_0(t) = f_0(n\tau) = \left(\Phi_0, e^{-in\tau\hat{\Omega}}\Phi_0\right), \tag{3}$$

where $(\cdot,\cdot)$ denotes the complex symmetric inner product, and Φ_0 is a $K \times 1$ size vector representing the initial state, which does not need to be known explicitly. At this stage the exact form of the inner product is not important; rather, it is the complex symmetric properties that are most pertinent. Extracting spectral information from the signal, $f_0(t)$, is therefore equivalent to diagonalising the evolution operator, $\hat{U} \equiv e^{-i\tau\hat{\Omega}}$. FDM is then used to extract the eigenvalues, or harmonics, of $\hat{\Omega}$, given as

$$u_k = e^{-i\tau\omega_k}. \tag{4}$$

The method is particularly well suited for spectra modelled as sums of discrete frequencies. It is summarised below, with more detailed treatment found in Neuhauser (1990), Wall and Neuhauser (1995), Mandelshtam and Taylor (1997), Hu et al. (1998), and Mandelshtam (2001).

3.1.2. *The FDM algorithm*

In common with many other techniques, including STFT, the purpose of FDM is to decompose the original time series, $f_0(t)$, into a sum of sinusoids, as described in equation (2).[2] As the musical signal can be quite complex, with a number of harmonics, we restrict our search to extracting the frequency and amplitude of the sinusoid with the largest amplitude. This gives sufficient accuracy at reduced computational cost, but we note that the technique may be generalised to output further harmonics.

The selection of a primitive Krylov basis for equation (3) reduces the problem to that of a linear prediction algorithm such as ESPRIT. However the use of a rectangular window Fourier basis significantly improves computational efficiency as the resultant matrix is almost automatically diagonalised, see e.g. Hu et al. (1998). Essentially, the idea is to split up the frequency range of interest into a discretised frequency grid over which the measured signal $f_0(t)$ can be analysed. Selecting the Fourier basis set,

$$\Psi(\phi) = \sum_{n=0}^{M} e^{in\phi}\Phi_n \equiv \sum_{n=0}^{M} e^{in(\phi-\hat{\Omega}\tau)}\Phi_0, \tag{5}$$

the set of ϕ values defines the grid of frequency components that represent an estimation of the location of the spectral peaks of interest. An important result of this basis selection is that the matrix elements of the operator $\hat{U}^p = \exp(-ip\tau\hat{\Omega})$, $p = 0, 1, 2, \ldots,$ of any two functions $\Phi(\phi)$ and $\Phi(\phi')$ can be evaluated purely in terms of the elements of the signal $f_0(n)$.[3] The matrix

[2] Here, the Direct Component (DC) of $f_0(t)$ for a frame is removed by subtracting the mean.

[3] For the purpose of neat representation, we use $f_0(n)$, which is short for $f_0(n\tau)$.

operator can be calculated to be

$$\mathbf{U}^{(p)}(\phi',\phi) = (\Psi(\phi'), \hat{U}^p\Psi(\phi)) = \sum_{n'=0}^{M}\sum_{n=0}^{M} e^{in\phi}e^{in'\phi'}f_0(n+n'+p), \tag{6}$$

which may be recognised as a 2D discrete Fourier transform.

To evaluate the operator $\mathbf{U}^{(p)}$ it is first necessary to define the grid of frequency components. It is prudent to define a uniformly spaced grid, such that

$$\phi_k = -2\pi(k\Delta f + f_{\min})\tau; \quad \Delta f = \frac{f_{\max} - f_{\min}}{K}, \tag{7}$$

where $K+1 = (f_{\max} - f_{\min})N_s\tau/2$ represents a suitable selection for the number of frequency points on the evaluation grid, as it will give the maximum spectral resolution, which could give a unique fit to the specified signal length. On substitution of this equation into equation (6), we have

$$\mathbf{U}^{(p)}_{kk'} = \sum_{n'=0}^{M}\sum_{n=0}^{M} f_0(n+n'+p)e^{-2i\pi f_{\min}\tau(n+n')}e^{-2i\pi\Delta f\tau kn}e^{-2i\pi\Delta f\tau k'n'}. \tag{8}$$

By setting $M = K-1$, one may solve the former equation for the evolutionary operator very efficiently by taking a 2D Fast Fourier Transform (FFT) of the function $f_0(n+n'+p)e^{-2i\pi f_{\min}\tau(n+n')}$. This is the primary achievement of the FDM algorithm.

Once the evolutionary operator $\mathbf{U}^{(p)}$ has been determined, the generalised eigenvalue problem can be solved:

$$\mathbf{U}^{(1)}\mathbf{B}_k = u_k\mathbf{U}^{(0)}\mathbf{B}_k, \tag{9}$$

from which the k eigenvalues (each giving fundamental/harmonic resonant frequencies and damping coefficients) are determined from u using equation (4), and the complex amplitudes (giving amplitude and phase information) are determined using

$$d_k = \left(\sum_{n=0}^{M} f_0(n)B_{nk}\right)^2. \tag{10}$$

To summarise, the real benefit of the FDM algorithm lies in very efficient and accurate determination of harmonic information using short time series. A computationally efficient 2D FFT is performed over a small frequency window $[f_{\min}, f_{\max}]$ in which there are up to K harmonics. A generalised eigenvalue equation then gives all relevant information. This technique, which is highly suited to vibrato detection, can reduce the linear algebraic computational effort and round-off errors. As the vibrato rate is usually between 4 and 8 Hz, the frequency window can be set around this range to reduce the computational cost of the generalised eigenvalue decomposition. We set the frequency window as 2–20 Hz.

The basic steps for FDM in the feature extraction module are summarised by the pseudocode in Algorithm 1. We consider only the frequency and amplitude, denoted by $F_H = f_{d_{\max}}$ and $A_H = 2|d_{\max}|$, respectively, for the sinusoid having the largest amplitude.

3.2. *Deciding vibrato presence*

Following the application of FDM or FFT, a further decision-making step is required to determine vibrato existence. In this section, we propose two alternative methods: DT and BR. Both methods use frequency (F_H) and amplitude (A_H) information.

Algorithm 1: The FDM algorithm

```
**Input**: $f_0$
**Output**: $f_{d_{\max}}$, $d_{\max}$
$f_{\min} = 2; f_{\max} = 20;$
Filter
$\omega_{\min} = 2\pi f_{\min}, \omega_{\max} = 2\pi f_{\max};$
$K + 1 = (f_{\max} - f_{\min}) N_s \tau / 2$
$\phi_k = -2\pi(k\Delta f + f_{\min})\tau, k = 1 : K$
Diagonalisation
$N_{\text{iteration}} = 4;$
**for** $n = 1 : N_{\text{iteration}}$ **do**
    **for** $p = 0 : 2$ **do**
        obtain $\mathbf{U}^p$ through 2D FFT of $f_0(n + n' + p)e^{-i\omega_{\min}\tau(n+n')}$;
    Solve $\mathbf{U}^{(1)}\mathbf{B_k} = u_k\mathbf{U}^{(0)}\mathbf{B_k}$;
    Get $u_k = \mathrm{e}^{-\mathrm{i}\tau\omega_k}$ and $\mathbf{B_k}$;
    **if** $\left\|(\mathbf{U}^{(2)} - u_k^2\mathbf{U}^{(0)})\boldsymbol{B_k}\right\| < \varepsilon$ **then**
        accept $u_k$;
    $z_k = u_k$;
Calculate $d_k$;
Return $d_{\max}$ and corresponding $f_{d_{\max}}$;
```

3.2.1. *Decision tree*

A decision tree (DT) is constructed to support the vibrato detection process, as in Herrera and Bonada (1998) and Ventura, Sousa, and Ferreira (2012). In contrast to the previous methods, which use only frequency information, we use both frequency and amplitude information provided by FDM. The method requires the frequency range thresholds $F_{\text{thd}} = [f_{\min}, f_{\max}]$ (Hz) and the amplitude range thresholds $A_{\text{thd}} = [a_{\min}, a_{\max}]$ to be pre-determined.

The frequency range thresholds can be obtained from the reported vibrato rate in the literature: $f_{\min} = 4$ Hz, $f_{\max} = 12$ Hz for Western classical music (Desain and Honing 1996); $f_{\min} = 4$ Hz, $f_{\max} = 9$ Hz for a singing voice (Prame 1994); and, $f_{\min} = 5$ Hz, $f_{\max} = 8$ Hz for erhu music (Yang, Chew, and Rajab 2013). The amplitude range thresholds can be determined empirically: for instance, for voice and erhu we used $a_{\min} = 0.15$, $a_{\max} = +\infty$, and for violin $a_{\min} = 0.07$, $a_{\max} = +\infty$.

3.2.2. *Bayes' Rule*

The second technique applies Bayes' Rule (BR), which assigns a probability of vibrato existence, rather than a binary answer, to each frame. Again, we consider the frequency and amplitude, F_H and A_H, respectively, of the sinusoid with the largest amplitude. Let V indicate vibrato existence, $\neg V$ indicates no vibrato. Suppose that $\mathrm{P}(F_H) \neq 0$, the probability of vibrato existence given F_H is

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(V \cap F_H)}{\mathrm{P}(F_H)}, \tag{11}$$

and suppose $\mathrm{P}(A_H) \neq 0$, the probability of vibrato existence given A_H is

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(V \cap A_H)}{\mathrm{P}(A_H)}. \tag{12}$$

According to Bayes' theory, we can re-write equations (11) and (12) as

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(F_H|V)\mathrm{P}(V)}{\mathrm{P}(F_H)}, \tag{13}$$

and

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(A_H|V)\mathrm{P}(V)}{\mathrm{P}(A_H)}, \tag{14}$$

where $\mathrm{P}(F_H|V)$ and $\mathrm{P}(A_H|V)$ are the probabilities of observing F_H and A_H, respectively, given vibrato existence. $\mathrm{P}(F_H|V)$ and $\mathrm{P}(A_H|V)$ can be obtained from the estimated probability density function (PDF) for F_H and A_H, respectively. $\mathrm{P}(V)$ is the prior probability of vibrato.

Equations (13) and (14) lead to

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(F_H|V)\mathrm{P}(V)}{\mathrm{P}(F_H|V)\mathrm{P}(V) + \mathrm{P}(F_H|\neg V)\mathrm{P}(\neg V)}, \tag{15}$$

and

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(A_H|V)\mathrm{P}(V)}{\mathrm{P}(A_H|V)\mathrm{P}(V) + \mathrm{P}(A_H|\neg V)\mathrm{P}(\neg V)}, \tag{16}$$

respectively, using the Law of Total Probability in the denominators.

$\mathrm{P}(F_H|\neg V)$ and $\mathrm{P}(A_H|\neg V)$ can be obtained from the estimated PDF for F_H and A_H from non-vibrato frames. One such example, where the PDFs are estimated,[4] is given in Figure 3. The graph suggests that high values of $\mathrm{P}(F_H|V)$ lie between 5 and 9 Hz, which is the typical vibrato frequency range. $\mathrm{P}(A_H|V)$ is larger than $\mathrm{P}(A_H|\neg V)$ for amplitudes between around 0.2 and 1, and $\mathrm{P}(\neg V)$ is obtained using $\mathrm{P}(\neg V) = 1 - \mathrm{P}(V)$.

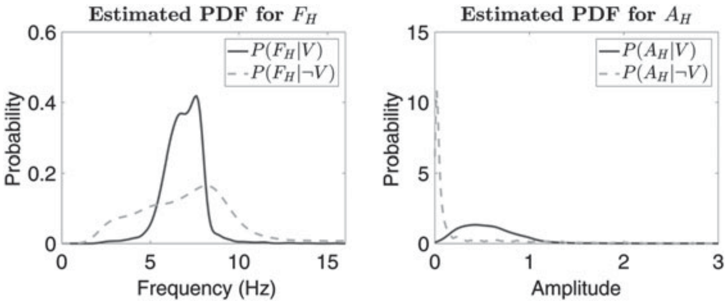


Figure 3. PDFs of $P(F_H|V)$, $P(F_H|\neg V)$, $P(A_H|V)$, and $P(A_H|\neg V)$ estimated using an erhu sample.

The five probabilities that need to be estimated from data are as follows: $\mathrm{P}(V)$, $\mathrm{P}(F_H|V)$, $\mathrm{P}(F_H|\neg V)$, $\mathrm{P}(A_H|V)$, and $\mathrm{P}(A_H|\neg V)$. For simplicity, we set the prior probability of vibrato[5] $\mathrm{P}(V) = 0.5$. Thus, the prior probability of non-vibrato is $\mathrm{P}(\neg V) = 0.5$. We then multiply equations (15) and (16) to get the probability of vibrato existence:

$$\mathrm{P}(V) = \mathrm{P}(V|F_H) \times \mathrm{P}(V|A_H). \tag{17}$$

A threshold needs to be set or tuned to determine vibrato presence. We assume no prior information, i.e. that the probability of vibrato given any frequency is 0.5 and that given any amplitude

[4] We estimated the density functions using the kernel density estimator in MATLAB. See for instance http://uk.mathworks.com/help/stats/kernel-distribution.html. We used data from one passage of *The Moon Reflected on the Second Spring* performed by Jiangqin Huang.

[5] This quantity can be tailored to specific performers, instruments, genres, and cultures.

is 0.5. Thus, empirically, for the experiments presented here, the threshold is set at 0.25 by assigning 0.5 each to $\mathrm{P}(V|F_H)$ and $\mathrm{P}(V|A_H)$, respectively.

4. Evaluation

This section describes the evaluation datasets, and reports on the comparison of vibrato detection methods, and evaluation of vibrato estimation.

4.1. *Evaluation datasets*

This section provides details on the two evaluation datasets.

4.1.1. *von Coler–Röbel and CMMSD datasets*

The first evaluation dataset consists of a combination of the existing von Coler and Röbel dataset (von Coler and Röbel 2011) and von Coler and Lerch's Classical Monophonic Music Segmentation Dataset (CMMSD) (von Coler and Lerch 2014). Both datasets consist of monophonic samples. The von Coler–Röbel dataset contains samples from 28 solo instrument passages of lengths ranging from 2 to 12 s. The samples are classified into four instrument groups: violin, voice, woodwind, and brass. Vibrato annotations were completed by two persons, each using the Audacity® software. The CMMSD dataset consists of 36 solo instrument (string, woodwind, and brass) excerpts. The vibrato annotations were created by the first author using Tony (Mauch et al. 2015).

4.1.2. *The Moon Reflected on the Second Spring dataset*

In contrast to the short excerpts of the previous evaluation dataset, we created another dataset featuring long passages of music, which readily allows different parts of the same passage to be used as training and held out (test) data, for example, for the FDM + BR method. This new dataset contains entire recordings of four performances of the traditional Chinese piece, *The Moon Reflected on the Second Spring*. Two performances were recorded on the Chinese erhu and the other two on the Western violin. See Table 2 for more details.

Table 2. *The Moon Reflected on the Second Spring* dataset.

No.	Instrument	Performer	Duration (s)	Number of vibratos
1	Erhu	Jiangqin Huang (Huang 2006)	445.83	170
2		Guotong Wang (Wang 2009)	387.53	168
3	Violin	Jiang Yang[a]	254.54	124
4		Laurel S. Pardue[a]	325.50	120

[a] Recorded by the performer.

Vibrato presence was annotated by the first author using Tony.[6] We divide each performance into contiguous training and test segments that are proportionally 70% (about 16.5 minutes of the

[6] In an effort to create a more robust dataset, we had another annotator generate a separate set of annotations. Because the second annotator was less experienced, the quality of the annotations were noticeably poorer (less uniform), and combining the two sets of annotations would have diluted the quality of the dataset. The perception of the vibrato onsets and offsets also varied from one person to another, and taking the average would not have produced a musically

recording or 31,700 consecutive 0.125 s frames) and 30% of the total length, respectively. We maintain a circular view of the recording so that the 70% training segment may begin at the tail end of the recording and loop back to the front. Figure 4 demonstrates the segmentation process. The process is iterated 10 times in order to obtain more stable results.

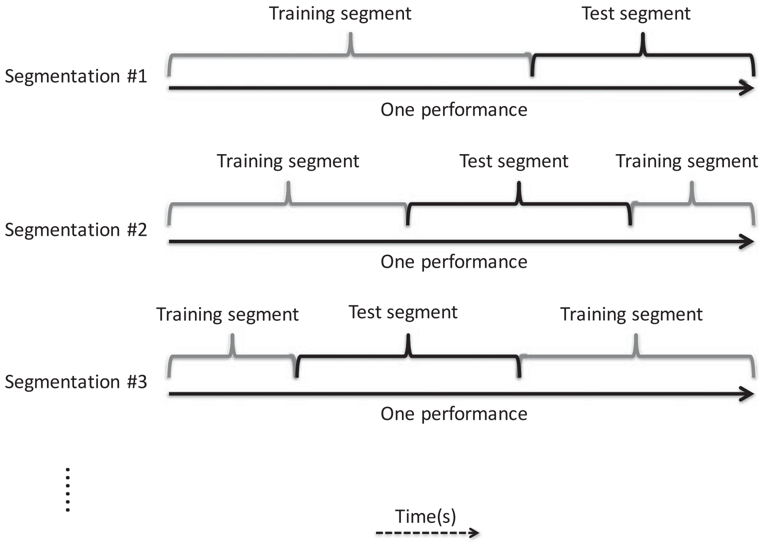


Figure 4. Demonstration of training/test data selection. The training segment spans 70% of the recording, with the remaining 30% held out as test data. The circular view of the recording is demonstrated in Segmentations #2 and #3.

4.2. *Vibrato detection comparison*

In this section, we compare the Herrera–Bonada (HB) (Herrera and Bonada 1998), Ventura–Sousa–Ferreira (VSF) (Ventura, Sousa, and Ferreira 2012), and von Coler–Röbel (CR) (von Coler and Röbel 2011) methods against our proposed FDM-based methods (with the two alternate decision mechanisms). The core components of the individual methods are outlined in Table 3. We re-implemented the HB and VSF methods; the CR method's MATLAB® code was provided by the authors.

Table 3. Experiment setup for comparison of candidate frame-wise vibrato detection methods.

Method	$f_0/\mathcal{A}$	Feature extraction	Decision-making
Herrera–Bonada	f_0	STFT(f_0) + Parabolic Interpolation	DT(F)
Ventura–Sousa-Ferreira	f_0	STFT(f_0) + RecSine Peak Estimation	DT(F)
von Coler–Röbel	f_0 and $\mathcal{A}$	Cross correlation of STFT(f_0_mod) and STFT($\mathcal{A}$_mod)	DT(corr)
FDM + DT (proposed)	f_0	FDM(f_0)	DT(F,A)
FDM + BR (proposed)	f_0	FDM(f_0)	BR

Note: DT(F) = decision tree using Frequency; DT(F,A) = decision tree using frequency and amplitude; BR = Bayes' Rule.

meaningful number. Thus, we chose a high-quality one-annotator dataset over a lower-quality two-annotator dataset, and decided to stick with only the original set of annotations. This highlights the difficulty in obtaining robust datasets. Additionally, we note that a single annotator may not be considered representative of all knowledgeable listeners; the model is thus of the perception of a single individual.

We use as input to the FDM-based techniques, and for the HB and VSF algorithms, the f_0 obtained using PYIN (Mauch and Dixon 2014), a probabilistic version of the original YIN method (de Cheveigné and Kawahara 2002). We left the CR f_0 and $\mathcal{A}$ (amplitude) extraction modules untouched as the method requires cross correlation of the STFT of both the frequency and amplitude modulation time series. The two variants of the FDM-based method used the DT (denoted as FDM + DT(F,A)) and BR (FDM + BR) decision mechanisms, respectively. For fine time resolution, we set the window size to $w = 0.125$ s and step size $s = w/4$. When two or more methods required the same kinds of thresholds – for example, the vibrato frequency range threshold, F_{thd}, employed by HB, VSF, and FDM + DT(F,A) – the threshold was made uniform across the methods. The CR method had different thresholds for different instruments, as published in von Coler and Röbel (2011); we used the thresholds as published for the CR method.

4.2.1. *Frame-level results*

For frame-by-frame evaluation, a ground truth vector was created using the same sampling rate as the detection vector. The performance was then evaluated using the F-measure

$$F = \frac{2PR}{P + R}, \tag{18}$$

where *precision*, P, is defined as the number of true positive vibrato frames divided by the total positive vibrato frames, and *recall*, R, is defined as the number of true positive vibrato frames divided by the total number of vibrato frames. The F-measure was calculated for each excerpt.

Figure 5 shows the precision, recall and F-measure results for each vibrato detection method. The subplot (a) shows the evaluation on the von Coler–Röbel and CMMSD datasets between our proposed FDM + DT(F,A) system and the HB, VSF, and CR methods. The FDM + BR system was omitted due to insufficient data for training the priors. Subplot (b) presents the evaluation performed using *The Moon Reflected on the Second Spring*.

The FDM-based methods perform significantly better with respect to the F-measure for both datasets, reflecting the better balance they strike between precision and recall. The statistical Bayes' Rule gives better results than the DT. The HB and VSF have higher recall values and lower precision values. This implies that these two FFT-based methods correctly identified most vibrato frames, but at the cost of a substantial number of false positives; in fact, almost all frames were classified as vibratos, likely due to the low frequency resolution of the FFT at the short window size. The alternate mechanism CR method obtains a slightly higher precision for the

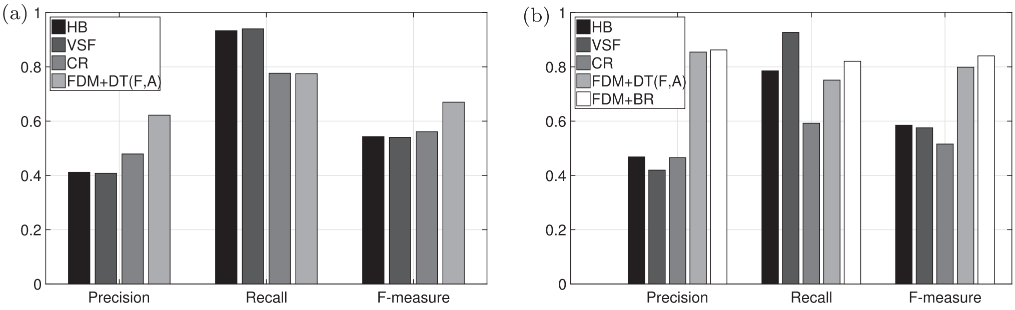


Figure 5. Frame-level evaluation. Results shown are averaged over all excerpts and iterations. (a) von Coler-Röbel and CMMSD dataset; (b) *The Moon Reflected on the Second Spring* dataset.

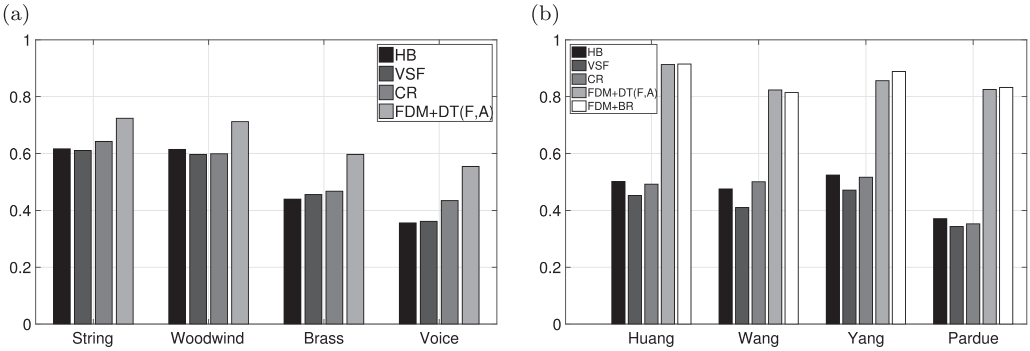


Figure 6. F-measure evaluation for (a) each instrument group – results shown are averaged over the von Coler–Röbel and CMMSD datasets – and (b) each performer in *The Moon Reflected on the Second Spring* dataset – results shown are averaged over all iterations.

von Coler–Röbel and CMMSD datasets but a lower recall for both datasets, resulting in F-measure values similar to those of HB and VSF. This may be due to the CR method using the cross correlation of the STFT of the frequency and amplitude modulations.

Figure 6 presents a further analysis. Subplot (a) shows the instrument-wise F-measure. There is an agreement amongst all methods that vibratos in string and woodwind instruments are more easily detected than those in brass instruments and voice. The data shows that vibratos in brass instruments have small frequency but high amplitude modulations; and those in voice, at least for the datasets we studied, are less well controlled and more irregular. Subplot (b) represents the performer-wise F-measure, showing little difference in vibrato detection between erhu and violin recordings. The vibrato detection can be improved by using performer-specific decision-making mechanisms.

4.2.2. *Note-level results*

Vibrato is a continuous phenomenon operating at the level of the musical note – see examples in Figure 1. To evaluate the accuracy of vibrato boundaries (onset and offset), we employ the note boundary evaluation metric described in Molina et al. (2014), originally applied to singing voice melody transcription. This is a more stringent evaluation than the frame-level evaluation described in the previous section. To the best of the authors' knowledge, this is the first note-level evaluation of vibrato detection.

We assume that a vibrato spans at least five consecutive frames, i.e. that it has duration greater than 0.28 s. A detected vibrato onset is considered to be correct if it is within ±100 ms of the ground truth onset. Note that this threshold is higher than that for music transcription, which is typically 50 ms. The detected vibrato offset is considered correct if it is within ±100 ms of the ground truth offset or no more than ±20% of the ground truth vibrato duration from the ground truth offset. By the *ground truth vibrato* we mean the real vibrato in the recordings, which is labelled by a human. This is in contrast to the *detected vibrato*.

We compute the F-measure for each excerpt, where the F-measure is defined as given inequation (18). Instead of using vibrato frames, the *precision*, P, is defined as the number of true positive vibratos divided by the total positive vibratos, and *recall*, R, is defined as the number of true positive vibratos divided by the total number of vibratos. A true positive (correctly identified) vibrato is defined as a detected vibrato for which both onset and offset information are correct.

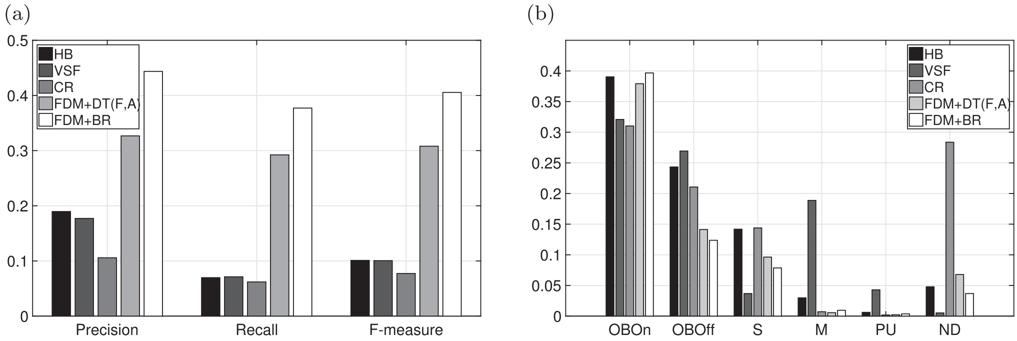


Figure 7. Note-level evaluation for *The Moon Reflected on the Second Spring* dataset. The results are the average values across all excerpts and iterations.

To analyse the errors, we use the six error types as defined in Molina et al. (2014). Only-Bad-Onset (OBOn) error refers to the case where an onset error occurs but the offset is correct. Only-Bad-Offset (OBOff) error refers to the case where the onset is correct but the offset is not. A Split (S) error is said to occur when the answer splits the ground truth vibrato into a number of consecutive detected vibratos. A Merge (M) error refers to the case where a number of consecutive ground truth vibratos are merged as one detected vibrato. A sPUrious (PU) error refers to the case where a transcribed note does not overlap with any ground truth note. A Non-Detected (ND) error occurs when the ground truth vibrato does not overlap with any detected vibrato. The error rates for these six error types are obtained by dividing each count by the number of ground truth vibratos; the PU error rate is divided by the number of detected vibratos.

Because the von Coler–Röbel and CMMSD datasets contain short excerpts having one, two, or several vibratos, one false positive or false negative will have significant impact on the results for each excerpt. Thus, we choose *The Moon Reflected on the Second Spring* dataset, which has sufficient numbers of vibratos for each passage, for the note-level evaluations.

The note-level evaluation results are shown in Figure 7. As expected, the note-level precision, recall, and F-measure results are lower than those for the frame level. The FDM-based methods produce higher F-measures than those of HB, VSF, and CR. More specifically, the FDM + DT(F,A) has an F-measure value of 0.31 and FDM + BR improves the F-measure to 0.41. Compared to the corresponding frame-level results shown in subplot (b) of Figure 5, BR has much more influence on note-level evaluation.

Focusing on the error types, OBOn rates are higher than OBOff rates for all methods, which points to better vibrato offset identification abilities. Even though HB and VSF are both FFT-based methods with similar precision, recall, and F-measure values, they have very different split, merged, spurious, and non-detected error rates. HB tends to split vibratos and fails to detect some vibratos, leading to higher S and ND values, and lower M and PU values. VSF stands at the opposite end of the spectrum. This may be due to its different peak-picking mechanisms. The CR has good performance on all types of error except ND, which negatively impacts the overall performance. One of the reasons may be due to its use of frequency–amplitude cross correlation, which may miss purely frequency-modulated or small amplitude-modulated vibratos. The proposed FDM-based methods successfully suppress these four types of error.

4.3. *Vibrato estimation evaluation*

The output of FDM also provides parameters for the vibratos detected. In this section, we evaluate the accuracy of these vibrato parameters, namely the vibrato rate and extent. F_H, the output frequency having the largest amplitude, is used directly as the vibrato rate for that frame. The vibrato extent is A_H (as described in Section 3). Each vibrato's rate and extent are aggregated from its consecutive vibrato frames. We maintain the assumption of a vibrato spanning at least five consecutive frames.

We manually annotated the vibrato rates and extents for *The Moon Reflected on the Second Spring* dataset using SonicVisualiser (Cannam, Landone, and Sandler 2010). The peaks and troughs of each vibrato were marked based on the spectrogram and f_0 information. The ground truth rate and extent for each vibrato were calculated from each half cycle. Assuming the interval between one peak and one trough is the duration of a half cycle, and the vibrato rate is the inverse of the cycle length, the vibrato extent is the difference between the peak and trough measured in semitones. The vibrato rate and extent for each note is the mean value over all half cycles.

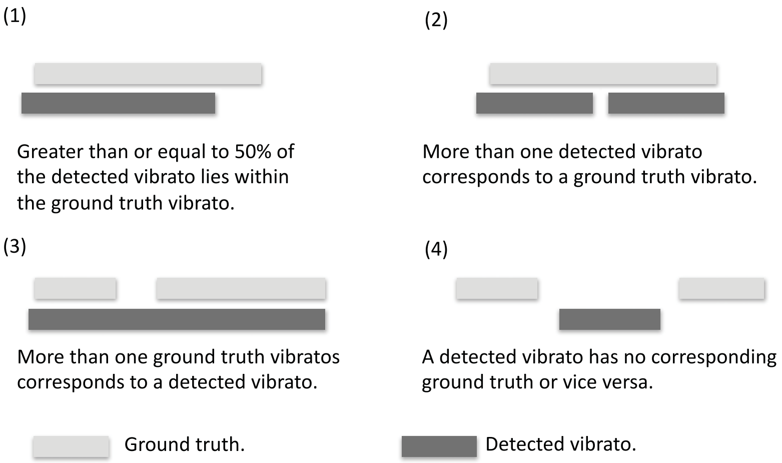
Vibrato parameter estimation accuracy is complicated by the fact that sometimes a ground truth vibrato is detected as two vibratos, and sometimes more than one ground truth vibrato are detected as one vibrato. So as systematically to determine corresponding ground truth and detected vibratos in order to assess parameter estimation accuracy, we apply the following rules, which are illustrated by Figure 8:

(1) for any ground truth vibrato, the corresponding detected vibrato is one for which at least half its interval lies within that of the ground truth vibrato;
(2) if there are more than one corresponding detected vibratos, the *average* of the parameters of the detected vibratos will be used for assessing accuracy;
(3) if more than one ground truth vibrato corresponds to a detected vibrato, the detected vibrato's parameters will be used for comparison with those of each ground truth vibrato; and,
(4) if a detected vibrato has no corresponding ground truth vibrato or vice versa, no comparisons will be done.

The accuracy percentage is defined as

$$A_p = \begin{cases} 1 - \dfrac{|\hat{p} - p|}{p} & \text{if } \hat{p} \leq 2p, \\ 0 & \text{if } \hat{p} > 2p, \end{cases} \qquad (19)$$

where $\hat{p}$ is the estimated vibrato parameter (rate or extent) and p is the corresponding ground truth value.



(1) Greater than or equal to 50% of the detected vibrato lies within the ground truth vibrato.

(2) More than one detected vibrato corresponds to a ground truth vibrato.

(3) More than one ground truth vibratos corresponds to a detected vibrato.

(4) A detected vibrato has no corresponding ground truth or vice versa.

Figure 8. Illustration of determining corresponding ground truth and detected vibratos.

Table 4. Vibrato rate accuracy for *The Moon Reflected on the Second Spring* dataset.

No.	Instrument	Performer	HB (%)	VSF (%)	FDM + DT(F,A) (%)	FDM + BR (%)
1	Erhu	Jiangqin Huang	92.68	86.08	93.84	93.61
2		Guotong Wang	87.99	76.86	90.64	90.79
3	Violin	Jiang Yang	90.73	85.30	93.27	93.03
4		Laurel S. Pardue	92.04	85.10	92.97	92.94
Average			90.86	83.33	92.68	92.59

Table 5. Vibrato extent accuracy for *The Moon Reflected on the Second Spring* dataset.

No.	Instrument	Performer	FDM + DT(F,A) (%)	FDM + BR (%)
1	Erhu	Jiangqin Huang	88.02	89.53
2		Guotong Wang	73.36	79.46
3	Violin	Jiang Yang	87.59	90.90
4		Laurel S. Pardue	87.23	90.49
Average			84.05	87.59

Tables 4 and 5 show the accuracy of the estimations of vibrato rate and extent, respectively. The accuracy values reported are the average over all iterations. For vibrato rate accuracy, we compare FDM + DT(F,A) and FDM + BR with the HB and VSF methods. The CR method is excluded here as it outputs neither vibrato rate nor extent. All methods achieve relatively high vibrato rate accuracies. FDM + DT(F,A) and FDM + BR obtained the highest vibrato rate accuracies, 92.68 and 92.59%, respectively. HB has a value of 90.86% and VSF a lower value at 83.33%. HB and VSF use DTs to determine vibrato existence from vibrato rates; thus, even though their vibrato detection performance may be lower, for the vibratos that were correctly detected, the vibrato rates have been reasonably accurately assessed.

For vibrato extent, we only report the results from our methods, FDM + DT(F,A) and FDM + BR, because the other three methods do not have a direct vibrato extent output. Extending these methods to give vibrato extent is beyond the scope of this paper. FDM + BR has better vibrato extent accuracy than FDM + DT(F,A), 87.59 versus 84.05%. For both proposed methods, the vibrato rate accuracy values are better than the vibrato extent accuracy values. This suggests that the FDM-based methods are better at determining vibrato rates than vibrato extents. This may due to the fact that they consider only the sinusoid with the largest amplitude.

5. Conclusions

We have presented a novel frame-wise vibrato detection and estimation method that uses the Filter Diagonalisation Method (FDM). FDM is able to extract sinusoid frequency and amplitude information for a very short time signal, making it possible to determine vibrato frequency and pinpoint vibrato boundaries over a short time span. Natural byproducts of the FDM algorithm are the vibrato parameters themselves (rate and extent); thus, no additional computation is necessary to obtain the vibrato parameters.

We have also created a new monophonic dataset consisting of erhu and violin performances of an entire piece of music, *The Moon Reflected on the Second Spring*, for vibrato detection and vibrato parameter estimation. The long sequences allow for both training and test data to be excerpted not only from different performances by the same player, but from the same performance. The performances on Chinese versus Western instruments also allow for cross-cultural style comparisons.

The proposed FDM-based methods outperform existing state-of-the-art methods when evaluated on monophonic datasets comprising string, wind, brass, and voice excerpts. The FDM method with DT had a significantly higher F-measure value than the results of Herrera and Bonada (1998), Ventura, Sousa, and Ferreira (2012), or von Coler and Röbel (2011) for vibrato detection when tested on the von Coler–Röbel + CMMSD dataset; furthermore, the FDM-based methods had more balanced precision and recall values. For all methods tested, vibratos produced on string and woodwind instruments were more easily identified than those on brass instruments or voice.

We also evaluated the FDM-based methods against other competing methods using frame-level and note-level vibrato detection metrics. The FDM-based methods performed best in both cases, with the BR decision mechanism achieving better results – F-measure 0.84 (frame-level) and 0.41 (note-level) – than DT – 0.80 (frame-level) and 0.31 (note-level) – when tested on *The Moon Reflected on the Second Spring* dataset. BR has the advantage (over DTs) of greater flexibility and ability to adapt. Future work includes applying other machine learning methods to vibrato existence classification.

We further evaluated the vibrato parameter estimation capabilities of FDM using *The Moon Reflected on the Second Spring* dataset. The accuracy of vibrato rate estimation is above 92.5%, and that of the vibrato extent estimation is on the order of 85% for both decision methods with FDM.

Finally, FDM can be applied not only to vibrato detection and estimation but also to other music research domains requiring the extracting of sinusoids from a short time signal; for instance, it may be worth exploring ways to adapt FDM to f_0 extraction. We have not completely exploited the full capabilities of FDM outputs; the imaginary component of FDM's sinusoid frequency output could be used to improve vibrato onset detection.

Acknowledgements

The authors would like to thank Jian Yang and Laurel S. Pardue for their violin recordings, the anonymous reviewers, and co-Editors in Chief Clifton Callender and Thomas Fiore for their helpful suggestions and comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by China Scholarship Council [201206470008].

References

Audacity®; software available at http://audacityteam.org.

Boersma, Paul. 1993. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound." In *Proceedings of the Institute of Phonetic Sciences* 17: 97–110. http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf.

Cannam, Chris, Christian Landone, and Mark Sandler. 2010. "Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files." In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, 25–29 October 2010, Firenze, Italy, 1467–1468. New York: Association for Computing Machinery. http://portal.acm.org/citation.cfm?id = 1873951.1874248.

Childers, Donald G., David P. Skinner, and Robert C. Kemerait. 1977. "The Cepstrum: A Guide to Processing." *Proceedings of the Institute of Electrical and Electronics Engineers* 65 (10): 1428–1443.

von Coler, Henrik, and Alexander Lerch. 2014. "CMMSD: A Data Set for Note-Level Segmentation of Monophonic Music." In *Proceedings of the 53rd International Audio Engineering Society Conference: Semantic Audio*, 27–29 January 2014, London. http://www.aes.org/e-lib/browse.cfm?elib=17099.

von Coler, Henrik, and Axel Röbel. 2011. "Vibrato Detection Using Cross Correlation Between Temporal Energy and Fundamental Frequency." In *Proceedings of the Audio Engineering Society Convention 131*, 19 October 2011. http://www.aes.org/e-lib/browse.cfm?elib=16002.

de Cheveigné, Alain and Hideki Kawahara. 2002. "YIN, a Fundamental Frequency Estimator for Speech and Music." *The Journal of the Acoustical Society of America* 111 (4): 1917–1930.

Desain, Peter, and Henkjan Honing. 1996. "Modeling Continuous Aspects of Music Performance: Vibrato and Portamento." In *Proceedings of the 4th International Music Perception and Cognition Conference (ICMPC)*, 11–15 August 1996, Montréal, edited by B. Pennycook and E. Costa-Giomi. Montréal: McGill University. http://hdl.handle.net/2066/74795.

Fabian, Dorottya, Renee Timmers, Emery Schuberteds. 2014. *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures*. Oxford: Oxford University Press.

Herrera, Perfecto, and Jordi Bonada. 1998. "Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework." In *Proceedings of the 1st Digital Audio Effects Workshop (DAFx-98)*, 19–21 November 1998, Barcelona, Spain. http://mtg.upf.edu/system/files/publications/Herrera-DAFX-1998.pdf.

Hu, Haitao, Que N. Van, Vladimir A. Mandelshtam, and A. J. Shaka. 1998. "Reference Deconvolution, Phase Correction, and Line Listing of NMR Spectra by the 1D Filter Diagonalization Method." *Journal of Magnetic Resonance* 134 (1): 76–87.

Huang, Jiangqin. 2006. "The Moon Reflected on the Second Spring, on The Ditty of the South of the Jiangsu." CD, Guangzhou, PR China. ISBN: 9787885180706.

Keiler, Florian, and Sylvain Marchand. 2002. "Survey on Extraction of Sinusoids in Stationary Sounds." In *Proceedings of the 5th Digital Audio Effects Conference (DAFx-02)*, 26–28 September 2002, Hamburg, Germany, 51–58. http://ant-s4.unibw-hamburg.de/dafx/paper-archive/2002/DAFX02_Keiler_Marchand_sine_extract_compare.pdf.

Liebman, Elad, Eitan Ornoy, and Benny Chor. 2012. "A Phylogenetic Approach to Music Performance Analysis." *Journal of New Music Research* 41 (2): 95–222.

Mandelshtam, Vladimir A. 2001. "FDM: The Filter Diagonalization Method for Data Processing in NMR Experiments." *Progress in Nuclear Magnetic Resonance Spectroscopy* 38 (2): 159–196.

Mandelshtam, Vladimir A., and Howard S. Taylor. 1997. "Harmonic Inversion of Time Signals and Its Applications." *The Journal of Chemical Physics* 107 (17): 6756–6769.

Martini, Beau R., Vladimir A. Mandelshtam, Gareth A. Morris, Adam A. Colbourne, and Mathias Nilsson. 2013. "Filter Diagonalization Method for Processing PFG NMR Data." *Journal of Magnetic Resonance* 234: 125–134.

Mauch, Matthias, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. 2015. "Computer-Aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency." In *Proceedings of the 1st International Conference on Technologies for Music Notation and Representation*, 28–30 May 2015, Paris, France, 23–30. http://matthiasmauch.de/_pdf/mauch2015computeraided.pdf.

Mauch, Matthias, and Simon Dixon. 2014. "PYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 4–9 May 2014, Florence, Italy, 659–663. http://matthiasmauch.de/_pdf/mauch_pyin_2014.pdf.

Mitchell, Helen F., and Dianna T. Kenny. 2010. "Change in Vibrato Rate and Extent During Tertiary Training in Classical Singing Students." *Journal of Voice* 24 (4): 427–434.

Molina, Emilio, Ana M. Barbancho, Lorenzo J. Tardón, Isabel Barbancho. 2014. "Evaluation Framework for Automatic Singing Transcription." In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 567–572. http://www.terasoft.com.tw/conf/ismir2014/proceedings/T102_298_Paper.pdf.

Neuhauser, Daniel. 1990. "Bound State Eigenfunctions from Wave Packets: Time Energy Resolution." *The Journal of Chemical Physics* 93 (4): 2611–2616.

Nwe, Tin Lay, and Haizhou Li. 2007. "Exploring Vibrato-Motivated Acoustic Features for Singer Identification." *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2): 519–530.

Özaslan, Tan Hakan, and Josep Lluis Arcos. 2011. “Automatic Vibrato Detection in Classical Guitar Recordings.” Technical Report IIIA 2011-05, October 2011. Universitat Autònoma de Barcelona, Spain. https://pdfs.semanticscholar.org/ff11/5794e5365d0cd5ce7d9612c25b8ae3fc6147.pdf.

Özaslan, Tan Hakan, Xavier Serra, and Josep Lluis Arcos. 2012. “Characterization of Embellishments in Ney Performances of Makam Music in Turkey.” In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 8–12 October 2012, Porto, Portugal. https://repositori.upf.edu/bitstream/handle/10230/22727/Ozaslan-Tan-ISMIR-2012.pdf?sequence = 1.

Palmer, Caroline, and Sean Hutchins. 2006. “What Is Musical Prosody?” In *Psychology of Learning and Motivation*, Vol. 46, 245–278. Amsterdam: Elsevier. http://dx.doi.org/10.1016/S0079-7421(06)46007-2.

Pang, Hee-Suk, and Doe-Hyun Yoon. 2005. “Automatic Detection of Vibrato in Monophonic Music.” *Pattern Recognition* 38 (7): 1135–1138.

Paulraj, A., Richard Roy, and Thomas Kailath. 1985. “Estimation of Signal Parameters via Rotational Invariance Techniques–ESPRIT.” In *19th Asilomar Conference on Circuits, Systems and Computers*, November 1985, Monterey, CA., 83–89.

Prame, Eric. 1994. “Measurements of the Vibrato Rate of Ten Singers.” *The Journal of the Acoustical Society of America* 96 (4): 1979–1984.

Regnier, Lise, and Geoffroy Peeters. 2009. “Singing Voice Detection in Music Tracks Using Direct Voice Vibrato Detection.” In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, 19–24 April 2009, Taipei, Taiwan, 1685–1688. http://dx.doi.org/10.1109/ICASSP.2009.4959926.

Rossignol, Stephane, Xavier Rodet, Joël Soumagne, Jean-Luc Collette, and Philippe Depalle. 1999. “Automatic Characterisation of Musical Signals: Feature Extraction and Temporal Segmentation.” *Journal of New Music Research* 28 (4): 281–915.

Schmidt, R.O. 1986. “Multiple emitter location and signal parameter estimation.” *IEEE Transactions on Antennas and Propagation* 34 (3): 276–280.

Sousa, Ricardo, and Anibal Ferreira. 2010. “Non-Iterative Frequency Estimation in the DFT Magnitude Domain.” In *Proceedings of the 4th IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP 2010)*, 3–5 March 2010, Limassol, Cyprus, 1–4. http://dx.doi.org/10.1109/ISCCSP.2010.5463483.

Ventura, Jose, Ricardo Sousa, and Anibal Ferreira. 2012. “Accurate Analysis and Visual Feedback of Vibrato in Singing.” In *Proceedings of the 5th IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP 2012)*, 1–6. http://dx.doi.org/10.1109/ISCCSP.2012.6217808.

Wall, Michael R., and Daniel Neuhauser. 1995. “Extraction, through Filter-Diagonalization, of General Quantum Eigenvalues or Classical Normal Mode Frequencies from a Small Number of Residues or a Short-Time Segment of a Signal. I. Theory and Application to a Quantum-Dynamics Model.” *The Journal of Chemical Physics* 102 (20): 8011–8022. http://dx.doi.org/10.1063/1.468999.

Wang, Guotong. 2009. “Track 4, Disk 2, An Anthology of Chinese Traditional and Folk Music – A Collection of Music Played on the Erhu.” CD, Shanghai, PR China. ISBN: 9787799919928.

Weninger, Felix, Noam Amir, Ofer Amir, Irit Ronen, Florian Eyben, and Böjrn Schuller. 2012. “Robust Feature Extraction for Automatic Recognition of Vibrato Singing in Recorded Polyphonic Music.” In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, 25–30 March 2012, Kyoto, Japan, 85–88. http://dx.doi.org/10.1109/ICASSP.2012.6287823.

Yang, Luwei, Elaine Chew, and Khalid Z. Rajab. 2013. “Vibrato Performance Style: A Case Study Comparing Erhu and Violin.” In *Proceedings of the 10th International Conference on Computer Music Multidisciplinary Research*, 15–18 October 2013, Marseille, France, 904–919.

Yang, Luwei, Khalid Z. Rajab, and Elaine Chew. 2016. “AVA: An Interactive System for Visual and Quantitative Analyses of Vibrato and Portamento Performance Styles.” In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 7–11 August, New York City. http://m.mr-pc.org/ismir16/website/articles/314_Paper.pdf.

Yang, Luwei, Mi Tian, and Elaine Chew. 2015. “Vibrato Characteristics and Frequency Histogram Envelopes in Beijing Opera Singing.” In *Proceedings of the 5th International Workshop on Folk Music Analysis*, 10–12 June 2015, Paris, France, 139–140. https://qmro.qmul.ac.uk/xmlui/handle/123456789/16062.