The Use of the Phase Vocoder in Computer Music Applications*

JAMES A. MOORER†

Center for Computer Research in Music and Acoustics, Artificial Intelligence Laboratory, Stanford University, Stanford, CA 94305

The phase vocoder is an analysis–synthesis system that has as intermediate data the time-variant discrete Fourier spectrum of the input signal. It can be formulated in such a way that the synthesized signal is identical to the original, both theoretically and practically. The intermediate data can be transformed, also with no loss of information, into the more conventional magnitude and frequency representation. These intermediate data can then be used to resynthesize the tone at different pitches or different rates than the original with the advantage that when no modification is made, the synthetic tone is absolutely identical to the original. This represents a significant advance over the Heterodyne filter, which placed severe restrictions on the amount of variation in pitch or amplitude that could be analyzed. The phase vocoder has no such restrictions and can just as easily deal with vibrato and inharmonic tones.

Since scaling the frequencies in the phase vocoder analysis data also scales the spectrum up, use of this modification with speech can produce altered vowel tones. If this method is combined with the linear predictor, using the phase vocoder to alter the pitch of the error signal, then the spectrum can be held constant while the pitch is changed, thus allowing independent control over time, pitch, and spectrum. Vowel quality can be preserved or altered at will. Again, if no modification is made, the combination of the linear predictor and the phase vocoder is an identity, both theoretically and practically.

This research is still in a very preliminary state, so only a few sound examples can be given at this time. A full theoretical explanation, however, can be given.

DEFINITION: A complete discussion of the mechanics of the implementation of the digital phase vocoder may be found in Portnoff [1]. We will only review the definitions here. Our contribution comes in the application of this technique. This definition is taken directly from Portnoff.

Let $x(n)$ represent samples of a speech waveform. The discrete short-time Fourier transform of $x(n)$ is defined as follows:

$$X_k(n) = \sum_{r=-\infty}^{\infty} x(r)h(n-r)W_N^{-rk}, \quad \text{for } k = 0, 1, \cdots, N-1 \tag{1}$$

* This paper was presented at the 55th AES Convention, October 29–November 1, 1976 in New York.

† Dr. Moorer is now working at IRCAM in Paris.

where $W_N = \exp[j(2\pi/N)]$ and $h(n)$ is an approximately chosen window. By properly choosing $h(n)$, it can be guaranteed that the original sequence $x(n)$ is exactly recoverable from its short-time transform as defined by Eq. (1). Furthermore, $x(n)$ is given in this case by

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_k(n) W_N^{nk}, \qquad \text{for all } n. \tag{2}$$

It is useful to consider Eqs. (1) and (2) in terms of a bank of digital bandpass filters with contiguous passbands. Consider a set of N complex bandpass filters $\{h_k(n)\}$ with passbands equally spaced about the unit circle and with unit-sample responses

$$h_k(n) = \frac{1}{N} h(n) W_N^{nk}, \qquad k = 0, 1, \ldots, N-1 \tag{3}$$

where $h(n)$ is a prototype low-pass filter with real unit-sample response. For the analysis–synthesis system defined in Eqs. (1) and (2) to be an identity, the only condition is that $h(n)=0$ for n being all integer multiples of N. This condition is precisely the constraint on the unit-sample response of a digital interpolating filter [2].

NOW WHAT HAVE WE GOT?

If we think of this system as a bank of bandpass filters, we can see that what comes out are the real and imaginary parts of the signal in each of the N equally spaced frequency bands. The difference between this and a conventional channel vocoder [3] is that the phase information is preserved in each channel and that we can guarantee that the output is exactly identical to the input. Notice that the output of each channel is band-limited by the filter so that each channel may be resampled at a lower rate, thus reducing the data involved. If the prototype low-pass filter $h(n)$ were perfect, we could resample the transform data $X_k(n)$ by just taking every Nth point. This means that the transform data would be exactly the same amount of the data as the input data. This is reassuring because it means that analyzing sound in this manner does not, by itself, produce a data explosion. In practice, however, we must resample somewhat more often than this. At Stanford we are resampling at twice the minimum rate, and this seems to be more than adequate.

We should say something about the filter involved. It should be a linear-phase filter because otherwise when modifications are made to the transform data, peculiar things happen. For example, when working with digitized speech this way, without a linear-phase filter, the synthesized speaker sounds drunk, presumably because the faster changing attributes happen at different times than the more slowly changing features. Discontinuities are not preserved because the phase relations are changed.

Already we can make one kind of modification to the data: we can accomplish the effect of filtering the original signal simply by multiplying the channel output so as to change the contribution from selected channels. We can give a "formant" effect by increasing the contribution from a particular region. There is, however, no particular advantage to performing filtering in this way rather than using conventional digital filters. The most interesting modifications can be made when the output of each channel is converted to magnitude-frequency form.

MAGNITUDE-FREQUENCY CONVERSION

If we run a perfectly periodic waveform through Eq. (1) with N set equal to (or greater than) the number of samples in a period, then each channel of the phase vocoder (to use Flanagan's original terminology [4]) will cover no more than one harmonic of the input waveform. This being the case, the real and imaginary parts of the output will then just be the cosine and sine of a frequency that corresponds to the difference of the frequency of the harmonic in that channel and the center frequency of the channel. The center frequencies are, of course, equally spaced around the unit circle. We can then see that we can get the amplitude of the harmonic by taking the square root of the sum of the squares of the real and imaginary parts of the channel output. Likewise, we can get the frequency of the harmonic by taking the derivative of the phase angle as defined by the arctangent of the imaginary part over the real part. To depict this in formulas, let $a_k(n)$ be the real part of the output of the kth channel and $b_k(n)$ be the imaginary part of the output of the kth channel.

$$A_k(n) = \sqrt{a_k(n)^2 + b_k(n)^2} \tag{4}$$

$$\theta_k(n) = \text{a tan}\left(\frac{b_k(n)}{a_k(n)}\right) \tag{5}$$

$$\dot{\theta} = \frac{a_k(n)\dot{b}_k(n) - b_k(n)\dot{a}_k(n)}{a_k(n)^2 + b_k(n)^2}. \tag{6}$$

Since we are working with sampled-data functions, the derivatives here should be evaluated with linear-phase finite impulse response band-limited differentiators. The only trouble with these formulas is that the functions described in Eqs. (4) and (6) are nonlinear and are thus non-band-limited. This lack of band-limiting means that the magnitude and frequency can not be resampled every Nth point like the real and imaginary parts of the channel output can. This is where the data explosion occurs. In fact, to do the magnitude-frequency conversion at all, $a_k(n)$ and $b_k(n)$ must be available at the original sampling rate. This means that they must be obtained by applying Eq. (1) at each point in time or by interpolating the resampled functions using a band-limited interpolator [2], [5]. We must be careful here, because the filter used to interpolate the functions must satisfy the same conditions as our prototype low-pass filter, which are that $h(n)=0$ at n equal to all integer multiples of N.

Even if we do the conversion as above, taking care of all the messy details, then we are no longer guaranteed that the resulting synthesis is identical to the original. For instance, the initial phase angles are lost. In most cases this does not make much difference. The difference comes in waveforms with discontinuities either in the signal or in its first derivative. To duplicate the discontinuity, the phase relations must be preserved. For this reason we have developed another way of doing the magnitude-frequency conversion that guarantees identity again. This can be done using simple trigonometric identities as follows:

$$\sin\{\theta_k(n) - \theta_k(n-1)\} = \sin\{\theta_k(n)\}\cos\{\theta_k(n-1)\} - \cos\{\theta_k(n)\}\sin\{\theta_k(n-1)\}$$
$$= b_k(n)a_k(n-1) - a_k(n)b_k(n-1) \quad (7)$$

$$\cos\{\theta_k(n) - \theta_k(n-1)\} = \cos\{\theta_k(n)\}\cos\{\theta_k(n-1)\} + \sin\{\theta_k(n)\}\sin\{\theta_k(n-1)\}$$
$$= a_k(n)a_k(n-1) + b_k(n)b_k(n-1) \quad (8)$$

$$\Delta\theta_k(n) = \text{atan}\left(\frac{\sin\{\theta_k(n) - \theta_k(n-1)\}}{\cos\{\theta_k(n) - \theta_k(n-1)\}}\right) \quad (9)$$

with the initial conditions $\theta_k(0) = 0, a_k(0) = 1, b_k(0) = 0$.

This gives us a sequence of angle differences for each channel that have the properties of a frequency with the added bonus that we can still recover the original signal exactly.

SOME RESULTS

Figs. 1–6 show the results of applying the aforementioned analysis technique [Eqs. (1), (4), (7), (8), and (9)] to two isolated musical tones (piano and tenor saxophone) and to a short segment of human voice. Again, tones synthesized from these data are indistinguishable, numerically, theoretically, and perceptually, from the original. We show the first and fifth harmonics of each tone. We see that when the tone is not present, the frequency trace goes crazy. This is because we are then analyzing tape hiss, and we can expect the frequency to be random there. Note, however, that the frequency is still quite active during the attack portion of these tones. This is a demonstration of the non-band-limited nature of the frequency due to the highly nonlinear transformation implied by Eqs. (7)–(9). If we artifically band-limit the frequency by, for instance, filtering it so that we can reduce the amount of data by resampling, then we lose fidelity on some kinds of tones. The tenor saxophone tone comes through ok but the attack of the piano tone becomes less abrupt. It is as if the hammer were made of putty. The human voice tone is even worse. Fig. 5 shows that the fundamental is pretty well behaved, but the fifth harmonic shows very erratic behavior of the frequency. It is hard to determine what this is telling us. It is a strong indication that the human voice is not particularly periodic, even during the more continuous regions, that there is considerable phase jitter on the upper harmonics. If this is true, it is not the least bit surprising then that conventional vocoders, such as the linear prediction vocoder, do not capture the "natural" voice sound, but invariably sound a bit "raspy." If we do not simulate this behavior, then we cannot hope to capture the tone quality accurately.

APPLICATIONS AND FUTURE DIRECTIONS

When this technique is combined with the linear predictor, a uniquely powerful combination results. The linear predictor can also be formulated as an identity system as follows. If you filter a signal by its optimum inverse filter, you get what is called the "error" signal. If you then take the error signal and filter it by the noninverse filter, you get the original signal back, both in theory and in practice—nothing has been done to the signal. Using the phase vocoder, however, we can now modify this error signal. We can change its pitch or its timing, then reimpose the spectral shape by applying the optimum filter again. This gives us independent control over the timing, spectrum, and pitch of the sound, a powerful combination indeed. The remaining problems, however, are considera-

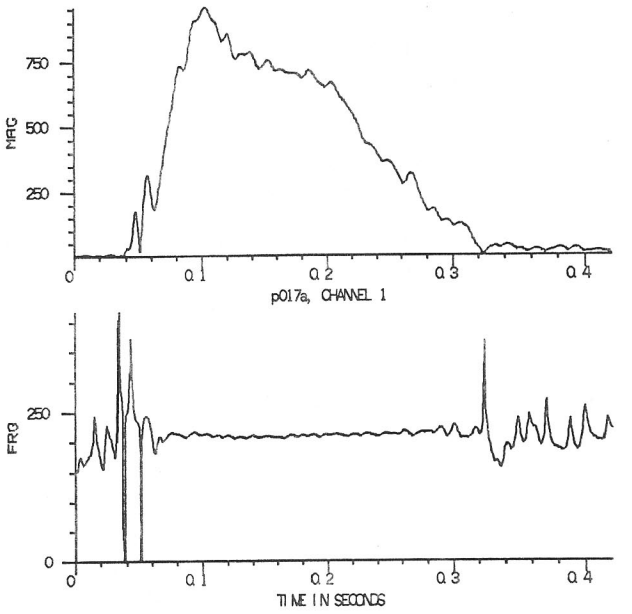

Fig. 1. Amplitude and frequency curves of the first harmonic of a piano tone (A3, 220 Hz).

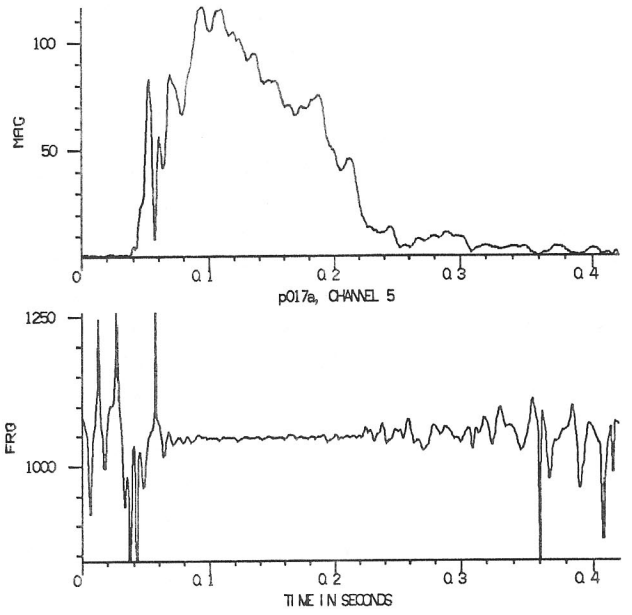

Fig. 2. Amplitude and frequency curves of the fifth harmonic of a piano tone (A3, 220 Hz).

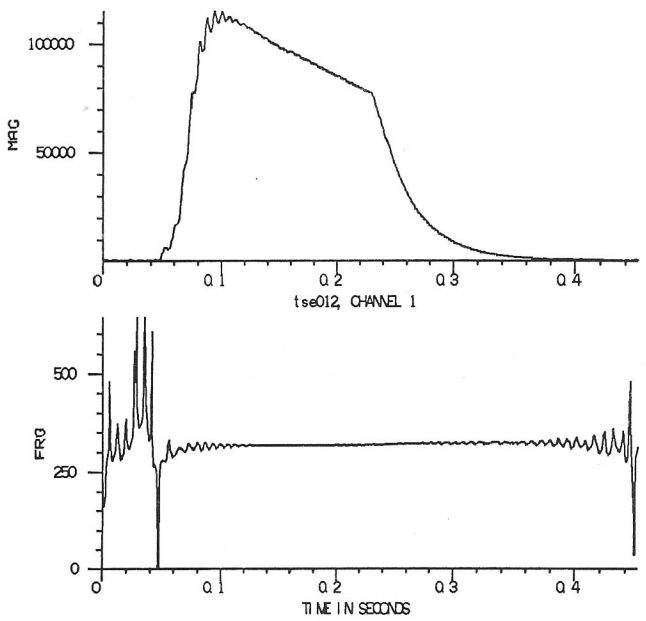

Fig. 3. Amplitude and frequency curves of the first harmonic of a note from a tenor saxophone.

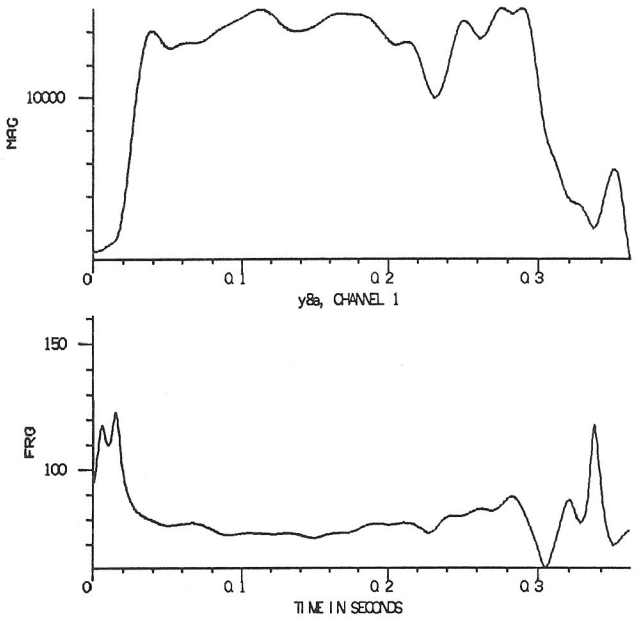

Fig. 5. Amplitude and frequency curves of the first harmonic of a vocal tone.

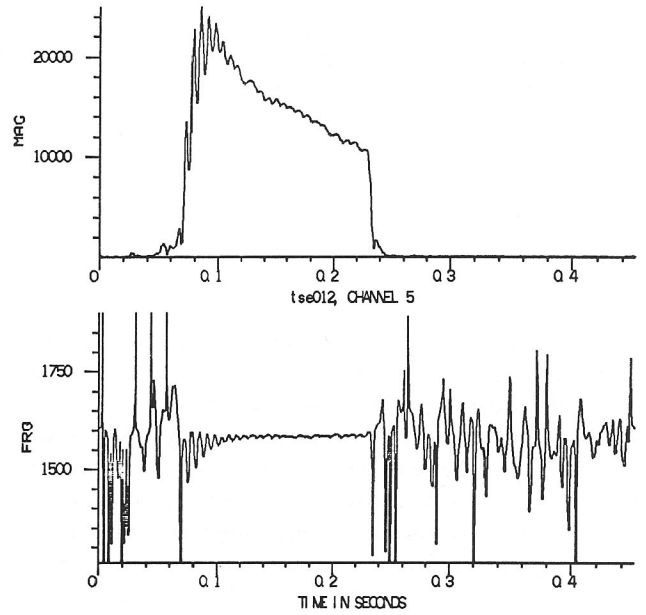

Fig. 4. Amplitude and frequency curves of the fifth harmonic of a note from a tenor saxophone.

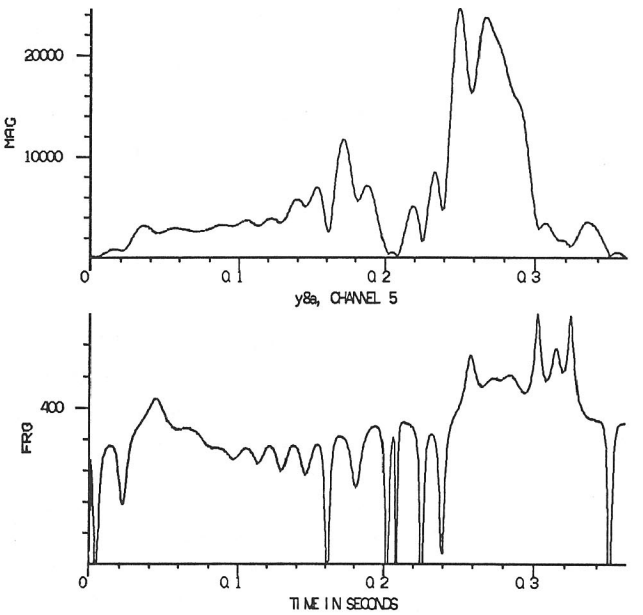

Fig. 6. Amplitude and frequency curves of the fifth harmonic of a vocal tone.

ble. We have to explore further what those erratic frequency traces mean and how to make modifications (such as pitch or timing changes) with less perceivable alteration to the sound. When pitch changes are made by blindly multiplying all the frequency traces by a fixed factor, we sometimes get a very strange "choral" effect. Although this may be useful, it is not what we expected, and we seek some way to control this effect to be able to turn it on or off at will.

REFERENCES

[1] M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Trans. Acous. Speech, and Signal Proc.*, vol. ASSP-24, pp. 243-248 (June 1976).

[2] R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," *Pro. IEEE*, vol. 61, pp. 692-702 (June 1973).

[3] H. Dudley, "The Vocoder," *Bell Labs. Re.*, vol. 17, pp. 122-126 (1939).

[4] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell Sys. Tech. J.*, vol. 45, pp. 1493-1509 (Nov. 1966).

[5] G. Oetken, T. W. Parks, and H. W. Schuessler, "New Results in the Design of Digital Interpolators," *IEEE Trans. Acoust. Speech, and Signal Proc.*, vol. ASSP-23 (June 1975).

[6] L. R. Rabiner, "Techniques for Designing Finite-Duration Impulse-Response Digital Filters," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 188-195 (Apr. 1971).

Dr. Moorer's biography appeared in the November 1976 issue.