Speech Analysis/Synthesis Based on a Sinusoidal Representation

ROBERT J. McAULAY, SENIOR MEMBER, IEEE, AND THOMAS F. QUATIERI, MEMBER, IEEE

Abstract—**A sinusoidal model for the speech waveform is used to develop a new analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. These parameters are estimated from the short-time Fourier transform using a simple peak-picking algorithm. Rapid changes in the highly resolved spectral components are tracked using the concept of "birth" and "death" of the underlying sine waves. For a given frequency track a cubic function is used to unwrap and interpolate the phase such that the phase track is maximally smooth. This phase function is applied to a sine-wave generator, which is amplitude modulated and added to the other sine waves to give the final speech output. The resulting synthetic waveform preserves the general waveform shape and is essentially perceptually indistinguishable from the original speech. Furthermore, in the presence of noise the perceptual characteristics of the speech as well as the noise are maintained. In addition, it was found that the representation was sufficiently general that high-quality reproduction was obtained for a larger class of inputs including: two overlapping, superposed speech waveforms; music waveforms; speech in musical backgrounds; and certain marine biologic sounds.**

Finally, the analysis/synthesis system forms the basis for new approaches to the problems of speech transformations including time-scale and pitch-scale modification, and midrate speech coding [8], [9].

I. Introduction

ONE approach to the problem of representation of speech signals is to use the speech production model in which speech is viewed as the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal tract. In many speech applications it suffices to assume that the glottal excitation can be in one of two possible states, corresponding to voiced or unvoiced speech. In attempts to design high-quality speech coders at the midband rates, generalizations of the binary excitation model have been developed. One such approach that is currently popular is multipulse [1]. In this paper the goal is also to generalize the model for the glottal excitation; but instead of using impulses as in multipulse, the excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitudes, frequencies, and phases.

A number of other approaches to analysis/synthesis that are based on sine-wave models have been discussed in the literature. Hedelin [3] proposed a pitch-independent sine-wave model for use in coding the baseband signal for speech compression. The amplitudes and frequencies of the underlying sine waves are estimated using Kalman filtering techniques, and each sine-wave phase is defined to be the integral of the associated instantaneous frequency. Another sine-wave-based speech compression system is being developed by Almeida and Silva [4]. In contrast to Hedelin's approach, their system uses a pitch estimate to establish a harmonic set of sine waves. The sine-wave phases are computed at the harmonic frequencies. To compensate for any errors that might be introduced as a result of the harmonic sine-wave representation, a residual waveform is coded along with the underlying sine-wave parameters.

In this paper a sinusoidal model for the speech waveform is derived that leads to a new analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. In Section II the glottal excitation is represented in terms of a sum of sine waves, which, when applied to a time-varying vocal tract filter, leads to the desired sinusoidal representation for speech waveforms. In Section III a parameter extraction algorithm is developed that shows that the amplitudes, frequencies, and phases of the sine waves can be obtained from the high-resolution short-time Fourier transform (STFT) by locating the peaks of the associated magnitude function. In order to perform speech synthesis the amplitudes, frequencies, and phases estimated on one frame must be matched and allowed to continuously evolve into the set of amplitudes, frequencies, and phases estimated on a successive frame. These issues are resolved in Sections IV and V where a frequency-matching algorithm is derived along with a solution to the phase unwrapping and phase interpolation problem. Experiments were performed with the resulting system, and the synthetic speech was judged to be of excellent quality, almost indistinguishable from the original. The results of some of these experiments are discussed in Section VI where pictorial comparisons of the original and synthetic waveforms are made. In addition, it has been found that the performance of the analysis/synthesis system did not degrade in the presence of environmental disturbances due to noise, multiple speakers, or music, and could be used to successfully reproduce certain marine biologic sounds.

II. The Sinusoidal Speech Model

In the speech production model, the speech waveform $s(t)$ is assumed to be the output of passing a glottal exci-

Manuscript received April 1, 1985; revised January 10, 1986. This work was supported by the U.S. Department of the Air Force. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

The authors are with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173-0073.

IEEE Log Number 8608125.

0096-3518/86/0800-0744$01.00 © 1986 IEEE

tation waveform $e(t)$ through a linear time-varying filter that models the characteristics of the vocal tract. If the time-varying impulse response of the vocal tract filter is $h(\tau; t)$, then

$$s(t) = \int_0^t h(t - \tau; t)\, e(\tau)\, d\tau. \tag{1}$$

As an alternative to the binary voiced/unvoiced excitation model and to the more general multipulse model, it is proposed that the excitation signal be represented in terms of a sum of sine waves of arbitrary amplitudes, frequencies, and phases. This model is written as

$$e(t) = \text{Re} \sum_{l=1}^{L(t)} a_l(t) \exp\left\{ j \left[\int_0^t \omega_l(\sigma)\, d\sigma + \phi_l \right] \right\} \tag{2}$$

where, for the lth sinusoidal component, $a_l(t)$ and $\omega_l(t)$ represent the amplitude and frequency and ϕ_l represents a fixed phase offset which accounts for the fact that the sine waves will generally not be in phase. This model leads to a particularly simple representation for the speech waveform. That this is so becomes apparent by letting

$$H(\omega; t) = M(\omega; t) \exp [j\Phi(\omega; t)] \tag{3}$$

represent the time-varying vocal tract transfer function, and, assuming that the glottal excitation parameters in (2) are constant over the duration of the impulse response of the vocal tract filter in effect at time t, then using (2) and (3) in (1) results in the speech model[1]

$$s(t) = \sum_{l=1}^{L(t)} a_l(t) M[\omega_l(t); t] \cdot \exp\left\{ j \left[\int_0^t \omega_l(\sigma)\, d\sigma + \Phi[\omega_l(t); t] + \phi_l \right] \right\}. \tag{4}$$

By combining the effects of the glottal and vocal tract amplitudes and phases, the representation can be written more concisely as

$$s(t) = \sum_{l=1}^{L(t)} A_l(t) \exp [j\psi_l(t)] \tag{5}$$

where

$$A_l(t) = a_l(t)\, M[\omega_l(t); t] \tag{6}$$

$$\psi_l(t) = \int_0^t \omega_l(\sigma)\, d\sigma + \Phi[\omega_l(t), t] + \phi_l \tag{7}$$

represent the amplitude and phase of the lth sine wave along the frequency track $\omega_l(t)$. The next step is to develop a robust procedure for extracting the amplitudes, frequencies, and phases of the component sine waves, a subject which will be discussed in the next section.

III. Estimation of Speech Parameters

The problem in analysis/synthesis is to take a speech waveform, extract the parameters that represent a quasi-stationary portion of that waveform, and use those parameters or coded versions of them to reconstruct an approximation that is "as close as possible" to the original speech. Furthermore, it is desirable to have a robust parameter extraction algorithm since the speech signal in many cases is contaminated by additive acoustic noise. The general identification problem in which the speech signal is to be represented by multiple sine waves is a difficult one to solve analytically. Therefore, the approach taken here will be pragmatic, in the sense that an estimator will be derived based on a set of idealized assumptions; then, once the structure of the ideal estimator is known, modifications will be made as the assumptions are relaxed to better model practical speech waveforms.

As a first step, the time line will be broken down into a contiguous sequence of frames, each of duration T. The center of the analysis window for the kth frame occurs at time t_k. Assuming that the vocal tract and glottal parameters are constant over an interval of time that includes the duration of the analysis window and the duration of the vocal tract impulse response, then (7) can be written as

$$\psi_l(t) = \omega_l^k(t - t_k) + \theta_l^k \tag{8}$$

where the superscript "k" is used to indicate that the parameters of the model may vary from frame to frame. As a consequence of (8) the synthetic speech waveform over frame k can be written as

$$s(n) = \sum_{l=1}^{L^k} \gamma_l^k \exp (jn\omega_l^k) \tag{9}$$

where $\gamma_l^k = A_l^k \exp (j\theta_l^k)$ represents the lth complex amplitude for the lth component of the L^k sine waves. Since the measurements are made on digitized speech, the sampled-data notation is used throughout this section. In this respect, the time index n corresponds to the uniform samples of $t - t_k$ so that n ranges from $-N/2$ to $N/2$, with $n = 0$ being reset to the center of the analysis window for every frame and where $N + 1$ is the duration of the analysis window. The problem now is to fit the synthetic speech waveform in (9) to the measured waveform, denoted by $y(n)$. A useful criterion for judging the goodness of fit is the mean-squared error,

$$\epsilon^k = \sum_n |y(n) - s(n)|^2 = \sum_n |y(n)|^2 - 2\, \text{Re} \sum_n y(n)\, s^*(n) + \sum_n |s(n)|^2. \tag{10}$$

Substituting the speech model of (9) into (10) leads to the error expression

$$\epsilon^k = \sum_n |y(n)|^2 - 2\, \text{Re} \sum_{l=1}^{L^k} (\gamma_l^k)^* \sum_n y(n) \exp (-jn\omega_l^k) + (N + 1) \sum_{l=1}^{L^k} \sum_{i=1}^{L^k} \gamma_l^k (\gamma_i^k)^* \operatorname{sinc} (\omega_l^k - \omega_i^k) \tag{11}$$

where $\operatorname{sinc}(x) = \sin [(N + 1)\, x/2]/[(N + 1) \sin (x/2)]$.

[1] The "real part" notation "Re" has been temporarily omitted.

The problem now is to try to identify a set of sine waves that minimizes (11), an identification problem that is, in general, difficult to solve. Insights into the development of a suitable estimator can be obtained by restricting the class of input signals to perfectly voiced speech, in which case (9) can be written as

$$s(n) = \sum_{l=1}^{L^k} \gamma_l^k \exp(jnl\omega_0^k) \tag{12}$$

where $\omega_0^k = 2\pi/\tau_0^k$ and where τ_0^k is the pitch period assumed to be constant over the duration of the kth frame. For the purpose of establishing the structure of the ideal estimator, it is further assumed that the pitch period is known and that the width of the analysis window is a multiple of τ_0^k. Under these highly idealized conditions, the sinc $(\cdot)$ function in the last term of (11) reduces to

$$\operatorname{sinc}(\omega_l^k - \omega_i^k) = \operatorname{sinc}[(l-i)\,\omega_0^k] = \begin{cases} 1 & \text{if } l = i \\ 0 & \text{if } l \neq i \end{cases} \tag{13}$$

where $\omega_l^k = l\omega_0^k$. Then the error expression reduces to

$$\epsilon^k = \sum_n |y(n)|^2 - 2(N+1)\operatorname{Re}\left[\sum_{l=1}^{L^k} (\gamma_l^k)^* \, Y(\omega_l^k)\right] + N \sum_{l=1}^{L^k} |\gamma_l^k|^2 \tag{14}$$

where

$$Y(\omega) = \frac{1}{N+1} \sum_n y(n) \exp(-jn\omega) \tag{15}$$

is the STFT of the measurement signal. By completing the square in (14), the error can be written as

$$\epsilon^k = \sum_n |y(n)|^2 + (N+1) \sum_{l=1}^{L^k} [|Y(\omega_l^k) - \gamma_l^k|^2 - |Y(\omega_l^k)|^2], \tag{16}$$

from which it follows that the optimum estimate for the amplitude and phase is

$$\hat{\gamma}_l^k = Y(l\omega_0^k), \tag{17}$$

which reduces the error to

$$\epsilon^k = \sum_n |y(n)|^2 - (N+1) \sum_{l=1}^{L^k} |Y(l\omega_0^k)|^2. \tag{18}$$

From this it follows that the error is minimized by selecting all of the harmonic frequencies in the speech bandwidth Ω (i.e., $L^k = \Omega/\omega_0^k$).

Equations (15), (17), and (18) completely specify the structure of the ideal estimator and show that the speech data are manifest in the optimum estimator through the DFT. Although these results are equivalent to a Fourier series representation of a periodic waveform, the above equation leads to an intuitive generalization to the practical case. This is done by considering the function $|Y(\omega)|^2$ to be a continuous function of ω. For the idealized voiced speech case, this function (referred to as the periodogram) will be pulselike in nature, with peaks occurring at all of the pitch harmonics. Therefore, the frequencies of the underlying sine waves correspond to the location of the peaks of the periodogram, and the estimates of the amplitudes and phases are obtained by evaluating the STFT at the frequencies of the peaks. The advantage of this latter interpretation of the estimator structure is that it can be applied when the ideal voiced speech assumption is no longer valid. That this is so can be seen by calculating the STFT for the general sinusoidal speech model in (9). In this case the STFT is simply

$$Y(\omega) = \sum_{l=1}^{L^k} \gamma_l^k \operatorname{sinc}(\omega_l^k - \omega). \tag{19}$$

Provided the analysis window is "wide enough" that

$$|\omega_i^k - \omega_l^k| \geq \frac{4\pi}{N+1}, \tag{20}$$

then the periodogram can be written as

$$|Y(\omega)|^2 \approx \sum_{l=1}^{L^k} |\gamma_l^k|^2 \operatorname{sinc}^2(\omega_l^k - \omega), \tag{21}$$

and as before, the location of the peaks of the periodogram corresponds to the underlying sine-wave frequencies and the STFT samples at these frequencies correspond to the complex amplitudes. Therefore, the structure of the ideal estimator applies to a more general class of speech waveforms provided (20) holds. Since, during steady voicing, neighboring frequencies are separated by the pitch fundamental, (20) suggests that the desired resolution can be achieved "most of the time" by requiring that the analysis window be at least two pitch periods wide. Of course, these properties are based on the assumption that the sinc $(\cdot)$ function is essentially zero outside of the region defined by (20). In fact, this is not a valid approximation and there will be sidelobes outside of this region which will lead to leakage that will compromise the performance of the estimator. These sidelobes are due to the rectangular window that is implicit in the definition of the STFT, a problem which is reduced but not eliminated by using the weighted STFT. Letting $\bar{Y}(\omega)$ denote the weighted STFT, i.e.,

$$\bar{Y}(\omega) = \sum_{n=-N/2}^{N/2} w(n)\, y(n) \exp(-jn\omega) \tag{22}$$

where $w(n)$ represents the temporal weighting due to the window function, then the practical version of the idealized estimator estimates the frequencies of the underlying sine waves as the locations of the peaks of $|\bar{Y}(\omega)|$ (i.e., the frequency at which the slope changes from positive to negative). Letting these frequency estimates be denoted by $\{\hat{\omega}_l^k\}$, then the corresponding complex amplitudes are given by

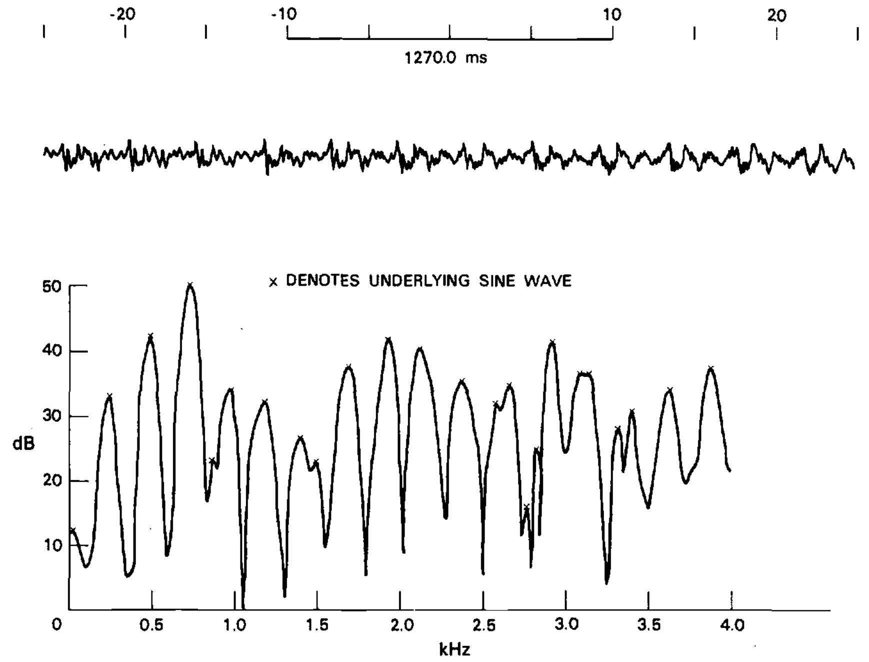


Fig. 1. Typical periodogram for a frame of voiced speech and the amplitude and frequency estimates of the underlying sine waves.

$$\hat{\gamma}_l = \bar{Y}(\hat{\omega}_l^k) = \hat{A}_l^k \exp(j\hat{\theta}_l^k). \qquad (23)$$

Assuming that the component sine waves have been properly resolved, then, in the absence of noise, $\hat{A}_l^k$ will yield the value of an underlying sine wave provided the window is scaled so that

$$\sum_{n=-N/2}^{N/2} w(n) = 1. \qquad (24)$$

The Hamming window was used in all of the experiments reported in this paper, and while this resulted in a very good sidelobe structure, it did so at the expense of broadening the mainlobes of the periodogram estimator. Therefore, in order to maintain the resolution properties that were needed to justify the optimality properties of the periodogram processor, the constraint implied by (20) is revised to require that the window width be at least $2\frac{1}{2}$ times the pitch period. Although the window width could be set on the basis of the instantaneous pitch, it is adequate to adapt it to the average pitch, as this makes the analyzer less sensitive to the performance of the pitch extractor. On adjusting the analysis window, the average pitch and the window width are continually being updated in real time using the pitch computed during strongly voiced frames and are averaged using a $\frac{1}{4}$ s time constant. During frames of unvoiced speech, the window is held fixed at the value obtained on the preceding voiced frame. Once the width for a particular frame has been specified, the Hamming window is computed, normalized according to (24), and the STFT of the input speech is taken using a 512-point FFT. Plotted in Fig. 1 is a typical periodogram for voiced speech, along with the amplitudes and frequencies that are estimated using the above procedure.

The purpose of the preceding analysis was to produce an estimator structure that was closely related to the optimal estimator derived on the basis of ideal voiced speech. The approximations that were introduced were based on properties that were more representative of realistic voiced speech. Nowhere have the properties of unvoiced speech been taken into account. To do this in an optimal way requires use of the Karhunen–Loève expansion for noiselike signals [2]. Such an analysis shows that a sinusoidal representation is valid, provided the frequencies are "close enough" that the ensemble power spectral density changes slowly over consecutive frequencies. In order to apply the sinusoidal model to unvoiced speech, therefore, it is necessary to assume that the frequencies corresponding to the periodogram peaks will be "close enough" to satisfy the requirement imposed by the Karhunen–Loève expansion. If the window width is constrained to be at least 20 ms wide, then, "on the average," this will lead to a set of periodogram peaks that will be approximately 100 Hz apart, and this should provide a sufficiently dense sampling to satisfy the constraints of the Karhunen–Loève sinusoidal representation for the unvoiced case. Plotted in Fig. 2 is a typical periodogram for a frame of unvoiced speech along with the amplitudes and frequencies that are estimated using the above procedure.

The above analysis provides a justification for the representation of the speech waveform in terms of the amplitudes, frequencies, and phases of a set of sine waves. This representation applies to one analysis frame. Different sets of these parameters will be obtained for each frame. The next problem to address then is the association of amplitudes, frequencies, and phases measured on one frame with those that are obtained on a successive frame. This is the subject addressed in the next section.

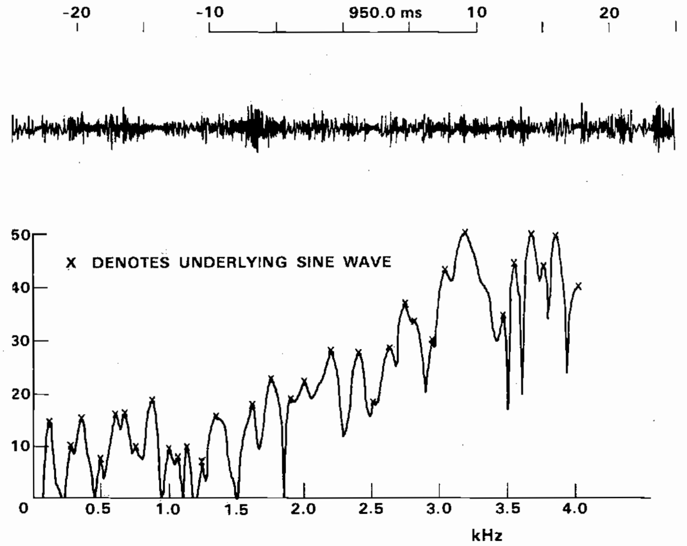


Fig. 2. Typical periodogram for a frame of unvoiced speech and the amplitude and frequency estimates of the underlying sine waves.

IV. Frame-to-Frame Peak Matching

If the number of peaks were constant from frame to frame, the problem of matching the parameters estimated on one frame with those on a successive frame would simply require a frequency-ordered assignment of peaks. In practice, however, there will be spurious peaks that come and go due to the effects of sidelobe interaction; the locations of the peaks will change as the pitch changes; and there will be rapid changes in both the location and the number of peaks corresponding to rapidly varying regions of speech, such as at voiced/unvoiced transitions. In order to account for such rapid movements in the spectral peaks, the concept of "birth" and "death" of sinusoidal components is introduced. The problem of matching spectral peaks in some "optimal" sense, while allowing for this birth–death process, is generally a difficult problem. One method, which has proven to be successful for signal reconstruction, is now described.

Suppose that somehow peaks up to frame k have been matched and a new parameter set for frame $k + 1$ is generated. Let the chosen frequencies on frames k and $k + 1$ be denoted by $\omega_0^k, \omega_1^k, \cdots, \omega_{N-1}^k$ and $\omega_0^{k+1}, \omega_1^{k+1}, \cdots, \omega_{M-1}^{k+1}$, respectively, where for convenience the "^" notation of the previous section has been dropped, and where N and M represent the total number of peaks selected on each frame ($N \neq M$ in general). The process of matching each frequency in frame k, ω_n^k, to some frequency in frame $k + 1$, ω_m^{k+1}, is given in the following four steps.

Step 1: Suppose that a match has been found for frequencies $\omega_0^k, \omega_1^k, \cdots, \omega_{n-1}^k$. A match is now attempted for frequency ω_n^k. Fig. 3(a) depicts the case where all frequencies ω_m^{k+1} in frame $k + 1$ lie outside a "matching interval" Δ of ω_n^k, i.e.,

$$|\omega_n^k - \omega_m^{k+1}| \geq \Delta \tag{25}$$

for all m. In this case the frequency track associated with ω_n^k is declared "dead" on entering frame $k + 1$, and ω_n^k is matched to itself in frame $k + 1$, but with zero amplitude. Frequency ω_n^k is then eliminated from further consideration, and step 1 is repeated for the next frequency in the list, ω_{n+1}^k.

If, on the other hand, there exists a frequency ω_m^{k+1} in frame $k + 1$ that lies within the matching interval about ω_n^k, and is the closest such frequency, i.e.,

$$|\omega_n^k - \omega_m^{k+1}| < |\omega_n^k - \omega_i^{k+1}| < \Delta \tag{26}$$

for all $i \neq m$, then ω_m^{k+1} is declared to be a candidate match to ω_n^k. A definitive match is not yet made since there may exist a better match in frame k to the frequency ω_m^{k+1}, a contingency which is accounted for in step 2.

Step 2: In this step, a candidate match from step 1 is confirmed. Suppose that a frequency ω_n^k of frame k has been tentatively matched to frequency ω_m^{k+1} of frame $k + 1$. Then, if ω_m^{k+1} has no better match to the remaining unmatched frequencies of frame k, the candidate match is declared to be a definitive match. This condition, illustrated in Fig. 3(c), is given by

$$|\omega_m^{k+1} - \omega_n^k| < |\omega_m^{k+1} - \omega_{i+1}^k| \qquad \text{for} \quad i > n. \tag{27}$$

When this occurs, frequencies ω_n^k and ω_m^{k+1} are eliminated from further consideration and step 1 is repeated for the next frequency in the list, ω_{n+1}^k.

If condition (27) is not satisfied, then the frequency ω_m^{k+1} in frame $k + 1$ is better matched to the frequency ω_{n+1}^k in frame k than it is to the test frequency ω_n^k. Two additional cases are then considered. In the first case, illustrated in Fig. 3(d), the adjacent remaining lower frequency ω_{m-1}^{k+1} (if one exists) lies below the matching interval, hence, no match can be made. As a result, the frequency track associated with ω_n^k is declared "dead" on entering frame $k + 1$, and ω_n^k is matched to itself with zero amplitude. In the second case, illustrated in Fig. 3(e), the frequency ω_{m-1}^k is within the matching interval about ω_n^k, and a definitive match is made. After either case, step 1 is repeated using the next frequency in the list ω_{n+1}^k. It should be noted that many other situations are possible in this step, but to keep the tracker alternatives as simple as possible, only the two cases discussed were implemented.

Step 3: When all frequencies of frame k have been tested and assigned to continuing tracks or to dying tracks, there may remain frequencies in frame $k + 1$ for which no matches have been made. Suppose that ω_m^{k+1} is one such frequency; then it is concluded that ω_m^{k+1} was "born" in frame k, and its match, a new frequency ω_m^{k+1}, is created in frame k with zero magnitude. This is done for all such unmatched frequencies. This last step is illustrated in Fig. 3(f).

An illustration of the effects of the birth–death procedure to account for extraneous peaks is shown in Fig. 4. The results of applying the tracker to a segment of real speech are shown in Fig. 5, which demonstrates the ability of the tracker to adapt quickly through transitory speech behavior such as voiced/unvoiced transitions and mixed voiced/unvoiced regions.

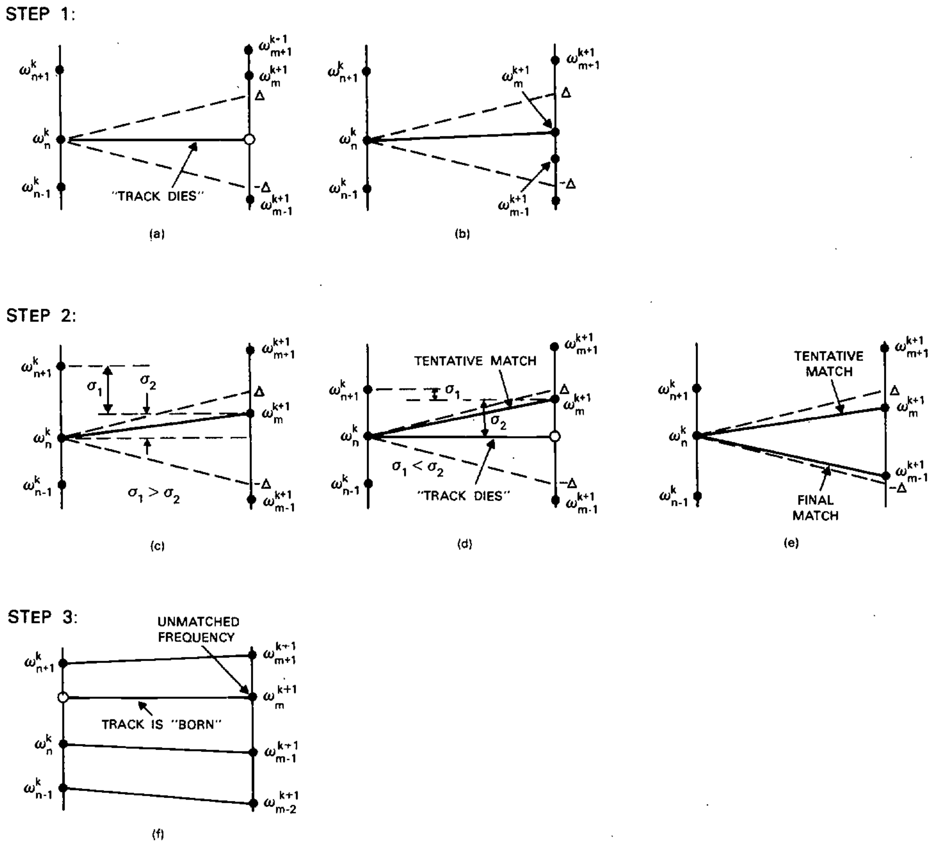


Fig. 3. Illustration of different modes used in the birth-death frequency tracker.

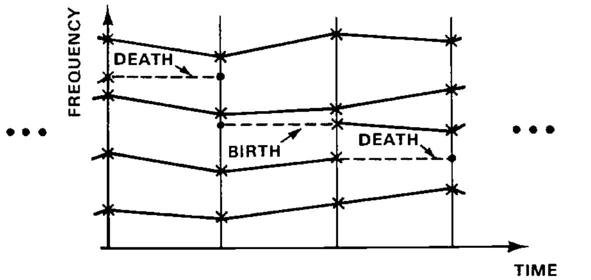


Fig. 4. Illustration of frequency tracks using the birth-death frequency tracker.

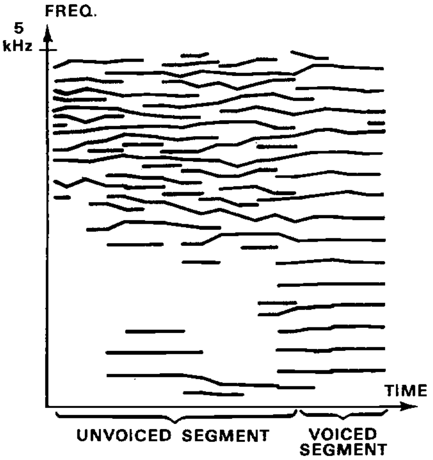


Fig. 5. Typical frequency tracks for real speech.

V. The Synthesis System

Since a set of amplitudes, frequencies, and phases are estimated for each frame, it might seem reasonable to estimate the original speech waveform on the kth frame by generating synthetic speech using the equation

$$\tilde{s}(n) = \sum_{l=1}^{L(k)} \hat{A}_l^k \cos\,[n\hat{\omega}_l^k + \hat{\theta}_l^k] \tag{28}$$

where $n = 0, 1, 2, \cdots, S - 1$ and where S is the length of the synthesis frame. Due to the time-varying nature of the parameters, however, this straightforward approach leads to discontinuities at the frame boundaries, which seriously degrades the quality of the synthetic speech. Therefore, a method must be found for smoothly interpolating the parameters measured from one frame to those that are obtained on the next.

The most straightforward approach for performing this interpolation is to overlap and add time-weighted segments of the sinusoidal components. This is done by using the measured amplitude, frequency, and phase (referenced to the center of the synthesis frame) to construct a sine wave, which is then weighted by a triangular window over a duration equal to twice the length of the synthesis frame. The time-weighted components corresponding to the lagging edge of the triangular window are added to the overlapping leading edge components that were generated during the previous frame. Two systems were implemented in real time, operating with frames separated by 11.5 and 23.0 ms, respectively. While the synthetic speech produced by the first system was quite good, almost indistinguishable from the original, the longer frame interval resulted in synthetic speech that was "rough" and, although very intelligible, was deemed to be of poor quality. Therefore, if a particular application can support a high frame rate, then the overlap-add synthesizer is a good system to use. However, there are many practical situations, such as midrate speech coding, where lower frame rates are necessary, for which an alternative to the overlap-add synthesizer must be developed. A method will now be described that interpolates the matched sine wave parameters directly.

As a result of the frequency-matching algorithm described in the previous section, all of the parameters measured for an arbitrary frame k are associated with a corresponding set of parameters for frame $k + 1$. Letting $(\hat{A}_l^k, \hat{\omega}_l^k, \hat{\theta}_l^k)$ and $(\hat{A}_l^{k+1}, \hat{\omega}_l^{k+1}, \hat{\theta}_l^{k+1})$ denote the successive sets of parameters for the lth frequency track, then an obvious solution to the amplitude interpolation problem is to take

$$\tilde{A}(n) = \hat{A}^k + \frac{(\hat{A}^{k+1} - \hat{A}^k)}{S} n, \tag{29}$$

where $n = 0, 1, \cdots, S - 1$ is the time sample, into the kth frame. (The track subscript "l" has been omitted for convenience.)

Unfortunately, such a simple approach cannot be used to interpolate the frequency and phase because the measured phase $\hat{\theta}^k$ is obtained modulo 2π. Hence, phase unwrapping must be performed to ensure that the frequency tracks are "maximally smooth" across frame boundaries. The first step in solving this problem is to postulate a phase interpolation function that is a cubic polynomial, namely,

$$\tilde{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3. \tag{30}$$

It is convenient to treat the phase function as though it were a function of a continuous time variable t, with $t = 0$ corresponding to frame k and $t = T$ corresponding to frame $k + 1$. Since the derivative of the phase is the frequency, it is necessary that the cubic phase function and its derivative equal the phases and frequencies measured at the frame boundaries. This idea of applying a cubic polynomial to interpolate the phase between frame boundaries was independently proposed by Almeida and Silva for use in their harmonic sine-wave synthesizer [4]. Since only the principal value of the phase can be measured, provision must also be made for unwrapping the phase subject to the above constraints on the cubic phase interpolation function. In this paper an explicit solution is obtained for interpolation and phase unwrapping by invoking an additional constraint which requires that the unwrapped cubic phase function be "maximally smooth." The mathematics leading to the complete solution are now presented.

Using the fact that the instantaneous frequency is the derivative of the phase, then

$$\dot{\tilde{\theta}}(t) = \gamma + 2\alpha t + 3\beta t^2, \tag{31}$$

and it follows that at the starting point, $t = 0$,

$$\tilde{\theta}(0) = \zeta = \hat{\theta}^k$$
$$\dot{\tilde{\theta}}(0) = \gamma = \hat{\omega}^k, \tag{32}$$

and at the terminal point, $t = T$,

$$\tilde{\theta}(T) = \hat{\theta}^k + \hat{\omega}^k T + \alpha T^2 + \beta T^3 = \hat{\theta}^{k+1} + 2\pi M$$
$$\dot{\tilde{\theta}}(T) = \hat{\omega}^k + 2\alpha T + 3\beta T^2 = \hat{\omega}^{k+1} \tag{33}$$

where again the track subscript "l" is omitted for convenience. Since the terminal phase $\hat{\theta}^{k+1}$ is measured modulo 2π, it is necessary to augment it by the term $2\pi M$ (M is an integer) in order to make the resulting frequency function "maximally smooth," a concept that will be quantified in the sequel. At this point M is unknown, but for each value of M, whatever it may be, (33) can be solved for $\alpha(M)$ and $\beta(M)$ (the dependence on M has now been made explicit). The solution is easily shown to satisfy the matrix equation

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \dfrac{3}{T^2} & \dfrac{-1}{T} \\ \dfrac{-2}{T^3} & \dfrac{1}{T^2} \end{bmatrix} \begin{bmatrix} \hat{\theta}^{k+1} - \hat{\theta}^k - \hat{\omega}^k T + 2\pi M \\ \hat{\omega}^{k+1} - \hat{\omega}^k \end{bmatrix}. \tag{34}$$

In order to determine M and ultimately the solution to the phase unwrapping problem, an additional constraint needs to be imposed that quantifies the "maximally smooth" criterion. Fig. 6 illustrates a typical set of cubic phase interpolation functions for a number of values of M. It seems clear on intuitive grounds that the best phase function to pick is the one that would have the least variation. This is what is meant by a maximally smooth frequency track. In fact, if the frequencies were constant and the vocal tract were stationary, the true phase would be linear. Therefore, a reasonable criterion for "smoothness" is to choose M such that

$$f(M) = \int_0^T [\ddot{\tilde{\theta}}(t; M)]^2 \, dt \tag{35}$$

is a minimum where $\ddot{\tilde{\theta}}(t; M)$ denotes the second derivative of $\tilde{\theta}(t; M)$ with respect to the time variable t.

Although M is integer valued, since $f(M)$ is quadratic in M, the problem is most easily solved by minimizing $f(x)$ with respect to the continuous variable x and then choosing M to be the integer closest to x. After straightforward but tedious algebra, it can be shown that the minimizing value of x is

$$x^* = \frac{1}{2\pi}\left[(\hat{\theta}^k + \hat{\omega}^k T - \hat{\theta}^{k+1}) + (\hat{\omega}^{k+1} - \hat{\omega}^k)\frac{T}{2}\right] \tag{36}$$

from which M^* is determined and used in (34) to compute $\alpha(M^*)$ and $\beta(M^*)$, and in turn, the unwrapped phase in-

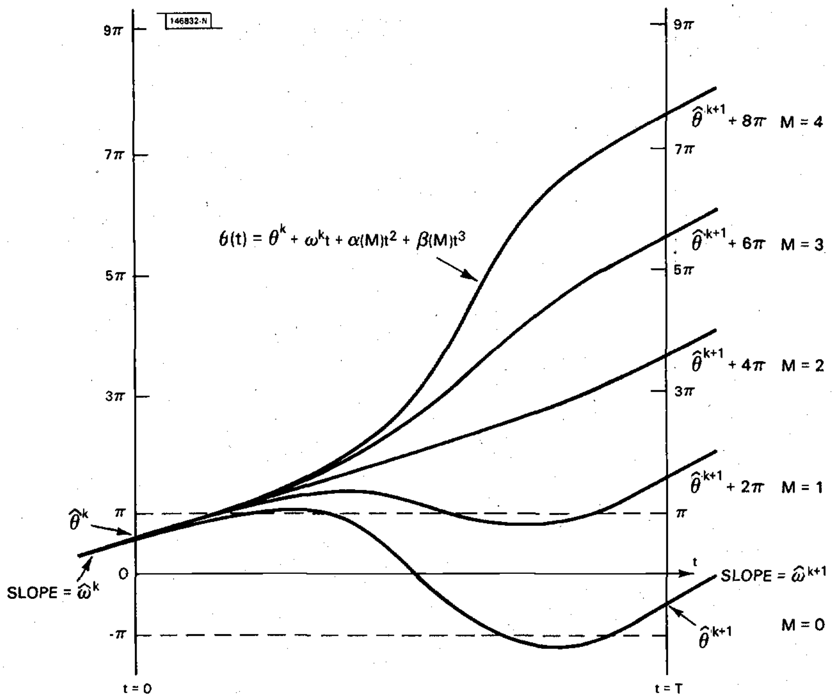


Fig. 6. Typical set of cubic phase interpolation functions.

terpolation function

$$\tilde{\theta}(t) = \hat{\theta}^k + \hat{\omega}^k t + \alpha(M^*)t^2 + \beta(M^*)t^3. \tag{37}$$

This phase function not only satisfies all of the measured phase and frequency endpoint constraints, but also unwraps the phase in such a way that $\tilde{\theta}(t)$ is maximally smooth.

Since the above analysis began with the assumption of an initial unwrapped phase $\hat{\theta}^k$ corresponding to frequency $\hat{\omega}^k$ at the start of frame k, it is necessary to specify the initialization of the frame interpolation procedure. This is done by noting that at some point in time the track under study was born. When this event occurred, an amplitude, frequency, and phase were measured at frame $k + 1$, and the parameters at frame k to which these measurements correspond were defined by setting the amplitude to zero (i.e., $\hat{A}^k = 0$) while maintaining the same frequency (i.e., $\hat{\omega}^k = \hat{\omega}^{k+1}$). In order to ensure that the phase interpolation constraints are satisfied initially, the unwrapped phase is defined to be the measured phase $\hat{\theta}^{k+1}$ and the startup phase is defined to be

$$\hat{\theta}^k = \hat{\theta}^{k+1} - \hat{\omega}^{k+1}S \tag{38}$$

where S is the number of samples traversed in going from frame $k + 1$ back to frame k.

As a result of the above phase unwrapping procedure, each frequency track will have associated with it an instantaneous unwrapped phase which accounts for both the rapid phase changes due to the frequency of each sinusoidal component and the slowly varying phase changes due to the glottal pulse and the vocal track transfer function. Letting $\tilde{\theta}_l(t)$ denote the unwrapped phase function for the lth track, then the final synthetic waveform will be given by

$$\tilde{s}(n) = \sum_{l=1}^{L^k} \tilde{A}_l(n) \cos [\tilde{\theta}_l(n)] \tag{39}$$

where $\tilde{A}_l(n)$ is given by (29), $\tilde{\theta}_l(n)$ is the sampled data version of (37), and L^k is the number of sine waves estimated for the kth frame.

This completes the theoretical basis for the new sinusoidal analysis/synthesis system. Although extremely simple in concept, the detailed analysis led to the introduction of the birth-death frequency tracker and the cubic interpolation phase unwrapping procedure. The usefulness with which these new procedures aid in the synthesis of speech will be discussed in the next section.

VI. Experimental Results

A block diagram description of the analysis/synthesis system is given in Fig. 7. A nonreal time floating-point simulation was developed in order to determine the effectiveness of the proposed approach in modeling real speech. The speech processed in the simulation was low-pass filtered at 5 kHz, digitized at 10 kHz, and analyzed at 10 ms frame intervals. A 512-point FFT using a pitch-adaptive Hamming window, having a width which was 2.5 times the average pitch, was found to be sufficient for accurate peak estimation. The maximum number of peaks that are used in synthesis was set to a fixed number (~ 80), and if excessive peaks were obtained only the largest peaks were used. A large speech database has been processed with this system, and it has been found that the synthetic speech was perceived to be essentially indistinguishable from the original. Visual examination of many of the reconstructed passages shows that the waveform

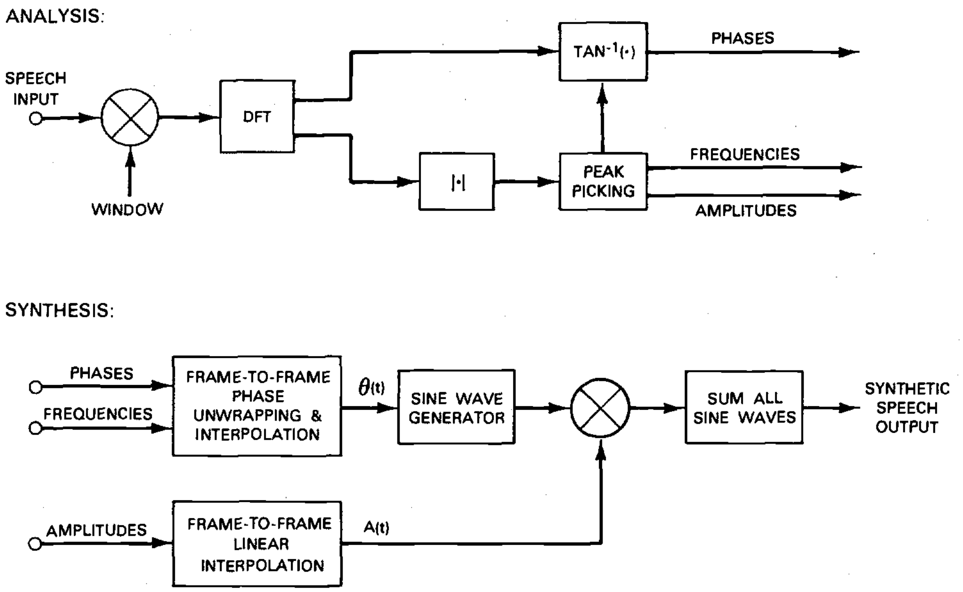


Fig. 7. Block diagram of the sinusoidal analysis/synthesis system.

structure is essentially preserved. An example of this property is shown in Fig. 8, which compares the waveforms for the original speech and the reconstructed speech during an unvoiced/voiced speech transition. This suggests that the quasi-stationarity conditions seem to be satisfactorily met and that the use of the parametric model based on the amplitudes, frequencies, and phases of a set of sine-wave components appears to be justifiable for both voiced and unvoiced speech.

Although the sinusoidal model was originally designed for a single speaker, it can represent any waveform consisting of a sum of sine waves with time-varying amplitudes and frequencies. Thus, the analysis/synthesis system should be capable of synthesizing a broader class of signals. This hypothesis was verified by successfully reconstructing multispeaker waveforms, music, speech in a musical background, and marine biologic signals such as whale sounds. Furthermore, it was found that the reconstruction does not break down in the presence of noise. The synthesized speech is perceptually nearly indistinguishable from the original noisy speech with essentially no modification of the noise characteristics. Illustrations depicting the performance of the system in the face of the above degradations are provided in [10].

Although high-quality analysis/synthesis of speech has been demonstrated using the amplitudes, frequencies, and phases of the peaks of the high-resolution STFT, it is often argued that the ear is insensitive to phase, a proposition which forms the basis of much of the work in narrowband speech coders. The question arises as to whether or not the phase measurements are essential to the sum of sine waves synthesis procedure. An attempt to explore this question was made by replacing each cubic phase track by a phase function that was defined to be the integral of the instantaneous frequency [3], [5]. In this case the instantaneous frequency was taken to be the linear interpolation of the frequencies measured at the frame boundaries, and the integration, which started from a zero value at the birth of the track, continued to be evaluated along that track until that track died. This "magnitude-only" reconstruction technique was applied to several sentences of speech, and, while the resulting synthetic speech was very intelligible and free of artifacts, it was perceived as being different from the original speech. Furthermore, the differences were more pronounced for low-pitched (i.e., pitch < ~100 Hz) speakers. An example of a waveform synthesized by the "magnitude-only" system is shown in Fig. 9(b). Compared to the original speech, shown in Fig. 9(a), the synthetic waveform is quite different owing to the failure to maintain the true sine-wave phases. In an

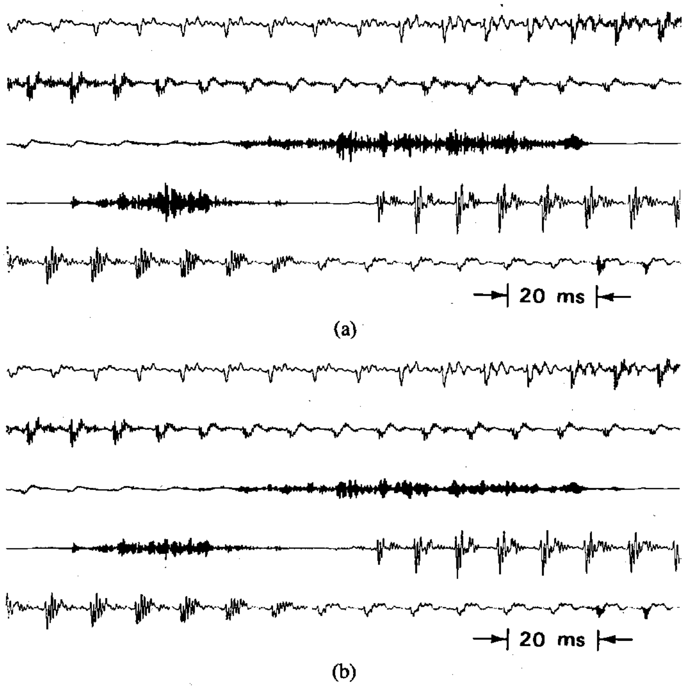
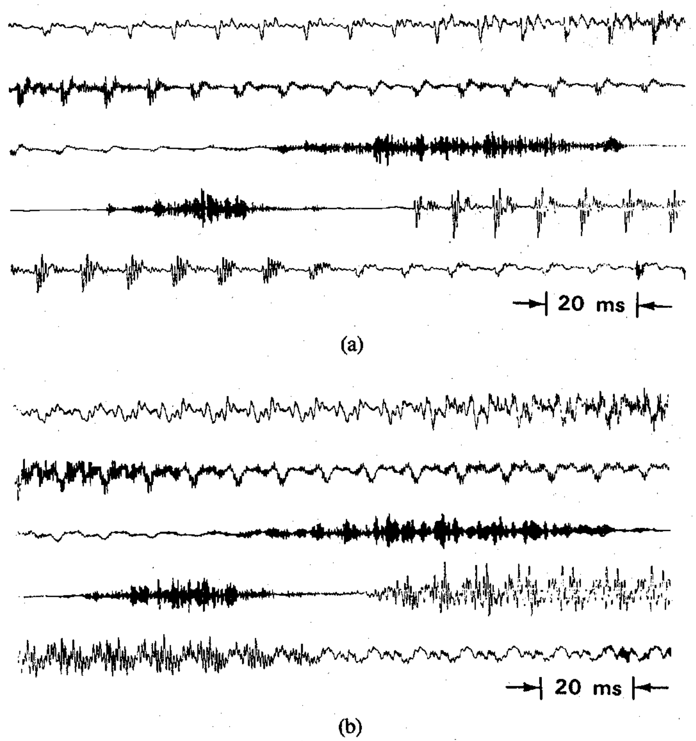


Fig. 8. Sinusoidal reconstruction of speech.



(a)

(b)

Fig. 9. Magnitude-only reconstruction of speech.

additional experiment the magnitude-only system was applied to the synthesis of noisy speech, and it was found that the synthetic noise took on a tonal quality that was unnatural and annoying.

VII. Conclusions

A sinusoidal representation for the speech waveform has been developed that extracts the amplitudes, frequencies, and phases of the component sine waves from the STFT (short-time Fourier transform). The parameter extraction routine is robust in noise in the sense that the parameters are obtained by coherently processing the data over the analysis window.

In order to account for spurious effects due to sidelobe interaction and time-varying voicing and vocal tract events, sine waves are allowed to come and go in accordance with a birth–death frequency-tracking algorithm. Once contiguous frequencies are matched, a smooth cubic phase interpolation function is obtained that is consistent with all of the frequency and phase measurements and is maximally smooth. This phase function is applied to a sine-wave generator, which is amplitude modulated and added to the other sinusoidal components to give the final speech output.

The analysis/synthesis system was applied to clear speech and speech that was subjected to various types of interference. Synthetic speech that was natural and of high quality was generated in every case. The system could also be used to develop a parametric representation for nonspeech sounds such as music and certain marine biologic sounds. Finally, it is important to note that, except in updating the average pitch used to adjust the width of the analysis window, no voicing decisions are used in the analysis/synthesis procedure.

In some respects the basic model has similarities to one that has been proposed by Flanagan [6]. Flanagan argues that because of the nature of the peripheral auditory system, a speech waveform can be expressed as the sum of the outputs of a *fixed* filter bank. The amplitude, frequency, and phase measurements of the filter outputs are then used in various configurations of speech synthesizers [7]. Although the present work is based on the discrete Fourier transform (DFT), which can be interpreted as a filter bank, the use of a high-resolution DFT in combination with peak picking renders a highly adaptive filter bank since only a subset of all of the DFT filters is used at any one frame. It is the use of the frequency tracker and the cubic phase interpolator that allows the filter bank to move with the highly resolved speech components. Therefore, the system fits into the framework described by Flanagan, but, whereas Flanagan's approach is based on the properties of the peripheral auditory system, the present system is designed on the basis of properties of the speech production mechanism.

Attempts to perform ''magnitude-only'' reconstruction were made by replacing the cubic phase tracks by a phase that was simply the integral of the instantaneous frequency. While the resulting speech was very intelligible and free of artifacts, it was perceived as being different in quality from the original speech; the differences were more pronounced for low-pitched (i.e., pitch $< \sim 100$ Hz) speakers. When the magnitude-only system was used to synthesize noisy speech, the synthetic noise took on a tonal quality that was unnatural and annoying. It was concluded that this latter property would render the system unsuitable for applications for which the speech would be subjected to additive acoustic noise.

While it may be tempting to conclude that the ear is not phase deaf, particularly for low-pitched speakers, it may be that this is simply a property of the sinusoidal analysis/synthesis system. No attempts were made to devise an experiment that would resolve this question conclusively. It was felt, however, that the system was well suited to the design and execution of such an experiment since it provides explicit access to a set of phase parameters that are essential to the high-quality reconstruction of speech.

It is important to note that the use of the frequency tracker and the cubic phase interpolation function resulted in a *functional* description of the time evolution of the amplitude and phase of the sinusoidal components of the synthetic speech. For the applications for which the system was developed (i.e., time-scale, pitch-scale, and frequency modification of speech and speech coding) such a functional model is essential. It should be noted, however, that if the system were to be applied simply to achieve synthesis using a set of sine waves, then the frequency tracking and phase interpolation procedures would be unnecessary. In this case a solution is achieved simply by overlapping and adding time-weighted segments of each of the sinusoidal components. The resulting synthetic speech was found to be essentially perceptually indistinguishable from the original speech, provided the frame rate was on the order of 100 Hz.

Finally, it should be noted that a fixed-point 16-bit realtime implementation of the system has been developed on the Lincoln Dıgital Signal Processors [11]. Diagnostic rhyme tests (DRT) have been performed, and it has been found that about one DRT point is lost relative to the unprocessed speech of the same bandwidth with the analysis/synthesis system operating at a 50 Hz frame rate. Currently, the system is being used in research aimed at the development of a midrate speech coder and has already been applied successfully to problems in time-scale, pitch-scale, and frequency modification of speech. Preliminary results on the application of the sinusoidal-based system to the speech transformation and coding problems have been reported in [8] and [9].

Acknowledgment

The authors would like to thank their colleague, J. Tierney, for his comments and suggestions in the early stages of this work. They also acknowledge the reviewers, whose comments resulted in significant improvements in the final draft.

References

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, 1982, p. 614.

[2] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I.* New York: Wiley, 1968, ch. 3.

[3] P. Hedelin, "A tone-oriented voice-excited vocoder," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, 1981, p. 205.

[4] L. B. Almeida and F. M. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, 1984, p. 27.5.1.

[5] R. J. McAulay and T. F. Quatieri, "Magnitude-only reconstruction using a sinusoidal speech model," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, 1984, p. 27.6.1.

[6] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, p. 412, 1980.

[7] J. L. Flanagan and S. W. Christensen, "Computer studies on parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, p. 420, 1980.

[8] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, 1985, p. 489.

[9] R. J. McAulay and T. F. Quatieri, "Mid-rate coding based on a sinusoidal representation of speech," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, 1985, p. 945.

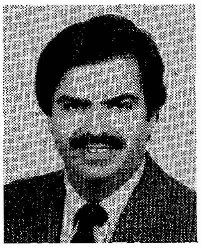
[10] —, "Speech analysis/synthesis based on a sinusoidal representation," M.I.T., Lincoln Lab., Rep. TR-693, May 1985, AD-A157023.

[11] P. E. Blankenship, "LDVT: High performance minicomputer for realtime speech processing," paper presented at EASCON'77, Sept. 1977.

Robert J. McAulay (S'63–M'67–SM'81) was born in Toronto, Ont., Canada, on October 23, 1939. He received the B.A.Sc. degree in engineering physics (honors) from the University of Toronto in 1962, the M.Sc. degree in electrical engineering from the University of Illinois, Urbana, in 1963, and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1967.

In 1967 he joined the Radar Signal Processing Group of the Massachusetts Institute of Technology (M.I.T.) Lincoln Laboratory, Lexington, where he worked on problems in estimation theory and signal/filter design using optimal control techniques. From 1970 until 1975 he was a member of the Air Traffic Control Division at Lincoln Laboratory and worked on the development of aircraft tracking algorithms, optimal MTI digital signal processing, and on the problems of aircraft direction finding for the Discrete Address Beacon System. On a leave of absence from Lincoln Laboratory during the winter and spring of 1974, he was a Visiting Associate Professor at McGill University, Montreal, P.Q., Canada. Since 1975 he has been a member of the Speech Systems Technology Group at Lincoln Laboratory where he has been involved in the development of robust narrow-band speech vocoders.

Dr. McAulay received the M. Barry Carlton award in 1978 for the best paper published in the IEEE Transactions on Aerospace and Electronic Systems for the paper "Interferometer design for elevation angle estimation."



Thomas F. Quatieri (S'73–M'79) was born in Somerville, MA, on January 31, 1952. He received the B.S. degree (summa cum laude) from Tufts University, Medford, MA, in 1973 and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (M.I.T.), Cambridge, in 1975, 1977, and 1979, respectively.

From 1973 to 1975 he was a Teaching Assistant and from 1975 to 1979 a Research Assistant in the area of digital signal processing, both within the Department of Electrical Engineering and Computer Science of M.I.T. His research for the Masters degree involved the design of two-dimensional digital filters and for the Sc.D. involved phase estimation with application to speech analysis/synthesis. He is presently a Research Staff Member at the M.I.T. Lincoln Laboratory where he is working on problems in digital signal processing with applications to speech communications and image processing.

Dr. Quatieri is the recipient of the 1982 Paper Award of the IEEE Acoustics, Speech, and Signal Processing Society for the best paper by an author under 30 years of age. He is a member of the IEEE Digital Signal Processing Technical Committee and has served on the steering committee for the 1984 Digital Signal Processing Workshop. He is also a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.