and panels in which frequent change of personnel is customarily experienced.

References

[1] H. M. Moser and J. J. Dreher, "Effects of training on listeners in intelligibility studies," *J. Acoust. Soc. Amer.*, vol. 27, pp. 1213-1219, Nov. 1955.

[2] C. W. Stuckey, "Investigation of the precision of an articulation testing program," *J. Acoust. Soc. Amer.*, vol. 35, pp. 1782-1787, Nov. 1963.

[3] I. J. Hirsch *et al.*, "Development of materials for speech audiometry," *J. Speech Hear. Disorders*, vol. 17, pp. 321-327, Sept. 1952.

[4] G. E. Peterson and I. Lehiste, "Revised CNC lists for audiometry tests," *J. Speech Hear. Disorders*, vol. 27, pp. 62-70, Feb. 1962.

[5] G. Fairbanks, "Test of phonemic differentiation: the rhyme test," *J. Acoust. Soc. Amer.*, vol. 30, pp. 596-600, July 1958.

[6] A. S. House *et al.*, "Articulation testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, pp. 158-166, Jan. 1965.

[7] B. B. Bauer, E. L. Torick, and R. G. Allen, "The measurement of loudness level," *J. Acoust. Soc. Amer.*, vol. 50, pp. 405-414, Aug. 1971.

[8] S. E. Gerber and P. Milner, "The transitivity of loudness level," *J. Audio Eng. Soc.*, vol. 19, pp. 656-659, Sept. 1971.

Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis

RONALD W. SCHAFER and LAWRENCE R. RABINER

Abstract—This paper discusses the theoretical basis for representation of a speech signal by its short-time Fourier transform. The results of the theoretical studies were used to design a speech analysis-synthesis system which was simulated on a general-purpose laboratory digital computer system. The simulation uses the fast Fourier transform in the analysis stage and specially designed finite duration impulse response filters in the synthesis stage. The results of both the theoretical and computational studies lead to an understanding of the effect of several design parameters and elucidate the design tradeoffs necessary to achieve moderate information rate reductions.

I. Introduction

The phase vocoder [1] is a system for representing a speech signal by its complex short-time Fourier transform. In its digital form, the short-time Fourier transform representation of a speech signal requires an information rate between that of conventional channel vocoders and waveform coding systems such as PCM and delta modulation. The short-time Fourier representation is of interest because it does not require any form of source coding such as pitch tracking or voiced-unvoiced detection and yet it offers considerable flexibility in manipulating the basic speech parameters. In addition, analysis-synthesis systems such as the phase vocoder may offer advantages for realization with multiplexed digital hardware.

In this paper we summarize some theoretical results [2] that have important implications for design of speech analysis-synthesis systems based on short-time Fourier analysis. We show how digital signal processing techniques such as the fast Fourier transform and interpolation using finite duration impulse response filters can be effectively employed in the simulation and realization of speech coding systems. These results indicate the important design parameters and also facilitate understanding of design tradeoffs necessary to achieve moderate information rate reductions.

Manuscript received April 30, 1972.
The authors are with Bell Laboratories, Murray Hill, N.J. 07974.

II. Discrete Short-Time Fourier Analysis and Synthesis

If a speech signal is passed through a bank of ideal bandpass filters whose passbands are contiguous and precisely cover the speech band, then the sum of the filter outputs is equal to the input. With careful design, realizable bandpass filters can similarly be used with small degradation in quality. This is the fundamental principle underlying the representation of speech by its complex short-time Fourier transform.

In this section we show how the short-time transform is related to a bank of bandpass filters and discuss some modifications and improvements in the conventional formulation of short-time Fourier analysis and synthesis.

Consider a set of equally spaced causal bandpass digital filters whose impulse responses are

$$h_k(nT) = h(nT) \cos [\omega_k nT] \tag{1}$$

where $h(nT)$ is the impulse response of a causal prototype low-pass filter, $\omega_k = \Delta\omega \cdot k$ for $k = 1, 2, \cdots, M$ (covering the desired speech band), and T is the sampling period. If $x(nT)$ is the sampled input speech

signal and $y_k(nT)$ is the output of the kth bandpass filter, then the sum of the outputs is[1]

$$y(nT) = \sum_{k=1}^{M} y_k(nT). \tag{2}$$

Using discrete convolution, the output of the kth filter is

$$y_k(nT) = \sum_{r=-\infty}^{n} x(rT)\, h(nT - rT) \cos [\omega_k (nT - rT)]$$

$$= \text{Re}\, \{e^{j\omega_k nT}\, X(\omega_k, nT)\} \tag{3}$$

where

$$X(\omega_k, nT) = \sum_{r=-\infty}^{n} x(rT)\, h(nT - rT)\, e^{-j\omega_k rT},$$

$$= a(\omega_k, nT) - jb(\omega_k, nT) \tag{4}$$

is the short-time Fourier transform of the input $x(rT)$ for the data window $h(nT - rT)$, evaluated at frequency ω_k and time nT. The real and imaginary parts of $X(\omega_k, nT)$ can be expressed as

$$a(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \cos [\omega_k rT] \tag{5}$$

and

$$b(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \sin [\omega_k rT]. \tag{6}$$

From (3), we can write

$$y_k(nT) = a(\omega_k, nT) \cdot \cos [\omega_k nT] + b(\omega_k, nT) \cdot \sin [\omega_k nT]. \tag{7}$$

Equations (4), (5), and (6) define the operations required for short-time Fourier analysis, and (2) and (7) define the method of synthesis from the short-time Fourier transform representation of the speech signal. For a single channel, Fig. 1(a) depicts the method of computing the short-time Fourier transform suggested by (5) and (6). Fig. 1(b) shows the method of synthesis from the real and imaginary parts of the short-time transform as given by (7). It can be shown that if the systems of Fig. 1(a) and (b) are connected back-to-back, the effective impulse response of the kth channel is $h_k(nT) = h(nT) \cos [\omega_k nT]$.

The fidelity with which the synthesized output $y(nT)$ matches the input $x(nT)$ is a question of basic importance. This question is equivalent to inquiring as to how closely the composite frequency response relating $x(nT)$ and $y(nT)$ approaches the ideal of flat amplitude response and linear phase response, or how closely the composite impulse response approaches a delayed discrete-time impulse. Since ideal filters with

[1] Note that we have omitted the channel centered at zero frequency which is not normally of interest in speech coding.

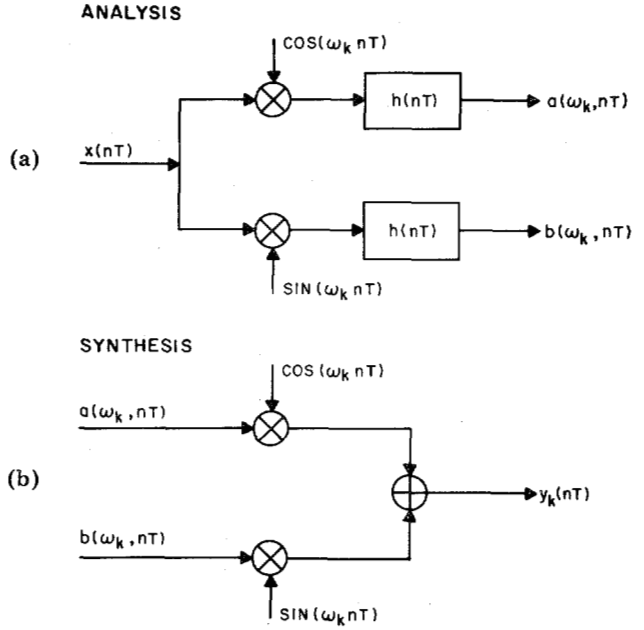

Fig. 1. Basic operations for short-time Fourier analysis (a) and synthesis (b).

infinite attenuation are not available in practice, the frequency responses of adjacent channels must overlap. This can lead to significant deviation from the ideal composite response—especially at the edges of the bands.

To see how to achieve a good approximation to the ideal composite response, it is convenient to study the composite impulse response $\tilde{h}(nT)$, which is simply the sum of the impulse responses of the individual channels. If M channels are used in the analysis, we obtain

$$\tilde{h}(nT) = h(nT) \sum_{k=1}^{M} \cos [\omega_k nT]$$

$$= h(nT)\, d(nT) \tag{8}$$

where

$$\omega_k = \Delta\omega \cdot k.$$

Thus we see that $\tilde{h}(nT)$ is the product of $h(nT)$, the low-pass filter impulse response, and a sequence denoted $d(nT)$ which is dependent only on the number of channels and the spacing of the channels. It can be shown [2], that if $\Delta\omega = 2\pi/NT$, where N is an integer, then the sequence $d(nT)$ is periodic with period N and is given by

$$d(nT) = \frac{\sin [(M + 1/2)\, \Delta\omega nT]}{[\sin \Delta\omega nT/2]} - 1. \tag{9}$$

This sequence has peaks at intervals of NT seconds. If $2\pi/(\Delta\omega T)$ is not an integer, the sequence $d(nT)$ is not periodic but still has peaks at intervals of NT seconds [2].

A particularly interesting choice of parameters is as

follows: let N be an odd integer[2] and $M = (N - 1)/2$. For $\Delta\omega = (2\pi)/NT$, it can easily be seen that this corresponds to evaluating the short-time Fourier transform at equally spaced frequencies in the range $0 < \omega < \pi/T$. If, in addition, we include a channel centered on zero frequency, it can be shown [2] that

$$d(nT) = \frac{\sin(\pi n)}{\sin(\pi n/N)}$$

$$= N, \qquad n = 0, \pm N, \pm 2N, \cdots$$

$$= 0, \qquad \text{elsewhere.}$$

Thus, for these conditions, $d(nT)$ is a periodic train of impulses, with a period NT which is inversely proportional to the frequency spacing between channels. Since $\tilde{h}(nT) = h(nT)\, d(nT)$, it is clear that the composite impulse response will also be an impulse train. Since the ideal composite impulse response is a delayed impulse, we must choose the prototype low-pass impulse response $h(nT)$ so as to eliminate all but one of the impulses in $d(nT)$. Suppose we fix T and N, corresponding to fixed frequency spacing $\Delta\omega$. Then if we choose a very narrow impulse response, e.g., of duration less than $2N$, the composite impulse response will appear as in Fig. 2(a). Here we have shown the prototype low-pass response or data window as a dotted curve superimposed on the impulse train that represents the composite response. Clearly there is only one impulse; however, such a narrow impulse response $h(nT)$ corresponds to a rather wide-band low-pass filter which would not give satisfactory frequency resolution. If we use a narrower bandwidth filter, the impulse response will become proportionately greater in duration as in Fig. 2(b) where we note that the composite impulse response consists of several impulses which would give rise to a reverberant quality in the output speech. Thus we see that good frequency resolution, i.e., narrow-band channels, seems to be at odds with low reverberation. However, Fig. 2(c) suggests one way in which, at least theoretically, the output can match the input exactly. Here we have used a wider filter but have constrained the values of $h(nT)$ to be zero at integer multiples of the period N. In this case the composite response is a single impulse delayed by $2N$. Thus the output is a delayed and scaled replica of the input. Such a data window can be designed [3]. Therefore, the short-time Fourier transform can theoretically represent the speech signal exactly.

In many practical systems it will not be convenient to choose the parameters so that the composite response is as depicted in Fig. 2(c). However, the analysis and synthesis equations can be modified to effect further improvements even if the optimum

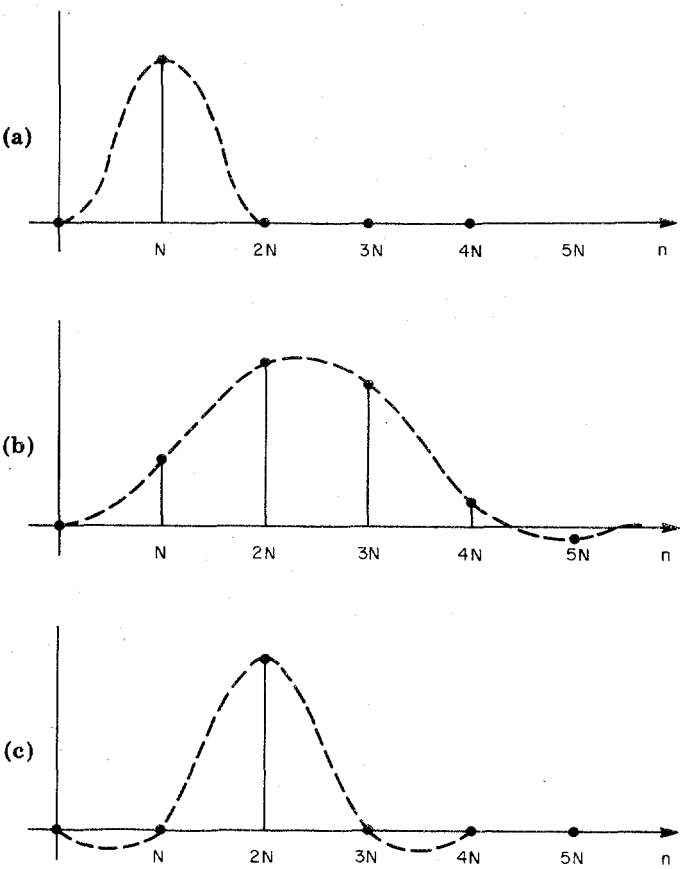

Fig. 2. Composite impulse responses for short-time Fourier analysis-synthesis for fixed frequency spacing $\Delta\omega$. (a) Wide-band filter. (b) Narrow-band filter. (c) Illustration of how the prototype filter can be designed to give perfect reconstruction of the input from the short-time transform.

response cannot be achieved. The example of Fig. 3 illustrates this point for an analysis-synthesis system of 30 channels spaced at 100 Hz intervals, with a sampling rate of 10 kHz [2]. The dotted curve on the left in Fig. 3(a) shows the low-pass impulse response $h(nT)$ (sixth-order Bessel filter), and the solid curve shows the composite impulse response $\tilde{h}(nT)$. From this latter curve it is possible to visualize the periodic pulse-like character of the sequence $d(nT)$ in the case when not all the channels are used in synthesis. We see that in addition to the main pulse at 10 ms, there is a significant echo at 20 ms. (The period of $d(nT)$ is 10 ms.) This echo manifests itself in the composite frequency response as an amplitude and phase ripple, (Fig. 3(b) and (c) on the left) and perceptually as a reverberant quality in the synthesized output. This example, together with the fact that $\tilde{h}(nT)$ is the product of $d(nT)$ and $h(nT)$, suggests two ways of improving the composite response.

As we have noted, for a given frequency spacing we could widen the bandwidth of the low-pass filter, thereby reducing the duration of $h(nT)$. As can be seen from Fig. 3(a) on the left, this would have the effect of increasing the amplitude of the first pulse and decreasing the amplitude of the second pulse. However, this means that we must effectively sacri-

[2] Similar results can be derived for N even [2].

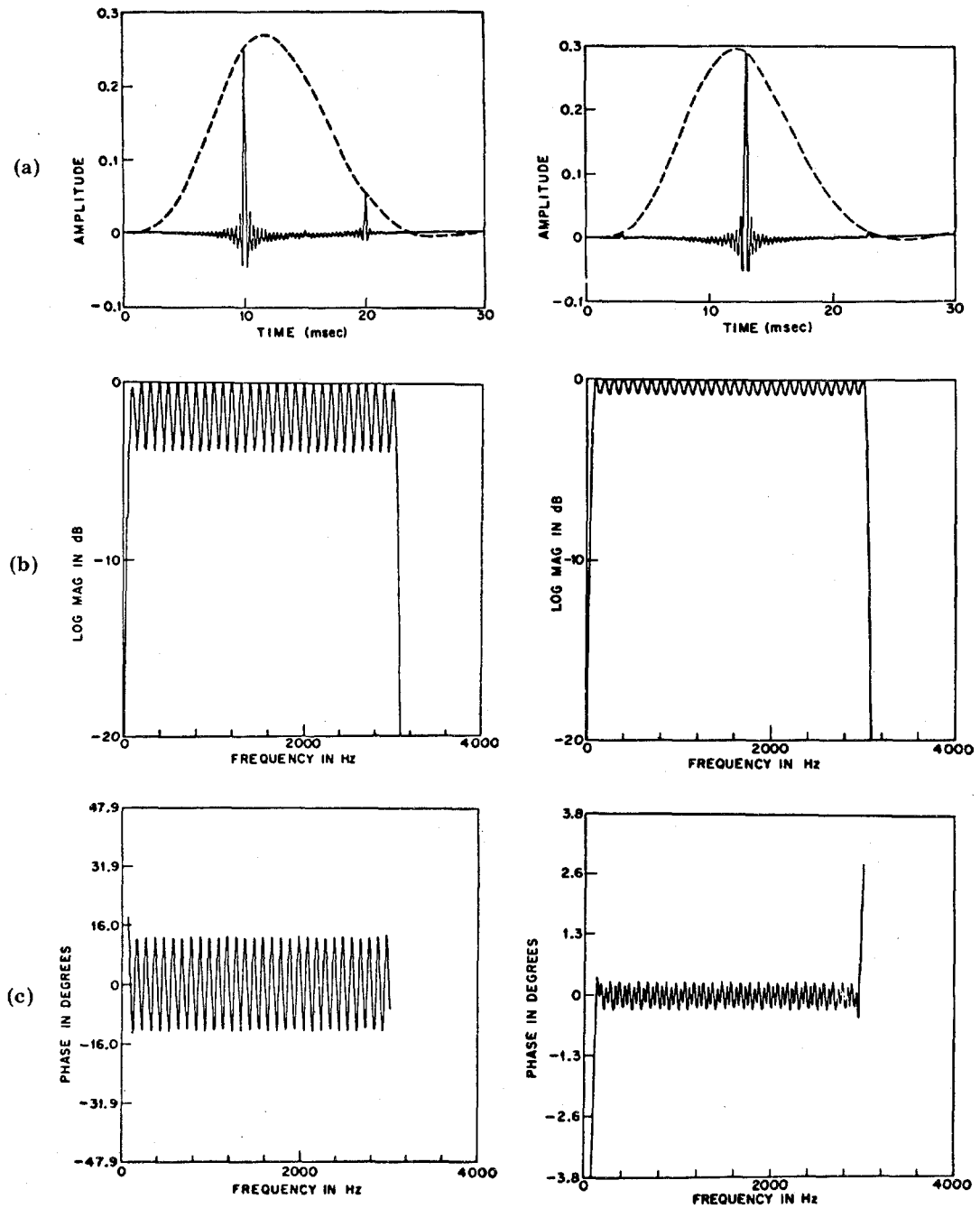

Fig. 3. Composite responses. Left column—no phase adjustment. Right column—best phase adjustment. (a) Impulse response [dotted curve is $h(nT)$]. (b) Magnitude response. (c) Phase response after subtracting linear phase corresponding to delay of main pulse.

fice frequency resolution. An alternative approach is suggested if we note that if $d(nT)$ could be shifted to the right relative to $h(nT)$ (the dotted curve), then the main pulse would grow in amplitude and the echo would be smaller. At the same time, however, the pulse in $d(nT)$ at $nT = 0$, which was completely suppressed by $h(nT)$, grows in amplitude as $d(nT)$ moves to the right. This is shown in Fig. 3(a) on the right. Thus, for a given frequency resolution there is an optimum delay of $d(nT)$ relative to $h(nT)$ for which $\tilde{h}(nT)$ consists of a large central pulse and two small pulses of equal size, one on each side of the main pulse. It can be shown [2] that this condition gives minimum amplitude and phase ripple for a given frequency resolution.

The mechanism for delaying $d(nT)$ relative to $h(nT)$ is available in either the analysis or the synthesis stage. If we change (5) and (6) to

$$a(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \cos\left[\omega_k (rT + n_a T)\right] \tag{10}$$

and

$$b(\omega_k, nT) = \sum_{r=-\infty}^{n} h(nT - rT)\, x(rT) \sin [\omega_k (rT + n_a T)] \tag{11}$$

where $n_a = n_0$, and we use (7) for synthesis, the effective impulse response of the kth channel will be

$$h_k(nT) = h(nT) \cos [\omega_k (nT - n_0 T)]$$

and the composite impulse response will be

$$\tilde{h}(nT) = h(nT) \cdot d(nT - n_0 T).$$

Alternatively, the same channel response will be obtained if we use (5) and (6) for analysis and substitute for (7),

$$y_k(nT) = a(\omega_k, nT) \cos [\omega_k (nT - n_s T)] + b(\omega_k, nT) \sin [\omega_k (nT - n_s T)] \tag{12}$$

where $n_s = n_0$. As a third possibility we can use (10) and (11) for analysis and (12) for synthesis if $n_a + n_s = n_0$. An interactive design program [2] facilitates the choice of the parameters of such systems so as to obtain composite responses similar to the right-hand side of Fig. 3.

III. Short-Time Fourier Analysis Using the FFT

Fig. 1(a) shows the basic analysis configuration for complex short-time Fourier analysis. The low-pass filters required for each channel can have an impulse response duration that is theoretically infinite, thus requiring a recursive realization, or we can use a finite duration impulse response filter realized either recursively or nonrecursively. In either case the computation required is quite time consuming, since all of the operations for a single channel must be repeated for each channel. Furthermore, when recursive realizations are employed, the channel outputs $a(\omega_k, nT)$ and $b(\omega_k, nT)$ must be computed at the sampling rate of the input speech signal even though it is clear that the channel signals have a bandwidth equal to the bandwidth of the low-pass filter. If we use a finite duration impulse response, however, this is not so. Clearly, we can easily compute the output of the low-pass filter at a lower sampling rate by simply skipping as many samples, at the input rate, as desired.

In this section, we show a method that employs the fast Fourier transform (FFT) [4] to compute the channel signals at sampling rates much lower than the input sampling rate. This scheme offers significant speed advantages for software realizations and it may be viewed as a simulation of a hardware realization which might employ multiplexing techniques to reduce the amount of required hardware.

The expressions for the short-time Fourier transform in (4) can be written

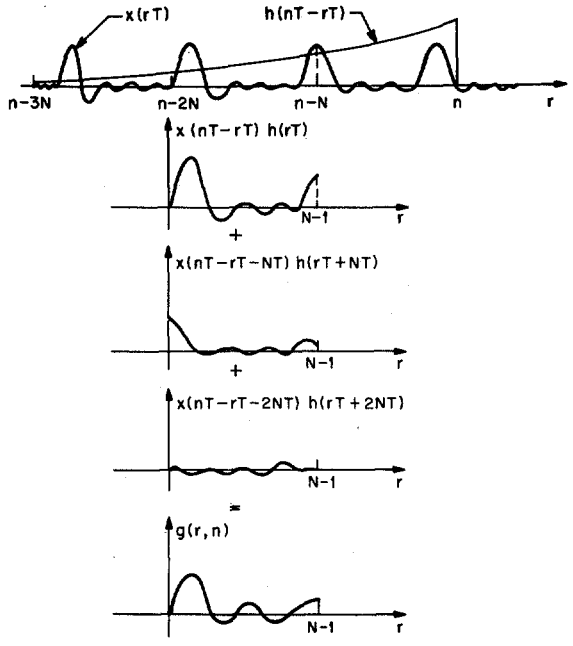

Fig. 4. Illustration of FFT method of computing the short-time Fourier transform.

$$X(\omega_k, nT) = \sum_{m=0}^{\infty} \sum_{r=n-(m+1)N+1}^{n-mN} x(rT) \cdot h(nT - rT)\, e^{-j\omega_k rT} \tag{13}$$

where we have rewritten the (possibly infinite) sum as a sum of finite sums over N samples. With a change of variable in the inner sum we can write

$$X(\omega_k, nT) = \sum_{m=0}^{\infty} \sum_{r=0}^{N-1} x(nT - rT - mNT) \cdot h(rT + mNT)\, e^{j\omega_k (r-n+mN)T}. \tag{14}$$

If we choose $\omega_k = (2\pi/NT)\, k$, i.e., equally spaced frequencies where $\Delta\omega = 2\pi/NT$, then we can take advantage of the periodicity of the complex exponential to write

$$X(\omega_k, nT) = e^{-j(2\pi/N)kn} \sum_{r=0}^{N-1} g(r, n)\, e^{j(2\pi/N)kr} \tag{15}$$

where

$$g(r, n) = \sum_{m=0}^{\infty} x(nT - rT - mNT)\, h(rT + mNT). \tag{16}$$

The factor

$$G(k, n) = \sum_{r=0}^{N-1} g(r, n)\, e^{j(2\pi/N)kr}, \qquad k = 0, 1, \cdots, N-1 \tag{17}$$

can be recognized as the discrete Fourier transform (DFT) [4] of the sequence $g(r, n)$ in (16). Fig. 4 aids in the interpretation of this method of computing $X(\omega_k, nT)$. At the top is shown a typical window

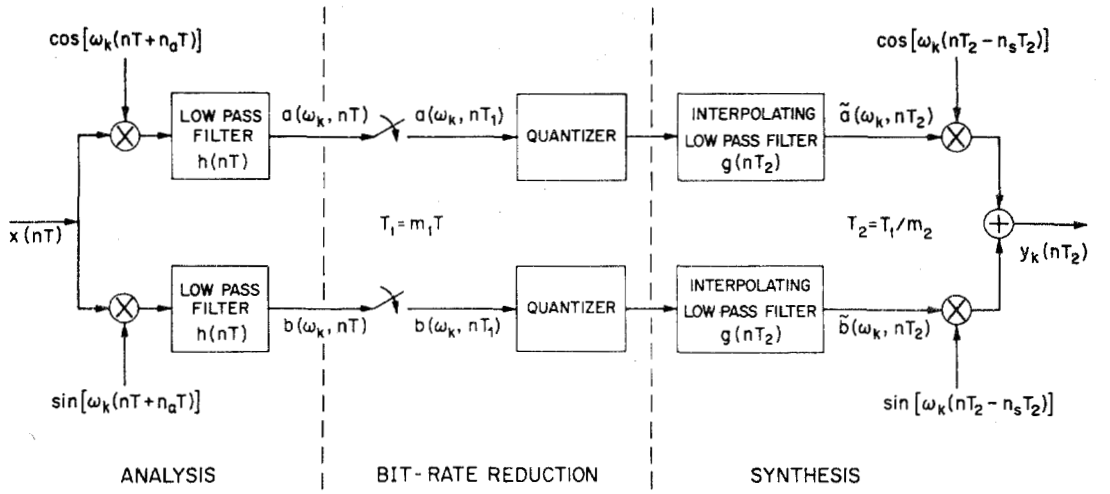

Fig. 5. Block diagram for one channel of a speech coding system based on short-time Fourier analysis.

$h(nT - rT)$ superimposed upon the input speech at a given time nT. The past values of the speech signal are weighted by the window. The resulting sequence is divided into segments of N samples and reversed in time, as shown in the next three plots. These segments are then added together to form the sequence $g(r, n)$ which is transformed using an FFT program. The number of segments required is dependent on the length of the impulse response $h(nT)$. If $h(nT)$ has finite length, only a finite number of segments are required. If the impulse response is infinite, an infinite number of segments are theoretically required; however, in practice the impulse response of a low-pass filter usually approaches zero rapidly so that an accurate simulation can still be performed using the above scheme with a truncated impulse response.

In order to use an FFT program to compute $G(k, n)$, N must be a power of two, or, if an appropriate program is available, it must at least be a composite number [4]. If the sampling rate is fixed, this places a strong limitation on the choice of analysis frequencies, since it was assumed that $\Delta\omega = 2\pi/NT$. If, however, we are free to adjust the input sampling rate (i.e., $1/T$), we can fix N and still obtain any desired $\Delta\omega$.

IV. Simulation of an Analysis-Synthesis System

The concepts of the previous two sections have been employed in a general purpose computer simulation of a complete analysis-synthesis system. In this section we discuss the signal processing techniques employed in the simulation.

The complete analysis-synthesis system is depicted in Fig. 5. This figure is conveniently segmented into three parts, an analysis section, a section for bit-rate reduction, and a synthesis section. In this section of the paper we will discuss the details of analysis and synthesis, leaving the details of the bit-rate reduction for the next section.

We begin with speech that is sampled at a rate of 12 195 Hz. With $N = 128$, this sampling rate allows us to analyze up to 65 channels with 94.273 Hz spacing using the technique of the previous section. An FFT program is used to compute $N = 128$ point transforms of windowed speech according to (16) and (17). The finite length lowpass impulse response $h(nT)$, designed by frequency sampling techniques [3], has a length of 731 samples and corresponds to a filter with precisely linear phase and amplitude response as shown in Fig. 6. The filter is down 6 dB at 50 Hz and has more than 60 dB attenuation above 83 Hz. Since the linear phase impulse response is symmetric about sample number 365, the composite response can be made to have precisely linear phase by incorporating a delay of 365 samples into the analysis as in (10) and (11). By an iterative procedure it was found that a frequency spacing of $\Delta\omega = 2\pi(94.273) = 2\pi/NT$ yields an impulse response that is very small in the vicinity of the pulses located at $\pm 128T$ in the shifted sequence $d(nT - 365T)$. In the FFT method, (15) is changed to

$$X(\omega_k, nT) = e^{-j(2\pi/N)k(n+n_a)}\, G(k, n)$$

where n_a is the desired analysis delay. Because the FFT processing restricts the value of N to be a power of two, the rather unusual sampling rate of 12 195 Hz is required. If the system were realized without this constraint on N, other sampling rates could of course be used.

Since the low-pass filter cuts off at about 50 Hz, the channel signals can be sampled at a much lower rate than the input speech rate. Thus we compute these signals with a sampling period $T_1 = m_1 T$ by moving the "window" $h(nT - rT)$ in steps of m_1 samples of the input waveform. For example, a value of $m_1 = 122$ samples yields a sampling rate of $1/T_1 = 99.96$ Hz.

At each sample time, we perform the operations called for by (16), (17), and (15), obtaining as a re-

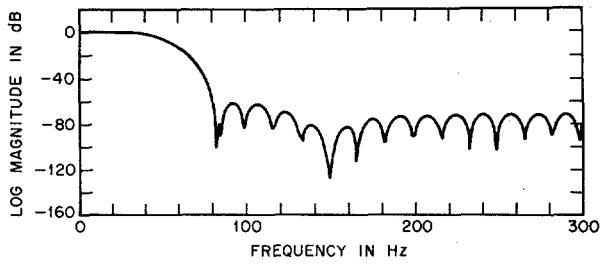

Fig. 6. Frequency response of finite duration linear phase low-pass filter used in simulation of analysis-synthesis system.

sult, all the channel signals required to completely cover the frequency range $0 < \omega < 2\pi/T$. The channels corresponding to frequencies in the range $\pi/T < \omega < 2\pi/T$ are complex conjugates of corresponding channels in the band $0 < \omega < \pi/T$, and can thus be discarded. Furthermore, to insure that no aliasing occurs in sampling the speech signal, we generally sample at a higher rate than is necessary to preserve the speech information. Therefore, some of the higher channels in the range $0 < \omega < \pi/T$ may also be ignored in synthesis by simply choosing $M < N/2$. This can lead to significant computational savings if the input is greatly over sampled.

In performing the synthesis at any reasonable speech sampling rate, the channel signals, having been computed at a low sampling rate, must be interpolated to the output speech sampling rate $(1/T_2)$. In the present simulation, we have only retained channels 1-28 ($M = 28$), and synthesis is done at a rate of $1/T_2 = 10\,004$ Hz. If the parameters are computed at a rate of 99.96 Hz, the interpolation is achieved by filling in 99 zero samples between every sample of each of the channel signals. The resulting 10 004 Hz sequences are interpolated by filtering with a low-pass filter whose amplitude response is identical to Fig. 6. The resulting signals are used in (12), with $n_s = 299$. This delay is necessary since the analysis filter and the interpolation filter are effectively cascaded.

It is interesting to note that finite impulse response filters are particularly attractive for interpolation of signals in this way. For example, if 99 samples out of 100 are zero, we only need to perform one multiplication for each 100 samples of the impulse response. Thus, in this example, where the total length of the impulse response was 599 samples, we need only perform five multiplications and four additions to compute each sample of the interpolated channel signals.

The composite response for the system described above is shown in Fig. 7. Fig. 7(a) shows the composite impulse response $\tilde{h}(nT)$ and Fig. 7(b) shows the amplitude response. The phase is precisely linear, with delay equal to 59.9 ms. As can be seen, there is less than 1 dB of ripple across the output speech band.

Speech has been processed with the above system (without quantization) at a sampling rate of 99.96 Hz for the channel signals. If the phase compensation is not used, a definite reverberant quality is perceived

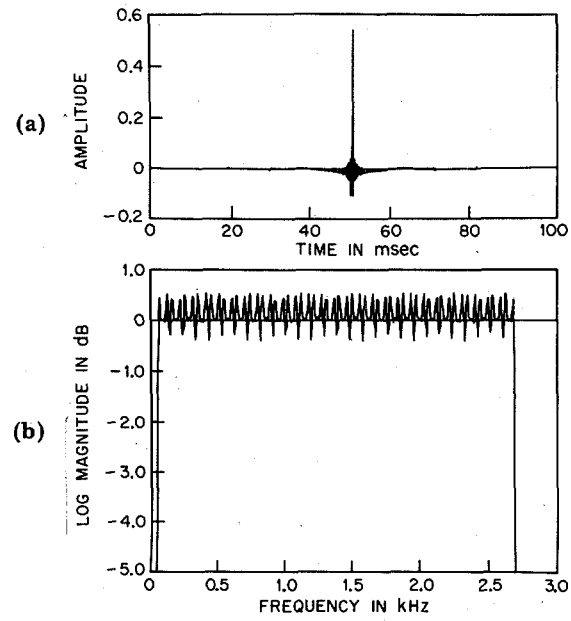

Fig. 7. Composite response of analysis-synthesis system using filter of Fig. 6. (a) Composite impulse response. (b) Composite amplitude response.

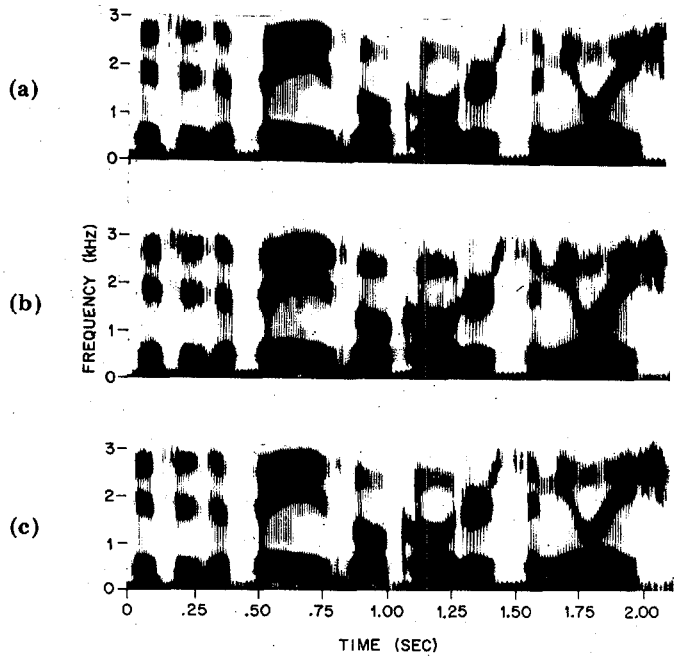

Fig. 8. Illustration of unquantized operation ($1/T_1 = 160$ Hz). (a) Input speech. (b) Output speech with no phase adjustment. (c) Output speech with best phase adjustment.

when the synthesized output is compared to the input. However, for the phase compensated system (corresponding to the response of Fig. 5), careful listening indicates essentially no perceptible difference between the synthetic output and the natural speech input. Fig. 8 shows wide band spectrograms of the input speech [Fig. 8(a)], the results of analysis and synthesis for no phase correction [Fig. 8(b)], and best phase correction [Fig. 8(c)]. The fuzziness of the spectrogram in Fig. 8(b) indicates the reverberant

nature of this signal, while the spectrogram in Fig. 8(c) compares very favorably to Fig. 8(a). In this realization, the channel signals were sampled at a rate of 160 Hz.

V. Information Rate Considerations

We have demonstrated that the complex short-time Fourier transform can provide an accurate digital representation of the speech signal. However, if such a representation is to be useful, it should also be superior in flexibility and information rate to a PCM representation of the waveform. Previous investigations have demonstrated that the short-time Fourier transform representation of speech (phase vocoder) affords significant flexibility in manipulating the time and frequency dimensions of speech. For example, it is possible to speed up or slow down a speech utterance by an arbitrary factor while leaving the frequency scale unaltered, or alternatively, the frequency scale can be changed while keeping the time scale fixed [1]. Another example of the flexibility of short-time Fourier representation is its application in a scheme for reducing the effect of multipath distortion [5].

There remains the question of the necessary bit rate of the short-time Fourier representation, and to what extent it may be reduced while maintaining an acceptable output signal. The bit rate is proportional to the number of channels, and from Fig. 5 we see that the bit rate of each channel depends on the sampling rate and quantization (i.e., number of bits per sample) of the spectrum parameters a and b. The two approaches to coding of the spectrum signals that we have studied are adaptive delta modulation (ADM) and PCM.

ADM Coding

We have used a 1-b ADM system as described by Jayant [6]. In this method, the parameters are represented by only 1 b/sample, and the hardware for encoding the signals is extremely simple. For the 28 channel system we have discussed, the ADM bit rate is $56/T_1$ b/s, where T_1 is the sampling rate of the channel signals. The ADM system requires a sampling rate on the order of 5 to 10 times the Nyquist rate for good performance. Thus we can expect that bit rates on the order of 20 to 30 kb/s would be required for good results. Examples of ADM coding at several rates are shown in Fig. 9. As can be seen, even at 28 kb/s, there is some distortion apparent on the spectrogram although the perceptual quality is quite good. However, at lower bit rates it can be seen that there is considerable degradation. Although 28 kb is certainly not a low bit rate, the speech quality is comparable with ADM coding of the waveform at this rate. By coding the short-time Fourier representation at this rate we can take advantage of the flexibility for manipulating the time and frequency dimensions of a speech signal. Thus it would appear that because of its simplicity, ADM coding is an interesting possibility in spite of its rather high bit-rate requirements.

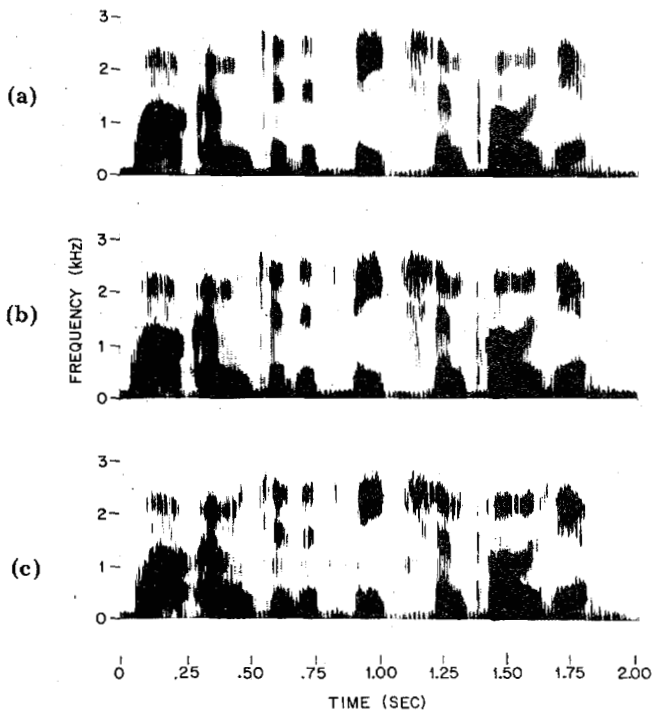

Fig. 9. Adaptive delta modulation coding of the spectrum parameters. (a) 28 kb/s ($1/T_1$ = 500 Hz). (b) 21 kb/s ($1/T_1$ = 375 Hz). (c) 14 kb/s ($1/T_1$ = 250 Hz).

PCM Coding

The bit rate for PCM coding of the spectrum parameters can be lowered by reducing the sampling rate and reducing the number of bits per sample. The sampling rate is chosen on the basis of the bandwidth of the channel signals a and b. Thus, in the system we have been discussing where the filter was down 60 dB above 80 Hz, we can be quite certain that negligible aliasing will occur if we sample at about $1/T_1$ = 160 Hz. Lower sampling rates can be used without aliasing only if we reduce the bandwidth of the analysis filter. If we make a corresponding reduction in the spacing of the channels, we will need more channels to cover a given bandwidth and thus we will achieve no saving. If we decrease the channel bandwidth but leave the channel spacing the same, we have seen that the composite response of the system becomes more reverberant. Thus, to lower the sampling rate below the Nyquist rate (160 Hz in this case), we can either choose aliasing with minimum reverberation or we can reduce the bandwidth and accept a more reverberant output. These effects are illustrated in Figs. 10 and 11.

In Fig. 10, we show the effect of reducing the channel bandwidth while keeping the channel spacing con-

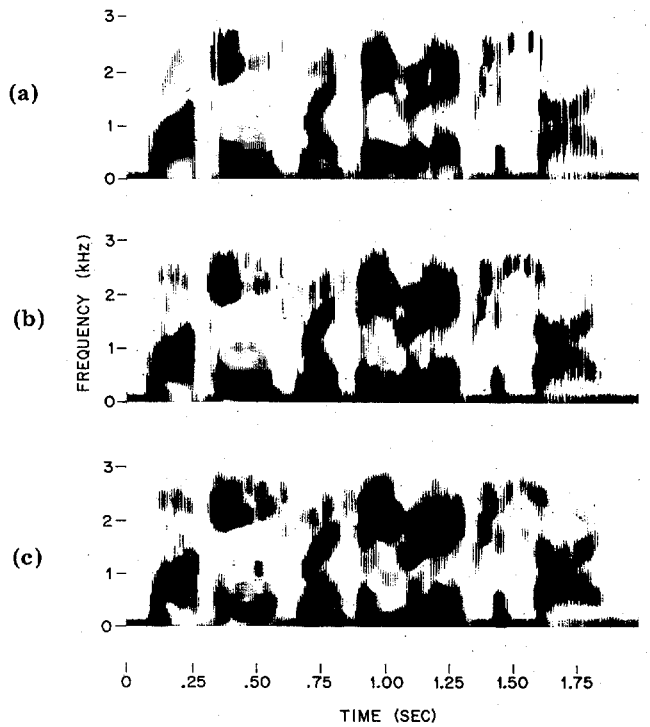

Fig. 10. Illustration of effect of narrow-band analysis filters for PCM coding: no quantization; $1/T_1$ = 160 Hz; low-pass cutoff = the following: (a) 80 Hz; (b) 53 Hz; (c) 11 Hz.

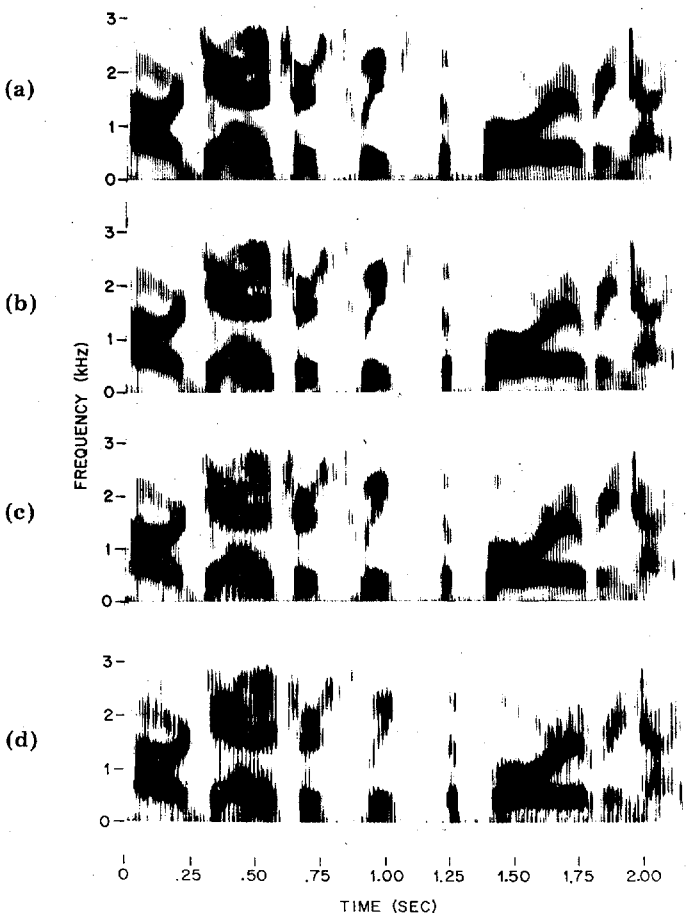

Fig. 11. Illustration of effect of aliasing for PCM coding: low-pass cutoff = 80 Hz; no quantization: (a) $1/T_1$ = 160 Hz; (b) $1/T_1$ = 100 Hz; (c) $1/T_1$ = 80 Hz; (d) $1/T_1$ = 60 Hz.

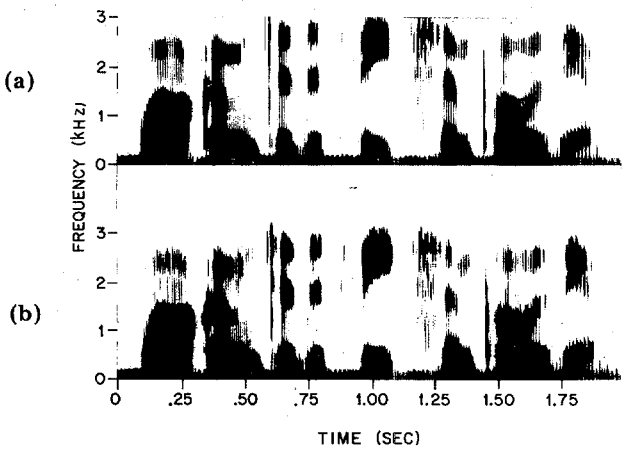

Fig. 12. Quantized operation. (a) Input speech. (b) ($1/T_1$ = 100 Hz). Total bit rate = 16 kb/s.

stant. The filter bandwidth is 80 Hz in (a), 53 Hz in (b), and 36 Hz in (c), and no quantization was performed. The reverberant nature of the speech is readily apparent in the 36 Hz case, and noticable but not severe in the 53 Hz case.

In Fig. 11 we show the effect of aliasing. No quantization was performed but in this case the bandwidth is held fixed at 80 Hz while the sampling rate of the channel signals is (a) 160 Hz, (b) 100 Hz, (c) 80 Hz, and (d) 60 Hz. There is considerable distortion in the cases (c) and (d) while the effect is much less severe in the case (b). Note that aliasing produces a type of distortion that is distinctly different from reverberation. As can be seen by comparing Figs. 10 and 11, aliasing tends to distort the pitch while leaving the formant frequency information relatively undistorted, while the opposite is true for reverberation. Thus it seems that aliasing distortion should have less effect on intelligibility than reverberation. The other factor in determining the bit rate is the quantization of the samples of the spectrum parameters. Because of the decreasing sensitivity of the ear at high frequencies, fewer bits should be allotted to the higher channels. Based on some earlier unpublished work by Carlson [7], we have found that for the 28 channel system that we have considered throughout this paper, the following distribution of bits results in speech of acceptable quality.

	Log Magnitude (Bits)	Phase (Bits)
channels 1-10	3	4
channels 11-28	2	3

Note that the quantization is applied to log $[(a(\omega_k, nT_1))^2 + (b(\omega_k, nT_1))^2]$ and $\tan^{-1}[b(\omega_k, nT_1)/a(\omega_k, nT_1)]$. Fig. 12 shows an example of 16 kb/s PCM coding of the spectrum parameters. The input speech is shown in Fig. 12(a) and the output speech for the above bit distribution, 80 Hz bandwidth, and 100 Hz sampling rate is shown in Fig. 12(b). (Note that some alisasing distortion occurs in achieving this bit rate.)

VI. Conclusion

We have discussed short-time Fourier analysis and synthesis as a means for representing a sampled speech signal. We have reviewed some theoretical results from an earlier paper [2] and shown how these results can be used to design a speech analysis-synthesis system. We have discussed the simulation of such systems on a general purpose digital computer. A novel feature of the simulation is a technique for using the fast Fourier transform to efficiently compute the short-time spectral parameters using arbitrary data windows and at arbitrary sampling rates. We have also illustrated the factors influencing the bit-rate of systems of this type.

It is clear that short-time Fourier analysis and synthesis is not an extremely efficient representation of speech. It nevertheless is interesting because no pitch tracking is required and because it allows a fair degree of flexibility in altering the speech parameters. It should be possible to lower the bit-rate for acceptable quality by taking advantage of our knowledge of hearing through the use of fewer channels spaced nonuniformly in frequency. Unfortunately a theoretical basis comparable to the theory presented here does not as yet exist for the nonuniform case. Future efforts should be directed toward understanding the nonuniform case, and toward applications of systems using uniform spacing in reducing multipath distortion [5] and in aids to the handicapped [1].

References

[1] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493-1509, Nov. 1966.
[2] R. W. Schafer and L. R. Rabiner, "Design of digital filter banks for speech analysis," *Bell Syst. Tech. J.*, vol. 50, pp. 3097-3115, Dec. 1971.
[3] L. R. Rabiner, "Techniques for designing finite-duration impulse-response digital filters," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 188-195, Apr. 1971.
[4] B. Gold and C. M. Rader, *Digital Processing of Signals*. New York: McGraw-Hill, 1969.
[5] J. L. Flanagan and R. C. Lummis, "Signal processing to reduce multipath distortion in small rooms," *J. Acoust. Soc. Amer.*, vol. 47, part 1, pp. 1475-1481, June 1970.
[6] N. S. Jayant, "Adaptive Delta Modulation with a one-bit memory," *Bell Syst. Tech. J.*, vol. 49, pp. 321-342, Mar. 1970.
[7] J. P. Carlson, "Digitalization of the phase vocoder," M.S. thesis, Dep. Elec. Eng., Mass. Inst. Technol., Cambridge, 1968.

Speech Processing With Walsh–Hadamard Transforms

F. YING Y. SHUM, A. RONALD ELLIOTT, and W. OWEN BROWN

Abstract—High-speed algorithms to compute the discrete Hadamard and Walsh transforms of speech waveforms have been developed. Intelligible speech has been reconstructed from dominant Hadamard or Walsh coefficients on a medium sized computer in a non-real-time mode. Degradation of some phonemes was noted at low bit rates of reconstruction, but the reconstruction could be improved by varying the position of the sampling window. A digital processor, which allows real-time analysis of speech to be conducted on the system, is described.

Manuscript received April 30, 1972. This work was supported by grants from the National Research Council of Canada.

F. Y. Y. Shum was with the Department of Electrical Engineering, McMaster University, Hamilton, Ont., Canada. He is now with the Department of Computer Science, University of Western Ontario, London, Ont., Canada.
A. R. Elliott is with the Department of Electrical Engineering, McMaster University, Hamilton, Ont., Canada.
W. O. Brown is with Bell-Northern Research Limited, Ottawa, Ont., Canada.

Introduction

Several classes of orthogonal systems of varying complexity have been used in digitized audio and video signal processing [1]-[7]. The general approach is to choose an arbitrary sample block that offers effective spectral resolution as well as computational efficiency. The performance of discrete Karhunen-Loève (K-L), Slant, Fourier, Walsh-Hadamard, and Haar orthogonal transformations when applied to bit-rate reduction has been measured, in a mean-square error sense or signal-to-noise ratio, and reported by independent researchers [8]-[10]. Their results tend to indicate that the K-L transform offers the most efficiency in terms of data compression, while the rest may be ranked in the order that appears above. However, the Walsh-Hadamard transform, which involves only additions and subtractions, has distinctive advantages in adaptability to computer analysis and digital implementation, particularly when small computers with limited facilities are used. Also, the results of speech synthesis presented by Boesswetter [6], and Campanella and Robinson [7] had demonstrated that dominant term synthesis from the Walsh domain was possible, and with good quality.

In the analysis given here, an audio signal was sampled at a rate of 8 kHz. The window size has been assigned a value of 64 which corresponds to a time interval of 8 ms. One reason for this assignment is that 64 is an integral power of 2. This condition