Performance of Transform and Subband Coding Systems Combined with Harmonic Scaling of Speech

DAVID MALAH, MEMBER, IEEE, RONALD E. CROCHIERE, SENIOR MEMBER, IEEE, AND RICHARD V. COX, MEMBER, IEEE

Abstract—In this study an approach for improving the performance of waveform coders, based on coding a frequency scaled speech signal, is examined and subjectively evaluated for specific subband and transform coding systems. The recently developed simple and efficient time-domain harmonic scaling (TDHS) algorithms are used to frequency scale the speech signal. The underlying frequency-domain model of the pitch-adaptive TDHS algorithms provides insight and guidelines for their use in this application, as outlined in this work. The subjective evaluation is based on an A-B comparison test involving 12 listeners and shows a meaningful improvement in quality for the waveform coders used at low bit rates. In particular, subband coding (SBC) combined with TDHS (SBC/HS) at 9.6 kbits/s was found to provide a quality equivalent to that of SBC alone at 16 kbits/s, i.e., a bit-rate advantage of about 7 kbits/s was realized. For the speech specific adaptive transform coder (ATC) used, the combined system (ATC/HS) achieves a bit-rate advantage of 4 kbits/s at 7.2 kbits/s. The SBC/HS system emerges as a particularly attractive method for speech encoding at the data rate of 9.6 kbits/s since its quality is comparable to that of ATC/HS (or SBC at 16 kbits/s). Yet, its complexity is lower than ATC and the system is amenable to real-time hardware implementation using current technology.

I. Introduction

WAVEFORM coding techniques attempt to reproduce encoded signals at lower bit rates than PCM encoding by utilizing the temporal and spectral properties of the signal. Speech-specific coders have achieved sizeable reductions in bit rate by incorporating in the coder design known properties of both short-time and long-time (pitch period) correlations which are related to spectral envelope (formants) and fine structure (pitch harmonics) of speech, respectively.

However, because they attempt to replicate the speech waveform, even the more complex forms of waveform coders, such as adaptive-predictive coders (APC) [1], [2] and adaptive transform coders (ATC) [3], [4], require bit rates typically at or above 9.6 kbits/s in order to have acceptable communication quality [5]. This range of bit rates is several times higher than the bit-rate range of vocoders which typically encode speech at 2.4 kbits/s and below [5]. Vocoders are based on a speech production model and achieve low bit rates by analyzing and then resynthesizing a speech signal which *sounds* like the original signal but does not necessarily replicate the original signal waveform. Because of this tight adherence to a speech production model, vocoders are found to be far less robust than waveform coders and lack naturalness in quality. While naturalness can be improved by adding more bits to better represent the excitation signal [6], the high sensitivity of vocoders to voicing decision errors, environmental conditions (background noise, simultaneous speakers), and channel errors make vocoders unattractive in many applications.

In this work we examine the performance of a speech encoding system which appears to meaningfully extend the range of waveform coder operation at the low bit-rate end, with only a modest loss of robustness. In the system under consideration the chosen waveform coder is used to encode a frequency scaled (compressed) speech signal which is rescaled (expanded) at the receiver—following waveform decoding. The frequency scaling operations are based on reducing (for compression) or increasing (for expansion) the interharmonic spectral gaps of the pitch by a factor of up to three using frequency shifting of the pitch harmonics. The actual scaling operations are done, however, in the time domain by means of the recently developed time-domain harmonic scaling (TDHS) algorithms [7], [8]. These algorithms are pitch-adaptive and perform a time-varying weighting of adjacent speech segments, with an appropriate weighting (window) function, in a way which assures continuity. Once the pitch period is known, the remaining operations required in the TDHS algorithms are typically only one multiplication and two additions per output sample. Furthermore, since the frequency compressed signal is decimated by a factor equal to the frequency compression factor, the computational load on the waveform coder is usually reduced by the same factor. A general block diagram which illustrates the way the waveform coders used in this study were combined with TDHS is given in Fig. 1.

The combination of a waveform coder with this particular method of frequency scaling can be viewed as an approach for exploiting the harmonic structure (pitch) of voiced speech signals in a different way than is currently done in waveform coders (e.g., a pitch prediction loop around the quantizer [5] or pitch dependent bit allocation [4]). Another point of view is to consider the combined system as a form of "soft vocoding," since unlike waveform coders the reconstructed

Manuscript received July 9, 1980; revised October 10, 1980.

D. Malah is with the Department of Acoustics Research, Bell Laboratories, Murray Hill, NJ 07974, on leave from the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel.

R. E. Crochiere and R. V. Cox are with the Department of Acoustics Research, Bell Laboratories, Murray Hill, NJ 07974.

0096-3518/81/0400-0273$00.75

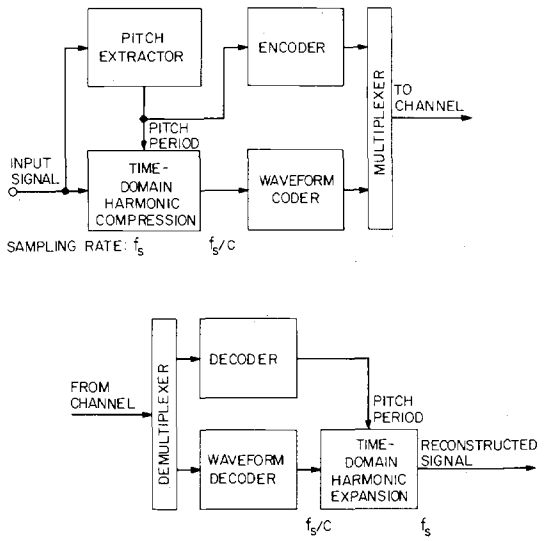


Fig. 1. General block diagram of combined TDHS–waveform coding system. C is the compression factor.

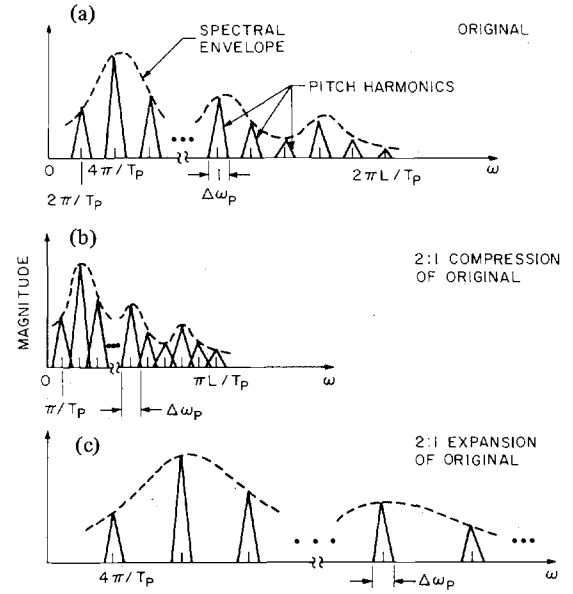


Fig. 2. Schematic spectral representation of (a) input voiced speech, (b) frequency compressed signal by pitch harmonic shifting, (c) frequency expanded signal. T_p is the pitch period duration and $\Delta\omega_p$ is the spectral width of a pitch harmonic.

signal may not replicate the input signal due to interharmonic aliasing caused by the frequency compression method used. However, unlike vocoders, the combined system is found to be robust and has a natural sounding quality. This robustness stems from the avoidance of voicing decisions (voiced-unvoiced), from the way the TDHS algorithms use pitch information, and the surprising ability of the system to encode speech signals of several simultaneous speakers (to be reported in the sequel).

Informal subjective tests suggest that by combining TDHS with waveform coders (particularly those which do not exploit pitch information), a reduction in bit rate by up to the scaling factor used (typically two) can be obtained. The performance advantage is found to be gained primarily at the low end of the bit-rate range of the coder, i.e., the performance "knee" (the bit rate below which the quality of the encoded signal starts falling rapidly) is pushed to lower bit rates. The more efficiently the waveform coder exploits pitch information the smaller is the expected improvement. However, it appears from the results of this work that even with a waveform coder which exploits pitch and with a scaling factor of two, the combined system achieves a meaningful improvement at the lower end of the bit-rate range of the waveform coder.

There is much interest to date in robust speech encoders which are capable of operating at or below 9.6 kbits/s with adequate quality to allow transmission of voice on digital data links [5]. For this reason, we have chosen in this study to examine the combination of TDHS with subband coding (SBC) [4], [9] and with adaptive transform coding (ATC) [3], [4]. The SBC technique efficiently exploits the formant structure in the speech spectrum, whereas the speech specific ATC technique takes advantage of both formant and pitch information in the way the bits are allocated. These coders offer high communications quality at 16 kbits/s and appear to have the potential of encoding speech at 9.6 kbits/s, or below, with adequate quality if combined with TDHS. Indeed, from informal listening and from an A-B comparison test, SBC combined with TDHS (SBC/HS) at 9.6 kbits/s was found to provide a quality equivalent to that of SBC alone at 16 kbits/s, i.e., a bit-rate advantage of about 7 kbits/s was realized. As could be expected, since pitch information is exploited by the particular ATC system used, the improvement gained by ATC/HS is smaller than for SBC/HS but still a 4 kbits/s bit-rate advantage was obtained at 7.2 kbits/s. The above results are in agreement with an earlier, less exhaustive, study in which a simple CVSD (continuously variable slope delta modulation) coder was combined with TDHS [10] and was found to be a useful approach for encoding at bit rates below the performance "knee" of the CVSD coder used.

In the following sections we briefly describe the TDHS algorithms and the waveform coders used in this study. We then discuss issues involved in combining TDHS with waveform coders and present results of subjective tests performed for comparing the various coding schemes.

II. Time-Domain Harmonic Scaling

The time-domain harmonic scaling (TDHS) algorithms [7], [8] provide simple and efficient means for frequency scaling of speech signals. The development of the algorithms is based on a frequency domain model of voiced speech as schematically shown in Fig. 2(a). The method used for frequency scale compression is to frequency shift the pitch harmonics to lower frequencies as shown in Fig. 2(b). The presence of relatively wide interharmonic gaps in the original spectral representation [Fig. 2(a)] allows for compression factors of two to three with tolerable interharmonic aliasing. Similarly, frequency scale expansion is achieved by shifting the pitch harmonics upwards as shown in Fig. 2(c). This approach for frequency scaling is different from the earlier phase-vocoder frequency division technique [11] which in addition to relocating the pitch harmonics also attempts to scale the width $\Delta\omega_p$ of each pitch "tooth" by the frequency scaling factor. The advantage of the frequency shifting approach is that it

can be performed simply and efficiently in the time domain, if pitch information is known, as explained below.

In principal, the frequency shifting operations can be performed by using a filter bank analysis to isolate the different pitch harmonics and then modulate (frequency shift) each harmonic to its designated new location. A convenient mathematical framework for a uniform filter bank analysis, modification, and synthesis is provided by the short-time Fourier transform (STFT) [12]-[14]. In this framework the filter bank analysis can be performed by frequency shifting each desired frequency band centered at $\omega = \omega_k$ to baseband, using complex modulation, filtering the baseband signal with the prototype lowpass filter $h(t)$ of the filter bank, and then remodulating back to the appropriate frequency band. Hence, if no modification is performed, the reconstruction $y(t)$ of the input signal $x(t)$ is given by

$$y(t) = \sum_{k=-L}^{L} X(\omega_k, t)\, e^{j\omega_k t} \tag{1}$$

where $X(\omega_k, t)$ is the STFT of $x(t)$ evaluated at $\omega = \omega_k$ [11], i.e.,

$$X(\omega_k, t) = \int_{-\infty}^{\infty} x(\tau)\, h(t-\tau)\, e^{-j\omega_k \tau}\, d\tau \tag{2}$$

and for a uniform filter bank the center frequencies satisfy $\omega_k = k\Delta\omega$, $k = 0, \pm 1, \cdots, \pm L$ (where $\Delta\omega$ is the bandwidth of each band).

If the pitch period, T_p, is known, the center frequencies of the filter bank can be chosen to coincide with the pitch harmonics (i.e., $\Delta\omega = 2\pi/T_p$). In that case frequency scaling by a factor q ($q < 1$ for compression, $q > 1$ for expansion), by means of frequency shifting the pitch harmonics, results in an output signal $y^q(t)$ given by

$$y^q(t) = \sum_{k=-L}^{L} X(\omega_k, t)\, e^{jq\omega_k t}, \tag{3}$$

i.e., the kth pitch harmonic, which was originally located at ω_k, is shifted to $q\omega_k$.

Substituting (2) into (3), interchanging the order of summation and integration, and changing the variable $(t - \tau)$ to τ, one obtains [8]

$$y^q(t) = \int_{-\infty}^{\infty} x(t-\tau)\, h(\tau)\, K_N((q-1)\, t + \tau)\, d\tau \tag{4}$$

where $K_N(t)$ is a kernel function given by [8]

$$K_N(t) = \sum_{k=-L}^{L} e^{-j\omega_k t} = \frac{\sin(N\pi t/T_p)}{\sin(\pi t/T_p)} \tag{5}$$

with $N = 2L + 1$ and T_p being the pitch period of the input speech signal $x(t)$. This kernel function is periodic and has $N - 1$ zeroes in each period T_p. To be useful in a digital implementation (4) must be discretized. It turns out that if q is assumed to be rational ($q = \mu/\delta$, where μ and δ are relatively prime integers); the output signal is sampled at a rate corresponding to its bandwidth; and the kernel function $K_N(t)$ is sampled at its zeroes, the following discrete-time approximation to (4) results [8]

$$\hat{y}^q(nT/q) = \sum_{l=-\infty}^{\infty} x(nT - lT_p)\, \hat{h}(lT_p - n(\mu - \delta)\, T'). \tag{6}$$

In (6), T is the Nyquist interval (or shorter) of the input signal $x(t)$ (so that $T_p = NT$), $T' = T/\mu$ (recall, $q = \mu/\delta$), and $\hat{h}(t) = h(t)\, T_p$.

Since speech signals are not stationary, the prototype filter impulse response (or window function) $h(t)$ is chosen to be of finite duration (FIR), avoiding the weighting of speech segments which are not within the same quasi-stationary interval. If we assume the finite duration of $h(t)$ to be mT_p, m an integer, then (6) appears to involve only m multiplications and $(m - 1)$ additions. However, due to constraints that $h(t)$ should satisfy [7], the computations can be rearranged into $(m - 1)$ multiplications and m additions.

Finally the TDHS algorithms presented in [7], [8] (one for compression and one for expansion) are obtained from (6) by using an FIR filter and letting $q = 1/C < 1$ for compression and $q = S > 1$ for expansion.

In the current work, as in most of our previous work with TDHS [7], [10], the simple triangular window (corresponding to $m = 2$) is used. Other suitable window functions can be applied [7], [8] but the triangular window is particularly convenient and simple to implement. It also provides adequate results with a scaling factor of two which is used in this work.

The way the compression algorithm is applied for 2:1 compression ($q = \frac{1}{2}$), using a triangular window, is shown in Fig. 3. It is seen that each two adjacent speech segments of pitch period duration are weighted and overlapped to provide one

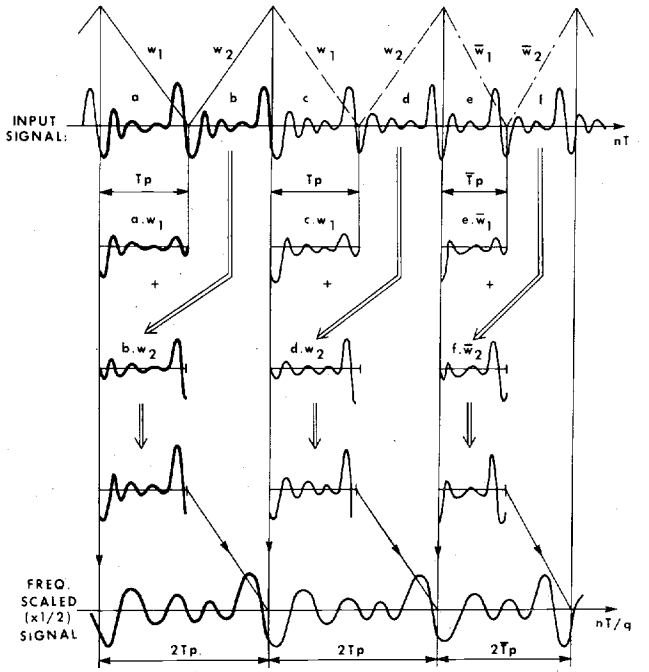


Fig. 3. Illustration of the time-domain operations for 2:1 compression ($q = \frac{1}{2}$) using a triangular window. T_p and $\overline{T}_p$ are two different pitch period durations.

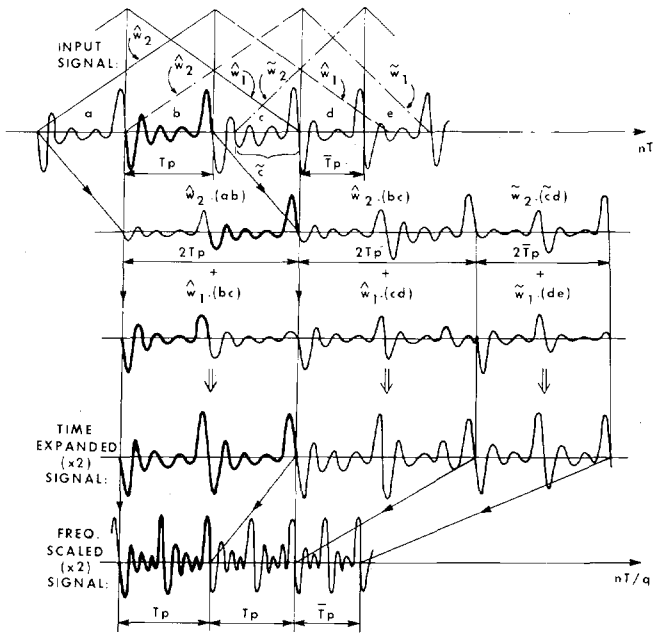


Fig. 4. Illustration of the time-domain operations for 2:1 expansion ($q = 2$) using a triangular window.

output segment of pitch period duration. If the output segments are concatenated and played out at the original sampling rate ($1/T$) a time compressed signal results. If, however, each segment is output at the decimated rate (i.e., with a sampling interval of $T/q = 2T$) a frequency compressed signal, which occupies the original time slot, results—as shown at the bottom of Fig. 3. Note that due to the way the input signal is weighted and due to the properties of the window, the last point in each output segment and the first point in the next segment are actually adjacent points in the original waveform. Hence, the output waveform suffers no discontinuities at segment boundaries. The procedure can, therefore, be viewed as an extension of earlier time-domain methods which are based on splicing the speech waveform, i.e., replacing each two speech segments with one segment. Such a crude approach corresponds to actually using a rectangular window instead of the triangular window used here. The deficiency of the rectangular window is quite clear from both time-domain (discontinuities) and frequency-domain (high sidelobes) considerations [7].

The expansion process for $q = 2$ is illustrated in Fig. 4, again using a triangular window. As before, adjacent speech segments are weighted to produce an output segment. However, this time for each input segment of pitch period duration, an output speech segment of two pitch periods duration is generated. This way a time-expanded signal is obtained if the original sampling rate is retained. If the output sampling interval is $T/q = T/2$, a frequency expanded signal is obtained. Obviously, if the expansion is applied to the compressed signal the reconstructed signal occupies the original frequency and time spans. Again, this procedure can be considered as an extension of the simple reiteration technique which corresponds to using a rectangular window instead of a better window function, such as the triangular window used here.

III. Subband Coding

Subband coding (SBC) is a waveform coding technique in which the speech band is partitioned into typically four to eight subbands by bandpass filters [5], [9], [15]-[17]. Each subband is effectively low pass-translated to *dc*, sampled at its Nyquist rate (twice the width of the subband), and then digitally encoded using adaptive PCM (APCM) as shown in Fig. 5. By carefully selecting the number of bits/sample used for encoding the subbands, each band can be preferentially coded to give a maximum overall perceptual quality with a minimum overall bit rate. The step-sizes in each band adapt independently in proportion to the rms speech level in their respective bands. In this way subband coding can take advantage of the properties of temporal nonstationarity, spectral formant structure, and auditory masking in speech production and perception.

The subband coder used in this experiment is based on the octave band approach of Barabell and Crochiere [16]. It utilizes quadrature mirror filters (QMF) in a "tree structure" to achieve an efficient filter bank framework [17]. The actual filter designs are selected from the family of designs given by Johnston [18]. The first QMF pair splits the initial speech band into two equally spaced bands, such that the aliasing or "leakage" of energy between bands is canceled in the reconstruction process. A second QMF pair in the "tree structure" then subdivides the lower band into two bands in a similar manner. This process is continued in the "tree structure" until the desired number of bands is obtained as shown in Fig. 6 for a four-band split. Tables I and II show the parameters of the resulting designs that are obtained from this approach for five-band and four-band systems at 16 and 9.6 kbits/s, respectively. The bit allocations for the combined systems will be detailed in a later section. Note that the initial sampling rates for the above designs are 6400 Hz and 5760 Hz, respectively. Since, typically, the input and output sampling rates are in the range of 8 to 10 kHz, digital decimation-interpolation techniques for rate conversion need to be applied.

IV. Adaptive Transform Coding

Adaptive transform coding of speech is a frequency domain technique in which a high resolution transform is applied to the speech signal on a block by block basis [3], [4]. The resulting transform coefficients (frequency components) are quantized using both step-size adaptation and dynamic bit assignment for each transform coefficient, i.e., the number of bits used to encode each transform coefficient is dynamically varied from block to block. The adaptive quantization is based on a smoothed spectral estimate of each speech block. This estimate is parametrized and encoded for transmission as "side information." A general block diagram of an ATC system is shown in Fig. 7. In this work a "speech-specific" coder was used similar to that of Tribolet and Crochiere [4]. It uses, however, the homomorphic model for parametrizing the smoothed spectral estimate of each block, as reported by Cox and Crochiere [19] instead of the LPC model used in [4]. Yet, as in [4], the spectral model also includes pitch information

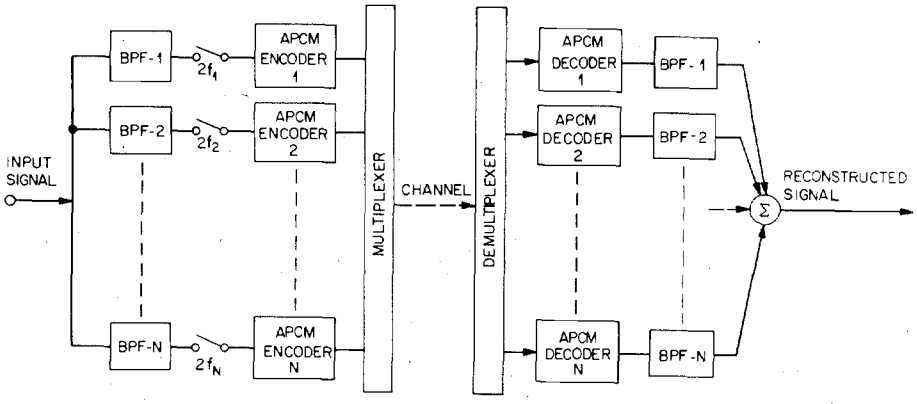


Fig. 5. General block diagram of a subband coding (SBC) system.

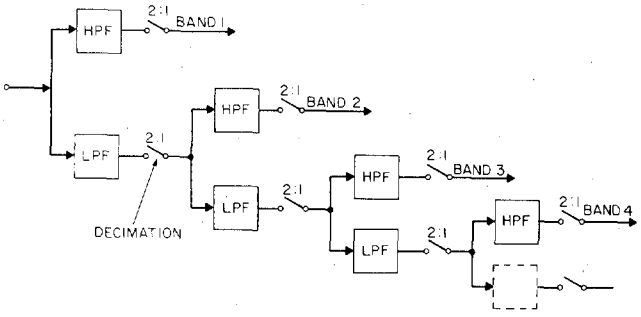


Fig. 6. Illustration of a four-band split using quadrature mirror filters (QMF).

TABLE I
FIVE-BAND SUBBAND CODER PARAMETERS FOR 16 kbits/s ENCODING*

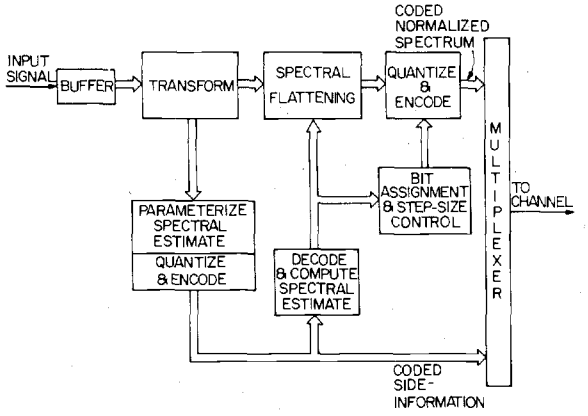
Band No.	1	2	3	4	5
Freq. Range [Hz]	3200-1600	1600-800	800-400	400-200	200-100
No. of Taps in QMF Splitting Filter	32	16	16	16	8
Bit Allocation	2	2	4	5	5

*Input Sampling Rate: 6400 Hz

TABLE II
FOUR-BAND SUBBAND CODER PARAMETERS FOR 9.6 kbits/s ENCODING*

Band No.	1	2	3	4
Freq. Range [Hz]	2880-1440	1440-720	720-360	360-180
No. of Taps in QMF Splitting Filter	32	16	16	8
Bit Allocation	1-1/3	2	2	3

*Input Sampling Rate: 5760 Hz



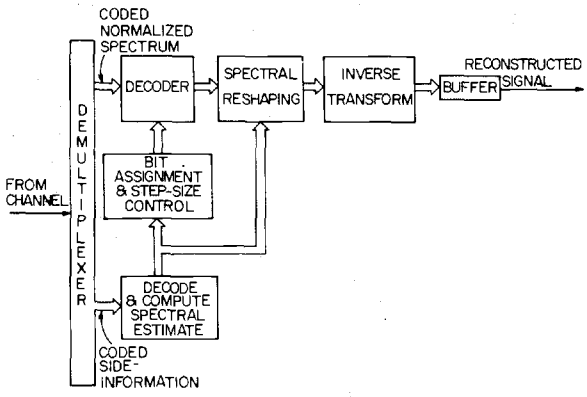


Fig. 7. General block diagram of an adaptive transform coding system (ATC).

which results in a more efficient bit assignment by assigning more bits to represent the spectral peaks at pitch harmonics.

Fig. 8 describes in more detail the system used to simulate adaptive transform coding of speech [19]. The transform used is a symmetric discrete Fourier transform, a close relative of the discrete cosine transform. It is obtained by forming a symmetric sequence of length $2M$ from a sequence of $M+1$ time samples

$$y(n) = \begin{cases} x(n), & n = 0, 1, \cdots M \\ x(2M-n), & n = M+1, \cdots, 2M-1 \end{cases} . \quad (7)$$

The $2M$ point DFT of this input sequence is taken via an FFT routine on an array processor. Since the input to the DFT is both real and symmetric, the output will also be both real and symmetric. In this way only $M+1$ values are needed to represent the $2M$ values in the transform domain.

The overall operation of the coder proceeds as follows. A block of $M+1=257$ samples of speech (at an 8 kHz sampling

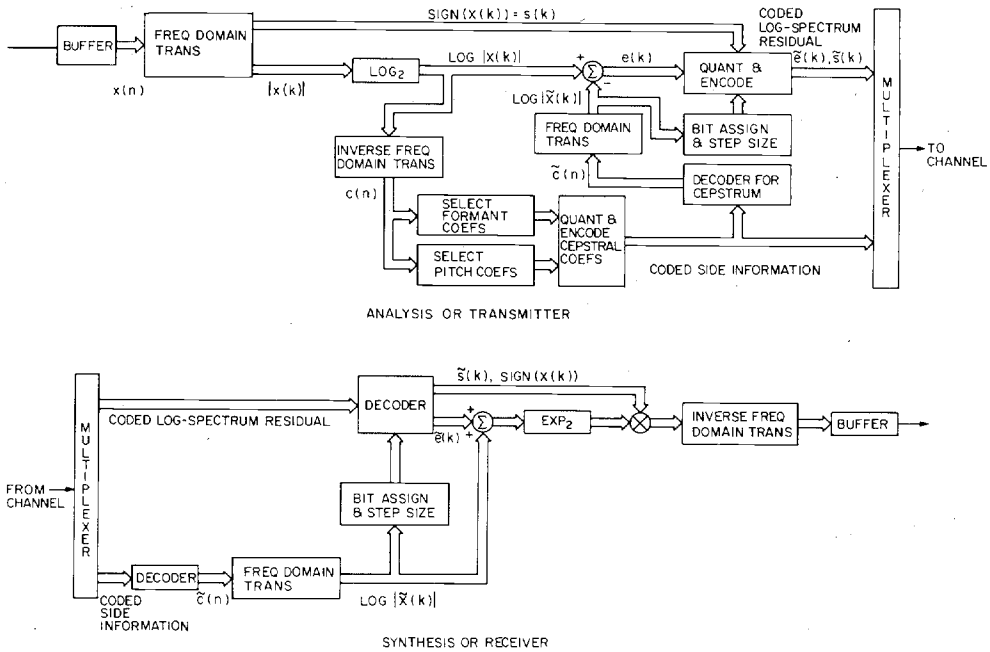


Fig. 8. Block diagram of the homomorphic-based ATC system used in the simulations with adaptive log-PCM quantization of the transform coefficients.

rate) is assembled in the input buffer. No analysis window is applied to the block; however, a trapezoidal synthesis window is used later at the output. Consequently, successive blocks have a small amount of overlap, typically about ten samples. The block is then transformed by the symmetric DFT (SDFT) described above yielding a real symmetric spectrum $X(k)$. The sign of each frequency domain sample $X(k)$ is saved for encoding and the base two logarithm of the magnitude of each sample is taken, yielding $\log |X(k)|$. The log magnitude spectrum is used for log-PCM quantization (to be described later). It is also used to generate a "pseudocepstrum," $c(n)$, by inverse SDFT transformation. This pseudocepstrum can be used to model the combined pitch and formant structure of the speech segment with relatively few parameters. The first 8 to 14 cepstral coefficients are used to generate the smooth spectral envelope estimate and the peaks further from the origin in the cepstrum provide the pitch information. This information is quantized, coded, and transmitted as side information, and also decoded to provide a rough estimate of the cepstrum, $\tilde{c}(n)$. Taking the SDFT of $\tilde{c}(n)$ then gives an estimate of the log magnitude spectrum, $\log |\tilde{X}(k)|$ of the speech segment. This estimate is used for the bit assignment in the ATC algorithm. It is also subtracted from the actual log magnitude spectrum $X(k)$ to provide a residual log magnitude spectrum $e(k)$. This residual is quantized and encoded according to the bit assignment (i.e., linear quantization of the difference, $e(k)$, in the log domain is equivalent to an adaptive step-size PCM of the transform coefficients using a quantizer with a logarithmic characteristic). At the receiver the quantized residual is decoded, the estimated log magnitude spectrum is added back and the frequency samples are exponentiated back to the linear frequency domain. Their signs are then reassociated and an inverse SDFT is taken.

TABLE III
BIT RATES FOR SIMULATED ATC (ALONE) SYSTEM

Overall Bit-Rate [kb/s]	Spectral-Residual Rate [kb/s]	Side-Information Rate [kb/s]	No. of Cepstral Coeff.
16	12.6	3.4	14
9.6	7	2.6	12
7.2	5	2.2	8

The resulting output is then windowed with a trapezoidal window and the overlap added to the previous output block.

The bit rates simulated (for ATC alone) and the breakdown of the overall bit rate into spectral residual and side information transmission is detailed in Table III. The corresponding information for the combined ATC/HS system is given in the next section.

V. COMBINING TDHS WITH WAVEFORM CODING

In the previous sections the TDHS algorithms and the waveform coding techniques were described. In this section we discuss issues involved in combining TDHS with waveform coding, with particular emphasis on the operation of the TDHS algorithms and the simulations performed.

The combination of TDHS with waveform coding can be done in two ways. The first is shown in Fig. 1, according to which the original pitch-period data is encoded and transmitted to the receiver. A second way is to extract the pitch information at the receiver from the decoded frequency-compressed signal. This approach was used in our earlier work which involved combining CVSD with TDHS [10] and it was chosen to keep the simplicity of CVSD transmission. However, since the number of pitch periods per unit time, in the frequency-compressed signal, is reduced by a factor equal to the scaling

factor used (see Fig. 3), the task of pitch extraction is more difficult. This results in a somewhat higher degradation in the reconstructed signal. In this work, since relatively complex coders are used anyway, the configuration of Fig. 1 was used.

It is evident from the earlier discussion on the TDHS algorithms that the most fundamental and crucial operation is the pitch extraction. Many pitch extraction algorithms were reported in the literature [20] and most of them are probably adequate for this application. However, the choice of the most suitable one would typically depend on its amenability to hardware implementation, its robustness to environmental conditions, the computational load involved, etc. To help in the choice of a suitable pitch extractor the following observations are made.

1) A simple mathematical analysis performed in [7], for a stationary periodic signal, shows that even in the presence of a pitch-period error the reconstructed signal has the correct pitch if the relative error ϵ in the fundamental Frequency $F_p = 1/T_p$ is limited according to

$$\epsilon \triangleq \Delta F/F_p < 1/(2MC) \tag{8}$$

where M is the number of harmonics desired to be exactly reconstructed, ΔF is the error in estimating F_p, and C is the frequency compression factor.

Since the number of harmonics present in a given frequency band is dependent on the pitch period itself we have derived from (8) the following upper limit on the pitch-period error ΔT_p.

$$\Delta T_p < 1/(2F_M C) \tag{9}$$

where F_M is the highest frequency in the band for which the voiced speech signal is to be reconstructed exactly (theoretically). Typically, it is required that F_M be at least 1.5 to 2 kHz. With $C = 2$ this results in a permitted pitch-period error of 0.125 to 0.16 ms. Practically, however, even errors of twice as high are tolerated with a modest increase in the output signal degradation. This range of allowed pitch-period error usually complies with the performance of most known pitch extractors.

2) The difficult task of voiced-unvoiced decision is practically avoided since any random value (within a limited range) provided by the pitch extractor at unvoiced segments can be used without noticeable perceptual effects. This is found to be true in spite of the spectral distortion caused by the TDHS algorithms to unvoiced signals due to their noiselike characteristics.

3) The effect of a double pitch-period decision by the pitch extractor is far less detrimental than for vocoders. Actually, if it does not occur at voiced-unvoiced transitions its effect is quite small. This can be understood from the underlying frequency domain model since using a double pitch-period corresponds to performing the filter bank analysis with twice as many filters. This means that only every other filter contains a pitch harmonic. However, since the filters are now narrower the tolerated pitch measurement error is now also reduced.

In the simulations performed in the course of this study, a homomorphic pitch extractor based on a cepstral analysis [21] of the speech signal was used. This rather complex pitch extractor was applied because of its high performance and the availability of an array processor, on the computer system used for the simulations, which enables pitch extraction—in almost real time. In our earlier work [10] we have used a simpler time-domain autocorrelation type pitch extractor [22] which was found to be adequate for that application.

Since the simulations were not performed in real time, we have found it convenient to first prepare the pitch data of the tested utterance (up to six sentences) and then use these data for experimenting with the TDHS algorithms.

The pitch data file is prepared using a data window of 32 ms (256 samples) and a fixed update of 8 ms (64 samples at the 8 kHz sampling rate used). No smoothing of the raw pitch data was needed. Since the TDHS algorithms are pitch-adaptive it is extremely important that the pitch data pointer be carefully matched to the input signal pointer which is in itself pitch dependent. In particular, since the compressed signal has less pitch periods per time unit and may also be delayed this matching of pitch and signal pointers is particularly important at the reconstruction phase (expansion).

As stated earlier, the TDHS algorithms were applied in this work with the simple triangular window. If a scaling factor which is higher than two is attempted, the choice of more complex window functions [7] can be advantageous. Alternatively, the application of the TDHS algorithms in cascade using lower scaling factors in each stage (e.g. 1.5 and 2 for achieving a scaling factor of 3) was also found to be effective. The use of an adaptive window[1] as suggested in [8] improves somewhat the resulting signal at transitions and also can reduce the spectral distortion in unvoiced segments. However, when combined with waveform coding this improvement appears to be masked.

Although not part of this study, we have examined the effect of some environmental conditions on the TDHS algorithms performance. In particular, we have found that in spite of being a pitch-dependent system, the TDHS algorithms were able to perform exceptionally well on speech of several simultaneous speakers (the test performed used speech of three speakers). This result could perhaps be attributed to the tracking of the dominant speaker at each short-time interval, and to the masking properties of the ear. The TDHS system was also found to be relatively robust to noisy or degraded speech (e.g. room reverberations) and performed adequately, as long as the pitch detector did not break down. With a cepstral pitch detector the system was able to operate down to a signal-to-noise ratio of 0 dB. At high noise levels, however, structuring (coloration) of the noise (due to the pitch synchronous processing) can be perceived.

Another issue of interest is the effect of channel errors. Preliminary simulations indicate that the TDHS system is quite insensitive to channel errors in the pitch data up to a bit-error-rate of 10^{-2}. At higher error rates, error protection of the

[1] E.g., a trapezoidal window with variable slopes such that it becomes a triangular window at sustained voiced segments, whereas at unvoiced or transition (voiced-unvoiced) segment a rectangular window with a small amount of tapering is used.

TABLE IV
RATES AND BIT ALLOCATIONS FOR SIMULATED COMBINED SBC/HS SYSTEMS

SBC/HS Rate [kb/s]	SBC Rate [kb/s]	Pitch Transmission [kb/s]	Bit Allocation Bands: 1	2	3	4	5
16	31	0.5	5	5	5	5	5
9.6	18.2	0.5	2	3	5	5	5
7.2	13.2	0.6	2	2	4	5	-

TABLE V
BIT RATES FOR SIMULATED COMBINED ATC/HS SYSTEMS

ATC/HS Rate [kb/s]	ATC Rate [kb/s]	Additional Pitch Transmission [kb/s]	Spectral-Residual Rate [kb/s]	Side-Information Rate [kb/s]
16	31	0.5	27.1	3.9
9.6	18.2	0.5	14.8	3.4
7.2	13.9	0.25	10.5	3.4
4.8	9.1	0.25	6.5	2.6

pitch data might be necessary. If the additional bits needed for pitch protection cannot be afforded, the reextraction of pitch at the receiver can be considered.

In the remainder of this section we will discuss the simulations of TDHS with the particular waveform coders used in this study, namely SBC and ATC, and the issue of exploiting pitch information within the waveform coder.

Again, since the simulations were not performed in real time it was convenient first to frequency scale (or time compress) the whole tested utterance, and then to encode the compressed signal by the waveform coders at different bit rates. Furthermore, since we used available waveform coding programs, it was most convenient to assume that the input sampling rate is the original 8 kHz and consider the input utterance to the waveform coders as a time compressed signal (see Fig. 3). Thus, to simulate, for example, the combined TDHS-SBC system at 9.6 kbits/s, the time-compressed (2:1) signal was encoded by the subband coder at the rate of 18.2 kbits/s which is equivalent to a 9.1 kbits/s transmission rate. Adding 500 kbits/s for the pitch information results in the overall desired bit-rate of 9.6 kbits/s. Table IV shows the different bit rates simulated with the combined SBC/HS system and the number of bits allocated to each band. (For SBC alone, the bit rates and bit allocations shown in Tables I and II were used.) Table V gives the corresponding information for the combined ATC/HS system.

Finally, the issue of exploiting pitch information within the waveform coder is discussed. Since TDHS already exploits pitch to reduce the redundancy present in speech signals, the effectiveness of using pitch information within the waveform coder itself is quite reduced. For systems which use a pitch prediction loop this can be explained by the fact that following the compression operation the correlation between adjacent pitch periods is reduced, since each pitch period in the compressed signal represents two pitch periods of the original signal. For systems which exploit the harmonic structure in the frequency domain (like ATC), the relative broadening of the pitch teeth reduces the gain obtained through the dynamic bit allocation (which assigns more bits at the pitch teeth and less bits at the interharmonic gaps). Yet, since TDHS requires pitch extraction due to the harmonic structure, the use of pitch information within the waveform coder can be useful if it is also found to be economic. In this study, with SBC and ATC, we found it to be uneconomical to use a pitch loop within the SBC due to its complexity [16] and the extremely small improvement. However, we continued to use pitch information within the ATC, since it is naturally embedded in the homomorphic model used.

VI. SUBJECTIVE EVALUATION OF THE COMBINED SYSTEMS

Due to the nature of the TDHS algorithms, which do not necessarily replicate the speech waveform, we could not use signal to noise measurements as performance indicators for the combined systems. To evaluate the performance of the combined systems an informal A-B comparison test for quality, similar to the one reported in [23], was performed. In the test, two groups of six listeners each compared the quality of the different coding systems by listening to pairs of sentences processed by the different systems. The preparation of the source material was done from six sentences (three sentences spoken by males and three sentences by females) and the material was presented to the listeners in a randomized order. Each of the 14 systems in the test was compared against all other systems, using both A-B and B-A comparisons (in random order). In all, each listener compared 182 pairs, and a total of 24 comparisons was done for each pair of coding systems. From the test results, the total number of "votes" given to each system (i.e., the number of times each system was preferred) was computed and used for rank ordering the different systems. The maximum number of votes that any coding system could get was 312. This number was used to present the results on a percentage basis. Since two different waveform coding systems were combined with TDHS, we found it useful to partition the test results and show separately the performance of the two different combined systems. Fig. 9 presents the results obtained for SBC alone in comparison to SBC combined with TDHS (denoted by SBC/HS), for bit rates of 16, 9.6, and 7.2 kbits/s (SBC alone at 7.2 kbits/s was not included due to its extremely low quality at this rate), and in comparison to the original signal and the reconstructed signal by TDHS only (with a scaling factor of two). The corresponding results for ATC and ATC/HS are shown in Fig. 10. Here the combined system was also operated at 4.8 kbits/s. To show how the two different systems relate to each other (subjectively) we present in Fig. 11 the overall test results. Before we turn to a discussion of the results we would like to mention that we have noted a certain variability in individual results due to the type of speaker (male or female). Since the test was balanced, we could extract the separate results for male and female speakers and these are presented in Figs. 12 and 13, respectively, for comparison with Fig. 11.

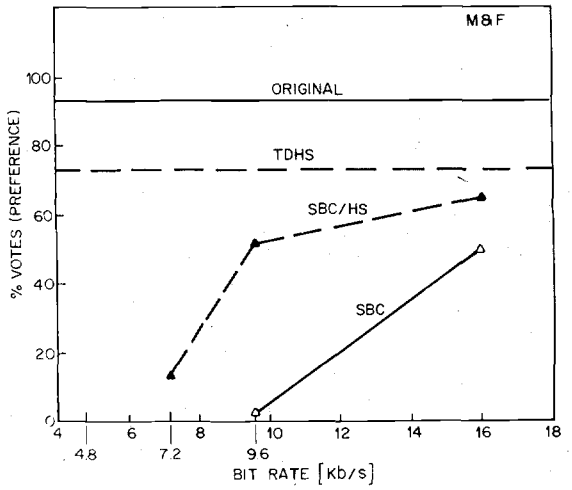


Fig. 9. A-B comparison test results for the combined subband coding system (SBC/HS) in comparison to SBC alone as well as to the original and TDHS (alone).

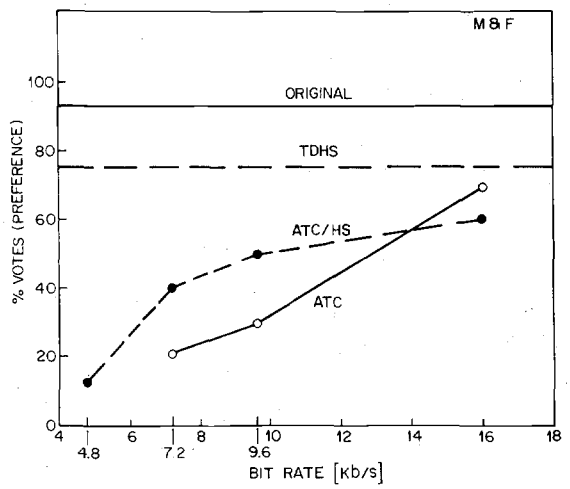


Fig. 10. A-B comparison test results for the combined ATC/HS system in comparison to ATC alone, as well as to the original and TDHS.

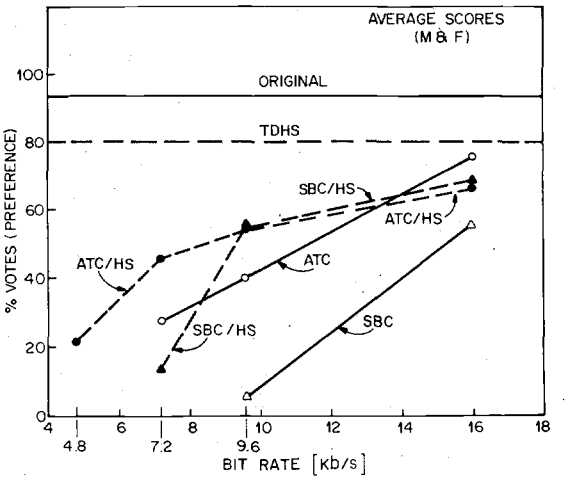


Fig. 11. Overall results of the A-B comparison test (averaged for male and female speakers).

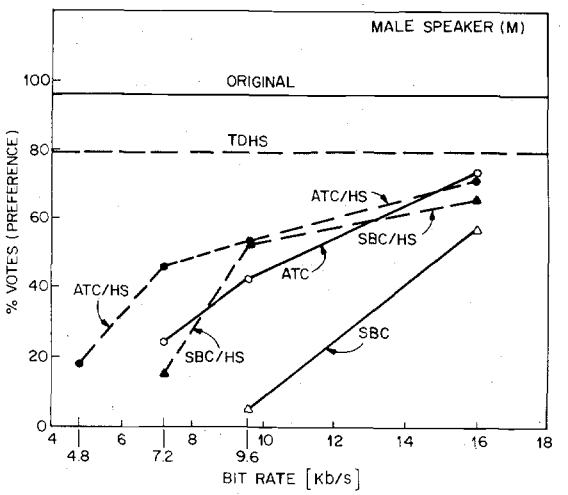


Fig. 12. Overall A-B comparison test results for male speaker.

A. Discussion of Results

We refer first to Fig. 9 which presents the results of the A-B comparison for SBC and SBC/HS. The rapid decrease in the quality of SBC alone, from 16 to 9.6 kbits/s, is quite evident. The combined SBC/HS system offers here a substantial improvement at 9.6 kbits/s, which corresponds to a bit-rate advantage of 7 kbits/s, i.e., the quality of the combined system at 9.6 kbits/s is equivalent to that of SBC alone at 16 kbits/s. Below 9.6 kbits/s the combined system degrades rapidly due to the rapid degradation of SBC below 16 kbits/s, and the added degradation by the TDHS operations which are not transparent. At 7.2 kbits/s the quality of the combined system is judged to be still acceptable for narrow band communication applications and the bit-rate advantage of the combined system at this rate is found to be 4 kbits/s.

For ATC, one observes from Fig. 10 that the improvement offered by the combined ATC/HS system is lower than for SBC, as the maximum bit-rate advantage is only 4 kbits/s (at 7.2 kbits/s), in comparison to a bit-rate advantage of 7 kbits/s for SBC/HS. From our earlier discussion, this is expected due to the fact that the ATC system simulated already exploits pitch information, leaving less redundancy to be removed by the TDHS system. Thus, the use of TDHS with a scaling factor of 2 results in a maximum improvement factor of 1.73 for SBC (at 9.6 kbits/s) and 1.55 for ATC (at 7.2 kbits/s). This appears to be a worthwhile improvement in light of the relatively small increase in the complexity of these systems.

Examining the overall test results in Fig. 11, one observes that the subjective qualities of SBC/HS and ATC/HS at 9.6 kbits/s and above are quite close to each other, whereas the complexity of SBC/HS is considerably lower than even ATC alone. As noted in the previous section, and shown in Figs. 12 and 13, the results for male and female speakers differ somewhat locally, but not globally, so that the above more general conclusions are not affected.

An additional interesting observation from the results for ATC is that at 16 kbits/s the combined system quality is actually slightly lower than the quality of ATC alone. This can be attributed to the relatively high quality of ATC at 16 kbits/s,

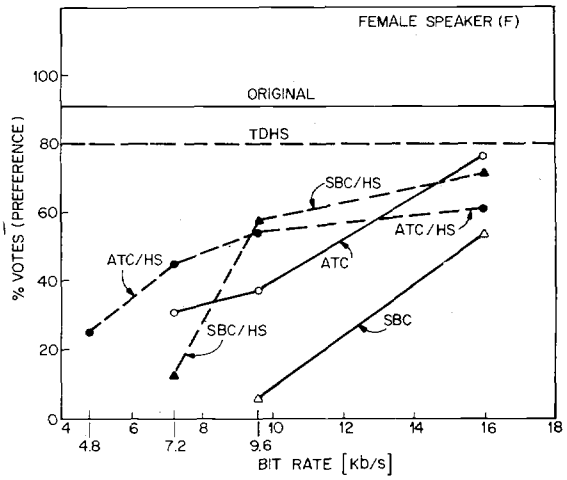


Fig. 13. Overall A-B comparison test results for female speaker.

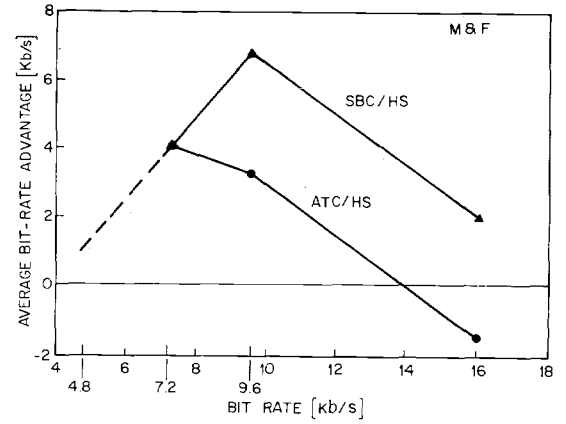


Fig. 14. Average bit-rate advantage obtained by the combined systems at different bit rates.

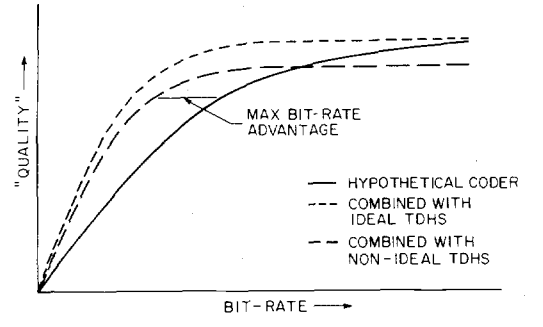


Fig. 15. Hypothetical coder characteristics showing the expected improvements by combining with TDHS.

and the nontransparency of the TDHS operations. For SBC, which has a lower quality at 16 kbits/s than ATC, there is some gain by using the combined SBC/HS system at this rate. As a summary the bit-rate advantages of the two systems at different bit rates is presented in Fig. 14.

The particular characteristic of the combined systems to improve the quality of the waveform coders mainly at rates which are below the performance knee of the coder, and to cause a degradation in quality at higher bit rates, is qualitatively explained by Fig. 15. In this figure the solid line is the performance curve of a hypothetical waveform coder. If an ideal TDHS system were available, one would expect the combined system to have the performance given by the short dashed-line, which is always above the solid line and is obtained by assuming a scale factor of two (each point on the original curve is moved to half the corresponding bit rate assuming no change in perceived quality). However, as is evident from the test results shown in the previous figures, the TDHS system introduces a certain degree of degradation (but is still judged to be of higher quality than ATC at 16 kbits/s). This causes the performance curve of the ideal combined system to be lowered, as shown by the long dashed-line in Fig. 15. It is clearly seen that the highest improvement is below the performance knee of the waveform coder and that for high bit rates one might actually obtain a loss in quality, as was found in our experimental results and discussed above.

VII. Conclusion

Waveform coding of the frequency scaled speech signal by means of time-domain harmonic scaling (TDHS) is shown to meaningfully improve the perceived performance of the waveform coding systems used—at their lower bit-rate end. In particular, bit-rate advantages of 7 and 4 kbits/s were obtained over subband coding (SBC) and adaptive transform coding (ATC), by the respective combined systems (SBC/HS and ATC/HS), at 9.6 and 7.2 kbits/s, respectively.

In spite of the "soft-vocoding" nature of the frequency scaling operations, the combined waveform coding-harmonic scaling system is almost as robust as the waveform coder used. The robustness is due to the avoidance of the voicing decisions and smoothing of the raw pitch-period data, the tolerance of the TDHS algorithms to limited pitch errors and to double pitch period decisions, and the ability to operate under different environmental conditions. In particular, it was found that the TDHS algorithms perform adequately on the speech of several simultaneous speakers.

The requirement for pitch extraction to implement the TDHS algorithms appears to increase the complexity of the system, but since the waveform coder operates on decimated samples, the overall computational load might even be reduced. However, the additional delay (of up to 100 ms) caused by the pitch extraction and harmonic scaling operations (transmitter and receiver) could be a limiting factor in some applications—but certainly not in many others.

The subjective evaluation which was performed by means of an A-B comparison test singles out the combined SBC/HS system at 9.6 kbits/s as having the highest perceived quality of all of the other systems examined at this rate. Since this system is much less complex than ATC, and is amenable to real-time hardware implementation using current technology, it provides an attractive system for speech coding at 9.6 kbits/s. The quality offered by this system is perceptually equivalent to that of 16 kbits/s SBC, which was judged in an earlier experiment [24] to be perceptually equivalent to 24 kbits/s ACPCM. An efficient implementation of this system could be based on the recently announced digital processing (DSP) chip by Bell Laboratories [25], [26]. Moreover,

the SBC/HS system could also be attractive at 16 kbits/s since its quality is only somewhat lower than ATC at this rate but is sufficiently less complex.

The approach of combining waveform coding with TDHS is not limited to specific waveform coders. However, the improvement to be achieved is dependent on the extent that the pitch structure is exploited by the waveform coder used. It appears, however, that even waveform coders which do quite efficiently exploit the pitch structure, such as the speech-specific ATC used in this study, can still be meaningfully improved at low bit rates by combining them with TDHS.

References

[1] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, pp. 1973–1986, Oct. 1970.

[2] J. Makhoul and M. Berouti, "Predictive and residual encoding of speech," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1633–1641, Dec. 1979.

[3] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299–309, Aug. 1977.

[4] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, Oct. 1979.

[5] J. L. Flanagan, *et al.*, "Speech coding," *IEEE Trans. Commun. Technol.*, vol. COM-27, pp. 710–746, Apr. 1979.

[6] B. S. Atal and N. David, "On synthesizing natural-sounding speech by linear prediction," in *Proc. 1979 IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 44–47.

[7] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 121–133, Apr. 1979.

[8] —, "Harmonic scaling of speech signals by linear time-varying spectral modifications," to be published.

[9] R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital coding of speech in subbands," *Bell Syst. Tech. J.*, vol. 55, pp. 1069–1085, Oct. 1976.

[10] D. Malah, "Combined time-domain harmonic compression and CVSD for 7.2 kbit/s transmission of speech signals," in *Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1980, pp. 504–507.

[11] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, Nov. 1966.

[12] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, pp. 1558–1566, Nov. 1977.

[13] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99–102, Feb. 1980.

[14] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 55–69, Feb. 1980.

[15] R. E. Crochiere, "On the design of sub-band coders for low bit-rate speech communication," *Bell Syst. Tech. J.*, vol. 56, pp. 747–770, May-June 1977.

[16] A. J. Barabell and R. E. Crochiere, "Sub-band coder design incorporating quadrature filters and pitch prediction," in *Proc. 1979 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979, pp. 530–533.

[17] D. Esteban and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," in *Proc. 1977 IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1977, pp. 191–195.

[18] J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1980, pp. 291–294.

[19] R. V. Cox and R. E. Crochiere, "Real-time simulation of adaptive transform coding," *IEEE Trans. Acoust., Speech, Signal Processing*, this issue, pp. xx–xx.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[21] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.

[22] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2–8, Feb. 1976.

[23] J. M. Tribolet and R. E. Crochiere, "A modified adaptive transform coding scheme with post-processing-enhancement," in *Proc. 1980 Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1980, pp. 336–339.

[24] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A comparison of the performance of four low-bit-rate speech waveform coders," *Bell Syst. Tech. J.*, vol. 58, pp. 699–712, Mar. 1979.

[25] J. S. Thompson and J. R. Boddie, "An LSI digital signal processor," in *Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1980, pp. 383–385.

[26] R. E. Crochiere, "Sub-band coding using the DSP," *Bell Syst. Tech. J.*, to be published.

David Malah (S'67–M'71) was born in Poland on March 31, 1943. He received the B.Sc. and M.Sc. degrees in 1964 and 1967, respectively, from the Technion–Israel Institute of Technology, Haifa, Israel, and the Ph.D. degree in 1971 from the University of Minnesota, Minneapolis, MN, all in electrical engineering.

During 1971–1972 he was an Assistant Professor in the Department of Electrical Engineering, University of New Brunswick, Fredericton, N.B., Canada. In 1972, he joined the Department of Electrical Engineering, Technion–Israel Institute of Technology and was engaged in teaching courses in digital electronic circuits and systems, linear systems and signal analysis, digital signal processing, and in research in digital signal processing. During 1975–1979 he was also in charge on a newly established signal processing laboratory which was active in speech and image communication research, and real time hardware developments. In 1979 he left on sabbatical to the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ. His main research interests are presently in digital speech coding, digital signal processing techniques, and adaptive filtering.

Ronald E. Crochiere (S'66–M'67–SM'78), for a photograph and biography, see this issue, p. 155.

Richard V. Cox (S'69–M'70), for a photograph and biography, see this issue, p. 155.