Applied Acoustics **30** (1990) 219–245

An Investigation of Vocal Vibrato for Synthesis‡

Robert Maher* & James Beauchamp

Computer Music Project, School of Music, University of Illinois at Urbana-Champaign, 2136 Music Bldg, 1114 W. Nevada, Urbana, Illinois 61801, USA

ABSTRACT

High-fidelity singing synthesis requires careful consideration of the properties and character of natural vibrato. The research reported here evaluates some of the characteristics of vocal vibrato and presents a new panned-wavetable *synthesis method. The singing tone 'a' was recorded digitally and analyzed for bass, tenor, alto and soprano voices, providing time-variant measurements of amplitude, frequency and phase for each of the voice partials. Formant 'tracings' provide a novel method for examining vocal spectra during vibrato. The significance of vibrato waveform parameters and the role of* spectrum modulation *(due to partial amplitude fluctuations) during the vibrato cycle was investigated by resynthesis of the singing tones from modified analysis data. Informal listening to examples produced by the panned-wavetable synthesis model indicated that inclusion of typical random fluctuations of vibrato rate, vibrato depth and nominal sung frequency resulted in no quality preference over examples with constant values. Inclusion of vibrato-induced spectrum modulation resulted in a substantial improvement over examples having constant spectra.*

1 INTRODUCTION

Vocal vibrato is an essential aspect of trained singing. Vibrato is employed for emphasis and timbral variety in many singing situations and provides a valuable added dimension for expressive control. Indeed, the degree of

‡(PACS) Subject Classification numbers: 43.75.Rs 43.75.Bc 43.75.Wx.

* Present address: Department of Electrical Engineering, University of Nebraska, Lincoln, Nebraska 68588-0511, USA.

Applied Acoustics 0003-682X/90/$03·50 © 1990 Elsevier Science Publishers Ltd, England.
Printed in Great Britain

control over vibrato may be a good indication of a singer's skill and flexibility. Other observations suggest that vibrato may help mask any small unintentional errors in singing pitch.[1] The importance of vibrato has many implications for singing synthesis methods: vibrato must be treated with care if a convincing result is desired.

Our work on vocal vibrato began as an offshoot of a more general investigation of time-variant spectral analysis methods for musical instrument sounds. Most techniques employed for time-variant analysis use a bank of fixed-frequency bandpass analysis filters to separate an input signal into its presumably harmonic partials. The presence of vibrato in the analyzed tone may cause unwanted 'cross-talk' between the analysis filters, i.e. a partial may appear in the passband of two or more analysis filters during one vibrato cycle (see Fig. 1). Additional problems appear if a valid estimate of the time-varying amplitude, phase and frequency (phase derivative) of each partial is desired. The fundamental time-bandwidth product constraining the minimum analysis filter bandwidth to be inversely proportional to the observation time interval must also be confronted.

In an attempt to circumvent some of the problems inherent in the fixed-band approach we have adopted a time-variant analysis procedure based on a speech analysis/synthesis method of McAulay and Quatieri.[2] This method has already been found to provide useful results for the analysis of musical signals.[3,4] The method assumes *a priori* that the input signal can be represented adequately as the sum of a finite number of possibly inharmonic sinusoids with appropriate time-varying amplitude and frequency (or phase) modulation. This model has proved to be accurate for representing most

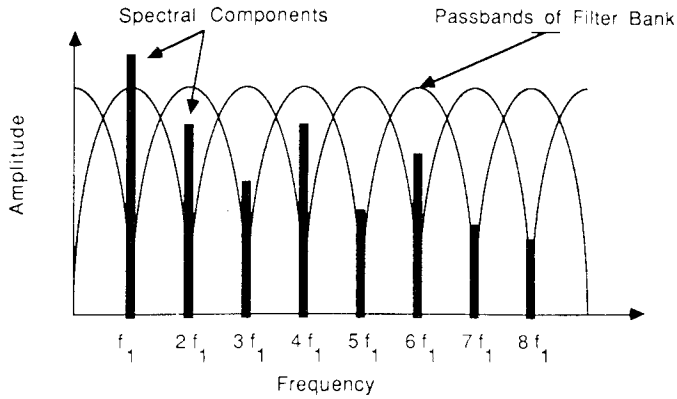


Fig. 1. A fixed-band harmonic analyzer employs a bank of bandpass filters centered at the harmonic frequencies. Vibrato causes the spectral components to shift back and forth in frequency, resulting in a misalignment ('cross-talk') between the analysis filters.

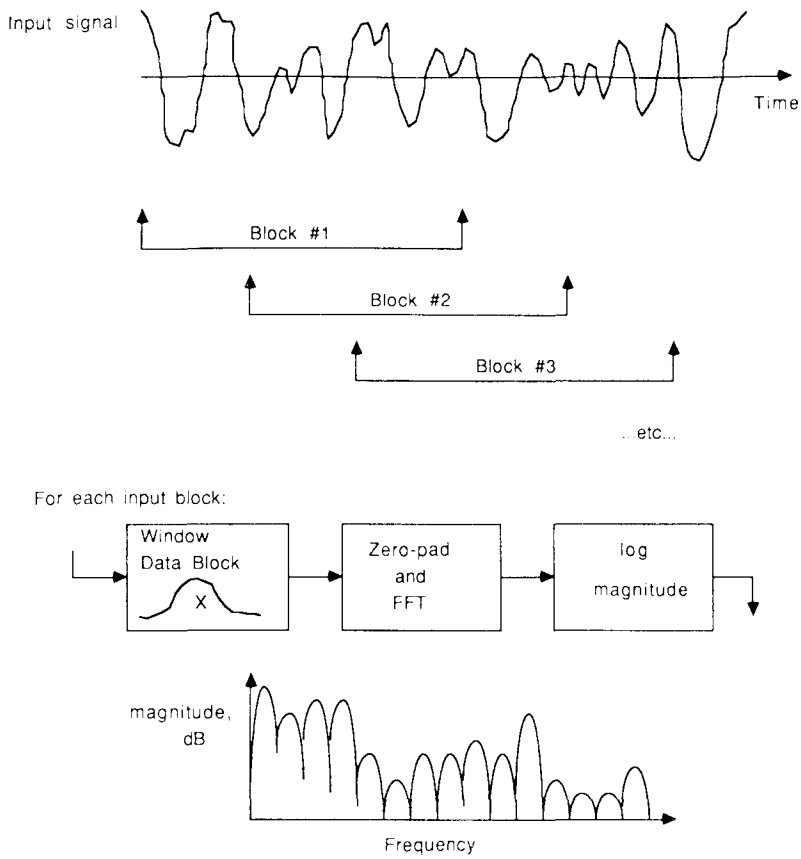
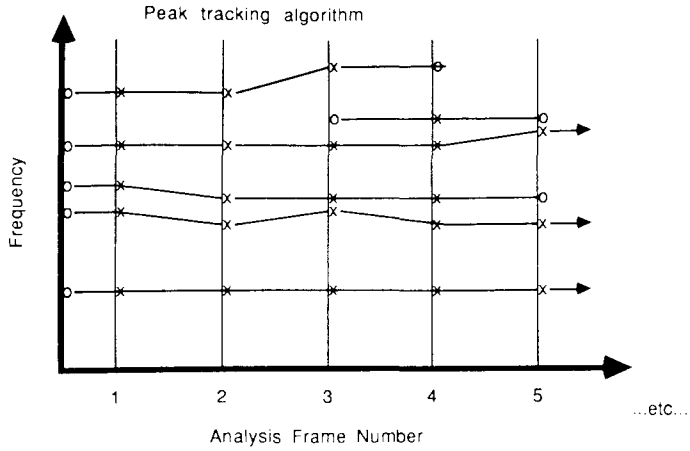


Fig. 2. Summary of the McAulay–Quatieri analysis procedure. The input signal is segmented into overlapping analysis blocks, or 'frames'. The log-magnitude spectrum of each block is computed, and the location and height of spectral peaks are identified.

musical sounds. Moreover, we have found it to be an improvement over fixed-band methods for the analysis of vibrato tones since it is possible to track changing frequencies, thereby avoiding the inter-band cross-talk problem.

In our version of the McAulay–Quatieri procedure (Fig. 2), the digitized input signal is divided into windowed, overlapping time segments or 'frames'. For each frame (zero-padded by a factor of two or more) we compute the complex spectrum and the log-magnitude spectrum, from which all spectral peaks are identified. The model assumes that each spectral peak is due to the presence of an underlying sinusoidal partial at the location of the peak. We then estimate the magnitude, frequency and phase of each partial from the log-magnitude and complex spectral data. The peaks are tracked from frame to frame and connected into frequency 'tracks'. Because



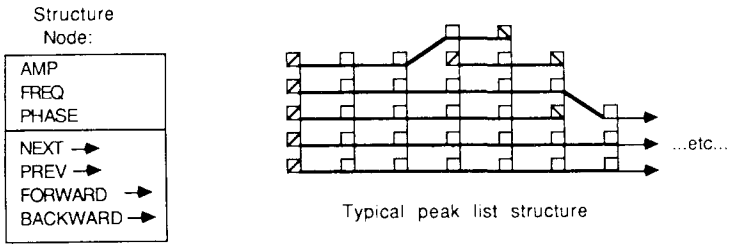


Fig. 3. The peaks from one analysis frame are connected with corresponding peaks in previous and subsequent frames. A data structure is constructed containing the amplitude, frequency and phase measured at each peak, and a set of pointers to each adjacent element (linked-list).

only those partials whose amplitudes exceed a certain threshold are retained, the number of tracks present can vary considerably from frame to frame. Thus, frequency tracks are continually being born and dying as the spectral content of the input signal changes from instant to instant. The collection of frequency tracks provides a connected mesh of amplitude, frequency and phase estimates for each partial as a function of time (Fig. 3).

The corresponding synthesis method is essentially a sum-of-sinusoids additive synthesis procedure, but polynomial interpolation of the analyzed phase data is employed to provide a smoothly varying phase function from frame to frame for each partial component. The technique is well suited for inharmonic sounds.[3,5] We have obtained useful results for tones containing vibrato and even for analysis/synthesis of polyphonic music (Maher[4]). Our implementation of the McAulay–Quatieri procedure will be referred to as the MQ technique for the remainder of this paper.

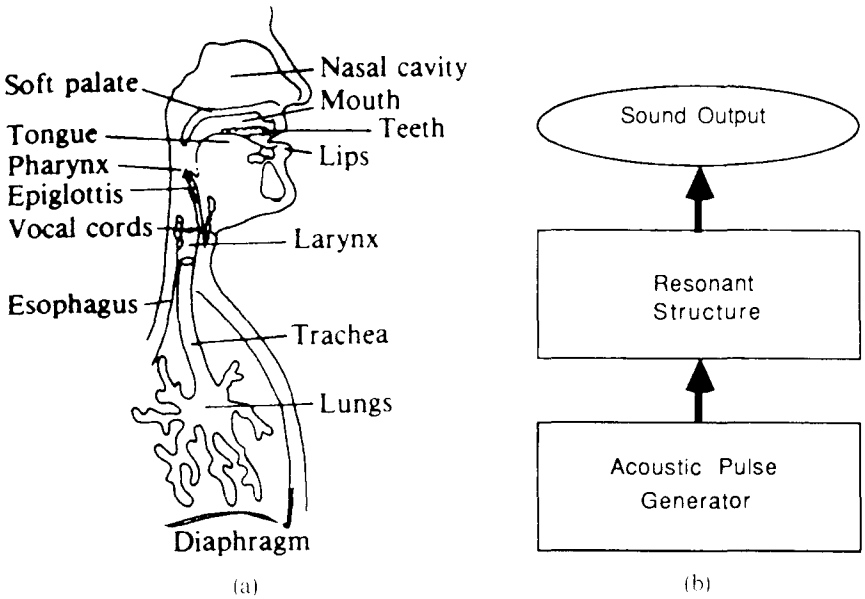


Fig. 4. A schematic model for the vocal apparatus. (a) Cutaway diagram of the human vocal system (after Rossing *et al.*[9]). (b) Simple block diagram consisting of an acoustic pulse generator (lungs and vocal cords) and a resonant structure (pharynx, oral cavity, etc.).

Several approaches to singing synthesis have been reported (cf. Refs 1 and 6–8). All of the methods attempt to capture the formant properties of the singing voice, either by direct simulation of the human vocal apparatus or by mimicking the voice spectrum itself. In general, the human vocal system has been treated as a multi-resonant structure (the pharynx and oral cavity) excited by an acoustic pulse generator (the vocal folds and breathing apparatus), as shown in Fig. 4. In this model the output of the vocal system is characterized by regions of spectral emphasis (formants) due to the implicit convolution of the vocal fold signal with the resonant impulse response of the vocal tract. Thus the goal has been to identify and match the salient spectral properties of the human singing voice.

In the case of vibrato most reports indicate, perhaps tacitly, that the resonant character of the vocal tract remains FIXED while the excitation from the vocal folds changes frequency in some quasi-sinusoidal manner.[1,9,10] If the amplitude of each partial in the source spectrum of the vocal folds is considered to be nearly constant throughout the vibrato cycle, the output spectrum of the singing voice will show frequency modulation due to activity of the vocal folds and amplitude modulation due to the source partials being swept back and forth through the vocal tract resonances. Although trained singers, particularly sopranos, are able to shift

the spectral location of certain formants to coincide with a source partial,[11] it is reasonable to assume that formant positions remain fixed as partial frequencies fluctuate according to a vibrato pattern, causing the amplitudes of the partials to also vary in time.

This vibrato-induced amplitude modulation of the partials leads to a time-varying magnitude spectrum, or 'spectrum modulation', occurring at the vibrato rate. We did not know in advance whether spectrum modulation might be a vital ingredient for a convincing synthesis, because the effects of natural vibrato on timbre perception have not been evaluated by psychoacoustic testing.[12] Therefore, it seemed prudent to investigate this aspect of natural vibrato in our experimental synthesis models. Also, since previous reports have stressed the importance of random variations in vibrato,[1,7] we decided to include in our study an analysis of the deviations from strict periodic behavior found in vocal vibrato waveforms.

The work reported here is a preliminary evaluation of various parameters of vibrato in singing. The results are intended to guide further work in this area, particularly in economical synthesis techniques. In several respects the approach and conclusions of this paper can be viewed as independent extensions and refinements of the work reported by Bennett.[1]

2 RESEARCH PLAN

Four singers (soprano, alto, tenor and bass) were recorded individually while singing the vowel 'a' at a comfortable output level and range of pitches. All singers had received some formal training, but none was a professional vocalist. No special instructions were given to the singers, except to relax and sing naturally.

The recordings were made in a nonreverberant room with a high-quality professional microphone. The soprano recordings were obtained using a Sony PCM-501ES digital tape system, while the alto, tenor and bass recordings were obtained from a previous study.[13] Several example tones from each singer were then rerecorded digitally at a 20 kHz sample rate using the Sound Conversion and Storage System (SCSS), built at the University of Illinois.[14] Once on the SCSS, the digitized tones were available as 'sample files' for processing and analysis using software on a general purpose digital computer (an IBM Model 125 RT PC workstation).

The sample files were analyzed first using the MQ procedure. Next, a fundamental frequency tracking technique was employed to identify and extract those partials which were closest to being the harmonics of the signal within each frame. The frequency domain histogram method reported by Schroeder[15] was chosen for fundamental frequency tracking since the

method operates on a list of partial frequencies, automatically provided by the MQ analysis procedure.

Since the fundamental frequency $f_1(t)$ varied in time, we assumed that the frequency of the kth partial would vary according to

$$f_k(t) \approx k f_1(t) \tag{1}$$

We should emphasize that this frequency extraction process does not inherently restrict partials to be perfect harmonics, but we expect they would be for most nonpercussive musical instrument and vocal sounds. An example of the raw output of the MQ analysis procedure is shown in Fig. 5(a). The quasi-harmonic form extracted after tracking the fundamental is given in Fig. 5(b).

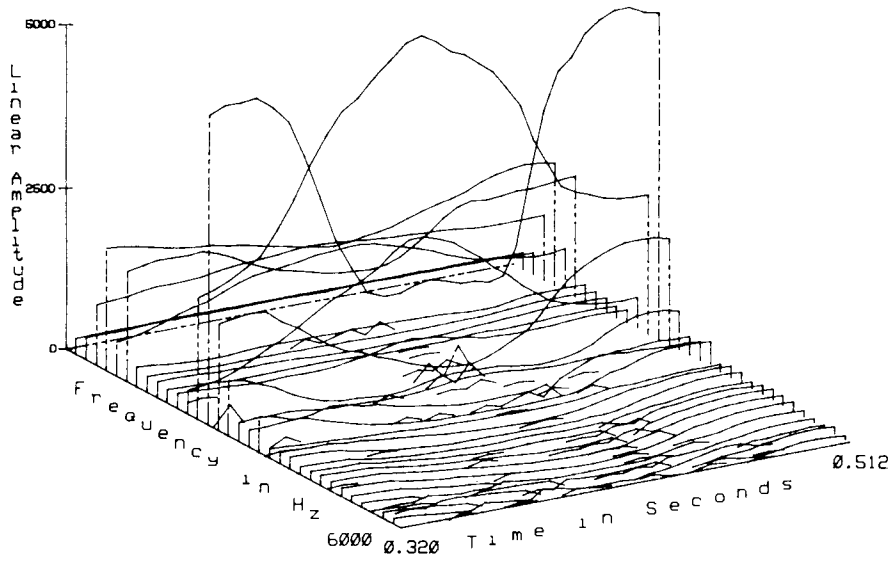
Finally, the quasi-harmonic voice partials were assembled as a data file consisting of the initial phase of each partial followed by the amplitude and frequency change measurements obtained for each partial at each analysis frame. A time domain signal could be resynthesized directly from the original quasi-harmonic data or from modified analysis data.

We investigated the significance of various time-varying vibrato parameters by generating several synthetic tones based on the MQ analysis data. A list of resynthesis alterations is given in Table 1. The MQ procedure supplies information on the amplitude and frequency evolution of each partial, so it was convenient to perform independent modifications of these domains. For example, frequency variations could be omitted while retaining amplitude variations, and vice versa.

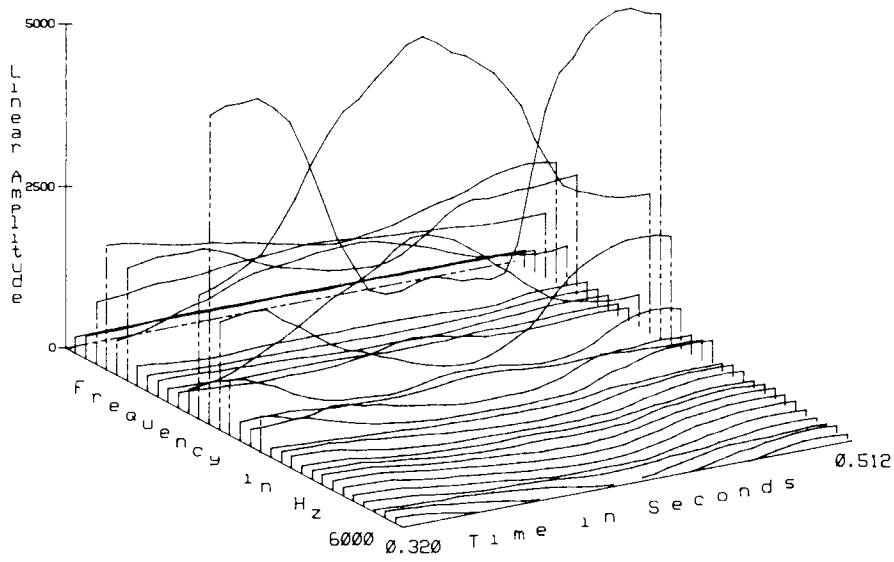
The importance of spectrum modulation due to vibrato was investigated by resynthesizing tones with (1) measured amplitude fluctuations for each partial, but with constant partial frequencies replacing the measured frequency oscillations, and (2) measured frequency oscillations, but with constant (time average) partial amplitudes.

TABLE 1
Resynthesis Alterations Applied to Analysis Data for Each Sung Vowel

(a)	No modifications (direct resynthesis from the MQ analysis)
(b)	Replacement of frequency vibrato of each partial with a 'composite' vibrato waveform
(c)	Additive resynthesis with fixed number of quasi-harmonic partials (partials between 0 and 4 500 Hz)
(d)	Elimination of partial amplitude fluctuations (synthesis with fixed partial amplitudes and measured individual partial vibratos)
(e)	Elimination of partial frequency vibratos (synthesis with fixed perfect harmonic frequencies and measured modulation on each partial)
(f)	Combination of (d) and (e) (fixed waveform synthesis)
(g)	Combination of (b) and (d)



(a)



(b)

Fig. 5. Example of the conversion from a full MQ analysis to quasi-harmonic form. (a) A portion of the MQ analyzer output for a sung tenor tone. (b) The example of (a) following quasi-harmonic extraction based on fundamental frequency tracking.

The effect of possible inharmonicity in vibrato tones was examined by replacing the individual partial vibratos with a 'composite' frequency vibrato waveform. The composite vibrato waveform was obtained by averaging the frequency waveforms measured for each partial. The contribution of each partial to the composite vibrato was weighted by its amplitude; thus, strong partials contributed more to the composite than weak partials. For a given analysis frame, the composite frequency is given by

$$f_{\text{composite}} = f_0 + \frac{\sum_{k=1}^{P} a_k (f_k - k f_0)/k}{\sum_{k=1}^{P} a_k} \tag{2}$$

where

P = the number of partials
a_k = measured amplitude of partial k
f_k = measured frequency of partial k
f_0 = assumed fixed target frequency.

The results were examined via several graphic display routines and audio playback. Evaluation was informal and intended to guide further research in computer music synthesis. Emphasis was on determining guidelines for vibrato generation that would be appropriate for inclusion in various synthesis algorithms.

To test the utility of the measured parameters we developed a simple software synthesis algorithm having a favorable tradeoff between computational efficiency and synthesis quality.

3 RESULTS AND DISCUSSION

We found direct resynthesis from the original MQ analysis data (where all tracks are retained and phase interpolation is used) to be of excellent quality. The results were nearly indistinguishable from the original recordings, based on informal auditions by several critical listeners who were not biased by familiarity with the analysis method nor the research goals. Examples were also generated from the analysis data by simple additive resynthesis employing a fixed number of quasi-harmonic, time-varying partials. While we found the initial attack quality to be inferior to the result obtained by the direct resynthesis method, the simple resynthesis provided good results for the sustained portions of the sung vowels.

The tones were further altered by removing partial amplitude fluctuations, frequency fluctuations (or both), and by replacing individual partial

TABLE 2
Rating of Resynthesis Examples in Order of Decreasing Fidelity

(1)	Original recording
(2)	Unaltered MQ synthesis
(3)	Simple additive synthesis of quasi-harmonic partials
(4)	Resynthesis of (3) with the composite frequency vibrato used on all partials
(5)	Resynthesis of (3) without amplitude fluctuations on the partials
(6)	Resynthesis of (4) without amplitude fluctuations on the partials
(7)	Resynthesis of (3) without frequency vibrato
(8)	Resynthesis of (3) without frequency vibrato or amplitude fluctuations

frequencies with frequencies harmonically related to the composite fundamental frequency (defined by eqn (2)). We used informal listening to rank the several resynthesis examples in order of decreasing fidelity, as shown in Table 2. The largest degradation in quality was noted between methods 6 and 7. The differences between methods 1, 2 and 3 were very slight, and methods 5 and 6 were essentially indistinguishable. However, methods 5 and 6 (constant amplitude partials) provided substantially different quality from the similar methods 3 and 4, which included amplitude fluctuations.

Table 3 summarizes several measured characteristics of vibrato for the vocal tones used in this investigation. We note the reasonable consistency of these parameters. Vibrato rate is most consistent, being between 5·0 and 5·7 Hz. The alto tone had the smallest vibrato depth ($\pm 2{\cdot}5\%$) and the smallest dB ripple over all partials. The ripple of the RMS amplitude (given by the square root of the sum of the squares of the partial amplitudes) is substantially smaller than the maximum for the individual partials because the maxima and minima of different partials are frequently out of phase and tend to compensate for one another.

3.1 Vibrato differences among partials

One aspect of interest to us was the possibility of time-varying inharmonicity during the vibrato cycle. Due to the frequency-dependent phase response of the vocal tract, the partials of a singing voice could undergo a frequency-dependent group delay (i.e. phase derivative) relative to one another. As the fundamental and other partial frequencies oscillate during the vibrato cycle, they could be inharmonic with respect to each other in direct relation to the product of the vibrato rate and the difference of the group delays for each two frequencies compared. In fact, some of our analyses of musical instrument tones with vibrato (e.g. violin and oboe tones) have revealed significant momentary inharmonicities. On the other hand,

TABLE 3
Some Typical Measurements of Vibrato Characteristics in Sung Vowels. Vowel Sound 'a'

Voice	*Musical pitch*	*Vibrato rate (Hz)*	*Vibrato depth*	*Max. ripple of any partial (in dB)*	*Max. ripple of RMS amplitude (in dB)*
Soprano	C♯5 (554·4 Hz)	5·7	±5%	10	3
Alto	D4 (293·7 Hz)	5·6	±2·5%	4	1
Tenor	G3 (196·0 Hz)	5·3	±4·2%	11	0·5
Bass	C3 (130·8 Hz)	5·0	±4·5%	8	0·5

assuming that the vocal system can be represented by a standard source/fixed-filter model having about five fairly broad formant resonances,[16] we expect that the group delay variation will be minimal (unlike the case of the violin, which is characterized by numerous overlapping narrow band resonances), indicating that inharmonicity effects due to vibrato should be insignificant in vocal tones. However, another effect, due to the possibility for modulation of the vocal tract parameters during vibrato, thereby affecting the individual partial frequencies differently, could not be ruled out. Thus, we chose to perform an analysis of time-varying inharmonicity. The results were that the analyses revealed some subtle inharmonicities, but the effects were of the same order as our analyzer's inherent frequency accuracy limitation of approximately 1 Hz and appeared to be too small to be perceivable. Indeed, informal listening tests of tones synthesized with and

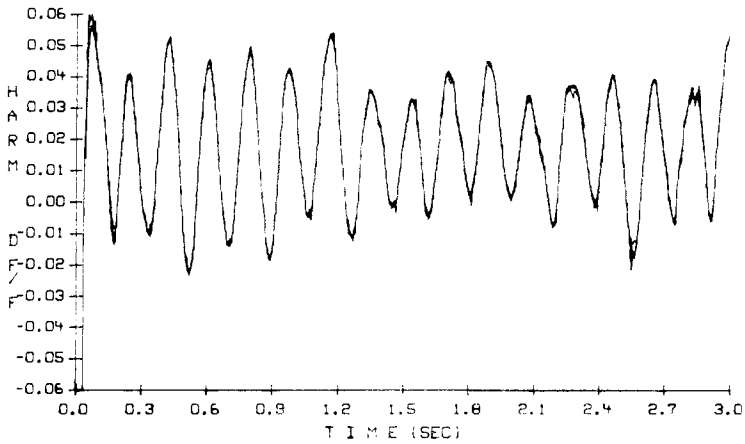


Fig. 6. Normalized measured frequency deviation for each of the partials of a typical alto tone, shown overlapped. Normalized deviations calculated by $(f - kf_0)/(kf_0)$, where f = measured frequency in Hz, f_0 = assumed fundamental frequency in Hz and k = partial number ($k > 0$, integer).

without the detected inharmonicities indicated that the differences were barely distinguishable. Thus, we conclude that inharmonicity effects can be ignored in vocal vibrato tones. Visually, this seems obvious from Fig. 6, which shows superimposed normalized frequency measurements for a number of partials of a typical alto tone.

The results also clearly show a complex relationship between the time-varying fundamental frequency and the individual partial envelopes. The prominence and qualitative complexity of the amplitude fluctuations is

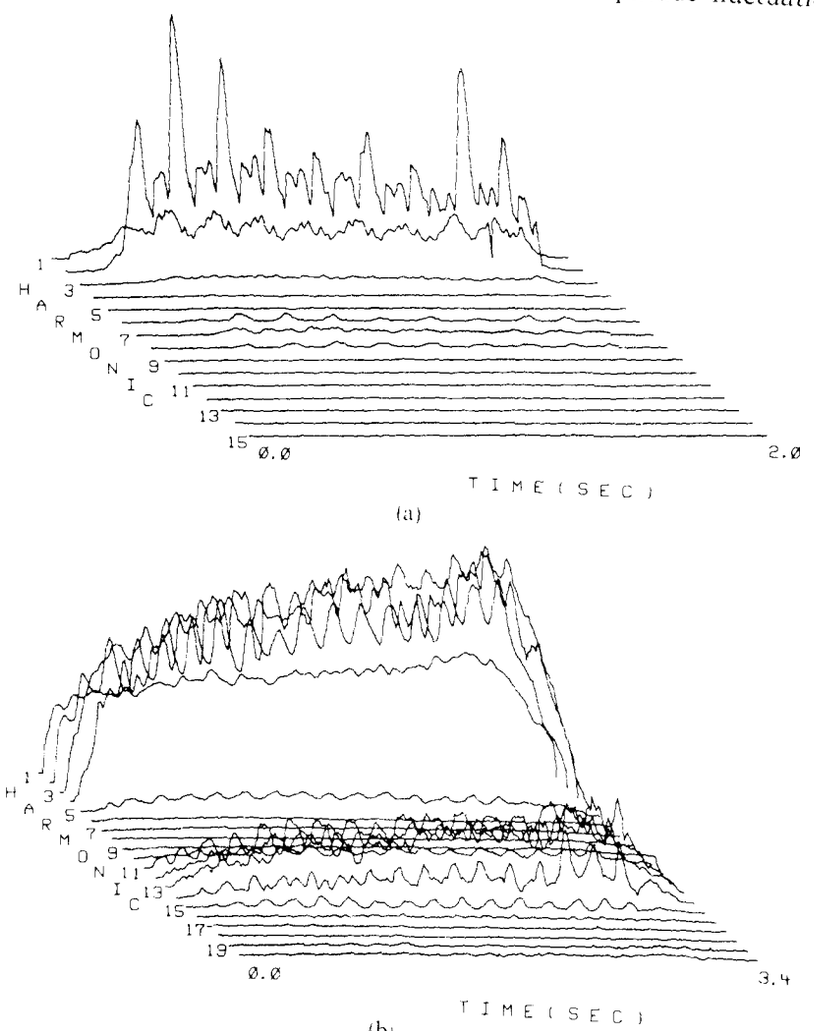
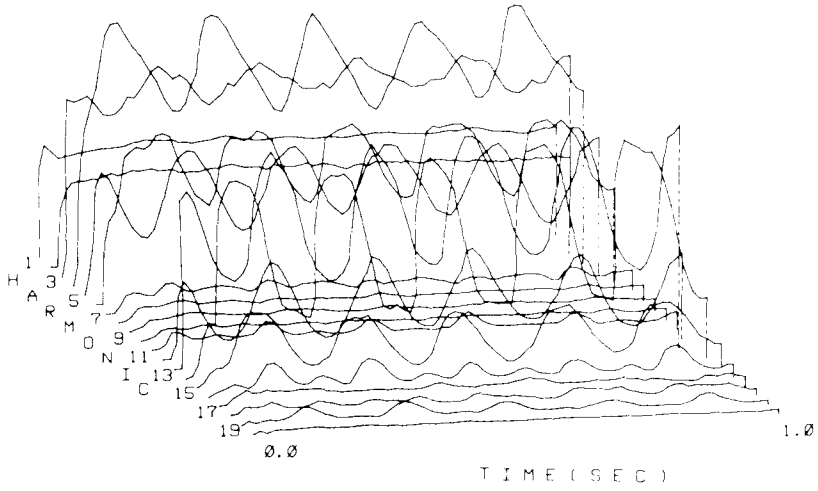


Fig. 7. Typical time-variant spectra for four singers using vibrato: (a) soprano, C5; (b) alto, D4; (c) tenor, G3; (d) bass, C3. Amplitude (linear scale) is the vertical dimension. Quasi-harmonic data were extracted from the full MQ analysis.

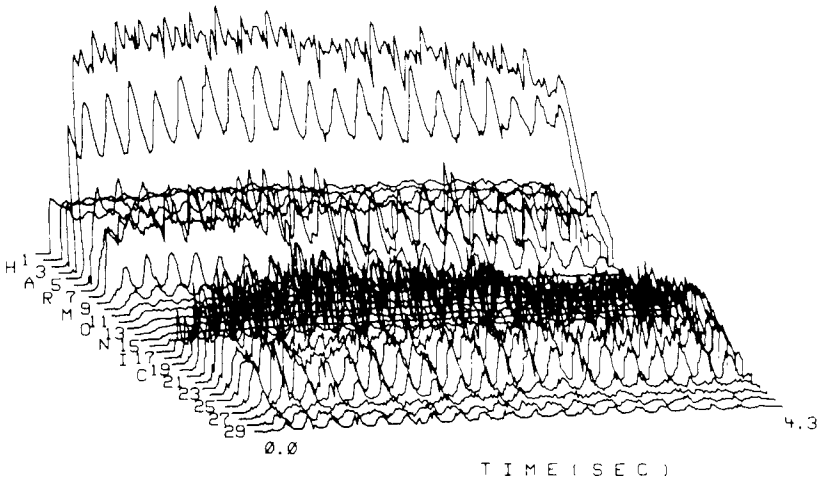
illustrated by the time-variant analysis data for four typical vocal tones shown in Fig. 7.

The phase relationship between the time-varying fundamental frequency and the amplitude fluctuation of an individual partial can be used to identify the position of that partial relative to a vocal tract resonance in the following ways:[17]

(i) If the amplitude of a partial is *in phase* with the frequency vibrato (i.e. it increases in amplitude as the frequency increases), the partial is on the low side of a resonance peak (see Fig. 8(a)).



(c)



(d)

Fig. 7 *contd.*

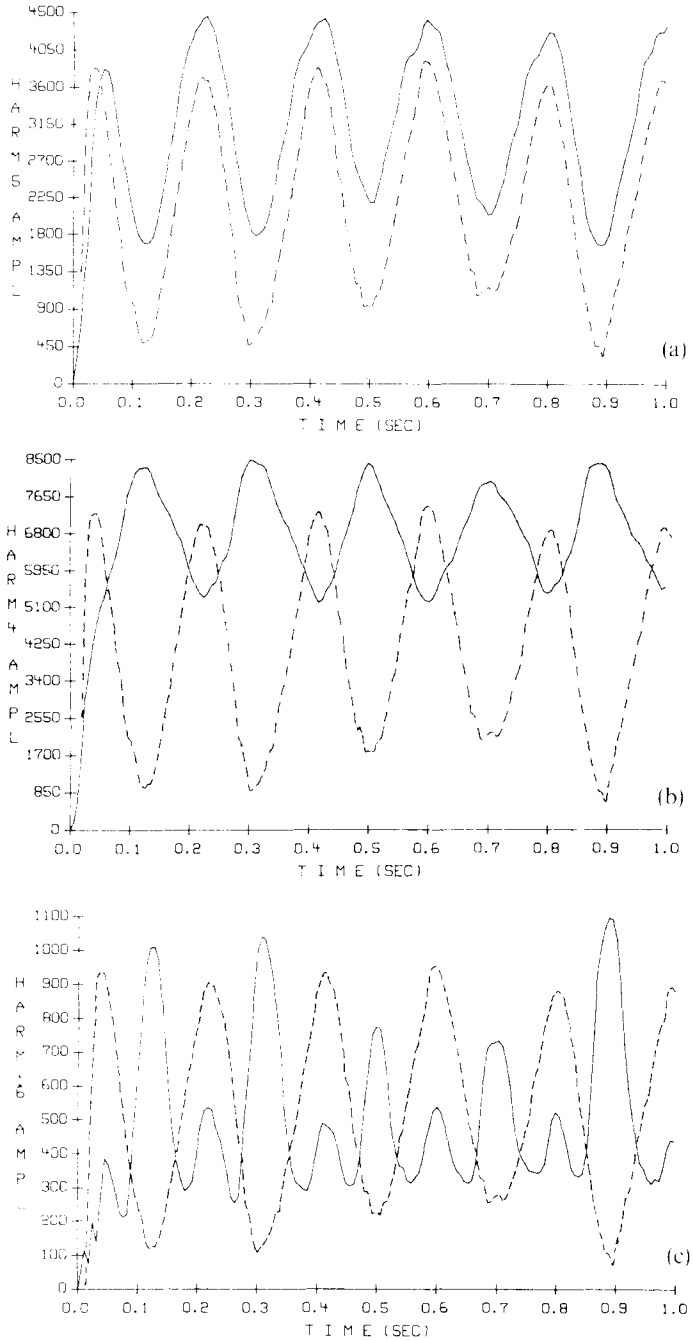


Fig. 8. Relationship between frequency of partial and a vocal resonance. Solid line is amplitude, dotted line is frequency. (a) Partial located on low-frequency shoulder of resonance peak; amplitude fluctuations *in phase* with vibrato. (b) Partial located on high-frequency shoulder of resonance peak; amplitude fluctuations 180 degrees *out of phase* with vibrato. (c) Partial located near resonance peak or trough; amplitude fluctuations at *twice* the vibrato rate.

(ii) If the amplitude of a partial is *out of phase* with the frequency vibrato, the partial is on the high side of a resonance peak (see Fig. 8(b)).
(iii) If the amplitude of a partial contains a component *changing at twice the vibrato frequency*, the partial is swept across a resonance peak or trough (see Fig. 8(c)).

The amplitude versus frequency behavior can be observed more clearly when the data are plotted as the locus of [frequency, amplitude] coordinates of each partial during vibrato. Figure 9 shows the spectral envelopes of the sung 'a' vowels for each of the four singers as traced by the partials during several vibrato cycles. The striking feature of these 'scribble-plots' is the ease with which the spectral position and shape of several formants may be identified. Although the excitation spectrum from the vocal folds and the frequency response of the vocal tract have *not* been separated here, the multi-resonant character of the output spectrum is obvious. Unfortunately, with increasing fundamental frequency, the spacing of the voice partials increases. This provides a coarser sampling of the underlying spectral envelope and a corresponding decrease is formant information available from the plots, as evidenced by the data from the female singers in Fig. 9(a) and (b).

What is the perceptual importance of the amplitude fluctuations during the vibrato cycle? Although the tracing of formant characteristics due to vibrato might indicate a means for a listener to identify more features of the spectral envelope—thereby increasing the ease of vowel identification—this has not been fully confirmed.[10,18] Perhaps most significant is that timbral variation occurs throughout the vibrato cycle. That periodic timbral fluctuation is perceptually obvious (in the absence of frequency vibrato) was demonstrated in our resynthesis examples where frequency vibrato had been removed, leaving only the spectrum modulation to be heard.

The perceptual importance of the amplitude fluctuations for convincing synthesis was also noted from our resynthesis examples where the partial amplitudes were held constant. In general, it was apparent that a certain warmth and natural quality found in the resyntheses including amplitude fluctuations was missing from the frequency-modulation-only examples. Also, some listeners noted that the inclusion of spectrum modulation seemed to aid perceptual fusion. With the amplitude fluctuations omitted, the upper partials produced a 'whistling effect' and did not blend well with the lower partials. It is possible that some of the unnatural character of certain economical synthesis methods—such as complex frequency modulation (FM)[7]—might be attributed to the unnaturally constant partial amplitude behavior of the synthesized tones.

We conclude that the time variation of vowel timbre due to frequency

vibrato provides a richness and sonic variety characteristic of good singing quality and can profitably be included in models for vocal vibrato synthesis.

3.2 Analysis of vibrato waveforms

The vibrato waveforms obtained from the singing examples were analyzed to evaluate their statistical properties. The long-term (1–2 s) average

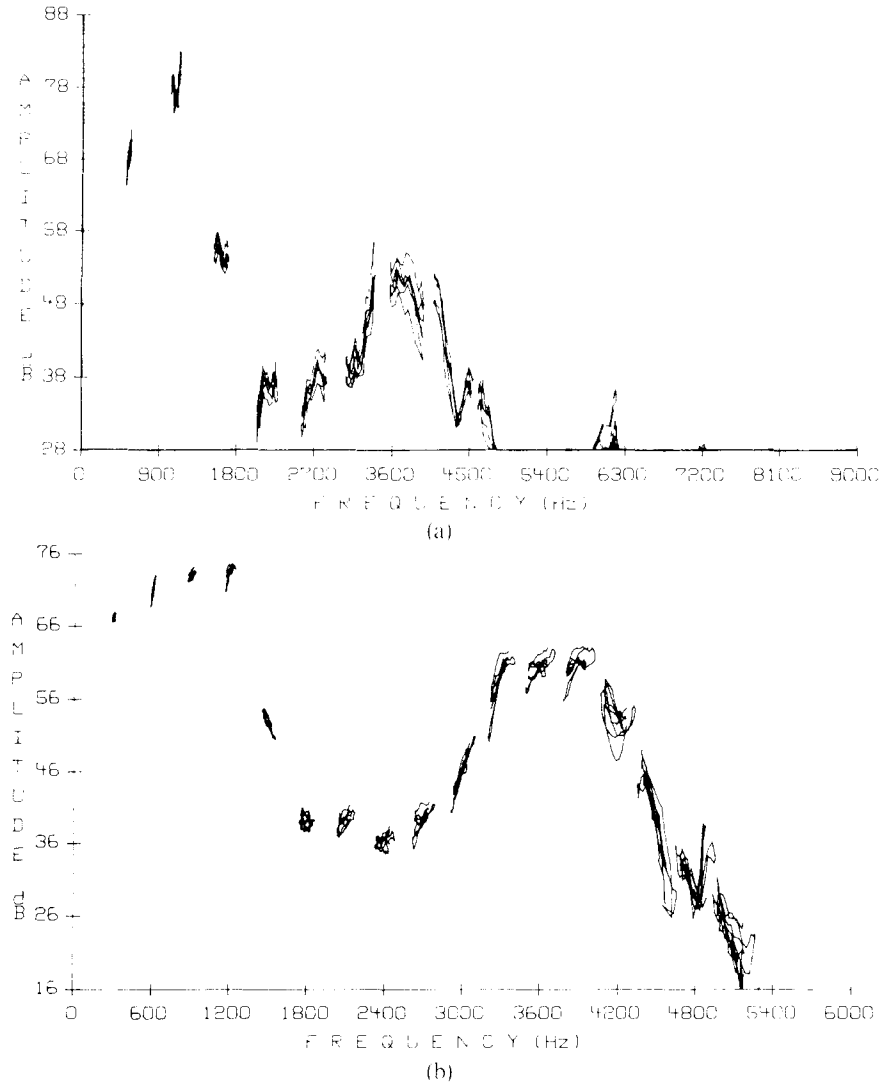
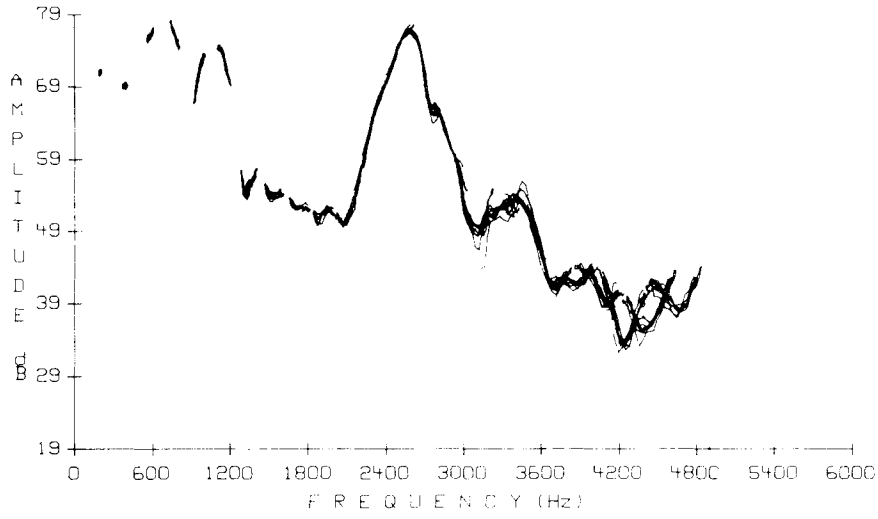


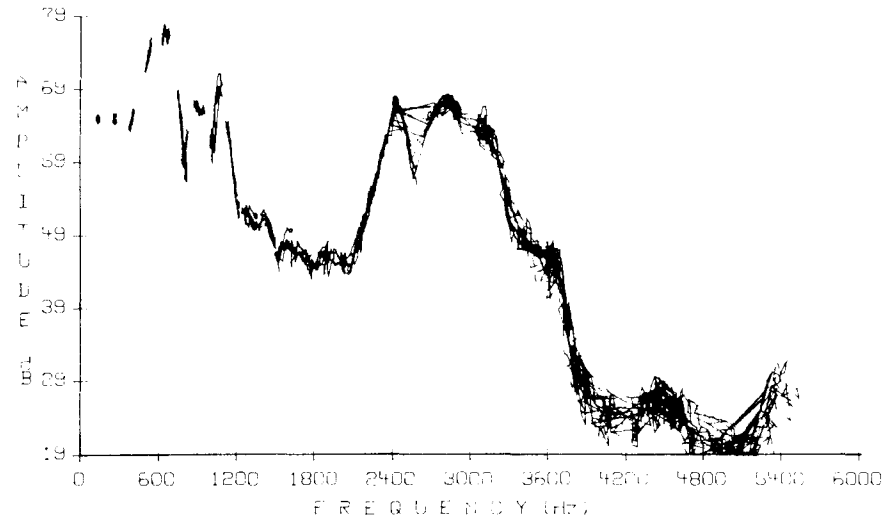
Fig. 9. Spectral tracings for four singers using vibrato. Each partial traced its amplitude versus frequency value over several vibrato cycles: (a) soprano, C5; (b) alto, D4; (c) tenor, G3; (d) bass, C3.

spectrum of each vibrato waveform was computed, and a time-variant analysis using fixed-band filters centered at the nominal vibrato frequency and its harmonics was performed.

Analyses of the vibrato waveforms (see examples, Fig. 10) showed nearly sinusoidal frequency changes, with amplitudes (vibrato depths) varying between $\pm 2\%$ and $\pm 5\%$ of the fundamental frequencies. The vibrato waveforms also exhibited a low frequency 'drift' above and below the *target*



(c)



(d)

Fig. 9 *contd.*

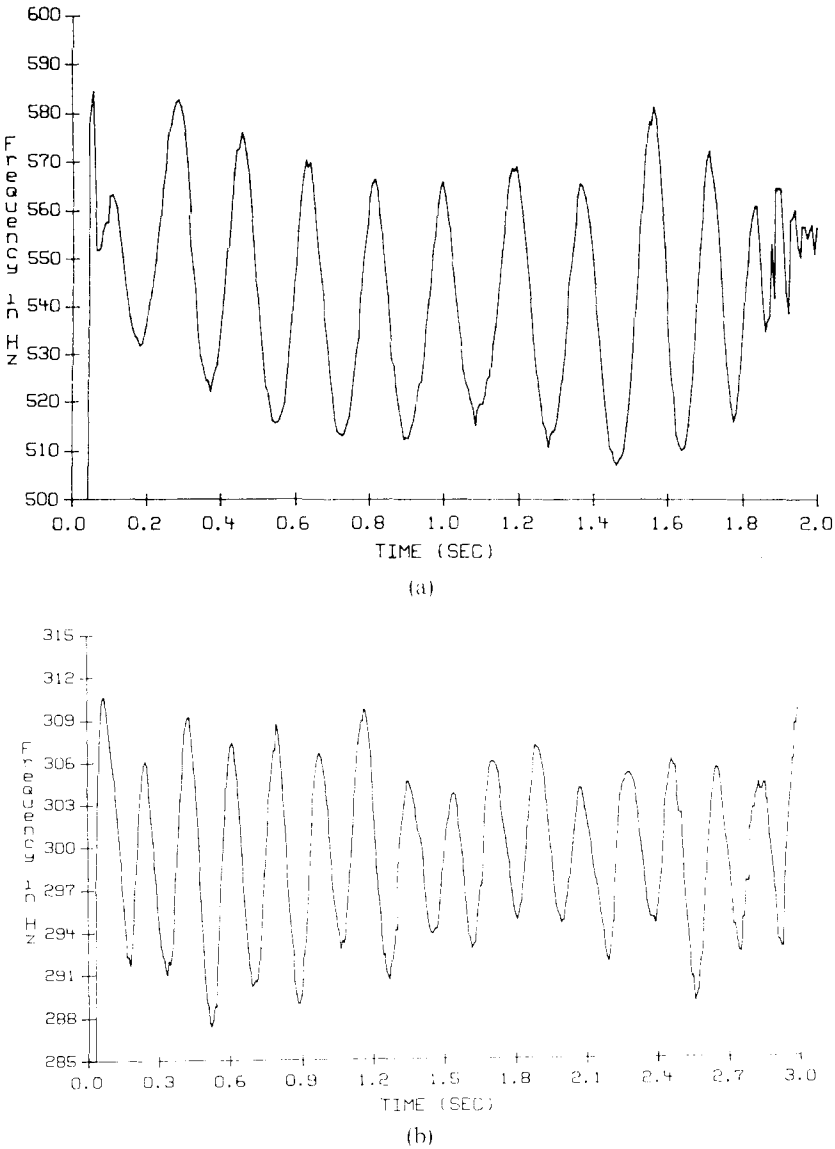


Fig. 10. Vibrato waveforms for four singers: (a) soprano; (b) alto; (c) tenor; (d) bass; (e) long-term average spectrum for tenor vibrato waveform.

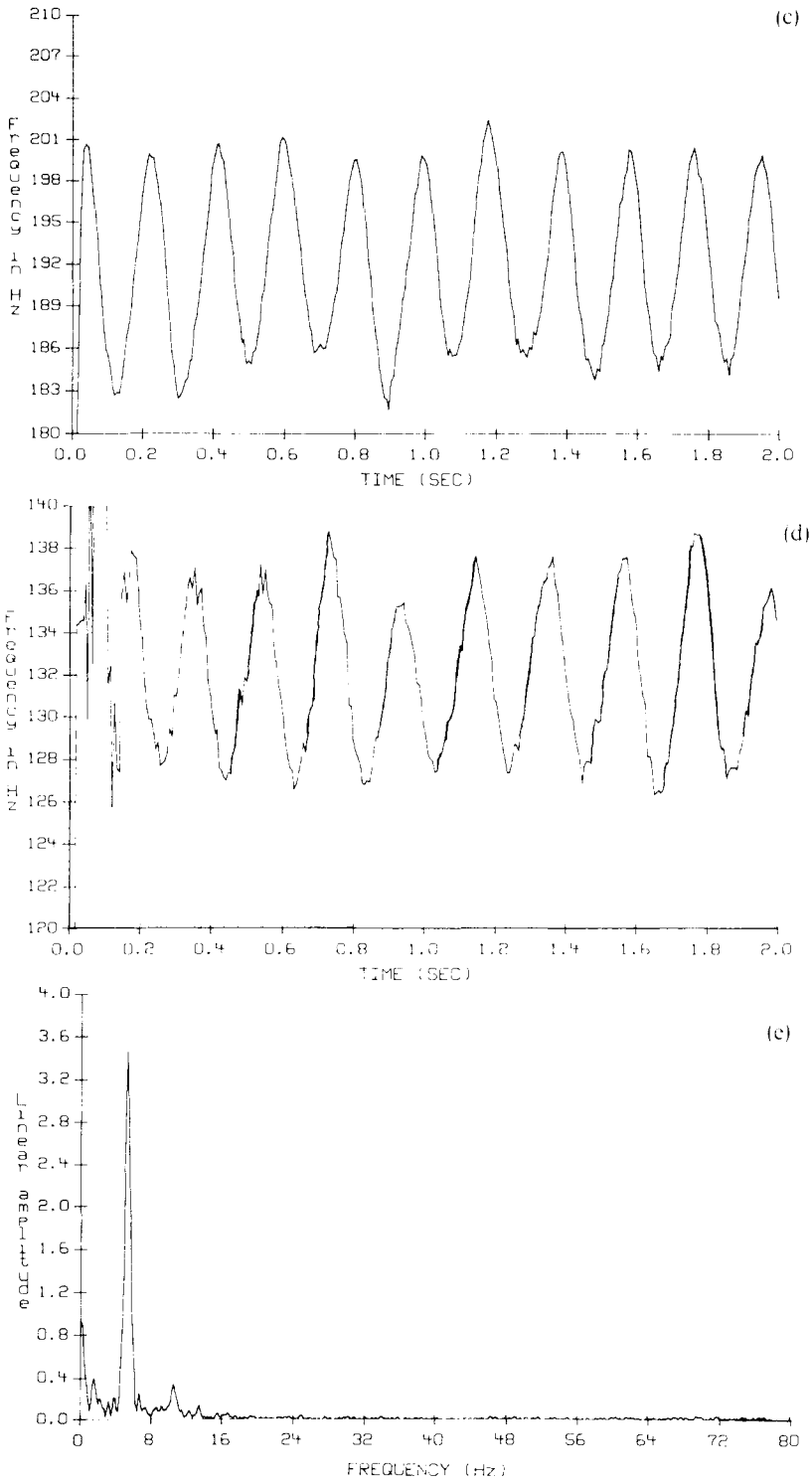

(c)
FREQUENCY IN HZ
TIME (SEC)
(d)
FREQUENCY IN HZ
TIME (SEC)
(e)
AMPLITUDE
FREQUENCY (Hz)

Fig. 10 *contd.*

fundamental frequency with an extent on the order of $\pm 1\%$ of the fundamental ($1\% \approx 17$ cents). The drift appeared to be random and, as indicated by the long-term spectral average of the vibrato waveform (example shown in Fig. 10(e)), had a bandwidth of one-third the nominal vibrato rate for the male singers and approximately one-half the nominal rate for the female singers. The vibrato rate itself varied over the duration of the note, with maximum deviation of $\pm 10\%$ of the nominal vibrato rate. An expression for the fundamental frequency of a sung tone with vibrato is

$$
\begin{aligned}
&f_1(t) = f_0 + d_f(t) \qquad d(t) = a_1 r_1(t) \\
&d_f(t) = d(t) + \Delta f(t) \sin(\theta_v(t) + \phi_0) = \text{frequency deviation} \\
&\theta_v(t) = 2\pi \int_0^t f_{v0}[1 + a_3 r_3(\tau)]\, d\tau = \text{phase of fundamental} \qquad (3) \\
&\Delta f(t) = f_\Delta[1 + a_2 r_2(t)] = \text{vibrato depth in Hz} \\
&\theta_1(t) = 2\pi \int_0^t f_1(\tau)\, d\tau
\end{aligned}
$$

where

r_k = random numbers in range -1 to $+1$, with specified bandwidth
a_k = extent scaling of random variations
f_Δ = nominal vibrato depth
f_{v0} = nominal vibrato rate
f_0 = target fundamental frequency in Hz.

Equation (3) could also be written with the frequency parameters expressed as fractions of the target fundamental frequency f_0. Parameter measurements for four vowel tones are given in Table 4.

It is important to realize the significance of the *variations* in vibrato rate,

TABLE 4
Typical Vibrato Parameters and Random Variations

Voice	*Random depth variation* a_2		*Random drift* a_1		*Random vibrato rate variation* a_3	
	Extent (% of f_Δ)	*Bandwidth* (*Hz*)	*Extent* (% of f_0)	*Bandwidth* (*Hz*)	*Extent* (% of f_{v0})	*Bandwidth* (*Hz*)
Soprano	30	2	0·6	5·0	8	5
Alto	30	3·5	0·7	2·5	8	3
Tenor	15	5·5	1	3·5	10	3·5
Bass	25	5	1	3·5	10	3

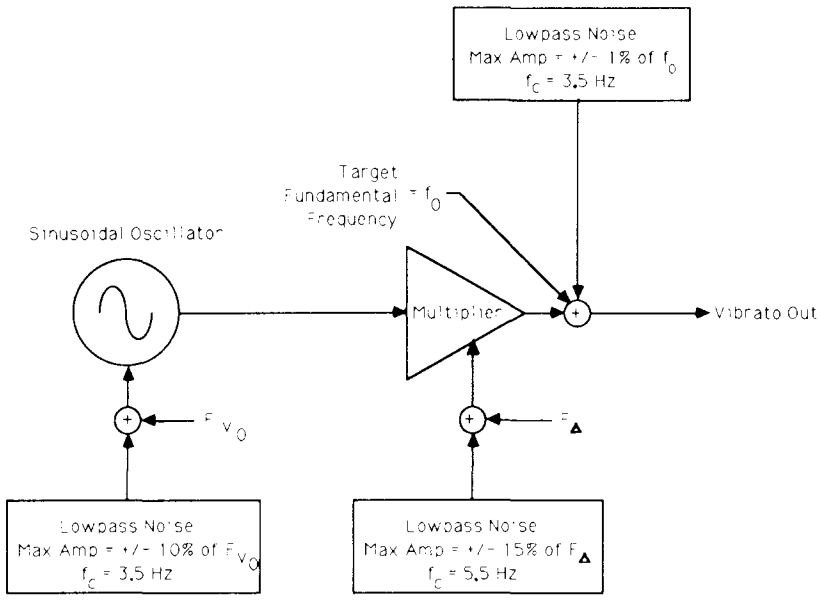
vibrato depth and fundamental frequency. Although a trained singer can adjust the *target* values of these parameters, the 'random' variations about the desired values do not appear to be controlled by the singer in any conscious manner.[1] The maximum measured deviation from the target values may give some indication of the 'looseness' of the vocal control system. The singer evidently uses feedback from the sensory and hearing systems to adjust the vocal fold apparatus in the direction of target values whenever the random drift is greater than some tolerable amount.[19] The ± 17 cents drift extent measured here for solo vocal tones with vibrato is comparable to the ± 15 cents frequency deviations between singers in a choir reported by Ternstrom and Sundberg.[20] It is unclear whether this agreement represents a typical limitation of the vocal/auditory feedback system. In any case, it has been reported that inclusion of appropriate random fluctuations in singing synthesis is vital to avoid a mechanical, unnatural sound.[1,7]

3.3 Computer model for singing synthesis with vibrato

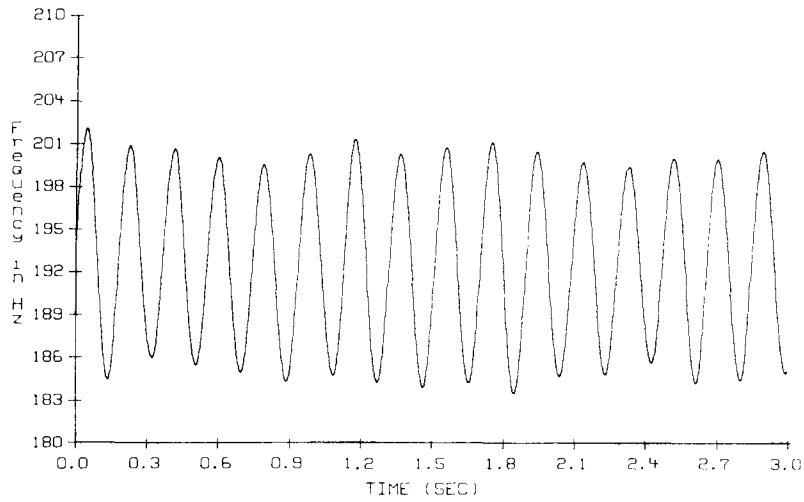
A proposed block diagram for 'natural' vocal vibrato is given in Fig. 11. The model has two main parameters: vibrato rate f_{v_0} and vibrato depth f_Δ. These parameters are augmented by lowpass random fluctuations, and a final low-frequency drift term is added. The design parameters of the random variations are based on the results of our analysis effort. The vibrato block diagram is appropriate for inclusion in many existing synthesis models.

Our final experiment involved a vibrato and singing synthesis algorithm incorporating the important features of the measured data. Although several sophisticated singing synthesis techniques have been reported (cf. References), those methods allowing control over partial amplitudes require substantial setup and computation time. Instead, we chose to reduce the computation load in a manner appropriate for a general-purpose software synthesis program (e.g. Music 4C[21]). Thus, what the algorithm lacked in sophistication, it gained in simplicity and economy.

The computation required to synthesize a sustained signal can be reduced by the use of a *wavetable*. A wavetable is typically a pre-calculated, single cycle of the desired waveform which is repetitively read from memory. The table lookup operation has been used in many software synthesis methods, and is very efficient in terms of computation load.[22] The fundamental frequency of the output signal can be varied by changing the effective rate at which the wavetable is accessed. Also, the overall amplitude envelope of the signal can be varied with multiplicative scaling of the data obtained from the wavetable. However, a single, fixed wavetable does not allow independent amplitude control of the individual partials comprising the stored waveform. In order to simulate the spectrum modulation present in vocal



(a)



(b)

Fig. 11. Block diagram for vocal vibrato simulation. (a) The nominal vibrato rate f_{v_0} and depth f_Δ vary in a quasi-random manner. A low-frequency drift of the target fundamental is also present. (b) Example of vibrato waveform generated by the process in (a).

vibrato, our approach was to use two wavetables, each being scanned at the same rate (frequency), but with different amplitude scaling. The output signal was computed as a time-varying weighted sum of the individual outputs of the two wavetables.

The pair of wavetables were generated using the time-varying spectrum obtained for one of the vocal vibrato tones as the source of two 'target' spectra. The amplitude of each partial measured at the extreme *positive* frequency excursion during a vibrato cycle was noted, and a fixed waveform table (one cycle) was computed using the partial amplitudes as Fourier series coefficients. The other table was generated corresponding to the extreme *negative* frequency excursion of the same vibrato cycle.

A 'blending' or 'panning' function $\alpha(t)$ can be defined as

$$\alpha(t) = \frac{d_f(t) - \min\{d_f(t)\}}{\max\{d_f(t)\} - \min\{d_f(t)\}} \tag{4}$$

where $d_f(t)$ is given by eqn (3), and $\min(\cdot)$ and $\max(\cdot)$ are the extreme negative and positive frequency excursions, respectively. For example, $\alpha(t)$ is equal to 1 when $d_f(t)$ is at its positive peak and equal to 0 when $d_f(t)$ is at its negative peak. Further, defining $w_1(\omega t)$ as the waveform corresponding to the peak positive frequency excursion during the vibrato cycle and $w_2(\omega t)$ corresponding to the peak negative frequency excursion, the resulting synthesis output becomes

$$\text{output} = \alpha(t)w_1(\theta_1(t)) + (1 - \alpha(t))w_2(\theta_1(t)) \tag{5}$$

where $\theta_1(t)$ is given by eqn (3).

Thus, by 'panning' between the wavetables in synchronous with a frequency vibrato waveform the output consists of a perfect spectral match to the original at the vibrato extremes and an approximation at frequencies in between. The spectral evolution created by this simple method cannot handle the situation in which a partial traverses a spectral peak or trough during a vibrato cycle, but the spectrum modulation is simulated in a reasonable way.

In our investigation, the main synthesis parameters were the fundamental (carrier) frequency and amplitude, and the vibrato (modulator) frequency and amplitude. The block diagram of Fig. 11 was used to synthesize the vibrato waveform and to control the blend between the two synthesis wavetables. A diagram for the complete synthesis algorithm is shown in Fig. 12.

This simple panned-wavetable synthesis method produced good-quality synthetic sung vowel sounds. We were surprised to find that syntheses of tones 2–3 s in duration containing the random vibrato variations of Fig. 11 did *not* sound superior to examples containing only perfect, sinusoidal

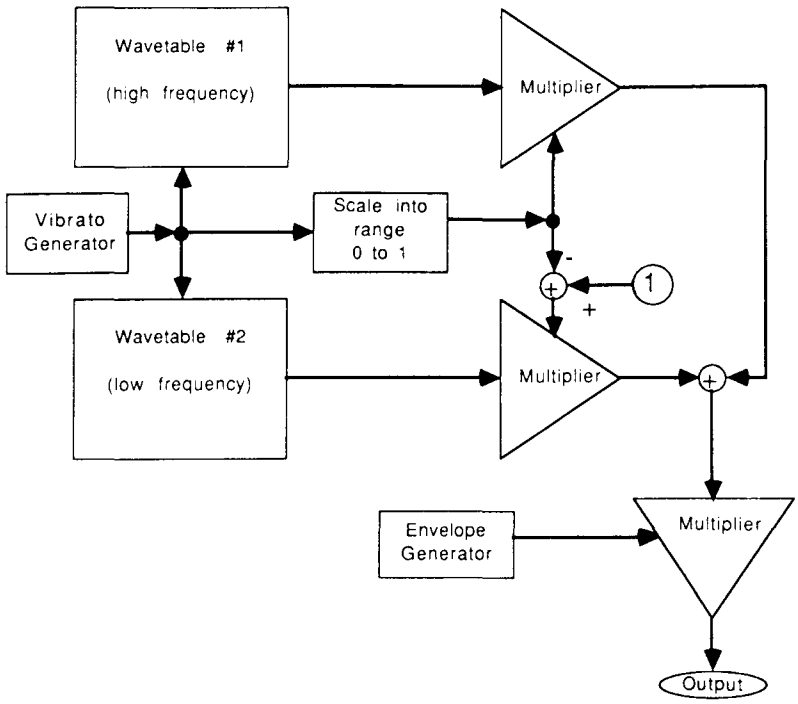


Fig. 12. Block diagram for panned-wavetable synthesis scheme. The contribution of each wavetable to the output is determined by the instantaneous vibrato extent. An overall amplitude envelope is applied to the signal.

vibrato. Our informal listening detected differences between 'straight' sinusoidal vibrato examples and examples with random vibrato variations, but did not express a strong preference. A satisfactory explanation of this result—given the reports of Bennett,[1] Chowning[7] and others regarding the importance of random vibrato fluctuations—will require further study. However, it is possible that the timbre fluctuations resulting from the panned-wavetable synthesis provide a sufficiently strong perceptual cue for singing that the assistance of simulated random variations of the vibrato waveform is not essential.

The synthesized vowel tones were only of useful quality for about two semitones above and below the target frequency, implying the need for at least three sets of wavetables per octave range. Also, the method did not include attack cues (e.g. portamento, onset noise, etc.), which may be necessary for a truly convincing solo synthesis. Thus, we may be satisfied with the simple method for use in synthesizing sustained vocal sounds, but not for rapid, exposed solo passages. We feel that the simple algorithm represents a useful computation/quality tradeoff, nonetheless, and is a good platform for further research.

4 AREAS FOR FURTHER RESEARCH

During the course of the work reported here, we came across several unanswered questions appropriate for further study.

(1) The behavior of vibrato in the vicinity of a transition from one note to another—particularly in legato singing—is not considered in our current model. However, some of our preliminary work in this area indicates that the singer systematically advances or delays the vibrato relative to the time of transition in order to perform a change from one note to the next so as to join particular points on successive vibrato cycles. A more extensive investigation will be necessary to confirm or refute this preliminary observation.

(2) The formant tracings (Fig. 9) obtained from the vibrato measurements suggest a means to design accurately a complex filter for use in the standard source–filter synthesis model.[8] The resulting filter could be designed to match the measured spectral contour instead of being based on approximate values for resonance frequencies, bandwidths, etc.

(3) Many composers using electronic media are not interested in merely recreating existing acoustic timbres. For these composers, synthesis models retaining certain qualities of a natural sound source are needed. For this purpose, the timbral characteristics of acoustic instruments, including the singing voice, must be catalogued and represented in a convenient, intuitive form.

(4) Vibrato generation based on musical context remains an *ad hoc* procedure. The standard guidelines for vocal vibrato are learned by singers without the benefit of written rules, so research is necessary if a useful descriptive model is desired. The need for random variations of the vibrato parameters and the importance of amplitude fluctuations also merit further study.

5 CONCLUSION

Based on our experiments, we draw the following conclusions:

(i) The amplitude fluctuation of a partial during the vibrato cycle varies in form, amplitude and phase according to the partial's position within the vocal tract resonances. The location and other characteristics of the formants can be aided by examination of this amplitude variation.

(ii) The partials of vocal tones are almost perfectly harmonic during

vibrato, i.e. a single vibrato waveform accurately characterizes the frequency variations of all partials when singing.

(iii) The vibrato waveform is typically nearly sinusoidal. However, its amplitude (vibrato depth), rate (vibrato frequency) and average value (nominal sung frequency) all appear to vary in a random fashion. From our measurements, the model and guidelines of Fig. 11 can be identified.

(iv) For singing synthesis, our informal evaluation indicates the perceptual importance of fluctuations of the partial amplitudes during vibrato. Inclusion of the fluctuations tends to add warmth to the sounds and improve their perceptual fusion.

(v) Although previous reports have mentioned the importance of random variations of the vibrato waveform, we were surprised to find that high-quality synthesis by the panned-wavetable algorithm of Fig. 12, using the vibrato generator of Fig. 11, did not reveal any notable improvement over panned-wavetable synthesis with fixed sine wave vibrato.

ACKNOWLEDGEMENTS

This work was supported in part by a National Science Foundation Graduate Fellowship and by the University of Illinois Research Board. The help and cooperation of the singers is greatly appreciated. The authors wish to acknowledge John C. O'Neill for collecting some of the vocal tones and the staff of the University of Illinois Computer Music Project for their technical assistance. We also would like to acknowledge the helpful comments of a reviewer in improving the clarity of several sections of this paper.

REFERENCES

1. Bennett, G., Singing synthesis in electronic music. *Proc. SMAC 1981* (Stockholm Music Acoustic Conference, 1981), Vol. 33. Royal Swedish Academy of Music, 1981, pp. 34–50.
2. McAulay, R. J. & Quatieri, T. F., Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-34**(4)(Aug.) (1986) 744–54.
3. Smith, J. & Serra, X., PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. *ICMC-1987: Proceedings of the International Computer Music Conference*. Computer Music Association, San Francisco, CA, 1987, pp. 290–7.

4. Maher, R. C., An approach for the separation of voices in composite musical signals. PhD dissertation, University of Illinois, Urbana, 1989.
5. Serra, X., A computer model for bar percussion instruments. *ICMC-1986: Proceedings of the International Computer Music Conference.* Computer Music Association, San Francisco, CA, 1986, pp. 257–62.
6. Rodet, X., Time-domain formant-wave-function synthesis. *Comput. Music J.*, **8** (1984) 9–14.
7. Chowning, J., Computer synthesis of the singing voice. *Proc. SMAC 1980* (Stockholm Music Acoustic Conference, 1980), Vol. 33. Royal Swedish Academy of Music, Stockholm, 1980, pp. 4–13.
8. Sundberg, J., Synthesis of singing. *Swed. J. Musicol.*, **60**(1) (1978) 107–12.
9. Rossing, T., Sundberg, J. & Ternström, S., Acoustic comparison of soprano solo and choir singing. *J. Acoust. Soc. Amer.*, **82**(3)(Sept.) (1987) 830–6.
10. Sundberg, J., Vibrato and vowel identification. *Arch. Acoust.*, **2** (1977) 257–66.
11. Sundberg, J., Singing and timbre. *Proc. SMAC 1977* (Stockholm Music Acoustic Conference, 1977), Vol. 17. Royal Swedish Academy of Music, Stockholm, 1977, pp. 57–88.
12. Bloothooft, G. & Plomp, R., The timbre of sung vowels. *J. Acoust. Soc. Amer.*, **84**(3)(Sept.) (1988) 847–60.
13. O'Neill, J. C., Computer analysis and synthesis of a sung vowel. DMA Dissertation, Department of Music, University of Illinois, Urbana-Champaign, Illinois, 1984.
14. Hebel, K., A machine-independent sound conversion/sound storage system. *ICMC-1985: Proceedings of the International Computer Music Conference.* Computer Music Association, San Francisco, CA, 1985.
15. Schroeder, M., Period histogram and product spectrum: new methods for fundamental-frequency measurements. *J. Acoust. Soc. Amer.*, **43**(4) (April) (1968) 829–34.
16. Rodet, X. & Bennett, G., Synthèse de la Voix Chantée par Ordinateur. Conf. des Journess d'Etudes 1980, Festival International du Son, 1981.
17. Oncley, P. B., Frequency, amplitude, and waveform modulation in the vocal vibrato. *J. Acoust. Soc. Amer.*, **49**(1)(Jan.) (1971) 136.
18. McAdams, S. & Rodet, X., The role of FM-induced AM in dynamic spectral profile analysis. In *Basic Issues in Hearing*, ed. H. Duifhuis, J. Horst & H. Wit. Academic Press, London, 1988.
19. Elyn, J. R., An analysis of the onset characteristics of the fundamental frequency of sung tones. PhD Dissertation, Department of Speech and Hearing Science, University of Illinois, Urbana-Champaign, 1977.
20. Ternström, S. & Sundberg, J., Intonation precision of choir singers. *J. Acoust. Soc. Amer.*, **84**(1)(July) (1988) 59–69.
21. Beauchamp, J. W., New computer music facilities at the University of Illinois at Urbana-Champaign. *ICMC-1985: Proceedings of the International Computer Music Conference.* Computer Music Association, San Francisco, CA, 1985, pp. 407–14.
22. Dodge, C. & Jerse, T., *Computer Music: Synthesis, Composition, and Performance.* Schirmer Books, New York, 1985.