ON VIBRATO AND FREQUENCY (DE)MODULATION IN MUSICAL SOUNDS

Jeremy Hyrkas and Tamara Smyth

Music Department
University of California San Diego
La Jolla, CA, USA
jhyrkas@ucsd.edu, trsmyth@ucsd.edu

ABSTRACT

Vibrato is an important characteristic in human musical performance and is often uniquely characteristic to a player and/or a particular instrument. This work is motivated by the assumption (often made in the source separation literature) that vibrato aids in the identification of multiple sound sources playing in unison. It follows that its removal, the focus herein, may contribute to a more blended combination. In signals, vibrato is often modeled as an oscillatory deviation from a center pitch/frequency that presents in the sound as phase/frequency modulation. While vibrato implementation using a time-varying delay line is well known, using a delay line for its removal is less so. In this work we focus on (de)modulation of vibrato in a signal by first showing the relationship between modulation and corresponding demodulation delay functions and then suggest a solution for increased vibrato removal in the latter by ensuring sideband attenuation below the threshold of audibility. Two known methods for estimating the instantaneous frequency/phase are used to construct delay functions from both contrived and musical examples so that vibrato removal may be evaluated.

1. INTRODUCTION

Musical performances often contain moments of musicians playing in unison or harmony to form complex timbres that may be perceived as a fused sonic source, as opposed to multiple instruments. Players may adjust their volume, pitch, vibrato or other performance techniques to more closely match those of the other performers. There are several analogous practices in electronic music where audio samples are overlaid in time to approximate the timbre of a target sound; examples of such practices include, but are not limited to: target-based concatenative synthesis [1], audio mosaicing [2], and automated orchestration [3]. Unlike in live performance, samples from an audio database exist without musical context and may need to be edited or processed to blend together more naturally into a more fused sound. One common approach is sound morphing, where multiple signals are analyzed and resynthesized into a single signal that contains some timbral qualities of each individual signal [4]. However, the fidelity and potential richness of multiple overlaid sounds may be lost due to the reduction of signals or artifacts of the analysis or resynthesis. We are therefore interested in methods for processing multiple signals in the interest of creating the perception of a more uniform tone while maintaining the presence of two or more separate sources.

Our methods are informed by the related inverse problem of audio source separation. Source separation algorithms take one signal comprised of several sources and attempt to separate the sources into individual signals. Our interest is in understanding how these algorithms perform the separation so that we can do the opposite: take separate signals and modify them so that they are less separable. Of particular interest is the source separation of unison signals, where separation is more difficult due to the many shared frequencies between the partials of each source. Unison source separation algorithms [5, 6] focus on the Common Fate principle of psychoacoustics, where sound partials from the same source have related perceived movement. These algorithms analyze the amplitude and frequency modulation patterns of the signal to determine vibrato patterns for each source that can be used to isolate their partials.

It is believed that humans also use vibrato patterns to distinguish instruments playing in unison. Humans were found to be remarkably good at counting the number of instruments in a mix, even when playing in unison, up to a limit of three to four instruments [7]. These results hold for controlled laboratory settings and more general online surveys [8]. Based on these results, the psychoacoustic theory (Gestalt grouping principle) of Common Fate and the success of unison source separation algorithms in isolating signals based on vibrato patterns, it stands to reason that two signals (particularly when playing in unison) could be made to sound more uniform and fused if their vibrato patterns were matched exactly.

In order to match the vibrato pattern of signals to blend them, we must characterize and analyze the signal properties that represent vibrato in a sound. Vibrato is typically associated with slow frequency modulation, where a periodic signal oscillates above and below a fundamental frequency with perceptible regularity. Classic analog and digital implementations of vibrato use a delay-line with time-varying delay to introduce a desired fundamental frequency deviation to an input signal. It is less common and, as will be shown, less straightforward, to use the same method for removing existing vibrato from a signal.

In acoustic instruments, vibrato techniques often impart both amplitude and frequency modulation to the sounding note. There are a handful of existing methods for analyzing amplitude and frequency deviation in signals with vibrato, including at the per-harmonic level [9, 10]. This work focuses on the frequency modulation pattern caused by vibrato and how it can be modeled through signal analysis. If the modulation patterns caused by vibrato can be accurately modeled, the patterns can be removed and/or transferred to another signal using a delay line.

Vibrato patterns cannot be directly transferred to a signal with its own vibrato. Therefore, methods for demodulating a signal are also required. Frequency and amplitude demodulation are a well

researched topic in communications [11], but there are challenges in directly applying their methods in a musical context. In communication systems, the modulating wave contains all relevant signal information and the carrier is a simple, high-frequency sinusoid; demodulation recovers the modulating signal and the carrier wave is discarded. When modeling vibrato as frequency modulation, the modulating wave contains the vibrato patterns of interest while the carrier wave is itself time-varying and contains the timbre and fundamental frequency of the sound. It is therefore necessary to develop demodulation algorithms appropriate for a musical context.

This work proposes methods for analyzing, modeling and demodulating signals with vibrato. Section 2 reviews the vibrato effect as it is implemented in discrete time using a delay line with a delay function (related to the time warping function [5]). A first demodulation delay function is constructed by its inversion and then a method is proposed to further reduce any modulation sidebands that may remain from the vibrato. The theory is demonstrated in Section 3 by using two separate methods for estimating instantaneous frequency and/or phase from musical signals and using the resulting estimates to form (de)modulation delay functions. Finally, practical considerations that arise when deriving and using delay functions are presented, along with possible solutions.

2. VIBRATO AND FREQUENCY (DE)MODULATION

Vibrato is a human performance characteristic whereby the player deviates from (by moving slightly above and below) a center pitch or sounding frequency in a regular oscillatory way. The nature of the vibrato is often distinct among different players (as well as the instruments themselves) and while it is often simulated in computer music as being sinusoidal, the reality can offer much greater diversity (e.g. the swing above and below the carrier frequency is not necessarily equal). For this reason, analysis/estimation of the vibrato signal from a sound is necessary before demodulation.

Because vibrato is a slow wavering of frequency (though not necessarily sinusoidal) it is often modeled as a frequency modulated (FM) carrier oscillator having instantaneous frequency given by

$$\omega_i(t) = \omega_c - d\cos(\omega_m t) \quad \text{rad/s}, \tag{1}$$

where d is the oscillator's peak frequency deviation (vibrato depth) from carrier/center frequency ω_c and ω_m is the frequency of modulation (vibrate rate). For numeric reasons, it is often preferable to implement FM as phase modulation (PM) with the corresponding instantaneous phase being given by the integral of (1) with respect to time:

$$\theta_i(t) = \int_0^t \omega_i(t)dt = \omega_c t - I\sin(\omega_m t) + \phi_c, \tag{2}$$

where the amplitude of the time-varying sinusoidal term

$$I = \frac{d}{\omega_m}. \tag{3}$$

is known in the FM synthesis literature as the *index of modulation* because of its influence on the sidebands that are produced at frequencies $\omega_c \pm k\omega_m$ in the resulting spectrum. If ω_m is at audio rates, changing I can result in a significant perceptual change in both tone quality and sounding frequency. If ω_m is at lower rates of oscillation, the time variation of frequency/pitch is more easily tracked by the listener and the vibrato effect results, with I influencing its depth.

If synthesizing a tone from sinusoids, the modulation may be implemented simply by applying the phase (2) to a carrier oscillator (e.g. $\cos(\theta_i)$). If, however, the vibrato is applied to an existing signal, other methods such as a time-varying delay line (or resampling according to a warping function) or sinusoidal modeling may be used.

2.1. Modulation with a Time-Varying Digital Delay Line

Sampling at rate f_s Hz involves, in part, effectively replacing the continuous variable t with integer multiples of the sampling period $T = 1/f_s$,

$$t \longrightarrow nT, \tag{4}$$

where $n = 0, 1, 2 \ldots N$ is the sample index for a signal of length of N samples.

If the discrete-time sinusoid with angular frequency ω_c rad/s

$$z(n) = e^{j\omega_c nT} \tag{5}$$

is the input to a time-varying delay line with delay function

$$D_m(n) = \frac{I}{\omega_c}\sin(\omega_m nT)f_s + \frac{I}{\omega_c}fs, \tag{6}$$

with DC offset in (6) ensuring a positive delay [12], then the output may be represented by the difference equation

$$z_m(n) = z(n - D_m(n)) = e^{j\omega_c(n - D_m(n))T}, \tag{7}$$

effectively substituting instances of n with warping function $n - D_m(n)$ so that (7) is a frequency-modulated sinusoid with instantaneous phase given by

$$\theta_m(n) = \omega_c nT - \omega_c D_m(n)T = \omega_c nT - I\sin(\omega_m nT) - I, \tag{8}$$

the discrete-time approximation to (2) (with initial phase $\phi_c = -I$). Notice that while demodulation (removing the modulation caused by $D_m(n)$) could be accomplished for this single sinusoid via the complex multiply

$$e^{j\omega_c D_m(n)T} z_m(n) = e^{j(\omega_c nT - \omega_c D_m(n)T + \omega_c D_m(n)T} = z(n), \tag{9}$$

the same would not hold for a more typical signal having multiple (K) harmonically-related sinusoidal components

$$z_k(n) = \sum_{k=1}^{K} e^{jk\omega_c nT} \tag{10}$$

similarly modulated by the delay function (6). That is

$$e^{j\omega_c D_m(n)T} \sum_{k=1}^{K} e^{j(k\omega_c nT - k\omega_c D_m(n)T)} \neq z_k(n) \tag{11}$$

because modulating $z_k(n)$ produces a factor-of-k increase in the index of modulation for the k^{th} harmonic resulting in a greater number of sidebands (higher-order Bessel functions having greater value) for higher harmonics (see Figure 1).

It is for this reason that much of the research in vibrato removal/modification treats harmonics separately. While this has been shown to be effective using a sinusoidal modeling approach [13], it may also be desirable for some applications to have a simpler solution—one that treats the signal and its partials as a whole in the time-domain (save any necessary vibrato analysis) so that removal of the vibrato has a computational simplicity akin to its application in (7).

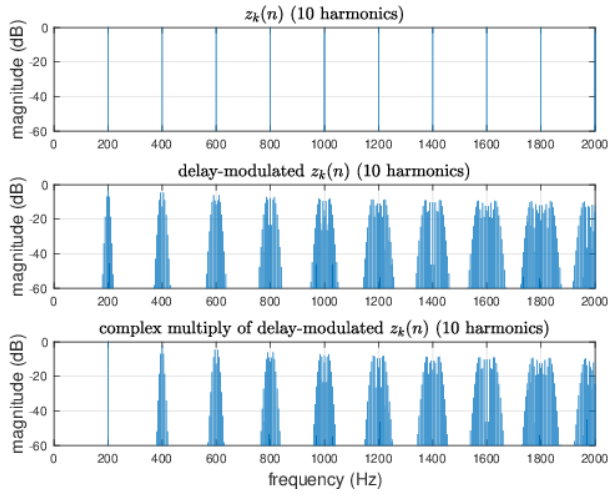

Figure 1: *The spectrum of signal* $z_k(n)$ (10) *having K=10 harmonics (top) is modulated by delay function* (6) *(middle) showing increased frequency spread at higher harmonics. While the complex multiply in* (11) *(bottom) may demodulate the vibrato (remove the sidebands) surrounding the first harmonic, it does not do so sufficiently for higher harmonics.*

2.2. Demodulation with a Time-Varying Digital Delay Line

Consider the modulated sinusoid $z_m(n)$ (7) as the input to a delay line with delay function $D_d(n)$ constructed by inverting (6), that is

$$D_d(n) = -\frac{I}{\omega_c}\sin(\omega_m nT)f_s + \frac{I}{\omega_c}fs \tag{12}$$

(again with an offset to ensure the delay-line delay is positive). The output of the delay line is given by

$$z_d(n) = z_m(n - D_d(n)) = e^{j\theta_m(n-D_d(n))} \tag{13}$$

and has instantaneous phase given by substituting instances of n in (8) with $n - D_d(n)$ to yield the difference equation

$$\begin{aligned}\theta_d(n) &= \omega_c(n - D_d(n))T - \omega_c D_m(n - D_d(n))T \\ &= \omega_c nT + I\sin(\omega_m nT) - 2I - I\phi(n),\end{aligned} \tag{14}$$

where the final expression has the FM sinusoidal term

$$\phi(n) = \sin(\omega_m nT + I_1\sin(\omega_m nT) - I_1), \tag{15}$$

for $I_1 = I\omega_m/\omega_c$. Because $I\phi(n) \neq I\sin(\omega_m nT)$, these two terms do not cancel in (14) and $z_d(n) \neq z(n)$, suggesting the delay function (12) does not *completely* demodulate the vibrato. It is worthwhile, however, to examine the extent to which $I\phi(n) \approx I\sin(\omega_m nT)$ and whether the vibrato is at least reduced (or audible) in $z_d(n)$ as this will motivate a subsequent strategy for further reduction.

A Fourier series expansion of (15) shows it expressed as a weighted sum of harmonically-related sinusoids

$$\phi(n) = \sum_{k=-\infty}^{\infty}\phi_k(n) = \sum_{k=-\infty}^{\infty}J_k(I_1)\sin((1+k)\omega_m nT - I_1), \tag{16}$$

where $J_k(I_1)$ is the k^{th}-order Bessel function of the first kind indexed by I_1. Assuming for vibrato that $\omega_m << \omega_c$ and I is a low integer value (with higher values more typical of synthesis applications), then I_1 will be small and very few harmonics (sidebands) of ω_m will have significant amplitude greater than 0.001 (-60 dB). For instance, the sideband of $\phi(n)$ at $k = -1$ given by

$$\phi_{-1}(n) = J_{-1}(I_1)\sin(-I_1) = J_1(I_1)\sin(I_1) \approx 0 \tag{17}$$

yields a DC component of negligible amplitude (less than 0.001) for small I_1. If the same approximation in (17) is made, then the sideband at $k = 1$ may be approximated by

$$\begin{aligned}\phi_1(n) &= J_1(I_1)(\cos(I_1)\sin(2\omega_m nT) - \sin(I_1)\cos(2\omega_m nT) \\ &\approx J_1(I_1)\cos(I_1)\sin(2\omega_m nT)\end{aligned} \tag{18}$$

and the sideband at $k = 0$ by

$$\begin{aligned}\phi_0(n) &= J_0(I_1)(\cos(I_1)\sin(\omega_m nT) - \sin(I_1)\cos(\omega_m nT)) \\ &\approx \sin(\omega_m nT) - J_0(I_1)\sin(I_1)\cos(\omega_m nT),\end{aligned} \tag{19}$$

where the unit-amplitude sinusoid arises since $J_0(I_1)\cos(I_1) \approx 1$ for small I_1. Adding (18) and (19) gives an approximation to (15) as the sum

$$\phi(n) \approx \sin(\omega_m nT) - \frac{I_2}{I}\cos(\omega_m nT) + \frac{I_3}{I}\sin(2\omega_m nT), \tag{20}$$

where

$$I_2 = IJ_0(I_1)\sin(I_1) \quad \text{and} \quad I_3 = IJ_1(I_1)\cos(I_1). \tag{21}$$

Substituting (20) into (14), yields

$$\theta_d(n) \approx \omega_c nT - 2I + I_2\cos(\omega_m nT) - I_3\sin(2\omega_m nT) \tag{22}$$

and a cancellation of the original modulating sinusoid in (14). Substituting (22) into (13) yields

$$z_d(n) \approx e^{j(\omega_c nT - 2I)}e^{jI_2\cos(\omega_m nT)}e^{-jI_3\sin(2\omega_m nT)} = \hat{z}_d(n), \tag{23}$$

showing that $\hat{z}_d(n)$ is an FM sinusoid with two very-low amplitude modulating sinusoids. For small I_2 and I_3 (having values much less than one), approximations may be made by considering only sidebands $k = -1, 0, 1$ to yield

$$\begin{aligned}e^{jI_2\cos(\omega_m nT)} &= \sum_{k=-\infty}^{\infty}J_k(I_2)e^{jk(\omega_m nT + \pi/2)} \\ &\approx 1 + J_1(I_2)j\left(e^{j\omega_m nT} + e^{-j\omega_m nT}\right), \\ e^{-jI_3\sin(2\omega_m nT)} &= \sum_{k=-\infty}^{\infty}J_k(I_3)e^{-jk2\omega_m nT} \\ &\approx 1 - J_1(I_3)\left(e^{j2\omega_m nT} - e^{-j2\omega_m nT}\right),\end{aligned}$$

so that their product in (23) yields the final approximation of the delay-line output given by

$$\begin{aligned}\hat{z}_d(n) \approx\ & e^{j(\omega_c nT - 2I)} \\ & J_1(I_2)je^{-j2I}\left(e^{j(\omega_c+\omega_m)nT} + e^{j(\omega_c-\omega_m)nT}\right) - \\ & J_1(I_3)e^{-j2I}\left(e^{j(\omega_c+2\omega_m)nT} - e^{j(\omega_c-2\omega_m)nT}\right),\end{aligned} \tag{24}$$

showing a unit-amplitude sinusoidal component at carrier frequency ω_c along with two upper and lower sidebands at frequencies $\omega_c \pm \omega_m$ and $\omega_c \pm 2\omega_m$ having amplitudes $J_1(I_2)$ and $J_1(I_3)$, respectively (see Figure 2). (Note that (24) omits a final combination term that contributes a negligible amplitude of $J_1(I_2)J_1(I_3)$ to the first and third upper and lower sidebands).

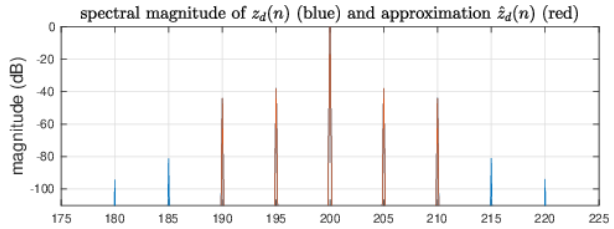

Figure 2: *Spectral magnitude of the delay line output* $z_d(n)$ (13) *and its approximation* $\hat{z}_d(n)$ (24) *for FM parameters* $I = 1$, $\omega_c = 2\pi 200$ *and* $\omega_m = 2\pi 5$. *Strong agreement is visible for 2 sidebands above and below the carrier frequency greater than -60dB: for* $k = \pm 1$ *the amplitude is* $J_1(I_2) = 0.0125 \approx -38$ *dB and for* $k = \pm 2$ *the amplitude is* $J_2(I_3) = 0.0062 \approx -44$ *dB.*

2.3. Further Vibrato Reduction

While the sidebands in (24) are small and associated vibrato barely audible, since they are within the audio range (their amplitude is greater than 0.001 or -60 dB) it may be desirable to further reduce their amplitude to ensure inaudibility.

Consider next a demodulating delay function in which (12) is delayed by itself, that is

$$\begin{aligned} D_{dd}(n) &= D_d(n - D_d(n)) \\ &= -\frac{I}{\omega_c}\sin(\omega_m(n - D_d(n)T)f_s + \frac{I}{\omega_c}f_s \\ &= -\frac{I}{\omega_c}\phi(n)f_s + \frac{I}{\omega_c}f_s, \end{aligned} \tag{25}$$

so the delay function is an FM sinusoid and the output of the delay line is

$$z_{dd}(n) = z_m(n - D_{dd}(n)) = e^{j\theta_m(n - D_{dd}(n))}, \tag{26}$$

having instantaneous phase given by substituting n with $n - D_{dd}(n)$ in (8) to yield

$$\begin{aligned} \theta_{dd}(n) &= \omega_c nT - \omega_c D_{dd}(n)T - \\ &\quad I\sin(\omega_m nT - \omega_m D_{dd}(n)T) - I \\ &= \omega_c nT - 2I + I\phi(n) - I\sin(\omega_m nT + I_1\phi(n) - I_1). \end{aligned} \tag{27}$$

Expressing the last term in the final expression of (27) as the imaginary part of an analytic signal,

$$I\sin(\omega_m nT + I_1\phi(n) - I_1) = I\Im\left\{e^{j(\omega_m nT - I_1)}e^{jI_1\phi(n)}\right\}, \tag{28}$$

and substituting the approximation for $\phi(n)$ in (20), yields intermediate factor

$$e^{jI_1\phi(n)} \approx e^{jI_1\sin(\omega_m nT)}e^{-jA_1\cos(\omega_m nT)}e^{jA_2\sin(2\omega_m nT)}, \tag{29}$$

showing again an FM sinusoid with very low-amplitude modulators. Since $A_1 = I_1 I_2/I$ and $A_2 = I_1 I_3/I$ are small, the approximations

$$\begin{aligned} e^{-jA_1\cos(\omega_m nT)} &= j\sum_{k=-\infty}^{\infty} J_k(A_1)e^{jk\omega_m nT} \approx J_0(A_1) \approx 1, \\ e^{jA_2\sin(\omega_m nT)} &= \sum_{k=-\infty}^{\infty} J_k(A_2)e^{jk2\omega_m nT} \approx J_0(A_2) \approx 1, \end{aligned} \tag{30}$$

may be made because the zeroth sideband ($k = 0$) is the only term in the infinite sum (30) having non-negligible amplitude (and it's amplitude is almost exactly equal to one). Substituting (30) into (29) yields the approximation

$$e^{jI_1\phi(n)} \approx e^{jI_1\sin(\omega_m nT)} \tag{31}$$

and substituting (31) into the analytic signal (28) in turn yields

$$\begin{aligned} \sin(\omega_m nT + I_1\phi(n) - I_1) &\approx \Im\left\{e^{j(\omega_m nT + I_1\sin(\omega_m nT) - I_1)}\right\} \\ &\approx \phi(n). \end{aligned} \tag{32}$$

Finally, when substituting the approximation of (32) into (27), there is a cancellation of the sinusoidal terms so that

$$\theta_{dd}(n) \approx \omega_c nT - 2I, \tag{33}$$

showing that the resulting delayed signal

$$z_{dd}(n) \approx z(n) \tag{34}$$

is nearly equal to the original unmodulated sinusoid (omitting a pure delay) with comparatively improved vibrato reduction and attenuation of sidebands that are well below the threshold of audibility (see Figure 3).

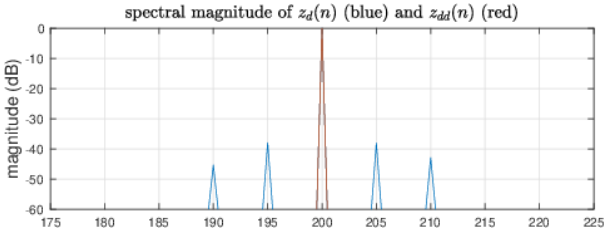

Figure 3: *Spectral magnitude of the delay line output* $z_{dd}(n)$ *(red) showing no sidebands greater than -60dB and thus greater vibrato reduction when compared to* $z_d(n)$ *(blue).*

3. OBTAINING THE DELAY FUNCTION

As with the phase-frequency relationship in (2), the instantaneous frequency of $z_m(n)$ modulated by delay function $D_m(n)$ is the derivative of the phase (8) with respect to time (or nT for the discrete-time signal here) and is given by

$$\begin{aligned} \omega_i(n) = \frac{d}{dnT}\theta_m(n) &= \omega_c - I\omega_m\cos(\omega_m nT)) \\ &= \omega_c\left(1 - \dot{D}_m(n)\right) \end{aligned} \tag{35}$$

where the derivative of the delay function with respect to n

$$\dot{D}_m(n) = \frac{d}{dn}D_m(n) = I\omega_m/\omega_c\cos(\omega_m nT) \tag{36}$$

is the relative frequency shift imparted by the delay function $D_m(n)$. From (35), it may be easily seen that the transposition factor (or momentary transposition [12]), i.e. that value which, when multiplying an input frequency ω_c, results in a transposition of sounding frequency $\omega_i(n)$, is given by

$$\frac{\omega_i(n)}{\omega_c} = 1 - \dot{D}_m(n) \tag{37}$$

and thus the relative frequency shift may be expressed as

$$\dot{D}_m(n) = 1 - \frac{\omega_i(n)}{\omega_c}, \tag{38}$$

showing (38) may be estimated given frequency $\omega_i(n)$ and carrier frequency ω_c. It follows from (36) that the delay function is obtained by integrating the relative frequency shift, an operation that may be approximated by the discrete-time cumulative sum of (38)

$$\hat{D}_m(n) \approx \sum_{l=0}^{n} \dot{D}_m(l) \approx n - \frac{\theta_m(n)}{\omega_c T}. \tag{39}$$

The (intermediate) demodulating function is then estimated as

$$\hat{D}_d(n) = \max\{\hat{D}_m(n)\} - \hat{D}_m(n), \tag{40}$$

which is then delayed according to (25) to obtain $\hat{D}_{dd}(n)$.

Obtaining either delay function is, then, a matter of first estimating either the instantaneous phase $\theta_m(n)$ or frequency $\omega_i(n)$. In the following section, two commonly-used such techniques, one that estimates frequency and one that estimates phase, are briefly described and evaluated for their accuracy in examples estimating vibrato delay functions for 1) a contrived broadband signal (sawtooth wave) with known vibrato, 2) a musical signal (clarinet) with known vibrato (prior modulation with $D_m(n)$) and finally 3) for a musical sound with unknown vibrato.

3.1. Estimating Instantaneous Frequency with Peak Picking

In this section two methods often used for obtaining ω_i are explored and evaluated. The first method uses the Short-Time Fourier Transform (STFT) followed by momentary peak tracking to obtain ω_i and the second uses a bandpass filter followed by the Hilbert Transform to obtain the analytic (complex) signal with angle θ_i (from which ω_i may be derived). Examples of a guitar and vocal sound are presented using these methods (see Figures 9 and 10).

In this method, the modulated (real) length-N signal $y(n)$ is analysed using the short-time Fourier Transform to obtain the $N \times N_f$ matrix of N_f spectral magnitude frames

$$\mathbf{Y} = \begin{bmatrix} |Y_0(\omega_0)| & |Y_1(\omega_0)| & |Y_2(\omega_0)| & \dots & |Y_{N_f-1}(\omega_0)| \\ |Y_0(\omega_1)| & |Y_1(\omega_1)| & |Y_2(\omega_0)| & \dots & |Y_{N_f-1}(\omega_1)| \\ |Y_0(\omega_2)| & |Y_1(\omega_2)| & |Y_2(\omega_0)| & \dots & |Y_{N_f-1}(\omega_2)| \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ |Y_0(\omega_N)| & |Y_1(\omega_N)| & |Y_2(\omega_0)| & \dots & |Y_{N_f-1}(\omega_N)| \end{bmatrix} \tag{41}$$

where the m^{th} column, $m = 0, 1, 2, \dots, N_f - 1$, is the magnitude of the DFT

$$|Y_m(\omega_k)| = \left| \sum_{n=0}^{N-1} x(n) w(n - mN_h) e^{-j2\pi kn/N} \right| \tag{42}$$

at time starting from sample mN_h for hopsize N_h and window $w(n)$. Because both time and frequency resolution are needed for estimating the vibrato signal, hopsize N_h is made small and, for the smallest hopsize of $N_h = 1$, a sliding DFT [14] may be used.

From each momentary spectral magnitude $|Y_m(\omega_k)|$ frame, the "fundamental" frequency, defined here as the lowest frequency with a significant peak amplitude, is tracked from frame to frame to obtain frequency vector $f_p(0), f_p(1), \dots, f_p(N_f - 1)$. To find the true peak, a quadratic interpolation [15] [1] is used (see Figure 4). Vector $f_p(\cdot)$ is then resampled at a rate equal to the hop size N_h to yield a length-N vector $f_i^p(n)$ such that

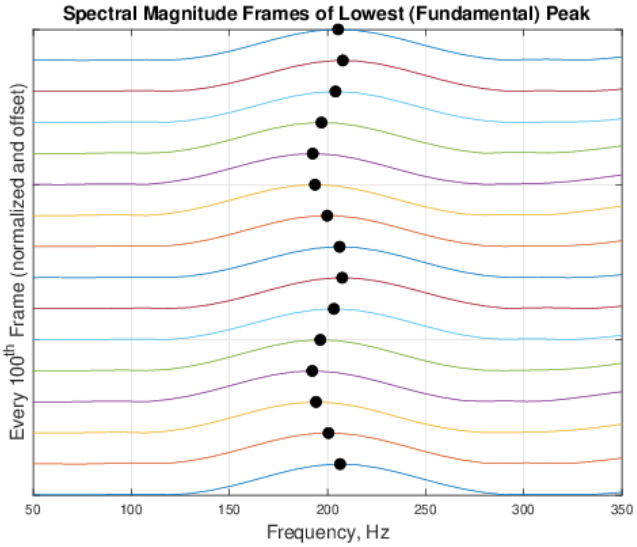

Figure 4: *First "fundamental" spectral magnitude peak of a sawtooth wave with a sinusoidal vibrato applied. Every 100th frame is offset to show time evolution (from bottom to top) of the peak position as it follows a sinusoidal vibrato track.*

$$f_i^p(mN_h) = f_p(m), \tag{43}$$

along with interpolated frequency values at intermediate indeces. The estimated instantaneous frequency $\omega_i(n)$ is then given by multiplying (43) by 2π. The p superscript indicates frequencies obtained using the 'peak picking' method, examples of which are plotted (along with the bandpass filter method described subsequently) for modulated sawtooth and clarinet signals in Figure 5.

3.2. Estimating Instantaneous Phase with a Bandpass Filter

Whether from $f_i^p(n)$ in (46) as was done here, or some external pitch detection tool, a narrow-band bandpass filter $H_b(\omega)$ may be designed to have bandwidth

$$B_w = (\max\{f_i^p(n)\} + f_e) - (\min\{f_i^p(n)\} - f_e), \tag{44}$$

and center frequency given by the mean of $f_i^p(n)$

$$f_c = \frac{1}{N} \sum_{n=0}^{N-1} f_i^p(n) \approx \frac{\max\{f_i^p(n)\} + \min\{f_i^p(n)\}}{2}. \tag{45}$$

The value f_e in (44) extends the lower and upper band edges to ensure all modulation sidebands are able to pass with minimal error while not overlapping with the second harmonic (and it's possible sidebands) or introducing any other unwanted signal noise. It should be noted that in real instrument sounds f_c may not be constant throughout the note but rather time varying with a (non-oscillating) upward or downward trend that is distinct from the vibrato (see clarinet Figure 5, bottom).

[1] See "Sinusoidal Peak Interpolation" in the chapter *Spectrum Analysis of Sinusoids*.

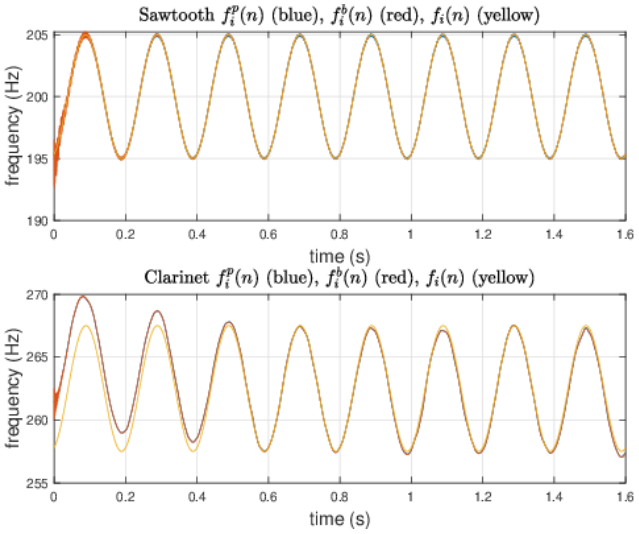

Figure 5: *Estimated instantaneous frequency for sawtooth (top) and clarinet (bottom) using peak picking and bandpass filter methods. Each is modulated with parameters $f_m = 5$ and $I = 1$ and compared to known frequency modulator $f_i(n) = I f_m \cos(\omega_m n T)$. For the sawtooth $f_c = 200$, but for the clarinet, f_c is time varying (with a slight downward trend).*

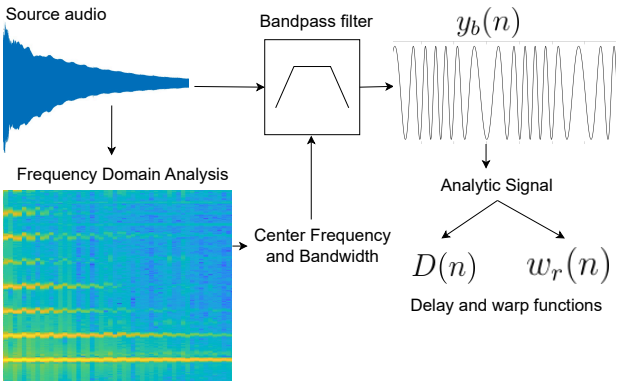

Figure 6: *A signal is processed using a bandpass filter to isolate the signal of one partial. Delay and warp functions are derived from the analytic signal.*

If the modulated signal is passed through bandpass filter H_b, it is assumed the output is nearly sinusoidal, having only the first "fundamental" harmonic along with any sidebands from the vibrato's low-frequency modulation. For this reason, the angle of the complex analytic signal of the filter output yields the instantaneous phase $\theta_m^b(n)$ (with the b superscript indicating the bandpass method) and instantaneous frequency by its derivative, approximated by the finite difference

$$\omega_i^b(n) = \frac{\theta_m^b(n+1) - \theta_m^b(n)}{T}, \quad \text{and} \quad f_i^b(n) = \frac{\omega_i^b(n)}{2\pi}. \quad (46)$$

Experiments with various bandpass filters for signal examples (Figure 5) showed using a sixth-order Butterworth[2]. resulted in $f_i^b(n)$ having the best agreement with $f_i^p(n)$.

Once the instantaneous frequency $\omega_i^p(n)$ or phase $\theta_m^b(n)$ is estimated, the corresponding delay function may be constructed by (38) and/or (39) and then (40) for demodulation. Since both peak and bandpass methods show good agreement, the latter may seem to show the advantage of estimating $\theta_m(n)$ directly (without numerically integrating $\omega_i(n)$). This is somewhere eclipsed, however, by the fact that $f_i^p(n)$ (or some external library) is needed to assign parameters of the bandpass filter.

3.3. (De)trending the Delay Function

The effect of (de)modulating actual instrument sounds relies on the assumption that 1) the sustained tone has significant amplitude and 2) the oscillation is stable and sounds at a center frequency ω_c which, with some tolerance, can trend up/downward.

The need for an accurate estimate of ω_c lies in the fact that any small value (even rounding error) above or below the actual value, will introduce an offset in $\omega_i(n)$ that, because of the integral relationship with phase θ_i, will compound over time and introduce an upward or downward trend in the estimated delay function $\hat{D}_m(n)$. To see this, consider the example of a constant transposition of 1/2 (down by one octave) a DC offset which from (37) and (39) yields a delay function

$$\sum_{l=0}^{n} \left(1 - \frac{1}{2}\right) = \sum_{l=0}^{n} \frac{1}{2} = \frac{1}{2} n,$$

having an upward linear trend with increasing sample index n. While in this example the upward trend is desired, error in estimating ω_c can lead to similar but undesirable growth and a delay function that is not sufficiently accurate to (de)modulate the vibrato. Further, such an upward (or downward) trend can make implementation using a time-varying (circular) delay line impractical due to the constraint of a maximum delay length. For this reason, delay lines are more suited to implementing time-varying delay that is oscillatory, with a defined maximum and minimum frequency, rather than constant transpositions or signals that chirp over time.

The delay functions in Figure 7 correspond to the estimated f_i^p from signals in Figure 5 with an upward trend that is particularly prominent in the clarinet, likely due to the fact that f_i^p is estimated in the presence of an ω_c that is itself time varying (upward and downward trend independent of the vibrato), a situation that is typical in many musical examples. Detrending (using either a linear or polynomial detrending function) on these delay functions as in Figure 7 (bottom), allows for separation of vibrato (an oscillating signal) from more gradual (natural) frequency changes (gliding upward to a pitch during the note onset and/or falling slightly flat in its release). This produces delay functions that are more accurate thus improving demodulation results (vibrato removal) and making time-varying delay lines more practical for implementation.

Finally, returning to the motivation of improved blending of combined signals by vibrato matching, Figure 8 demonstrates a vocal signal with vibrato, its vibrato transferred to a synthesized square wave, and the vocal signal after demodulation.

[2]Filters were created using the `butter` function available in both MATLAB and scipy.

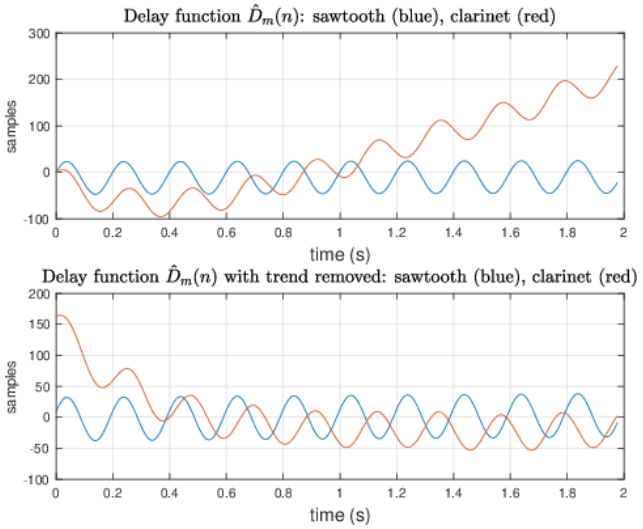

Figure 7: *Delay function (top) for sawtooth (blue) and clarinet (red), with the clarinet showing a prominent upward trend, likely due to the time dependent nature of its sounding frequency f_c (which is static for the sawtooth). The trending exists for both peak picking and bandpass methods.*

4. CONCLUSIONS

Given a method for estimating the instantaneous phase or frequency of a signal, it is possible to construct a time-varying delay line with a delay (or warping) function capable of imparting similar vibrato-rate frequency modulation onto another signal (if it is itself initially without vibrato). Once the delay function related to the modulation is obtained, a demodulation delay function is derived from its inversion and is shown herein to remove *most* audible effects of the vibrato. Because this demodulation approach results in an additional FM sinusoidal term however, very low-amplitude sidebands are introduced by the process. It is shown, however, that if the demodulating delay function is first delayed by itself (serving as its own delay function) before being used to demodulate vibrato from the signal, there is much greater reduction in the sidebands of the additional FM term, effectively removing any vibrato effects by attenuating them well below the point of inaudibility (defined here as -60 dB).

Delay functions derived from frequency and/or phase estimates (using peak picking and bandpass filter methods herein) may also introduce up/downward trends that should be removed (via detrending functions) before use. Finally, though beyond the scope of the work presented here, demodulation of any oscillating amplitude envelope (known to frequently accompany vibrato [11] and [16]) is also necessary for the sound to be truly void of any audible vibrato effects.

5. REFERENCES

[1] Benjamin Hackbarth, Norbert Schnell, and Diemo Schwarz, "Audioguide: a framework for creative exploration of concatenative sound synthesis," *Musical research residency report. Paris, Institut de Recherche et Coordination Acoustique-Musique*, 2010.

[2] Jonathan Driedger, Thomas Prätzlich, and Meinard Müller, "Let it bee-towards nmf-inspired audio mosaicing.," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 350–356.

[3] Carmine-Emanuele Cella, "Orchidea: A comprehensive framework for target-based computer-assisted dynamic orchestration," *Journal of New Music Research*, vol. 51, no. 1, pp. 40–68, 2022.

[4] Marcelo Caetano, "Morphing musical instrument sounds with the sinusoidal model in the sound morphing toolbox," in *International Symposium on Computer Music Multidisciplinary Research*. Springer, 2019, pp. 481–503.

[5] Fabian-Robert Stöter, Stefan Bayer, and Bernd Edler, "Unison source separation.," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2014, pp. 235–241.

[6] Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler, and Paul Magron, "Common fate model for unison source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 126–130.

[7] Fabian-Robert Stöter, Michael Schoeffler, Bernd Edler, and Jürgen Herre, "Human ability of counting the number of instruments in polyphonic music," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[8] Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, and Jürgen Herre, "An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results.," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 389–394.

[9] Mingfeng Zhang, Mark Bocko, and James Beauchamp, "Temporal analysis, manipulation, and resynthesis of musical vibrato," *Proceedings of Meetings on Acoustics*, vol. 22, no. 1, 2014.

[10] Mingfeng Zhang, Mark Bocko, and James Beauchamp, "Measurement and analysis of musical vibrato parameters," *Proceedings of Meetings on Acoustics*, vol. 23, no. 1, 2015.

[11] Petros Maragos, James F Kaiser, and Thomas F Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on signal processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[12] Miller S. Puckette, *The Theory and Technique of Electronic Music*, World Scientific Publ., 2007.

[13] A. Roebel and S. Maller, "Transforming vibrato extent in monophonic sounds," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2011.

[14] Russell Bradford, Richard Dobson, and John ffitch, "Sliding is smoother than jumping," in *Proceedings of the International Computer Music Conference (ICMC)*, 2005.

[15] Julius O. Smith, *Physical Audio Signal Processing*, Online book, 2010, Accessed March 06, 2024, .

[16] Tae Hong Park, Jonathan Biguenet, Zhiye Li, Conner Richardson, and Travis Scharr, "Feature modulation synthesis (fms)," in *Proceedings of the International Computer Music Conference (ICMC)*. International Computer Music Association, 2007, pp. 368–372.

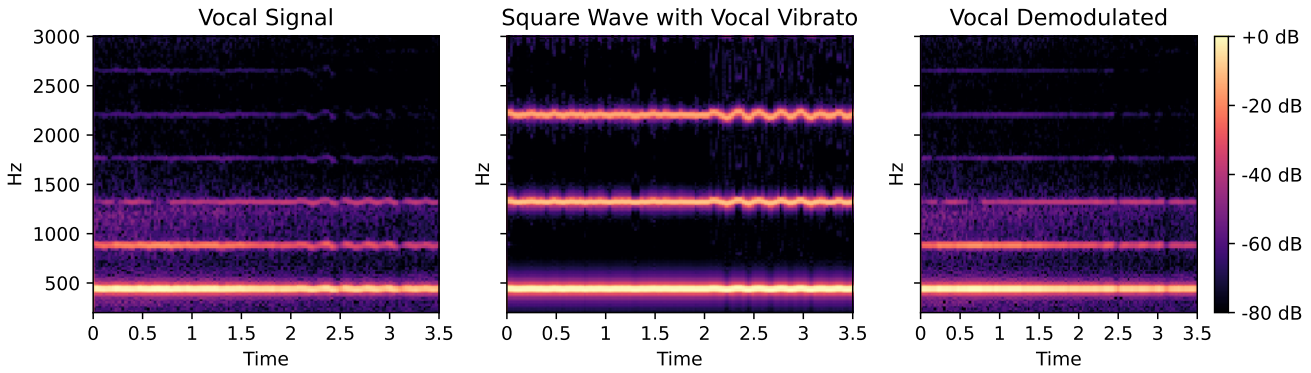

Figure 8: *The frequency modulation from a vocal (left) is transferred to a previously unmodulated square wave (center). The vocal signal can also be demodulated using the estimated functions (right).*

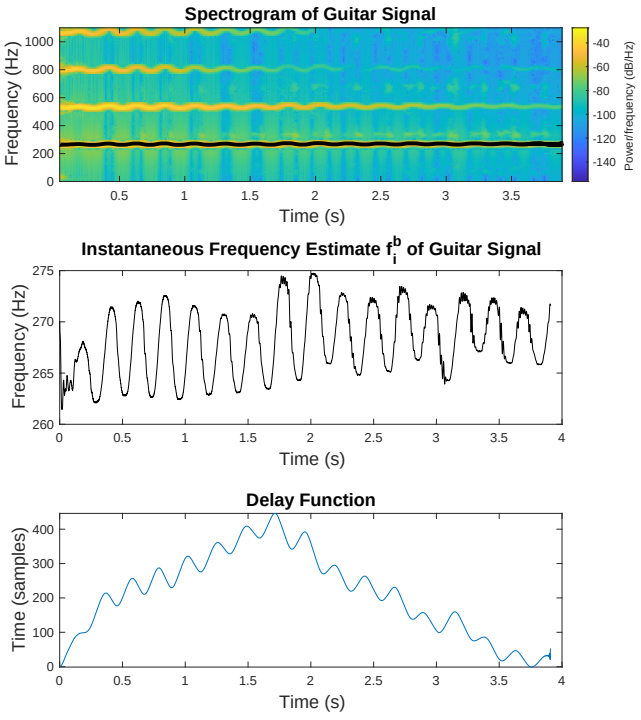

Figure 9: *The estimated instantaneous frequency f_i^b and modulating delay function calculated from a guitar signal with vibrato. This example shows a delay function that does not need detrending due to the lack of an ever-increasing delay time.*

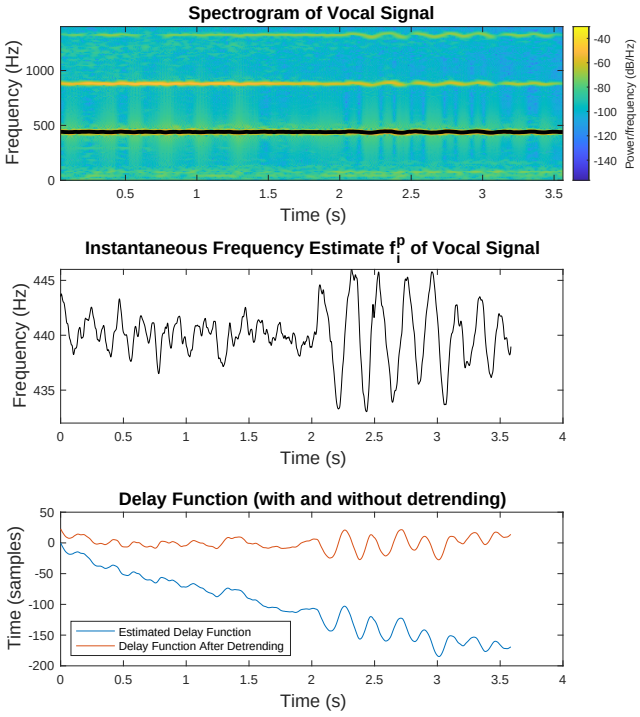

Figure 10: *The estimated instantaneous frequency f_i^p and modulating delay function calculated from a vocal signal with vibrato. The delay function is detrended due to its linear growth.*