Accuracy of Frequency Estimates Using the Phase Vocoder

Miller S. Puckette and Judith C. Brown

***Abstract*— The phase vocoder is a well-known technique for dividing an audio signal into time-varying sinusoidal components and estimating their frequencies and amplitudes. The accuracy of the frequency estimates is studied here by predicting, and then measuring experimentally, the magnitude of errors due to two factors: 1) interference between different components, and 2) interference due to the presence of noise in the signal. The magnitude of the error depends on the relative amplitudes of the component in question and the disturbing signal, on the size and spacing of the analysis windows, on the window function used, and, in the case where the disturbance is due to another sinusoidal component, on the phase difference between the two. The implications of these results for choosing analysis parameters are discussed. The case of a one-sample spacing between analysis windows is treated in detail. Finally, we compare the phase vocoder with the maximum likelihood frequency estimator.**

I. Introduction

THE PHASE vocoder has long been used to analyze and resynthesize speech and monophonic musical sounds. First introduced as a time-domain technique by Flanagan [1], its modern, fast Fourier transform (FFT) based implementation was worked out by Portnoff [2]. The sounds to be analyzed are assumed to consist (in part) of a sum of many sinusoidal components whose frequencies and amplitudes may change over time. When used as an analysis tool, the phase vocoder's output is a time-varying list of amplitudes and frequencies of the components; this might be used for estimating the pitch of a musical sound as in [3], or for obtaining an additive-synthesis model for a musical instrument as in [4]. As an analysis/resynthesis tool, the phase vocoder has been used to alter the time scale of recorded sounds, as reported in [5]–[8], among others. Here we will be primarily concerned with the phase vocoder as an analysis tool, although our results might be of use in analysis/resynthesis applications as well.

We will use the phase vocoder in its FFT-based, bandpass configuration, as shown in Fig. 1. Let $x[n], n = 0, 1, 2, \cdots$ be a discretely sampled signal containing a sinusoidal component of frequency ω. In effect, we use a bandpass FIR filter tuned at or near ω to isolate the component. We evaluate the output of the filter at two points, M and $M + H$, where H is called the *hop size*. These two samples of the filter's output are equal to the dot product of two windows of the signal with the filter kernel. We then find the component's frequency by measuring the phase difference of the filter outputs at the two points. The filter kernels we are interested in will be of the form $w[n]\exp(-2\pi jkn/N)$, where k is a bin number, N is the window size, and $w[n]$ is a window function. The filter outputs can be written in terms of the windowed short-time Fourier transform (WSTFT) of $x[n]$:

$$X_w[M,k] = \sum_{n=0}^{N-1} w[n]x[n+M]e^{-2\pi jnk/N}. \tag{1}$$

The phase vocoder's frequency estimate is the phase change over an interval of time H samples long, divided by H:

$$\omega_{\text{est}}(H,N,M) = \frac{\arg X_w[M+H,k] - \arg X_w[M,k]}{H} \tag{2}$$

for a suitably chosen bin number k.[1]

It is natural to ask how to choose N, H, and the window w to get the greatest possible accuracy in the face of constraints which may involve time resolution and/or computational expense. We will take the quantity $N + H$ as our measure of time resolution, since ω_{est} depends on $N+H$ successive points of the signal. Smaller values of $N + H$ might be better for two reasons. First, if a component's frequency and amplitude are changing with time, an accurate frequency measurement should be as local as possible; the phase vocoder's output for a given moment in time should not depend on values of the input signal except in a small neighborhood of that moment. There is no easy way to quantify the inaccuracies that an increased analysis window size would introduce, but it is clearly preferable to keep the analysis as local as possible. Second, if we happen to be designing a real-time system, the output will incur more delay as $N + H$ increases.

On the other hand, the accuracy of (2) improves with increasing values of N since the selectivity of the bandpass filter can be greater for larger values of N. The accuracy also tends to increase with H since we can write

$$\begin{aligned}\omega_{\text{est}}(H,N,M) &= \frac{1}{H}[\omega_{\text{est}}(1,N,M) + \omega_{\text{est}}(1,N,M+1) + \cdots \\ &\quad + \omega_{\text{est}}(1,N,M+H-1)]\end{aligned} \tag{3}$$

Manuscript received August 26, 1995; approved February 12, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

M. Puckette is with the Department of Music, University of California, San Diego, La Jolla, CA 92093-0326 USA (e-mail: msp@ucsd.edu).

J. Brown is with the Massachusetts Institute of Technology Media Laboratory, Cambridge, MA 02139 USA (e-mail: brown@media.mit.edu).

Publisher Item Identifier S 1063-6676(98)01741-6.

[1] Portnoff and others use the lowpass formulation of the phase vocoder instead of the bandpass one; this makes it easier to consider the effect of time decimation of the phase vocoder's output. Our use of the bandpass formulation here simplifies our calculations; our results are independent of the choice of formulation since ω_{est} is.

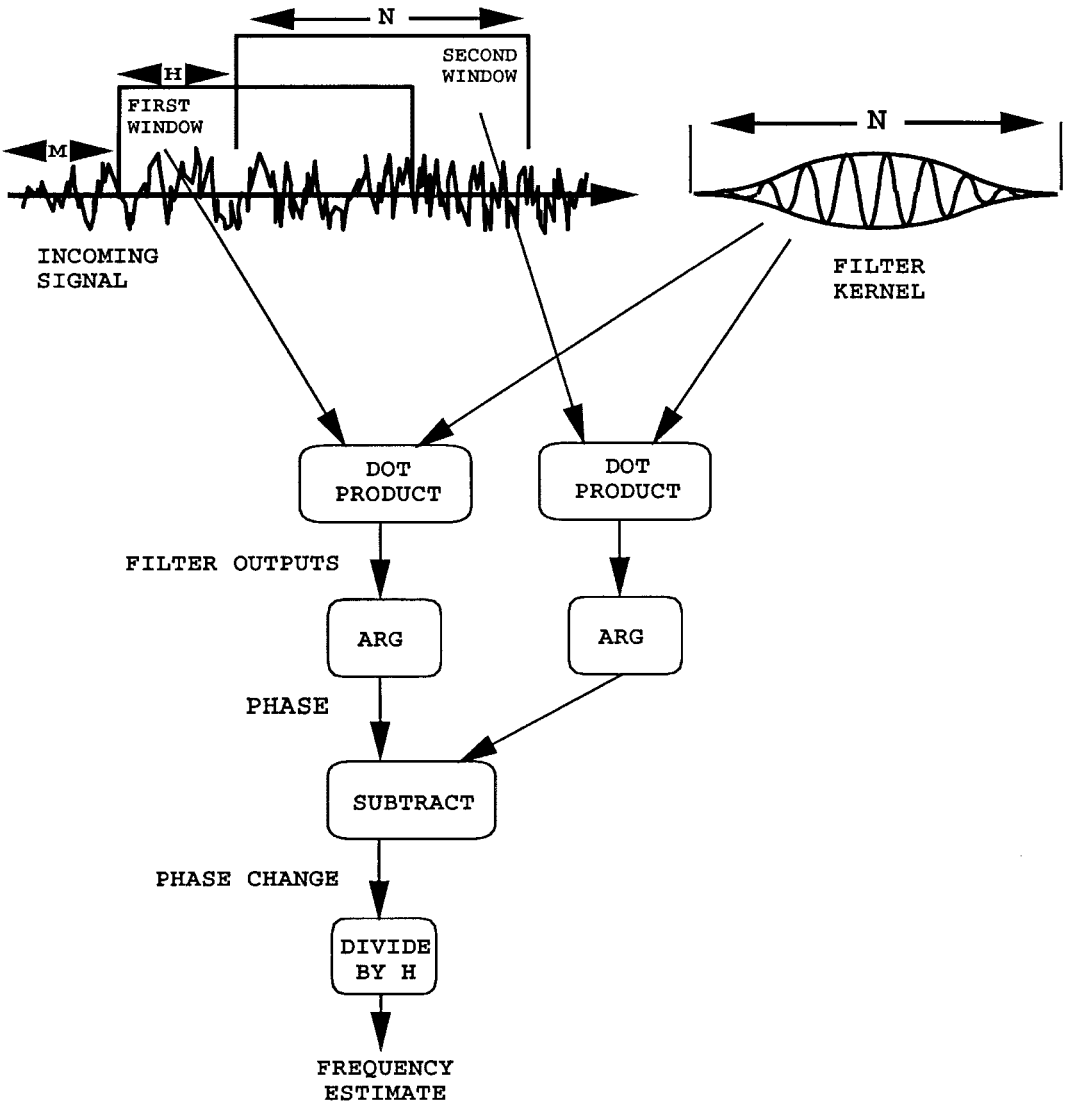

Fig. 1. Block diagram of the phase vocoder as used for analysis.

so that we can regard the frequency estimate for larger values of H as an average of estimates for smaller values of H. (We are neglecting phase-unwrapping effects for the moment; see Section VI.) Since our time resolution is $N + H$, and since uncertainty in frequency decreases with both N and H, we can trade H off with N to get the best frequency resolution for a given time resolution. The time/frequency tradeoff, that is, $N + H$ versus uncertainty in frequency, is a familiar one in signal processing.

To determine the accuracy of (2), we will study how ω_{est} is affected by the presence of noise or additional sinusoidal components in the input signal. We then verify the results on signals having known components.

In the discrete Fourier transform (DFT) implementation of the phase vocoder, the filter kernels shown in Fig. 1 are evaluated by multiplying the input signal by the window function $w[n]$ and then taking the DFT (thereby simultaneously evaluating N filter outputs). The computation time is often dominated by the time required to calculate the two DFT's. Two optimizations are known which in effect save half the DFT computation. First, if a series of results is desired, one can compute values of $\omega_{\text{est}}(H, N, mH)$ for $m = 0, 1, 2, \cdots$, thereby using each calculated value of X_w twice. This makes sense if the points at which results are desired correspond to a suitable hop size. The second possibility, reported independently by Brown [3] and Charpentier [9] and based on a technique of Goertzel [10], avoids the second DFT for the special case $H = 1$. (See also [6].) If computation time is of concern, the availability of these optimizations might affect our choice of N and H.

In applications where the number of sinusoids present is relatively small, the maximum likelihood (ML) estimator is usually preferred over the phase vocoder. Here we will be able to provide some insight into the relative performance of the two.

Another technique described by Kay [11] is to estimate the frequency of a single sinusoid in noise as a weighted average of phase differences between successive samples of a signal. Kay finds that the best weighting function is an inverted parabola, and that under suitable conditions his technique

reduces to the ML estimate for a single sinusoid in white noise. While Kay's technique is different from the phase vocoder, we will use his result as preliminary justification for including the parabolic window in our analysis, in addition to three windows traditionally used with the phase vocoder (Hanning, Hamming, Blackman–Harris).

A. Terminology

We will use the unnormalized DFT defined as

$$\mathcal{FT}\{x[n]\}[k] = \sum_{n=0}^{N-1} x[n]e^{-2\pi jnk/N}$$

where k is called the *bin number*. When using nonintegral values of k we will enclose k in parentheses as in "$\mathcal{FT}\{x[n]\}(k)$." We rewrite (1) as

$$X_w[M,k] = \mathcal{FT}\{w[n]x[n+M]\}[k]. \tag{4}$$

We will use the Hanning, Hamming, three-term Blackman–Harris, and parabolic window functions, denoted by $w[n]$, and with Fourier transform

$$W(k) \equiv \mathcal{FT}\{w[n]\}(k). \tag{5}$$

The Hanning, Hamming, and Blackman–Harris windows are all of the form

$$w[n] = \alpha - \beta\cos\frac{\pi(2n+1)}{N} + \gamma\cos\frac{2\pi(2n+1)}{N} \tag{6}$$

for $0 \le n < N$ and zero, otherwise. For the Hanning window we have $(\alpha,\beta,\gamma) = (0.5, 0.5, 0)$, for the Hamming window, (0.54, 0.46, 0), and for the minimum-sidelobe three-term Blackman–Harris window, (0.423 23, 0.497 55, 0.079 22) [12]. The Fourier transform of $w[n]$ is then

$$W(k) = e^{-j(N-1)\pi k/N}W_0(k) \tag{7}$$

where

$$W_0(k) = \alpha D_N(k) + \frac{\beta}{2}(D_N(k+1) + D_N(k-1)) + \frac{\gamma}{2}(D_N(k+2) + D_N(k-2)) \tag{8}$$

and

$$D_N(k) = \frac{\sin(\pi k)}{\sin\left(\frac{\pi k}{N}\right)}.$$

The parabolic window defined as

$$w[n] = 1 - \left[\frac{n-(N-1)/2}{N/2}\right]^2 \tag{9}$$

for $0 \le n < N$, also satisfies (7) with

$$W_0(k) = \frac{1}{N^2(\cos(4\pi k/N) - 4\cos(2\pi k/N) + 3)} \cdot \left[(2N-1)\cos\frac{(N+3)\pi k}{N} + (-2N-5) \cdot \cos\frac{(N+1)\pi k}{N} + (-2N+5)\cos\frac{(N-1)\pi k}{N} + (2N+1)\cos\frac{(N-3)\pi k}{N}\right]. \tag{10}$$

For all of these windows, the main lobe of the Fourier transform is at least two bins wide, so that when $|k| < 1$, (7) gives

$$\begin{aligned} \arg W(k) &= -(N-1)\pi k/N \\ |W(k)| &= W_0(k). \end{aligned} \tag{11}$$

II. General Error Formulas for the Phase Vocoder

We will consider a complex exponential signal with amplitude a, angular frequency ω, and initial phase δ, as follows:

$$x[n] = ae^{j(\omega n+\delta)}$$

with an additive perturbing signal $y[n]$

$$x'[n] = x[n] + y[n]. \tag{12}$$

Let $X_w[M,k], Y_w[M,k]$, and $X'_w[M,k]$ denote the corresponding WSTFT's as in (4). In particular, we have

$$X_w[M,k] = ae^{j((N-1)((\omega/2)-(\pi k/N))+\delta+M\omega)} \cdot W_0\left(k - \frac{N\omega}{2\pi}\right). \tag{13}$$

Here, the phase term comes from two sources: the terms in ω and δ give the phase of the signal $x[n]$ at the middle of the window, and the term in k is the window's phase term from (7).

We want to estimate the contribution of the disturbance $y[n]$ to ω_{est} given by (2) where k is the bin whose center frequency is closest to ω, as follows:

$$\left|k - \frac{N\omega}{2\pi}\right| \le \frac{1}{2}. \tag{14}$$

If $y[n] = 0$, plugging (13) into (2) gives exactly ω regardless of the choice of H and N. We will assume that X_w dominates Y_w in the kth bin

$$\begin{aligned} |Y_w[M,k]| &\ll |X_w[M,k]| \sim aN \\ |Y_w[M+H,k]| &\ll aN. \end{aligned} \tag{15}$$

In other words, we are assuming that the WSTFT resolves the sinusoid $x[n]$ from the disturbance $y[n]$.

We can now estimate the effect of $y[n]$ on ω_{est} by considering each term of (2) separately. Using the identity

$$\arg(A+B) = \arg(A\cdot(1+B/A)) = \arg(A) + \arg(1+B/A)$$

the first term of (2) becomes

$$\begin{aligned} \arg X'_w[M+H,k] &= \arg\{X_w[M+H,k]\} + \arg\left\{1 + \frac{Y_w[M+H,k]}{X_w[M+H,k]}\right\} \\ &= \arg\{X_w[M+H,k]\} + \arg\left\{1 + \frac{Y_w[M+H,k]}{e^{j\omega H}X_w[M,k]}\right\}. \end{aligned}$$

Here, we have written the total phase as the original phase plus a disturbance phase. To estimate the latter we use the assumption (15) to make an approximation: If z is a complex

number whose magnitude is small, then to order $|z|^2$ we have $\arg(1+z) \approx \mathrm{Im}(z)$. Applying this to the disturbance term gives

$$\begin{aligned}
&\arg\left\{1+\frac{Y_w[M+H,k]}{e^{j\omega H}X_w[M,k]}\right\} \\
&\quad\approx \mathrm{Im}\left\{\frac{Y_w[M+H,k]}{e^{j\omega H}X_w[M,k]}\right\} \\
&\quad= \frac{1}{|X_w[M,k]|}\,\mathrm{Im}\{Y_w[M+H,k] \\
&\qquad\cdot e^{-j(\omega H+\arg X_w[M,k])}\} \\
&\quad= \frac{1}{|X_w[M,k]|}\,\mathrm{Im}\sum_{n=0}^{N-1}(w[n]y[M+H+n] \\
&\qquad\cdot e^{-j((2\pi nk/N)+\omega H+\arg X_w[M,k])}) \\
&\quad= \frac{1}{|X_w[M,k]|}\,\mathrm{Im}\sum_{n=H}^{N+H-1}(w[n-H]y[M+n] \\
&\qquad\cdot e^{j((2\pi Hk/N)-(2\pi nk/N)-\omega H-\arg X_w[M,k])}).
\end{aligned}$$

This estimate can also be used to find the effect of $y[n]$ on the second term of (2) by substituting 0 for H. The total error is thus

$$\begin{aligned}
\epsilon &\equiv \omega_{\text{est}} - \omega \\
&\approx \frac{\mathrm{Im}\{Y_w[M+H,k]e^{-j(\omega H+\phi_1)} - Y_w[M,k]e^{-j\phi_1}\}}{H|X_w[M,k]|} \qquad (16)\\
&= \frac{1}{H|X_w[M,k]|}\,\mathrm{Im}\left\{\sum_{n=H}^{N+H-1} w[n-H]y[M+n]\right. \\
&\qquad\cdot e^{-j((2\pi nk/N)+\phi_2)} \\
&\qquad\left.-\sum_{n=0}^{N-1} w[n]y[M+n]e^{-j((2\pi nk/N)+\phi_1)}\right\} \qquad (17)
\end{aligned}$$

where for convenience we have defined

$$\begin{aligned}
\phi_1 &= \arg X_w[M,k] \\
\phi_2 &= \phi_1 + \left(\omega - \frac{2\pi k}{N}\right)H.
\end{aligned}$$

If we add several perturbing signals $y_1[n], \cdots, y_P[n]$ to $x[n]$, with each y_p satisfying (15), the errors introduced in ω_{est} are approximately additive. We can thus estimate the accuracy of the phase vocoder for complex signals by breaking them down into simple components. To this end, we will now calculate the contributions due to the presence of sinusoidal components at frequencies other than ω (see Section III) and white noise (see Section IV).

III. Disturbance by a Sinusoid

We first consider a disturbance by a complex exponential signal of amplitude a', frequency ω', and phase δ'

$$y[n] = a'e^{j(\omega' n+\delta')}$$

so that

$$\begin{aligned}
Y_w[M,k] &= a'e^{j((N-1)((\omega'/2)-(\pi k/N))+\delta'+M\omega')} \\
&\quad\cdot W_0\left(k-\frac{N\omega'}{2\pi}\right).
\end{aligned}$$

Equation (16) then becomes

$$\begin{aligned}
&\epsilon_{\text{sinusoid}} \\
&\quad= \omega_{\text{est}} - \omega \approx \frac{a'W_0\left(k-\frac{n\omega'}{2\pi}\right)}{H|X_w[M,k]|} \\
&\qquad\cdot \mathrm{Im}\{e^{j((N-1)((\omega'/2)-(\pi k/N))+(\omega'-\omega)H+\delta'+M\omega'-\phi_1)} \\
&\qquad- e^{j((N-1)((\omega'/2)-(\pi k/N))+\delta'+M\omega'-\phi_1)}\}.
\end{aligned}$$

Plugging in $X_w[M,k]$ and ϕ_1 from (11) and (13) and simplifying gives

$$\begin{aligned}
\epsilon_{\text{sinusoid}} &\approx \frac{2a'W_0\left(k-\frac{N\omega'}{2\pi}\right)}{aW_0\left(k-\frac{N\omega}{2\pi}\right)H}\cdot\sin\left(\frac{(\omega'-\omega)H}{2}\right) \\
&\quad\cdot\cos\left(\frac{(\omega'-\omega)(H+N-1)}{2}+\delta'-\delta\right. \\
&\quad\left.+M(\omega'-\omega)\right). \qquad (18)
\end{aligned}$$

The result is proportional to the relative strengths of the two signals' contributions to the kth bin of the WSTFT, and depends on two phase terms.

IV. Disturbance by White Noise

Here we consider the effect of adding real or complex-valued white noise. Suppose first that $y[n]$ is real-valued white noise with power σ^2. We will regard each sample of $y[n]$ as a random variable with mean zero and standard deviation σ, all of them uncorrelated. If $a[n]$ is a sequence of real numbers, the sum

$$\sum_n a[n]y[n]$$

is another random variable, whose mean is again zero and whose variance is the sum of the variances of the individual samples

$$\sigma^2_{\text{total}} = \sigma^2\sum_n a[n]^2. \qquad (19)$$

For the moment we will restrict our attention to the case $H < N$. Evaluating the imaginary part in (17), we can estimate the error as

$$\begin{aligned}
\epsilon_{\text{noise}} &= \omega_{\text{est}} - \omega \\
&\approx \frac{1}{H|X_w[M,k]|}\left[\sum_{n=0}^{N-1} w[n]y[M+n]\right. \\
&\qquad\cdot\sin\left(\frac{2\pi nk}{N}+\phi_1\right) - \sum_{n=H}^{H+N-1} w[n-H]y[M+n] \\
&\qquad\left.\cdot\sin\left(\frac{2\pi nk}{N}+\phi_2\right)\right].
\end{aligned}$$

Combining like terms gives

$$\epsilon_{\text{noise}} \approx \frac{1}{H|X_w[M,k]|}\Bigg[\sum_{n=0}^{H-1} w[n]\sin\left(\frac{2\pi nk}{N}+\phi_1\right) \cdot y[M+n] + \sum_{n=H}^{N-1}\Bigg\{w[n]\sin\left(\frac{2\pi nk}{N}+\phi_1\right) - w[n-H]\sin\left(\frac{2\pi nk}{N}+\phi_2\right)\Bigg\} y[M+n] - \sum_{n=N}^{H+N-1} w[n-H]\sin\left(\frac{2\pi nk}{N}+\phi_2\right) \cdot y[M+n]\Bigg].$$

The variance (19) becomes

$$\sigma^2_{\text{noise}} = \frac{\sigma^2}{H^2|X_w[M,k]|^2}\Bigg[\sum_{n=0}^{H-1}\left(w[n]\sin\left(\frac{2\pi nk}{N}+\phi_1\right)\right)^2 + \sum_{n=H}^{N-1}\Bigg\{w[n]\sin\left(\frac{2\pi nk}{N}+\phi_1\right) - w[n-H] \cdot \sin\left(\frac{2\pi nk}{N}+\phi_2\right)\Bigg\}^2 + \sum_{n=N}^{H+N-1} \cdot \left(w[n-H]\sin\left(\frac{2\pi nk}{N}+\phi_2\right)\right)^2\Bigg].$$

Carrying out the square of the middle term and regrouping gives

$$\sigma^2_{\text{noise}} = \frac{\sigma^2}{H^2|X_w[M,k]|^2}\Bigg[\sum_{n=0}^{N-1}\left(w[n]\sin\left(\frac{2\pi nk}{N}+\phi_1\right)\right)^2 - 2\sum_{n=H}^{N-1} w[n]w[n-H]\sin\left(\frac{2\pi nk}{N}+\phi_1\right) \cdot \sin\left(\frac{2\pi nk}{N}+\phi_2\right) + \sum_{n=H}^{H+N-1} \cdot \left(w[n-H]\sin\left(\frac{2\pi nk}{N}+\phi_2\right)\right)^2\Bigg].$$

Letting $h = H/N$ denote the relative window overlap and introducing the variable $t = n/N$, we can approximate the sums by integrals, as follows:

$$\begin{aligned}\sigma^2_{\text{noise}} &\approx \frac{N\sigma^2}{H^2|X_w[M,k]|^2}\Bigg[\int_0^1 [w(Nt)\sin(2\pi kt+\phi_1)]^2\,dt - 2\int_h^1 w(Nt)w(N(t-h))\sin(2\pi kt+\phi_1) \cdot \sin(2\pi kt+\phi_2)\,dt + \int_h^{1+h}[w(N(t-h)) \cdot \sin(2\pi kt+\phi_2)]^2\,dt\Bigg] \\ &= \frac{N\sigma^2}{2H^2|X_w[M,k]|^2}\Bigg[\int_0^1 w(Nt)^2(1-\cos(4\pi kt + 2\phi_1))\,dt - 2\int_h^1 w(Nt)w(N(t-h)) \cdot (\cos(\phi_1-\phi_2)-\cos(4\pi kt+\phi_1+\phi_2))\,dt + \int_h^{1+h} w(N(t-h))^2(1-\cos(4\pi kt+2\phi_2))\,dt\Bigg].\end{aligned}$$

We may reduce this expression to a concise value by adding the new assumption that k be large enough that we may ignore terms in $\cos(4\pi kt)$ in the above integrals (since the integral will have $4\pi k$ in the denominator for those terms.) Using this assumption and combining the first and last terms above, we get

$$\sigma^2_{\text{noise}} \approx \frac{N\sigma^2}{H^2|X_w[M,k]|^2} \cdot \Bigg[\int_0^1 w(Nt)^2\,dt - \cos(\phi_1-\phi_2) \cdot \int_h^1 w(Nt)w(N(t-h))\,dt\Bigg]. \tag{20}$$

In the case where $H \geq N$, the calculation is the same except that the cross term never appears. The result still holds if we set $h = 1$, which corresponds to $H = N$.

For the trigonometric window (6), this evaluates to

$$\sigma^2_{\text{noise}} \approx \frac{N\sigma^2}{H^2|X_w[M,k]|^2}\Bigg(\alpha^2 + \frac{\beta^2+\gamma^2}{2} - \cos(\phi_1-\phi_2) \cdot \Bigg[(1-h)\left(\alpha^2 + \frac{\beta^2\cos(2\pi h)}{2} + \frac{\gamma^2\cos(4\pi h)}{2}\right) + \frac{12\alpha\beta - 3\beta^2 - 4\beta\gamma}{12\pi}\sin(2\pi h) + \frac{-12\alpha\gamma + 16\beta\gamma - 3\gamma^2}{24\pi}\sin(4\pi h)\Bigg]\Bigg) \tag{21}$$

where α, β, and γ are the window parameters. For the parabolic window (9) we get

$$\sigma^2_{\text{noise}} \approx \frac{N\sigma^2}{H^2|X_w[M,k]|^2}\left(\frac{8}{15} - \cos(\phi_1-\phi_2) \cdot \left(\frac{8-40h^2+40h^3-8h^5}{15}\right)\right). \tag{22}$$

These results hold for real-valued white noise. If we add pure imaginary-valued white noise, we can multiply both $x[n]$ and $y[n]$ by j to get the same result. Complex-valued white noise therefore gives twice this variance, since the disturbances are additive.

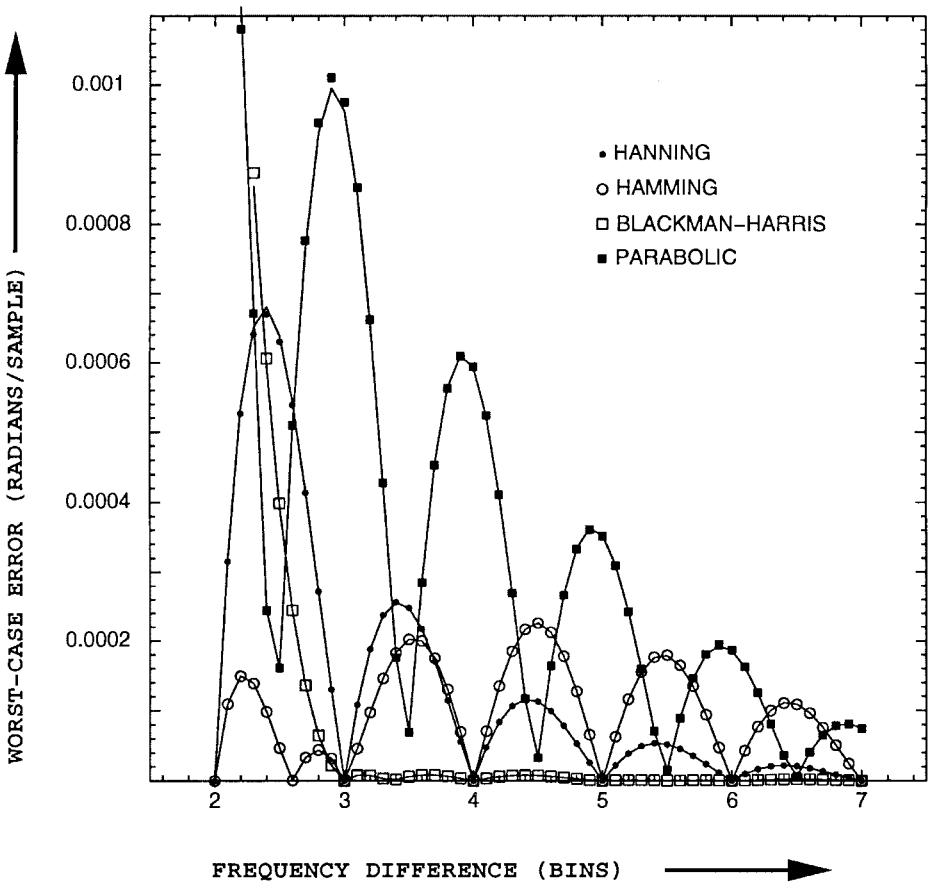

Fig. 2. Predicted worst-case interference of one sinusoidal component in the phase vocoder's frequency estimate for a second component (shown as curves) and the measured interference (shown as points). The X axis gives the frequency separation in bins between the original signal and the interfering one.

V. Measured Results

The results given above depend on three approximations: the replacement of $\arg(1+z)$ by $\mathrm{Im}(z)$ in Section II; and in Section IV, the replacement of sums by integrals and the suppression of high-frequency terms in the integral. The first step is straightforward, but it is difficult to bound the error contributed by the other two steps. To verify these results and also to investigate their qualitative behavior, we ran numerical simulations of the two situations considered above.

In the noise-free, interference-only case (18), the predicted error depends not only on the relative amplitudes of the signals and on H and N, but also on the frequencies and phases of both components. We took fixed values of $a = a' = 1, N = 512, M = 0, H = 64$, and $\omega = 200\pi/N$ (the middle of the 100th bin), and chose δ and δ' in such a way as to maximize the cosine term in (18). For each of the four windows under consideration, we investigated the dependence of the phase vocoder's frequency error on ω' as it ranged from two to seven bins away from ω, i.e., $204\pi/N \leq \omega' \leq 214\pi/N$, in increments of a tenth of a bin. Fig. 2 shows the predicted error (as curves) against the measured error (shown as points). The results show clearly the dominant effect of the W_0 term in the numerator of (18). Also, the predicted and measured results agree closely. For the Hanning window, for example, the greatest absolute deviation between the predicted and measured results occurs at 2.4 bins, at which they equal $6.6995 \cdot 10^{-4}$ and $6.8045 \cdot 10^{-4}$ rad/sample, respectively; the two differ by 1.6%.

We tested the white-noise influence estimates (21) and (22) using a signal equal to the sum of a complex sinusoid (with unit amplitude and angular frequency $\pi/16$) and uniformly distributed, pseudorandom, real-valued white noise in the range $[-1, 1]$; the power of the noise signal was thus 1/3. The window size N was allowed to range over the powers of two between 32 and 2048, inclusively, and the hop size H, from 1–2048. For each (N, H) pair we measured the root mean squared (RMS) error of the phase vocoder's frequency estimate over 10 000 trials. Fig. 3 shows the results plotted against the predictions. The two almost always agree to within 1%; the greatest deviation for a trigonometric window was 2.2%, and for the parabolic window 3.7%, both for $N = 32$.

VI. Choosing N, H, and w

We will now consider specific cases that illustrate the above results, taking a fixed time resolution of $N + H = 1000$ and letting H vary. First let

$$x'[n] = e^{j\omega n} + y[n]$$

where ω is taken to be an exact bin frequency and $y[n]$ is real white noise with unit power. The predicted RMS error is graphed in Fig. 4. The parabolic window with $H = 1$ gives the best results.

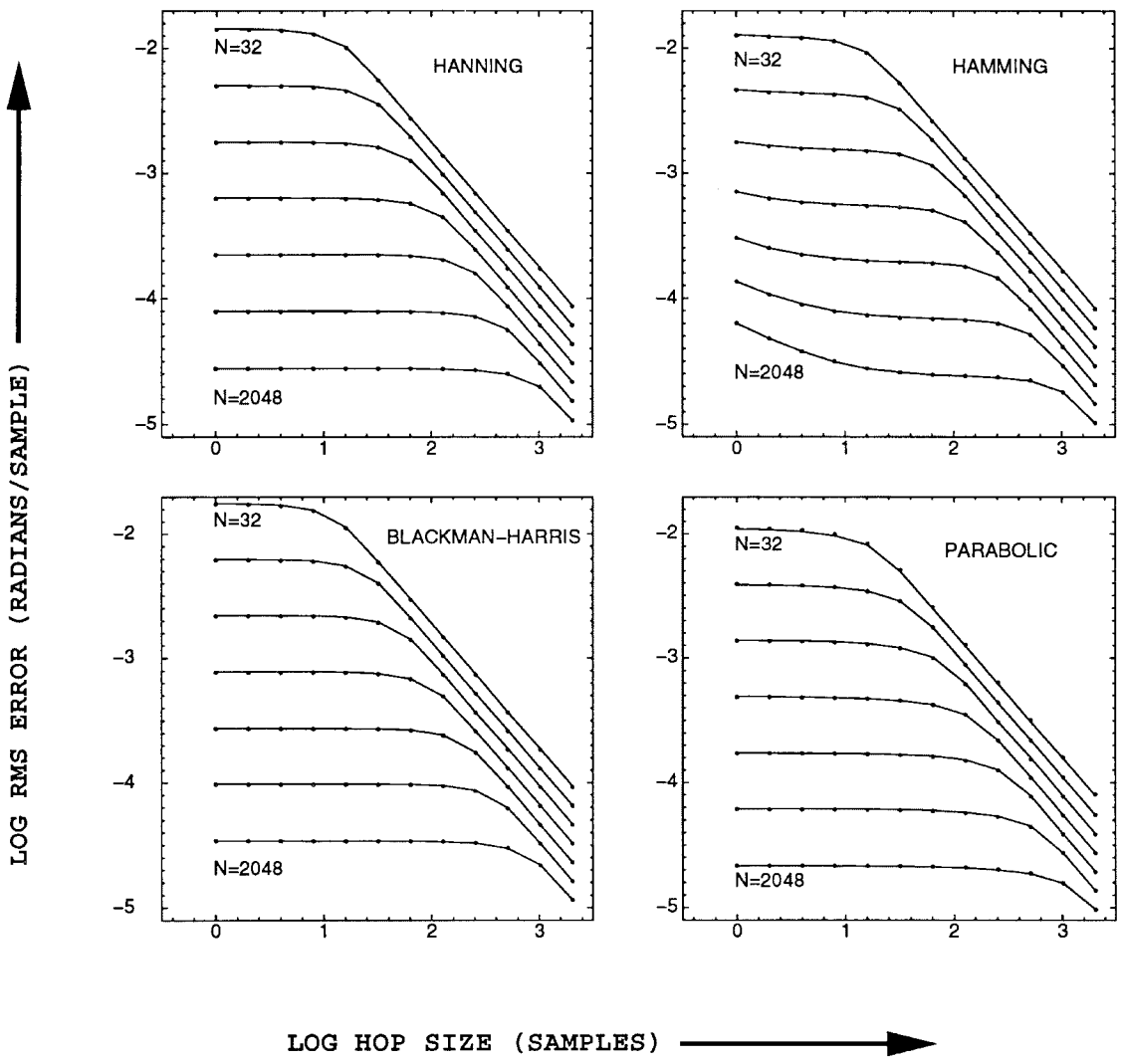

Fig. 3. Predicted RMS error of the phase vocoder's frequency measurement due to white noise (shown as curves), and the measured RMS error (shown as points), plotted against the hop size H.

We now simplify our estimates for $H = 1$ and the Hanning and parabolic windows. The value of (21) depends on H both explicitly and via the quantity h. We rewrite all occurrences of H in terms of h, set $\alpha = \beta = 0.5, \gamma = 0$, and, noting that $h = H/N$ is small when $H = 1$, we take the limit as h approaches zero, giving

$$\sigma^2_{\text{Hanning}} \approx \left[1+3\left(k-\frac{N\omega}{2\pi}\right)^2\right]\frac{\pi^2\sigma^2}{4N\left|W\left(k-\frac{N\omega}{2\pi}\right)\right|^2}.$$

Because k is the bin nearest to the peak at ω, the quantity $k-(N\omega/2\pi)$ varies between $-1/2$ and $1/2$; it is zero if the frequency of the sinusoid coincides with a bin frequency (the best case) and is $+$ or $-1/2$ if the sinusoid is halfway between two bin frequencies (the worst case.) Plugging in values of W from (8) gives a best-case variance of

$$\sigma^2_{\text{Hanning}} \approx \frac{\pi^2\sigma^2}{N^3}$$

and a worst-case variance approximately 2.43 times greater.

Doing the same for the parabolic window, we get

$$\sigma^2_{\text{Para}} \approx \left[1+\frac{2\pi^2}{5}\left(k-\frac{N\omega}{2\pi}\right)^2\right]\frac{8\sigma^2}{3N\left|W\left(k-\frac{N\omega}{2\pi}\right)\right|^2}.$$

The best case is

$$\sigma^2_{\text{Para}} \approx \frac{6\sigma^2}{N^3} \tag{23}$$

and the worse case is 3.32 times as great. For both windows, the variance is very sensitive to the bin difference $k-(N\omega/2\pi)$. We could reduce its value either by zero-padding the time-domain signal (so that the bins of the DFT are more closely spaced), or else by proceeding iteratively, using a first estimate of ω to suggest a fractional value of k at which we can evaluate X_w explicitly. (Here we are tacitly assuming that the error is already small enough that the iteration is closer than the original estimate.)

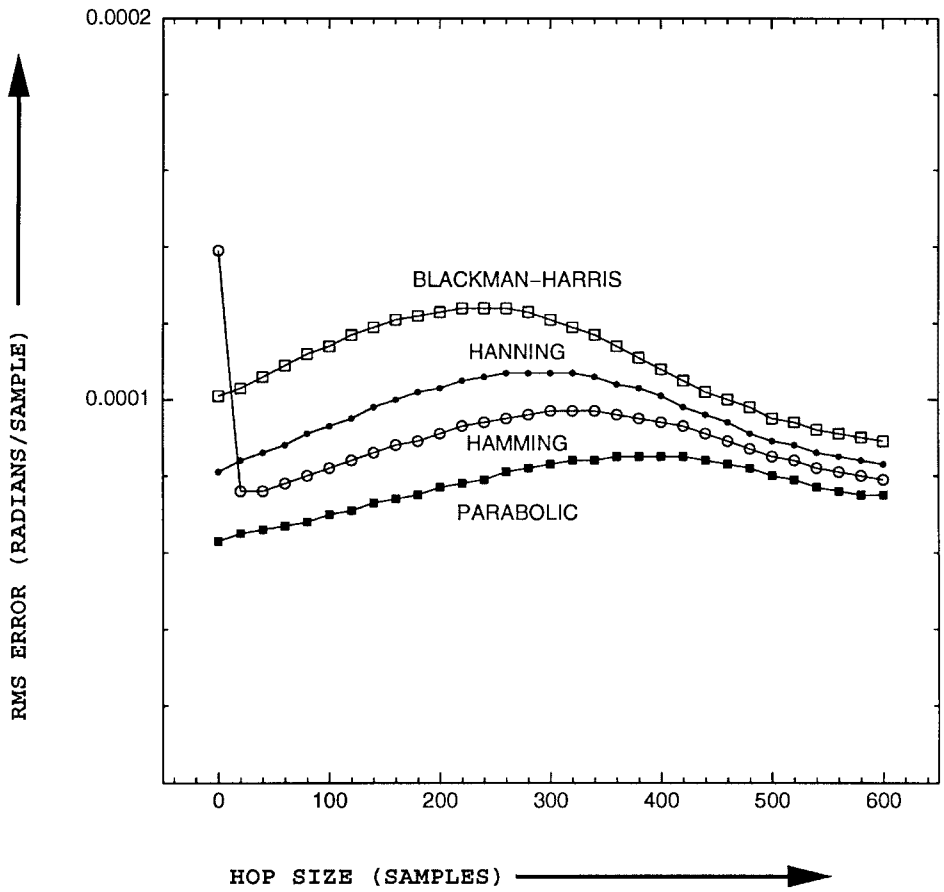

Fig. 4. Predicted RMS error due to white noise in rad/sample, as a function of hop size H, keeping $N + H = 1000$ constant. (Both the curves and the points show predicted values.)

Next we consider the case of a sinusoidal disturbance

$$x'[n] = e^{j\omega n} + e^{j(\omega' n + \delta')},$$

where ω' is sufficiently far from ω to be distinguished by the WSTFT. The result is highly sensitive to the difference $\omega' - \omega$ and to the relative phase δ'. The δ' dependence can be averaged out by replacing the cosine term of (18) by its RMS over all values of δ', equal to $2^{-1/2}$. The other terms could be estimated either in terms of an average or a worst-case value, over various possible ranges of the frequency difference $\omega' - \omega$. We have chosen to take the RMS average over two ranges of the frequency difference; this is intended to give a heuristic measure of the average disturbance by a sinusoid of unknown phase and frequency. The first range, from $4\pi/1000$ to $40\pi/1000$ rad/sample, entails situations where signal components might be closely spaced; the second, from $6\pi/1000$ to $40\pi/1000$ rad/sample, represents a wider and more comfortable spacing. The averages were obtained by numerically integrating the square of (18).

As seen in Fig. 5, the results are very different depending on the range of $\omega - \omega'$ used. In the first case, the optimum is the Hamming window function and $H = 140$; for the second, Blackman–Harris and $H = 60$.

In the second frequency range, we see that the Blackman–Harris window does not perform especially well for very small values of H (also, the Hamming window does badly for small values of H in all the situations we have considered.) The loss of accuracy arises because the two windows are different from zero at the window boundaries; thus, for $H = 1$ in particular, two single samples are disproportionately weighted in the analysis. Because $H = 1$ behaves well from the standpoint of phase unwrapping and calculation time (as we will see below), we have derived the minimum-sidelobe trigonometric window of the form (6) which attains zero at the boundaries. Setting $\beta = \alpha + \gamma = 0.5$, we numerically minimized the worst-case sidelobe strength to obtain $(\alpha, \beta, \gamma) = (0.4090, 0.5, 0.0910)$. This window gives a fifth trace shown in the lower plot of Fig. 5. Our proposed window gives an error of 5.81×10^{-6} rad/sample when $H = 1$, as compared to 3.89×10^{-6} for Blackman–Harris and $H = 60$.

As with interfering white noise, the error magnitude due to interfering sinusoids is greater if the measured sinusoid lies halfway between two bins than if it lies on a bin frequency, but here the ratios are smaller, coming from the falloff in W_0 between zero and 1/2.

Additional factors should be considered when choosing H and N. First, our calculations have all ignored the possibility of phase unwrapping errors, which are progressively harder to control with increasing H. Without an analysis of the probability of getting this type of error, the most prudent choice might sometimes be to insist on setting $H = 1$, which

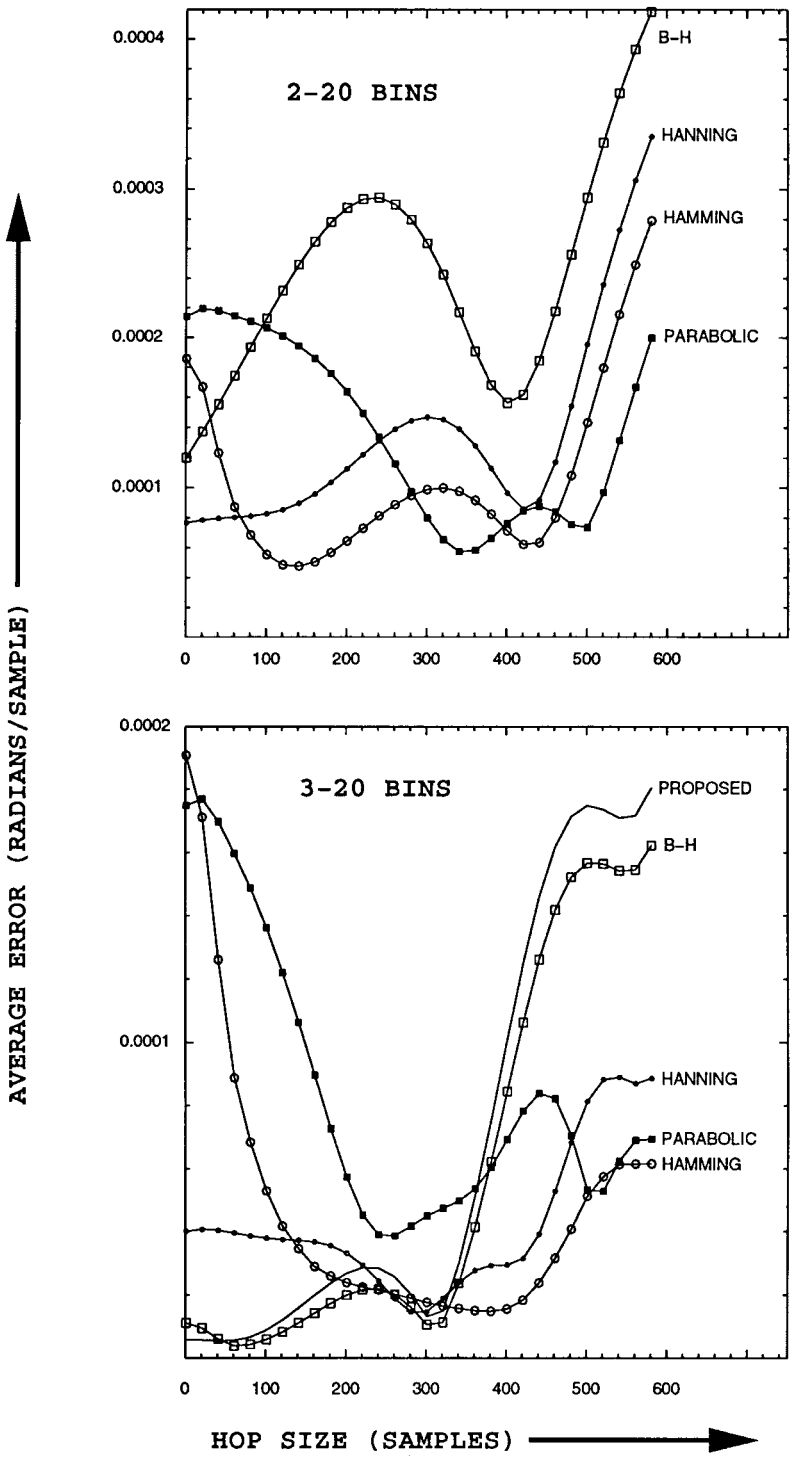

Fig. 5. Predicted RMS error in radians per sample, due to a disturbing complex exponential, as a function of H with $N + H = 1000$. The RMS average is taken over all possible phase differences and over a range of separation between the two complex exponentials (2–20 bins for the upper graph and 3–20 bins for the lower one).

in none of the cases we have considered gave worse than twice the error of the "optimal" choice. (Another possibility, if computation time is no constraint, is to use (3) to get large values of H from smaller ones.) Here we have even presented data where $H > N$, primarily for completeness; unless we have some *a priori* knowledge about the signal we are likely to get unwrapping errors there. We have also not considered the problem of reconstructing the signal after analysis; adding this requirement would also place an upper bound on H [13].

VII. Simplified Formula for the Case $H = 1$

In some situations, we wish to use $H = 1$ and a trigonometric window. Here we can use a variant of Goertzel's technique

[10] to simplify the calculation, as shown in [3] and [9]. (See also [14] for a different way to avoid redundant calculations when computing more than one DFT.) In essence, the Goertzel technique is to calculate the DFT on the points $1, 2, \cdots, N$ directly from the DFT on $0, 1, \cdots, N-1$. For this section we will let $x[n]$ denote an arbitrary real or complex signal, not necessarily a complex exponential. We have

$$\begin{aligned}
&\mathcal{FT}\{x[n+1]\}[k] \\
&\quad = \sum_{n=0}^{N-1} x[n+1]e^{-2\pi jkn/N} \\
&\quad = \sum_{n=1}^{N} x[n]e^{-2\pi jk(n-1)/N} \\
&\quad = e^{2\pi jk/N} \sum_{n=1}^{N} x[n]e^{-2\pi jkn/N} \\
&\quad = e^{2\pi jk/N} \left[\sum_{n=0}^{N-1} x[n]e^{-2\pi jkn/N} + x[N] - x[0] \right] \\
&\quad = e^{2\pi jk/N} [\mathcal{FT}\{x[n]\}[k] + x[N] - x[0]]. \qquad (24)
\end{aligned}$$

We can use this to calculate two WSTFT's, $X_w[M, k]$ and $X_w[M+1, k]$, using a single fast Fourier transform (FFT) calculation and applying the windows in the frequency domain via a three or five point convolution. In situations where the FFT calculation dominates the computation time, we can thereby reduce it almost in half. Letting

$$\begin{aligned}
X[k] &= \mathcal{FT}\{x[n]\}[k] \\
\eta &= e^{\pi j/N}
\end{aligned}$$

we convolve $X[k]$ with $W[k]$ from (7) and (8), giving

$$\begin{aligned}
&\mathcal{FT}\{w[n]x[n]\}[k] \\
&\quad = \eta^{-(N-1)k}[\alpha X[k] - \beta/2(\eta X[k-1] + \eta^{-1}X[k+1]) \\
&\qquad + \gamma/2(\eta^2 X[k-2] + \eta^{-2}X[k+2])].
\end{aligned}$$

Applying (24), the frequency estimate (2) becomes

$$\begin{aligned}
\omega_{\text{est}} = 2\pi k/N &+ \arg\,[2\alpha X[k] - \beta(\eta^{-1}X[k-1] \\
&+ \eta X[k+1]) + \gamma(\eta^{-2}X[k-2] + \eta^2 X[k+2]) \\
&+ 2w[0](x[N] - x[0])] - \arg\,[2\alpha X[k] \\
&- \beta(\eta X[k-1] + \eta^{-1}X[k+1]) + \gamma(\eta^2 X[k-2] \\
&+ \eta^{-2}X[k+2])].
\end{aligned}$$

If $w[0]$ is sufficiently small we can ignore the term in $x[N] - x[0]$. Proceeding as in Section II, we can further simplify the estimate as shown in the first equation at the bottom of the page.

In the numerator we now apply the approximations

$$\begin{aligned}
\eta - \eta^{-1} &\approx 2\pi j/N \\
\eta^2 - \eta^{-2} &\approx 4\pi j/N
\end{aligned}$$

and in the denominator we simply replace terms in η with unity, finally giving us the second equation at the bottom of the page.

VIII. Comparison with Maximum Likelihood Estimation of Frequency

In applications where the number of sinusoids present is relatively small, the ML estimator is often used to determine their frequencies. This was first done for a single sinusoid in white noise by Rife and Boorstyn [15]; see also [16]. They derive the Cramer–Rao (CR) lower bound for the variance of any possible unbiased estimator of the frequency (and also phase and amplitude.) Here, the noise component is taken to be complex Gaussian white noise whose real and imaginary parts each have power σ^2. Except when the signal to noise ratio is very poor, the maximum likelihood estimator's variance is shown to achieve the CR lower bound:

$$\sigma^2_{\text{CR}} = \frac{12\sigma^2}{N^3}.$$

For a single sinusoid in white noise as in Section IV, the phase vocoder's frequency estimate is also unbiased, so it must obey the CR bound. The estimate (23) shows that, if we take a parabolic window, $H = 1$, and ω centered on an FFT bin, the phase vocoder actually attains the CR bound; the factor of two difference comes from our having used real, not complex, white noise.

The ML technique has been tested on signals with two or three components [17] whose frequencies are low-order polynomial functions of time. Except in singular cases, the CR bound is still attained by the ML estimate. The literature does not indicate how this technique would scale to situations involving many sinusoids (typical of phase vocoder applications), either theoretically or in terms of numerical tractability. Thus we compare the two methods for small numbers of sinusoids.

$$\begin{aligned}
\omega_{\text{est}} &\approx 2\pi k/N + \arg\left[\frac{2\alpha X[k] - \beta(\eta^{-1}X[k-1] + \eta X[k+1]) + \gamma(\eta^{-2}X[k-2] + \eta^2 X[k+2])}{2\alpha X[k] - \beta(\eta X[k-1] + \eta^{-1}X[k+1]) + \gamma(\eta^2 X[k-2] + \eta^{-2}X[k+2])}\right] \\
&\approx 2\pi k/N + \text{Im}\left[\frac{\beta(\eta^{-1} - \eta)(X[k+1] - X[k-1]) + \gamma(\eta^2 - \eta^{-2})(X[k+2] - X[k-2])}{2\alpha X[k] - \beta(\eta X[k-1] + \eta^{-1}X[k+1]) + \gamma(\eta^2 X[k-2] + \eta^{-2}X[k+2])}\right].
\end{aligned}$$

$$\omega_{\text{est}} \approx 2\pi/N\left(k + \text{Re}\left[\frac{\beta(X[k-1] - X[k+1]) - 2\gamma(X[k-2] - X[k+2])}{2\alpha X[k] - \beta(X[k-1] + X[k+1]) + \gamma(X[k-2] + X[k+2])}\right]\right).$$

When more than one sinusoid is present, the ML estimator acts quite differently from the phase vocoder. For example, in the case where there is no noise component at all, the ML error is zero as long as the signal obeys the model and meets appropriate nonsingularity conditions. As we have seen above, even if the frequencies of the components of a sound do not vary at all with time, the phase vocoder will make errors in its measurement of their frequencies due to interference between components. Since these errors are deterministic and not random, we cannot even say that the phase vocoder provides an unbiased estimate of frequency.

On the other hand, the phase vocoder is usually used on signals which do not obey a low-dimensional model such as those on which ML has been tested. While many signal processing applications might adhere quite well to the assumptions made in the ML papers, music and speech do not really follow any known model at all. In the literature on phase vocoders, we never find explicit assumptions about the model of the signal; the closest thing to it that we find (in [5], for example) is that the frequencies of the components of a sound change slowly relative to the sample rate.

The model we have assumed here is a fairly loose one: The signal must differ from a sinusoid by a quantity (12) which does not interfere much with it in its spectral neighborhood (15). This model allows for sinusoidal components with time-varying amplitudes and frequencies as long as the deviation within the space of the analysis window is relatively small. In contrast, the frequency variations treated in [17] may have greater magnitude but fewer degrees of freedom.

Whereas the ML technique assumes that the disturbance is Gaussian white noise, in speech and music many other types of disturbing signals can be present: nonstationarity either in the desired component or in interfering ones; nonwhite noise; or time-modulated noise as occurs in some wind instruments. Here, the ML technique can not even be applied until a model is found for the disturbance. In these situations, the phase vocoder's generality is an important advantage.

REFERENCES

[1] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, Nov. 1966; also in *Speech Analysis*, R. W. Schaefer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[2] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 374–387, June 1981.

[3] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," *J. Acoust. Soc. Amer.*, vol. 94, pp. 662–667, Aug. 1993.

[4] J. M. Grey and J. A. Moorer, "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Amer.*, vol. 62, pp. 454–462, Aug. 1977.

[5] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 243–248, June 1976, also in *Speech Analysis*, R. W. Schaefer and J. D. Markel, Eds. New York: IEEE Press, 1979.

[6] J. A. Moorer, "The use of the phase vocoder in computer music applications," *J. Audio Eng. Soc.*, vol. 26, pp. 42–45, Jan./Feb. 1976.

[7] T. Wishart, "The composition of Vox-5," *Comput. Music J.*, vol. 12, pp. 21–27, Winter 1988.

[8] M. S. Puckette, "Phase-locked Vocoder," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.

[9] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 1986, pp. 113–116.

[10] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989, pp. 585–587.

[11] S. Kay, "A fast and accurate single frequency estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1987–1990, Dec. 1989.

[12] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," in *Proc. IEEE*, vol. 66, pp. 51–83, Jan. 1978.

[13] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 235–238, June 1977.

[14] D. E. Paneras, R. Mani, and S. H. Nawab, "STFT Computation using pruned FFT algorithms," *IEEE Signal Processing Lett.*, vol. 1, pp. 61–63, Apr. 1994.

[15] D. C. Rife and R. R. Boorstyn, "Single-tone parameter estimation from discrete-time observations," *IEEE Trans. Inform. Theory*, vol. 20, pp. 591–598, Sept. 1974.

[16] B. James and B. D. O. Anderson, "Characterization of threshold for single tone maximum likelihood frequency estimation," *IEEE Trans. Signal Processing*, vol. 43, pp. 817–821, Mar. 1995.

[17] B. Friedlander and J. M. Francos, "Estimation of amplitude and phase parameters of multicomponent signals," *IEEE Trans. Signal Processing*, vol. 43, pp. 917–926, Mar. 1995.

Miller S. Puckette received the B. S. degree from the Massachusetts Institute of Technology, Cambridge, in 1980 and the Ph.D. degree from Harvard University, Cambridge, MA, in 1986, both in mathematics.

He joined IRCAM Paris, France, and wrote the Max computer program. From 1979 to 1986, he also worked on real-time techniques for live music performance at the Massachusetts Institute of Technology (MIT) Media Laboratory, Cambridge. In September 1994, he joined the Department of Music, University of California at San Diego, La Jolla, CA, where he is now Professor of music. His current research interests include human-machine interaction strategies and real-time audio analysis, synthesis, and spatialization.

Dr. Puckette was the top scorer in the 1979–1980 William Lowell Putnam Mathematics Competition, and was awarded Putnam and NSF fellowships to study mathematics at MIT and Harvard. He also won *Keyboard* magazine's 1990 Software Innovation of the Year Award.

Judith C. Brown received the Ph.D. degreee in 1962 from the University of California, Berkeley. She continued her research as a postdoctoral fellow in France.

Since 1964, she has been a member of the Physics Department, Wellesley College, Wellesley, MA, becoming Professor in 1978. Beginning in 1970, she worked in the field of dynamic light scattering, publishing seminal papers on charge effects and polydispersity. Since 1986, she has been Visiting Scientist at the Massachusetts Institute of Technology Media Laboratory, Cambridge, where she has studied audio signal processing, emphasizing machine perception of musical signals. Her publications include work on rhythm, dynamics, vibrato perception, and pitch tracking. She has also been Visiting Scientist at the Royal Radar Establishment, Malvern, U.K. (1972–1974) and Visiting Researcher at IRCAM (1991 and 1994).

Dr. Brown is the recipient of Woodrow Wilson, NSF, NATO, and NIH fellowships.