R.J. McAulay and T.F. Quatieri

Speech Processing Based on a Sinusoidal Model

Using a sinusoidal model of speech, an analysis/synthesis technique has been developed that characterizes speech in terms of the amplitudes, frequencies, and phases of the component sine waves. These parameters can be estimated by applying a simple peak-picking algorithm to a short-time Fourier transform (STFT) of the input speech. Rapid changes in the highly resolved spectral components are tracked by using a frequency-matching algorithm and the concept of "birth" and "death" of the underlying sine waves. For a given frequency track, a cubic phase function is applied to a sine-wave generator, whose output is amplitude-modulated and added to the sine waves generated for the other frequency tracks, and this sum is the synthetic speech output. The resulting waveform preserves the general waveform shape and is essentially indistinguishable from the original speech. Furthermore, in the presence of noise the perceptual characteristics of the speech and the noise are maintained. It was also found that high-quality reproduction was obtained for a large class of inputs: two overlapping, superposed speech waveforms; music waveforms; speech in musical backgrounds; and certain marine biologic sounds.

The analysis/synthesis system has become the basis for new approaches to such diverse applications as multirate coding for secure communications, time-scale and pitch-scale algorithms for speech transformations, and phase dispersion for the enhancement of AM radio broadcasts. Moreover, the technology behind the applications has been successfully transferred to private industry for commercial development.

Speech signals can be represented with a speech production model that views speech as the result of passing a glottal excitation waveform through a time-varying linear filter (Fig. 1), which models the resonant characteristics of the vocal tract. In many speech applications the glottal excitation can be assumed to be in one of two possible states, corresponding to voiced or unvoiced speech.

In attempts to design high-quality speech coders at mid-band rates, more general excitation models have been developed. Approaches that are currently popular are multipulse [1] and code-excited linear predictive coding (CELP) [2]. This paper also develops a more general model for glottal excitation, but instead of using impulses as in multipulse, or code-book excitations as in CELP, the excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitudes, frequencies, and phases.

Other approaches to analysis/synthesis that are based on sine-wave models have been discussed. 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 were estimated using Kalman filtering techniques and each sine-wave phase was defined to be the integral of the associated instantaneous frequency.

Another sine-wave-based speech 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 from the STFT 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.

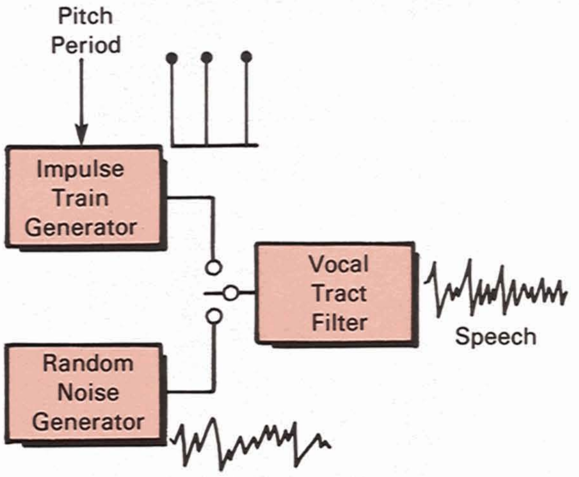

Fig. 1 — The binary model of speech production, illustrated here, requires pitch, vocal-tract parameters, energy levels, and voiced/unvoiced decisions as inputs.

This paper derives a sinusoidal model for the speech waveform, characterized by the amplitudes, frequencies, and phases of its component sine waves, that leads to a new analysis/synthesis technique. The glottal excitation is represented as a sum of sine waves that, when applied to a time-varying vocal-tract filter, leads to the desired sinusoidal representation for speech waveforms (Fig. 2).

A parameter-extraction algorithm has been 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. To synthesize speech, the amplitudes, frequencies, and phases estimated on one frame must be matched and allowed to evolve continuously into the set of amplitudes, frequencies, and phases estimated on a successive frame. These issues are resolved using a frequency-matching algorithm in conjunction with a solution to the phase-unwrapping and phase-interpolation problem.

A system was built and experiments were performed with it. The synthetic speech was judged to be of excellent quality — essentially indistinguishable from the original. The results of some of these experiments are discussed and pictorial comparisons of the original and synthetic waveforms are presented.

The above sinusoidal transform system (STS) has found practical application in a number of speech problems. Vocoders have been developed that operate from 2.4 kbps to 4.8 kbps providing good speech quality that increases more or less uniformly with increasing bit rate. In another area the STS provides high-quality speech transformations such as time-scale and pitch-scale modifications. Finally, a large research effort has resulted in a new technique for speech enhancement for AM radio broadcasting. These applications are considered in more detail later in the text.

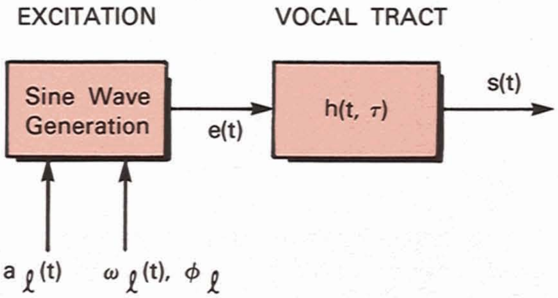

Fig. 2 — The sinusoidal speech model consists of an excitation and vocal-tract response. The excitation waveform is characterized by the amplitudes, frequencies, and phases of the underlying sine waves of the speech; the vocal tract is modeled by the time-varying linear filter, the impulse response of which is $h(t; \tau)$.

Speech Production Model

In the speech production model, the speech waveform, $s(t)$, is modeled as the output of a linear time-varying filter that has been excited by the glottal excitation waveform, $e(t)$. The filter, which models the characteristics of the vocal tract, has an impulse response denoted by $h(t; \tau)$. The speech waveform is then given by

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

If the glottal excitation waveform is represented as a sum of sine waves of arbitrary amplitudes, frequencies, and phases, then the model can be written as

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

where, for the lth sinusoidal component, $a_l(t)$ and $\omega_l(t)$ represent the amplitude and frequency (Fig. 2). Because the sine waves will not necessarily be in phase, ϕ_l is included to represent a fixed phase-offset. This model leads to a particularly simple representation for the speech waveform. Letting

$$H(\omega; t) = M(\omega; t) \exp\left[j\Phi(\omega; t) \right] \quad (3)$$

represent the time-varying vocal-tract transfer function and assuming that the excitation parameters given in Eq. 2 are constant throughout the duration of the impulse response of the filter in effect at time t, then using Eqs. 2 and 3 in Eq. 1 leads to the speech model given by

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

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

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

where

$$A_l(t) = a_l(t)\, M[\omega_l(t); t] \quad (6)$$

$$\psi_l(t) = \int_0^t \omega_l(\sigma)\, d\sigma + \Phi[\omega_l(t); t] + \phi_l \quad (7)$$

represent the amplitude and phase of the lth sine wave along the frequency track $\omega_l(t)$. Equations 5, 6, and 7 combine to provide a sinusoidal representation of a speech waveform. In order to use the model, the amplitudes, frequencies, and phases of the component sine waves must be extracted from the original speech waveform.

Estimation of Speech Parameters

The key problem in speech analysis/synthesis is to extract from a speech waveform the parameters that represent a quasi-stationary portion of that waveform, and to use those parameters (or coded versions of them) to reconstruct an approximation that is "as close as possible" to the original speech. The parameter-extraction algorithm, or estimator, should be robust, as the parameters must often be extracted from a speech signal that has been contaminated with acoustic noise.

In general, it is difficult to determine analytically which of the component sine waves and their amplitudes, frequencies, and phases are necessary to represent a speech waveform. Therefore, an estimator based on idealized speech waveforms was developed to extract these parameters. As restrictions on the speech waveform were relaxed in order to model real speech better, adjustments were made to the estimator to accommodate these changes.

In the development of the estimator, the time axis was first broken down into an overlapping 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 Eq. 7 can be written as

$$\psi_l(t) = \omega_l^k (t - t_k) + \theta_l^k \quad (8)$$

where the superscript k indicates that the parameters of the model may vary from frame to frame. Using Eq. 8, in Eq. 5 the synthetic speech waveform over frame k can be written as

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

where

$$\gamma_l^k = A_l^k \exp\left(j\theta_l^k \right) \quad (9A)$$

represents the lth complex amplitude for the lth component of the L^k sine waves. Since the measurements are made on digitized speech, sampled-data notation $[s(n)]$ is used. In this respect the time index n corresponds to the uniform samples of $t - t_k$; therefore n ranges from $-N/2$ to $N/2$, with $n = 0$ 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 Eq. 9 to the measured waveform, denoted by $y(n)$. A useful criterion for judging the quality of fit is the mean-squared error

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

Substituting the speech model of Eq. 9 into Eq. 10 leads to the error expression

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

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

The task of the estimator is to identify a set of sine waves that minimizes Eq. 11. Insights into the development of a suitable estimator can be obtained by restricting the class of input signals to the idealization of perfectly voiced speech, *ie*, speech that is periodic, hence having component sine waves that are harmonically related. In this case the synthetic speech waveform 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 Eq. 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

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

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 Eq. 14, the error can be written as

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

from which it follows that the optimal 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 calculation it follows, therefore, that the error is minimized by selecting all of the harmonic frequencies in the speech bandwidth, Ω (*ie*, $L^k = \Omega/\omega_0^k$).

Equations 15 and 17 completely specify the structure of the ideal estimator and show that the optimal estimator depends on the speech data through the STFT (Eq. 15). Although these results are equivalent to a Fourier-series representation of a periodic waveform, the results lead to an intuitive generalization to the more 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 (called a periodogram) will be pulse-like 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 associated with the peaks of the periodogram. This interpretation permits the extension of the estimator to a more generalized speech waveform, one that is not ideally voiced. This extension becomes evident when the STFT is calculated for the general sinusoidal speech model given in Eq. 9. In this case the STFT is simply

$$Y(\omega) = \sum_{l=1}^{L^k} \gamma_l^k \,\mathrm{sinc}\,(\omega_l^k - \omega) \; . \qquad (19)$$

Provided the analysis window is "wide enough" that

$$|\omega_i^k - \omega_l^k| \geqslant \frac{4\pi}{N+1} \; , \qquad (20)$$

then the periodogram can be written as

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

and, as before, the location of the peaks of the periodogram corresponds to the underlying sine-wave frequencies. The STFT samples at these frequencies correspond to the complex amplitudes. Therefore, provided Eq. 20 holds, the structure of the ideal estimator applies to a more general class of speech waveforms than perfectly voiced speech. Since, during steady voicing, neighboring frequencies are approximately seperated by the width of the pitch frequency, Eq. 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.

These properties are based on the assumption that the sinc (·) function is essentially zero outside of the region defined by Eq. 20. In fact, this approximation is not a valid one, because there will be sidelobes outside of this region due to the rectangular window implicit in the definition of the STFT. These sidelobes lead to leakage that compromises the performance of the estimator, a problem that is reduced, but not eliminated, by using the weighted STFT. Letting $\overline{Y}(\omega)$ denote the weighted STFT, *ie*,

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

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 $|\overline{Y}(\omega)|$. Letting these frequency estimates be denoted by $\{\hat{\omega}_l^k\}$, then the corresponding complex amplitudes are given by

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

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

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

Using the Hamming window for the weighted STFT provided a very good sidelobe structure in that the leakage problem was eliminated; it did so at the expense of broadening the main lobes of the periodogram. In order to accommodate this broadening, the constraint implied by Eq. 20 must be revised to require that the window width be at least 2.5 times the pitch period. This revision maintains the resolution features that were needed to justify the optimality properties of the periodogram processor. Although the window width could be set on the basis of the instantaneous pitch, the analyzer is less sensitive to the performance of the pitch extractor if the window width is set on the basis of the average pitch instead. The pitch computed during strongly voiced frames is averaged using a 0.25-s time constant, and this averaged pitch is used to update, in real time, the width of the analysis window. During frames of unvoiced speech, the window is held fixed at the value obtained on the preceding voiced frame or 20 ms, whichever is smaller.

Once the width for a particular frame has been specified, the Hamming window is computed and normalized according to Eq. 24, and the STFT of the input speech is taken using a 512-point fast Fourier transform (FFT) for 4-kHz bandwidth speech. A typical periodogram for voiced speech, along with the amplitudes and frequencies that are estimated using the above procedure, is plotted in Fig. 3.

In order to apply the sinusoidal model to unvoiced speech, the frequencies corresponding to the periodogram peaks must be close enough to satisfy the requirement imposed by the Karhunen-Loève expansion [5] for noiselike signals. If the window width is constrained to be at least 20 ms wide, on average, the corresponding periodogram peaks will be approximately 100 Hz apart, enough to satisfy the constraints of the Karhunen-Loève sinusoidal representation for random noise. A typical periodogram for a frame of unvoiced speech, along with the estimated amplitudes and frequencies, is plotted in Fig. 4.

This analysis provides a justification for representing speech waveforms in terms of the amplitudes, frequencies, and phases of a set of sine waves. However, each sine-wave representation applies only to one analysis frame; different sets of these parameters are obtained for each frame. The next problem to address, then, is how to associate the amplitudes, frequencies, and phases measured on one frame with those found on a successive frame.

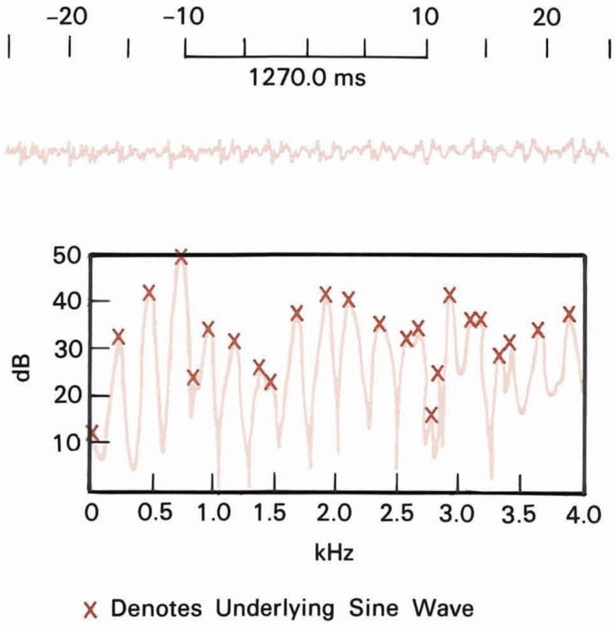

Fig. 3 — This is a typical periodogram of voiced speech. The amplitude peaks of the periodogram determine which frequencies are chosen to represent the speech waveform. The amplitudes of the underlying sine waves are denoted with an x.

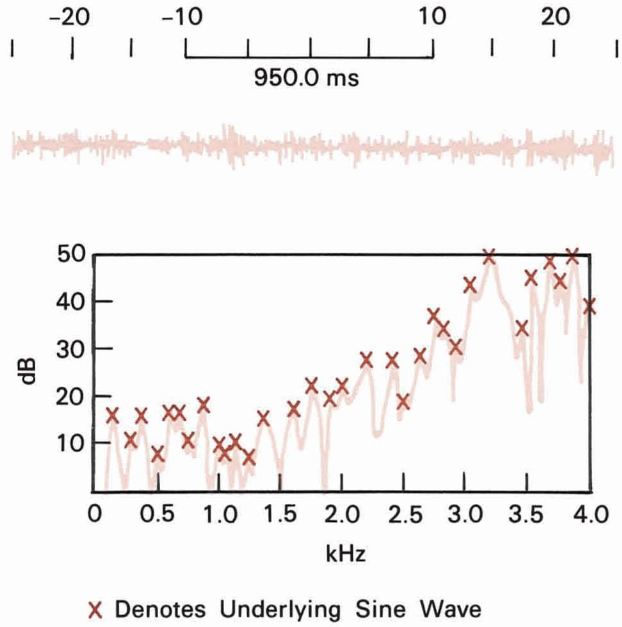

Fig. 4 — This periodogram illustrates how the power is shifted to higher frequencies in unvoiced speech. The amplitudes of the underlying sine waves are denoted with an x.

Frame-to-Frame Peak Matching

If the number of periodogram peaks were constant from frame to frame, the peaks could simply be matched between frames on a frequency-ordered basis. In practice, however, there are spurious peaks that come and go because of the effects of sidelobe interaction. (The Hamming window doesn't completely eliminate sidelobe interaction.) Additionally, peak locations change as the pitch changes and there are rapid changes in both the location and the number of peaks corresponding to rapidly varying regions of speech, such as at voiced/unvoiced transitions. The analysis system can accommodate these rapid changes through the incorporation of a nearest-neighbor frequency tracker and the concept of the "birth" and "death" of sinusoidal components. A detailed description of this tracking algorithm is given in Ref. 6.

An illustration of the effects of the procedure used to account for extraneous peaks is shown in Fig. 5. The results of applying the tracker to a segment of real speech is shown in Fig. 6, which demonstrates the ability of the tracker to adapt quickly through such transitory speech behavior as voiced/unvoiced transitions and mixed voiced/unvoiced regions.

The Synthesis System

After the execution of the preceding parameter-extraction and peak-matching procedures, the information captured in those steps can be used in conjunction with a synthesis system to produce natural-sounding speech. 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 k th frame by using the equation

$$\tilde{s}(n) = \sum_{l=1}^{L(k)} \hat{A}_l^k \cos\,[n\hat{\omega}_l^k + \hat{\theta}_l^k] \qquad (25)$$

where $n = 0, 1, 2, \ldots, S - 1$ and where S is the length of the synthesis frame. Because of the time-varying nature of the parameters, however, this straightforward approach leads to discontinuities at the frame boundaries. The discontinuities seriously degrade the quality of the synthetic speech. To circumvent this problem, the parameters must be smoothly interpolated from frame to frame.

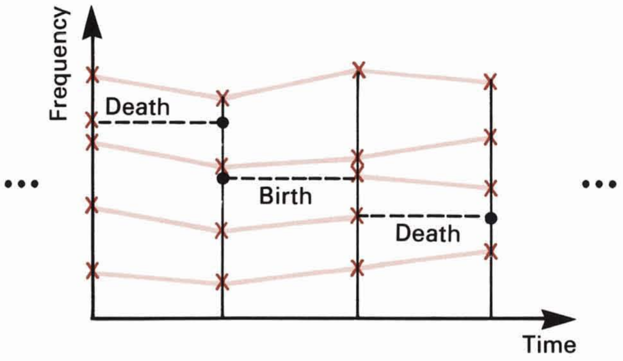

Fig. 5 — Different modes used in the birth/death frequency-track-matching process. Note the death of two tracks during frames one and three and the birth of a track during the second frame.

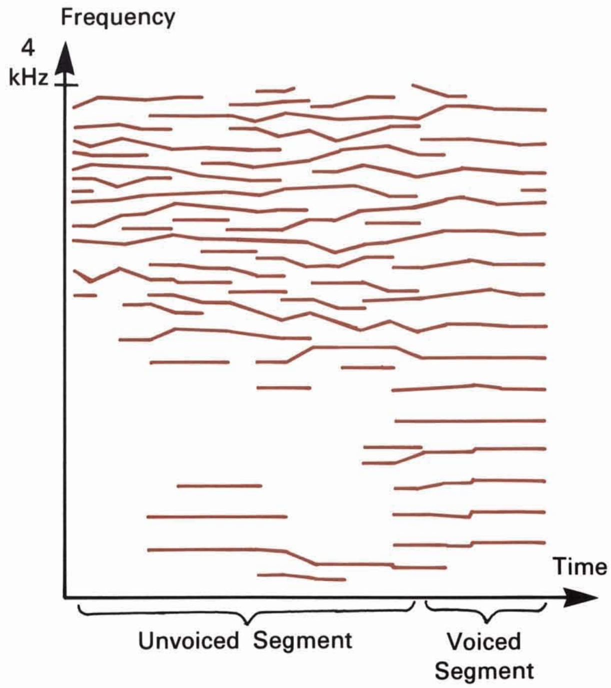

Fig. 6 — These frequency tracks were derived from real speech using the birth/death frequency tracker.

The most straightforward approach for performing this interpolation is to overlap and add time-weighted segments of the sinusoidal components. This process uses 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. Real-time systems using this technique were operated with frames separated by 11.5 ms and 20.0 ms, respectively.

While the synthetic speech produced by the first system was quite good, and almost indistinguishable from the original speech, the longer frame interval produced synthetic speech that was rough and, though quite intelligible, deemed to be of poor quality. Therefore, the overlap-add synthesizer is useful only for applications that can support a high frame rate. But there are many practical applications, such as speech coding, that require lower frame rates. These applications require an alternative to the overlap-add interpolation scheme.

A method will now be described that interpolates the matched sine-wave parameters directly. Since the frequency-matching algorithm associates all of the parameters measured for an arbitrary frame, k, with a corresponding set of parameters for frame $k + 1$, then letting

$$(\hat{A}_l^k, \hat{\omega}_l^k, \hat{\theta}_l^k) \text{ and } (\hat{A}_l^{k+1}, \hat{\omega}_l^{k+1}, \hat{\theta}_l^{k+1}) \tag{25A}$$

denote the successive sets of parameters for the lth frequency track, a solution to the amplitude-interpolation problem is to take

$$\tilde{A}(n) = \hat{A}^k + \frac{(\hat{A}^{k+1} - \hat{A}^k)}{S} n , \tag{26}$$

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

Unfortunately, this simple approach cannot be used to interpolate the frequency and phase because the measured phase, $\hat{\theta}^k$, is obtained modulo 2π (Fig. 7). Hence, phase unwrapping must be performed to ensure that the frequency tracks are "maximally smooth" across frame boundaries. One solution to this problem that uses a phase-interpolation function that is a cubic polynomial has been developed in Ref. 6. The phase-unwrapping procedure provides each frequency track with an instantaneous unwrapped phase such that the frequencies and phases at the frame boundaries are consistent with the measured values modulo 2. The unwrapped phase accounts for both the rapid phase changes due to the frequency of each sinusoidal component, and to the slowly varying phase changes due to the glottal pulse and the vocal-tract transfer function.

Letting $\tilde{\theta}_l(t)$ denote the unwrapped phase function for the l th track, then the final synthetic waveform will be given by

$$\tilde{s}(n) = \sum_{l=1}^{L^k} \tilde{A}_l(n) \cos[\tilde{\theta}_l(n)] \tag{27}$$

where $\tilde{A}_l(n)$ is given by Eq. 26, $\tilde{\theta}_l(n)$ is given by

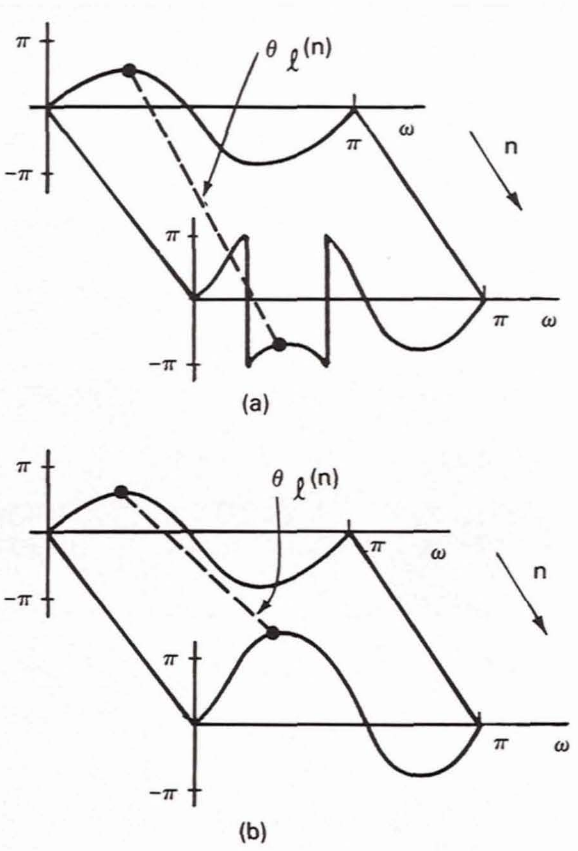

Fig. 7 — Requirement for unwrapped phase. (a) Interpolation of wrapped phase. (b) Interpolation of unwrapped phase.

$$\tilde{\theta}_l(n) = \hat{\theta}_l^k + \hat{\omega}_l^k n + \alpha_l^k n^2 + \beta_l^k n^3 \tag{28}$$

and L^k is the number of sine waves estimated for the k th frame.

Experimental Results

Figure 8 gives a block diagram description of the complete analysis/synthesis system. A non-real-time floating-point simulation was developed initially 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, with a width 2.5 times the average pitch, was used and found to be sufficient for accurate peak

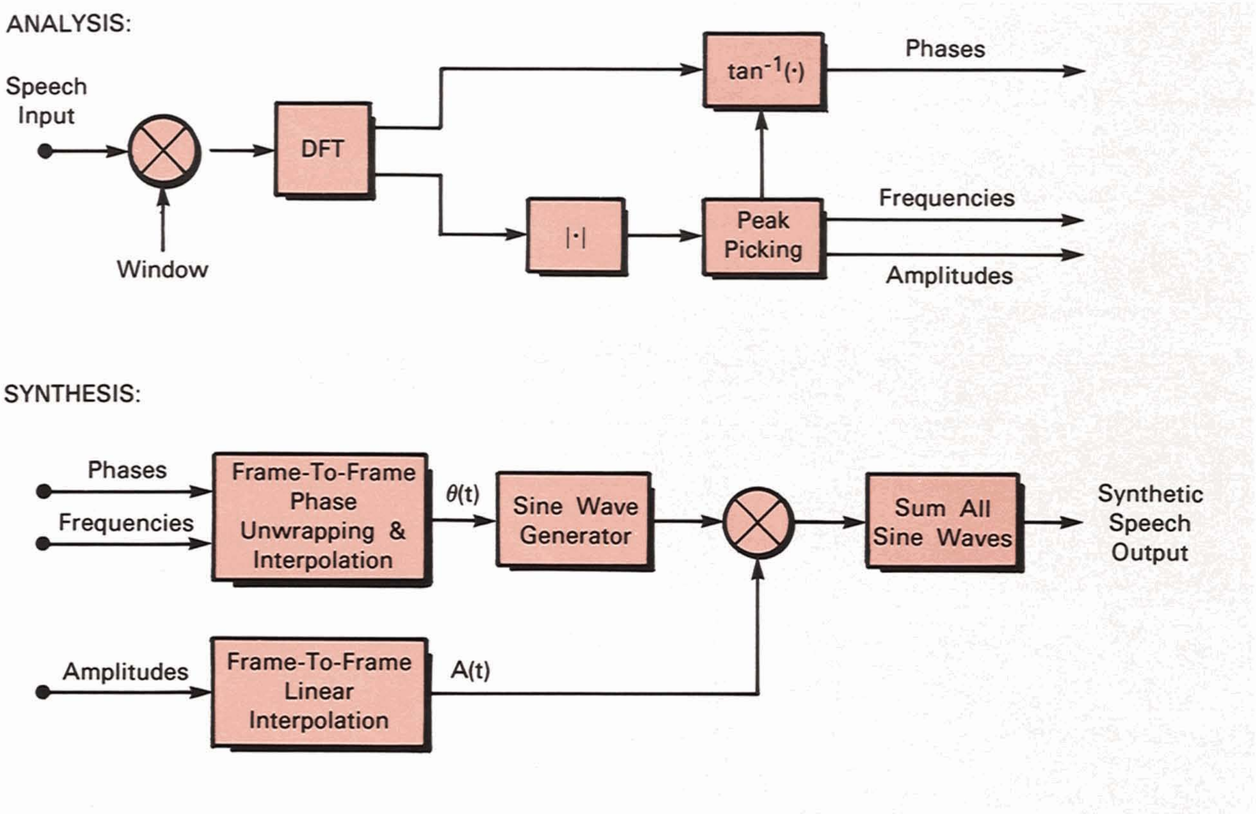

Fig. 8 — This block diagram of the sinusoidal analysis/synthesis system illustrates the major functions subsumed within the system. Neither voicing decisions nor residual waveforms are required for speech synthesis.

estimation. The maximum number of peaks used in synthesis was set to a fixed number (~80); if excess peaks were obtained only the largest peaks were used.

A large speech data base was processed with this system, and the synthetic speech was essentially indistinguishable from the original. A visual examination of the reconstructed passages shows that the waveform structure is essentially preserved. This is illustrated by Fig. 9, which compares the waveforms of the original speech and the reconstructed speech during an unvoiced/voiced speech transition. The comparison suggests that the quasi-stationary conditions imposed on the speech model are met and that the use of the parametric model based on the amplitudes, frequencies, and phases of a set of sine-wave components is justified for both voiced and unvoiced speech.

In another set of experiments it was found that the system was capable of synthesizing a broad class of signals including multispeaker waveforms, music, speech in a music background, and marine biologic signals such as whale sounds. Furthermore, the reconstruction does not break down in the presence of noise. The synthesized speech is perceptually 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 Ref. 6. More recently a real-time system has been completed using the Analog Devices ADSP2100 16-bit fixed-point signal-processing chips and performance equal to that of the simulation has been achieved.

Although high-quality analysis/synthesis of speech has been demonstrated using the amplitudes, frequencies, and phases of the peaks

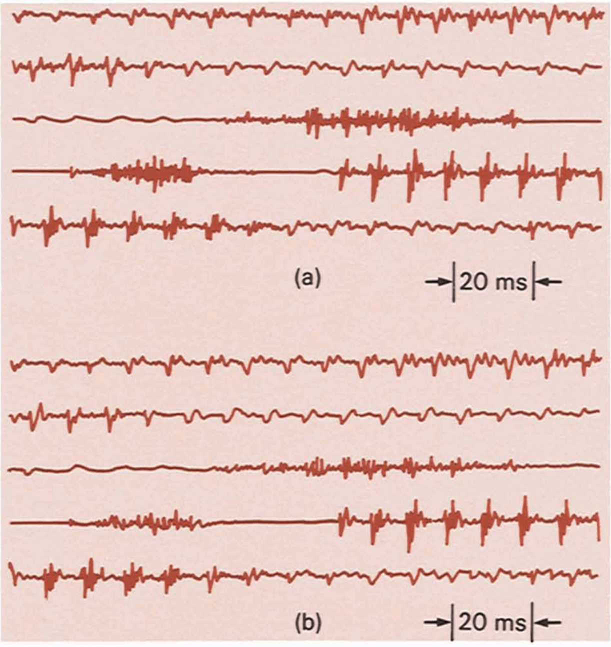

Fig. 9 — (a) Original speech. (b) Reconstructed speech. Both voiced and unvoiced segments are compared to the voiced/unvoiced segment of speech that has been reconstructed with the sinusoidal analysis/synthesis system. The waveforms are nearly identical, justifying the use of the sinusoidal transform model.

of the high-resolution STFT, it is often argued that the ear is insensitive to phase, a proposition that forms the basis of much of the work in narrowband speech coders. The question arises whether or not the phase measurements are essential to the sine-wave 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 [3,7]. 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 and continued to be evaluated along that track until the 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, having a somewhat mechanical quality. Furthermore, the differences were more pronounced for low-pitched (*ie*, pitch <~100 Hz) speakers. An example of a waveform synthesized by the magnitude-only system is shown in Fig. 10 (b). Compared to the original speech, shown in Fig. 10 (a), the synthetic waveform is quite different because of the failure to maintain the true sine-wave phases. In an additional experiment the magnitude-only system was applied to the synthesis of noisy speech; the synthetic noise took on a tonal quality that was unnatural and annoying.

Vocoding

Since the parameters of the sinusoidal speech model are the amplitudes, frequencies, and phases of the underlying sine waves, and since, for a typical low-pitched speaker there can be as many as 80 sine waves in a 4-kHz speech bandwidth, it isn't possible to code all of the parameters directly. Attempts at parameter reduction use an estimate of the pitch to establish a set of harmonically related sine waves that provide a best fit to the input speech waveform.

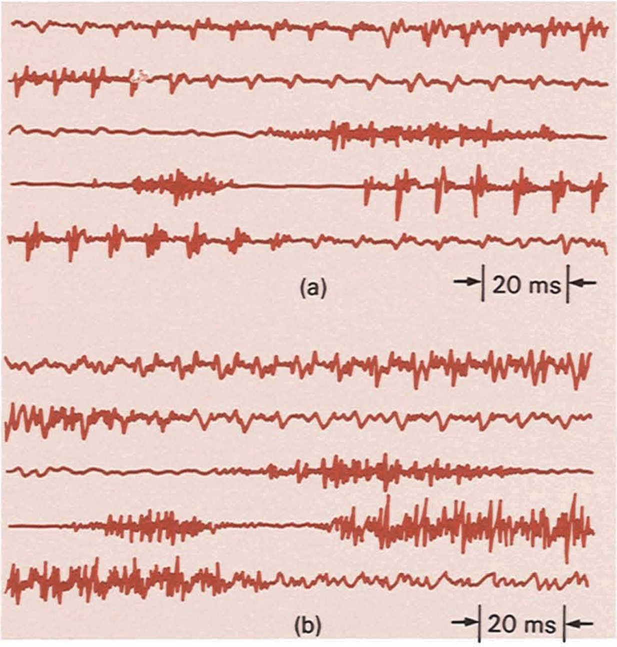

Fig. 10 — (a) Original speech. (b) Reconstructed speech. The original speech is compared to the segment that has been reconstructed using the magnitude-only system. The differences brought about by ignoring the true sine-wave phases are clearly demonstrated.

The amplitudes, frequencies, and phases of this reduced set of sine waves are then coded. Since the neighboring sine-wave amplitudes are naturally correlated, especially for low-pitched speakers, pulse-code modulation is used to encode the differential log amplitudes.

In the earlier versions of the vocoder all the sine-wave amplitudes were coded simply by allocating the available bits to the number to be coded. Since a low-pitched speaker can produce as many as 80 sine waves, then in the limit of 1 bit/sine-wave amplitude, 4000 bps would be required at a 50-Hz frame rate. For an 8-kbps channel, this leaves 4.0 kbps for coding the pitch, energy, and about 12 baseband phases. However, at 4.8 kbps and below, assigning 1 bit/ amplitude immediately exhausts the coding budget, so no phases can be coded. Therefore, a more efficient amplitude encoder had to be developed for operation at these lower rates.

The increased efficiency is obtained by allowing the channel separation to increase logarithmically with frequency, thereby exploiting the critical band properties of the ear. Rather than implement a set of bandpass filters to obtain the channel amplitudes, as is done in the channel vocoder, an envelope of the sine-wave amplitudes is constructed by linearly interpolating between sine-wave peaks and sampling at the critical band frequencies. Moreover, the location of the critical band frequencies was made pitch-adaptive whereby the baseband samples are linearly seperated by the pitch frequency and with the log-spacing applied to the high-frequency channels as required.

In order to preserve the naturalness at rates at 4.8 kbps and below, a synthetic phase model was employed that phase-locks all of the sine waves to the fundamental and adds a pitch-dependent quadratic phase dispersion and a voicing-dependent random phase to each sine wave [8]. Using this technique at 4.8 kbps, the synthesized speech achieved a diagnostic rhyme test (DRT) score of 94 (for three male speakers). A real-time, 4800-bps system that uses two ADSP2100 signal-processing chips has been successfully implemented. The technology developed from this research has been transferred to the commercial sector (CYLINK, INC), where a product is being produced that will be available in 1989.

Transformations

The goal of time-scale modification is to maintain the perceptual quality of the original speech while changing the apparent rate of articulation. This implies that the frequency trajectories of the excitation (and thus the pitch contour) are stretched or compressed in time and that the vocal tract changes at the modified rate. To achieve these rate changes, the system amplitudes and phases, and the excitation amplitudes and frequencies, along each frequency track are time-scaled. Since the parameter estimates of the unmodified synthesis are available as continuous functions of time, in theory any rate change is possible. Rate changes ranging between a compression of two and an expansion of two have been implemented with good results. Furthermore, the natural quality and smoothness of the original speech were preserved through transitions such as voiced/unvoiced boundaries. Besides the above constant rate changes, linearly varying and oscillatory rate changes have been applied to synthetic speech, resulting in natural-sounding speech that is free of artifacts [9].

Since the synthesis procedure consists of summing the sinusoidal waveforms for each of the measured frequencies, the procedure is ideally suited for performing various frequency transformations. The procedure has been employed to warp the short-time spectral envelope and pitch contour of the speech waveform and, conversely, to alter the pitch while preserving the short-time spectral envelope. These speech transformations can be applied simultaneously so that time- and frequency (or pitch)-scaling can occur together by simultaneously stretching and shifting frequency tracks. These joint operations can be performed with a continuously adjustable rate change.

Audio Preprocessing

The problem of preprocessing speech that is to be degraded by natural or man-made disturbances arises in applications such as attempts to increase the coverage area of AM radio broadcasts and in improving ground-to-air communications in a noisy cockpit environment. The transmitter in such cases is usually constrained by a peak operating power or the dynamic range of the system is limited by the sensitivity characteristics of the receiver or ambient-noise levels. Under these constraints, phase dispersion and dynamic-range compression, along with spectral shaping (pre-emphasis), are combined to reduce the peak-to-rms ratio of the transmitted signal. In this way, more average power can be broadcast without changing the peak-power output of the transmitter, thus increasing loudness and intelligibility at the receiver while adhering to peak-output power constraints.

This problem is similar to one in radar in which the signal is periodic and given as the output of a transmit filter the input of which consists of periodic pulses. The spectral magnitude of the filter is specified and the phase of the filter is chosen so that, over one period, the signal is frequency-modulated (a linear frequency modulation is usually employed) and has a flat envelope over some specified duration. If there is a peak-power limitation on the transmitter, this approach imparts maximal energy to the waveform. In the case of speech, the voiced-speech signal is approximately periodic and can be modeled as the output of a vocal-tract filter the input of which is a pulse train. The important difference between the radar design problem and the speech-audio preprocessing problem is that, in the case of speech, there is no control over the magnitude and phase of the vocal-tract filter. The vocal-tract filter is characterized by a spectral magnitude and some natural phase dispersion. Thus, in order to take advantage of the radar signal design solutions to dispersion, the natural phase dispersion must first be estimated and removed from the speech signal. The desired phase can then be introduced. All three of these operations, the estimation of the natural phase, the removal of this phase, and its replacement with the desired phase, have been implemented using the STS.

The dispersive phase introduced into the speech waveform is derived from the measured vocal-tract spectral envelope and a pitch-period estimate that changes as a function of time; the dispersive solution thus adapts to the time-varying characteristics of the speech waveform. This phase solution is sampled at the sine-wave frequencies and linearly interpolated across frames to form the system phase component of each sine wave [10].

This approach lends itself to coupling phase dispersion, dynamic-range compression, and pre-emphasis via the STS. An example of the application of the combined operations on a speech waveform is shown in Fig. 11. A system using these techniques produced an 8-dB reduction in peak-to-rms level, about 3 dB better than commercial processors. This work is currently being extended for further reduction of peak-to-rms level and for further improvement of quality and robustness. Overall system performance will be evaluated by field tests conducted by Voice of America over representative transmitter and receiver links.

Conclusions

A sinusoidal representation for the speech waveform has been developed; it extracts the amplitudes, frequencies, and phases of the component sine waves from the short-time Fourier transform. 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 maximally smooth phase-interpolation function is obtained that is consistent with all of the frequency and phase measurements. This phase function is applied to a sine-wave generator which is amplitude-modulated and added to the other sine waves to form the output speech. 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 Flanagan has proposed [11,12]. 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. 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 are used at any one frame. It is the use of the frequency tracker and the phase interpolator that allows the filter bank to move with the highly resolved speech components. Therefore, the system fits into the framework Flanagan described 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 with 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 (*ie*, pitch <~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.

Using 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 time-scale, pitch-scale, frequency modification of speech, and speech-coding applications such a functional model is essential. However, if the system is used merely to produce synthetic speech, using a set of sine waves, then the frequency-tracking and phase-interpolation procedures are unnecessary. In this case, the interpolation is achieved by overlapping and adding time-weighted segments of each of the sinusoidal components. The resulting synthetic speech is essentially indistinguishable from the original speech as long as the frame rate is at least 100 Hz.

A fixed-point 16-bit real-time implementation of the system has been developed on the Lincoln Digital Signal Processors [13] and using the Analog Devices ADSP2100 processors. Diagnostic rhyme tests have been performed and

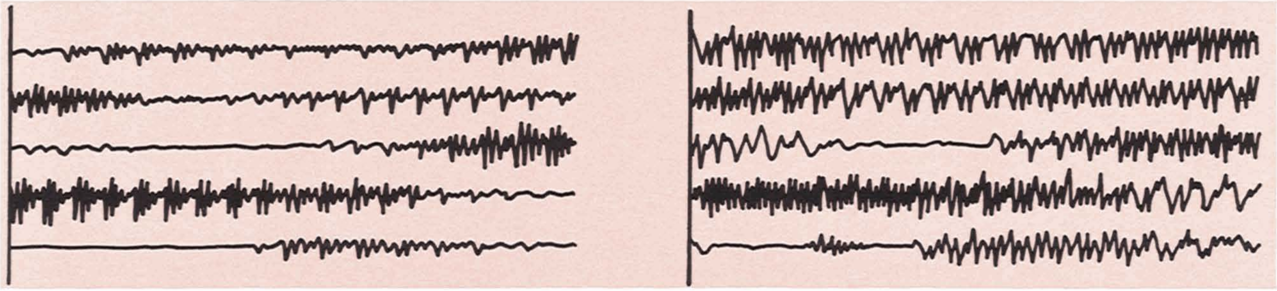

Fig. 11 — Audio preprocessing using phase dispersion and dynamic-range compression illustrates the reduction in the peak-to-rms level.

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. The system is used in research aimed at the development of a multirate speech coder [14]. A practical low-rate coder has been developed at 4800 bps and 2400 bps using two commercially available DSP chips. Furthermore, the resulting technology has been transferred to private industry for commercial development. The sinusoidal analysis/synthesis system has also been applied successfully to problems in time scale, pitch scale, and frequency modification of speech [9] and to the problem of speech enhancement for AM radio broadcasting [10].

Acknowledgments

The authors would like to thank their colleague, Joseph Tierney, for his comments and suggestions in the early stages of this work. The authors would also like to acknowledge the support and encouragement of the late Anton Segota of RADC/EEV, who was the Air Force program manager for this research from its inception in September 1983 until his death in March 1987.

This work was sponsored by the Department of the Air Force.

References

1. B.S. Atal and J.R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," *Int. Conf. on Acoustics, Speech, and Signal Processing 82* (IEEE, New York, 1982), p 614.
2. M.R. Schroeder and B.S. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," *Int. Conf. on Acoustics, Speech, and Signal Processing 85* (IEEE, New York, 1985), p 937.
3. P. Hedelin, "A Tone-Oriented Voice-Excited Vocoder," *Int. Conf. on Acoustics, Speech, and Signal Processing 81* (IEEE, New York, 1981), p 205 .
4. L.B. Almeida and F.M. Silva, "Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme," *Int. Conf. on Acoustics, Speech, and Signal Processing 84* (IEEE, New York, 1984), p 27.5.1.
5. H. Van Trees, *Detection, Estimation and Modulation Theory, Part I* (John Wiley, New York, 1968), Chap 3.
6. R.J. McAulay and T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34,** 744 (1986).
7. R.J. McAulay and T.F. Quatieri, "Magnitude-Only Reconstruction Using a Sinusoidal Speech Model," *Int. Conf. on Acoustics, Speech, and Signal Processing 84* (IEEE, New York, 1984), p 27.6.1.
8. R.J. McAulay and T.F. Quatieri, "Multirate Sinusoidal Transform Coding at Rates from 2.4 kbps to 8.0 kbps," *Int. Conf. on Acoustics, Speech, and Signal Processing 87* (IEEE, New York, 1987), p 1645.
9. T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34,** 1449 (Dec. 1986).
10. T.F Quatieri and R.J. McAulay, "Sinewave-based Phase Dispersion for Audio Preprocessing," *Int. Conf. on Acoustics, Speech, and Signal Processing 88* (IEEE, New York, 1988), p 2558.
11. J.L. Flanagan, "Parametric Coding of Speech Spectra," *J. Acoust. Soc. of America* **68,** 412 (1980).
12. J.L. Flanagan and S.W. Christensen, "Computer Studies on Parametric Coding of Speech Spectra," *J. Acoust. Soc. of America* **68,** 420 (1980).
13. P.E. Blankenship, "LDVT: High Performance Minicomputer for Real-Time Speech Processing," *EASCON '75* (IEEE, New York, 1975), p 214-A.
14. R.J. McAulay and T.F. Quatieri, "Mid-Rate Coding Based on a Sinusoidal Representation of Speech," *Int. Conf. on Acoustics, Speech, and Signal Processing 85* (IEEE, New York, 1985), p 945.

ROBERT J. McAULAY is a senior staff member in the Speech Systems Technology Group. He received a BASc degree in engineering physics (with Honors) from the University of Toronto in 1962, an MSc degree in electrical engineering from the University of Illinois in 1963, and a PhD degree in electrical engineering from the University of California, Berkeley, in 1967. In 1967 he joined Lincoln Laboratory and worked on problems in estimation theory and signal/filter design using optimal control techniques. From 1970 to 1975 he was a member of the Air Traffic Control Division and worked on the development of aircraft tracking algorithms. Since 1975 he has been involved with the development of robust narrow-band speech vocoders. Bob received the M. Barry Carlton award in 1978 for the best paper published in the IEEE Transactions on Aerospace and Electronic Systems. In 1987 he was a member of the panel on Removal of Noise from a Speech/Noise Signal for Bioacoustics and Biomechanics of the National Research Council's Committee on Hearing.

THOMAS F. QUATIERI is a staff member in the Speech Systems Technology Group, where he is working on problems in digital signal processing and their application to speech enhancement, speech coding, and data communications. He received a BS degree (summa cum laude) from Tufts University and SM, EE, and ScD degrees from Massachusetts Institute of Technology in 1975, 1977, and 1979, respectively. He was previously a member of the Sensor Processing Technology group involved in multidimensional digital signal processing and image processing. Tom received the 1982 Paper Award of the IEEE Acoustics, Speech and Signal Processing Society for the best paper by an author under thirty years of age. He is a member of the IEEE Digital Signal Processing Technical Committee, Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.