THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 5, May–June 1977
Printed in U.S.A.

A Variable-Band Coding Scheme for Speech Encoding at 4.8 kb/s

By R. E. CROCHIERE and M. R. SAMBUR

(Manuscript received November 5, 1976)

The standard fixed sub-band coding scheme has been modified to allow the center frequency of the two upper bands to vary in accordance with the dynamic movement of the vocal tract resonances F2 and F3. A relatively simple zero-crossing technique is used to measure the formants F2 and F3. Through the use of this variable band coder, it is possible to produce moderate-quality, intelligible speech at 4.8 kb/s (quality is slightly less than that of a 7.2-kb/s fixed sub-band coder and equal to that of about a 16-kb/s ADM *coder). The reasonably good intelligibility of the 4.8-kb/s variable-band coded speech can be attributed to the coder's attempt to capture and encode those spectral components of the signal that are perceptually most significant (the region around the formants). The major advantage of the variable-band scheme is that its implementation is considerably less complex than other waveform coding schemes or vocoder systems that can produce intelligible, narrowband speech.*

I. INTRODUCTION

Recently, a method for digitally coding speech signals in terms of sub-bands of the total spectrum was introduced that resulted in an improvement in quality of the coded signal over that obtained from a single full-band coding of the total spectrum.[1,2] The rationale for coding the signal in sub-bands is based upon the experimental fact that quantizing distortion is not equally detectable at all frequencies, and hence, the quality of the coded signal can be significantly improved by controlling the distribution of quantizing noise across the signal spectrum. Coding the signal in sub-bands offers the possibility of achieving this control.

In the recent work by Crochiere, Webber, and Flanagan,[1,2] the selection of the appropriate sub-bands was guided by the perceptual data contained in the so-called articulation index (AI).[3] The articulation index denotes, on the average, the contribution of each part of the spectrum to the overall perception of the spoken sound. For high-quality speech at moderate bit rates (16 kb/s and greater), the frequency range 200 to 3200 Hz was partitioned into four fixed contiguous bands that contributed equally to the AI. The transmission bit rate of the sub-band coder could be lowered gracefully by limiting these sub-bands in width and by tolerating some spectral gaps that did not contribute significantly to the AI. However, carried to excess, the noncontiguous bands produced an unpleasant, reverberant quality in the signal that finally resulted in unacceptable speech. A bit rate of approximately 7.2 kb/s was found to be about the lowest bit rate that still produced acceptable, intelligible speech. (The quality at this bit rate was judged about equal to that of 18-kb/s ADM speech.[2])

In using the AI in selecting sub-bands, it should be noted that this index only indicates the *average* perceptual contribution of each part of the spectrum. Since the speech spectrum is highly variable across a given utterance, it seems appropriate to select sub-bands that do not remain fixed but vary in accordance with the changing character of the speech. One way to achieve this goal is to allow the center frequency of the sub-bands to follow the variation of the formant frequencies across the utterance. The formant frequencies of a particular sound correspond to the resonance frequencies of its short time spectrum, and the frequency bands around these formants are perceptually the most significant regions of the spectrum. It is the purpose of this paper to show that by varying the sub-bands in accordance with the formant frequencies, it is possible to lower the bit rate to 4.8 kb/s and still maintain a speech quality that is approximately comparable to that of the 7.2-kb/s fixed sub-band coder. In addition, it is also shown that the formant frequencies can be sufficiently estimated for use in the variable-band scheme by a simple zero-crossing measurement technique. Thus, the variable-band coder can achieve very low data rates (4.8 kb/s) at considerably less expense than conventional vocoder systems, while still providing an intelligible signal.

II. VARIABLE-BAND CODER

The concept of the variable-band coder is illustrated in Fig. 1. The speech band is divided into four sub-bands and encoded separately in each sub-band. The two lower sub-bands are fixed bands that cover the frequency range from approximately 250 to 820 Hz. This represents the region of primary speech energy for voiced sounds. The two upper sub-bands are variable bands (with fixed bandwidths) centered about the F2 and F3 resonance peaks of the short-time speech spectrum (as illustrated by the dotted line). By varying the center frequencies of these two bands as the short-time spectrum changes, the encoder attempts to capture the maximum amount of speech energy and represent those

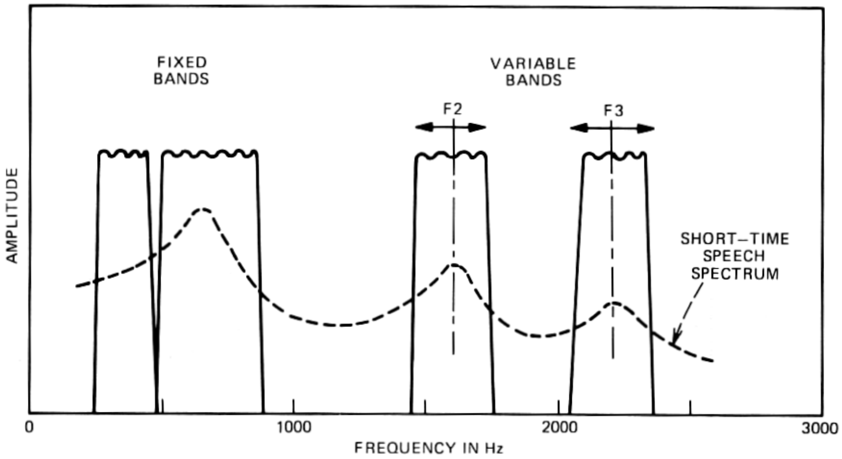

Fig. 1—Frequency domain interpretation of the variable-band coder.

frequencies that are perceptually most significant. Regions between the sub-bands are ignored to conserve bandwidth. While these gaps give a reverberant quality to the coder, the effect, as will be discussed, is not as pronounced as with a fixed-band scheme at the same bit rate.

The implementation of the sub-band coder can be achieved by any of the modulation schemes suggested in Ref. 1. In particular, the most efficient approach for implementing the fixed bands is the integer band sampling method. For the two upper sub-bands, a modulation scheme is required in which the center frequency of the band can be varied. This can be accomplished with the complex modulation method discussed in Ref. 1. In addition, a method for adaptively varying the center frequencies of these bands is required.

The overall configuration of the sub-band coder then takes the form shown in Fig. 2. The formant estimator determines the resonances F2 and F3 in the speech band. This information is encoded at a low bit rate and sent to the receiver. It is also decoded and used to control the variable-band center frequencies in the transmitter. In this way, the variable bands in the transmitter and receiver track identically.

The measurement of the resonances F2 and F3 is accomplished by a simple zero-crossing measurement technique. In this method, the individual resonances are first isolated by filtering the speech signal into frequency ranges appropriate to each formant.[4] After filtering, the resulting signal is ideally a damped sinusoid, and the formant frequency can then be estimated by measuring the axis-crossing rate of the filtered waveform. Figure 3 depicts the structure of the formant frequency extraction system.* To correct for isolated errors in the formant extraction

* The formants are measured 50 times per second and can be efficiently coded using less than 300 b/s by ADPCM techniques.[5]

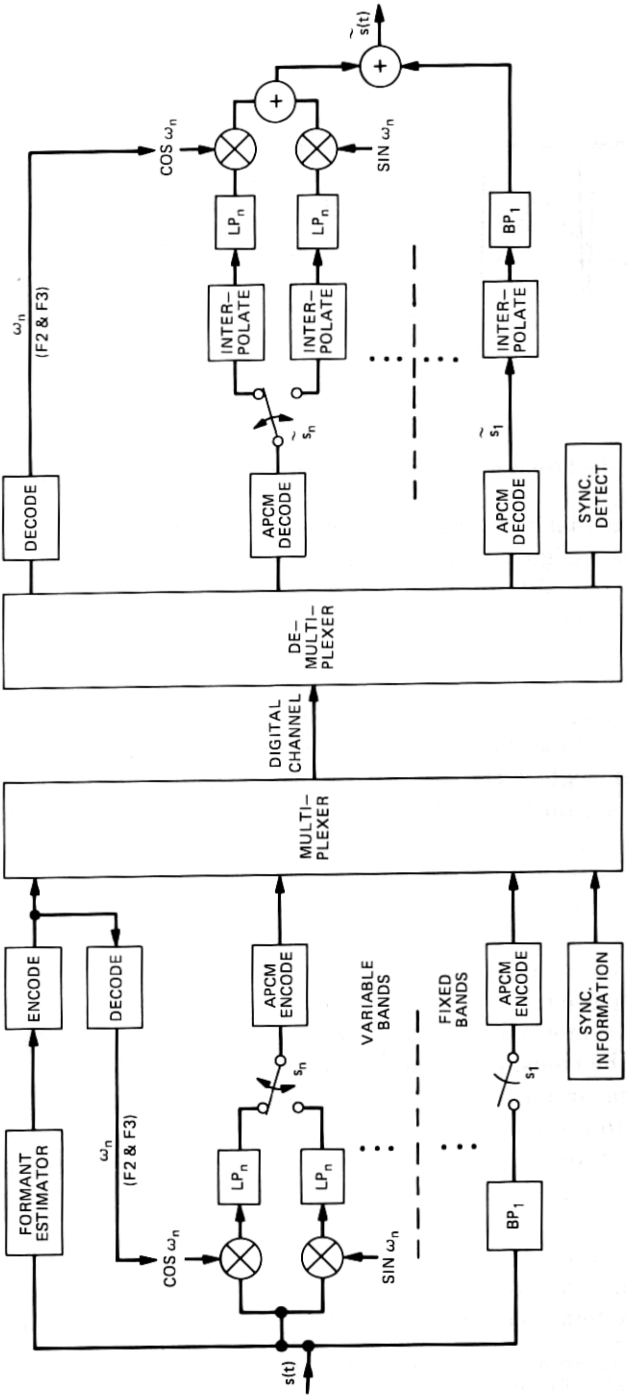

Fig. 2—Implementation of variable-band coder with integer-band sampling for the fixed bands and complex modulation for the variable bands.

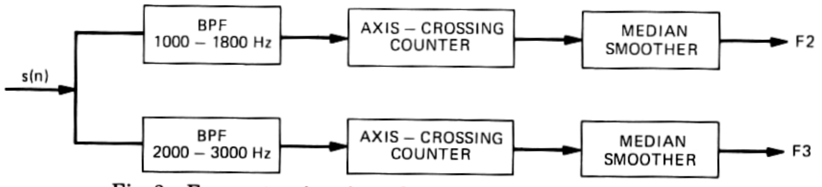

Fig. 3—Formant estimation scheme for estimating F2 and F3.

scheme and to insure that the measured formant trajectory is not excessively rough, a median smoother is employed.[6] It should be noted that although this system of formant measurement is not as accurate as the more elaborate methods of linear prediction[7] or spectral estimation,[4] the zero-crossing technique is computationally far less expensive than these schemes and, moreover, there is no enhancement in quality in the variable-band encoded signal when the more sophisticated measurements of F2 and F3 are used to control the variable sub-bands.

The sub-band signals are encoded with APCM encoders and the data are multiplexed together with the synchronization data and formant data, as illustrated in Fig. 2. Typically, more bits/sample are used for encoding lower sub-bands for the perceptual reasons explained in Refs. 1, and 2. Alternatively, a dynamic allocation of bits/sample can be employed in a manner similar to that used by Noll for transform coding.[7] Also, a slight amount of center-clipping can be used in sub-bands to reduce idle channel noise.

III. RESULTS OF COMPUTER SIMULATIONS

The sub-band coder system in Fig. 2 has been implemented by computer simulation for a transmission bit rate of 4.8 kb/s. Sub-band center frequencies and bandwidths corresponding to those in Table I were used. These bands also correspond to those shown in Fig. 1.

The formants were estimated by the method in Fig. 3 and were used to control the center frequencies of bands 3 and 4. Figure 4 shows the variation of these center frequencies as a function of time for the sentence

Table I—4.8-kb/s variable-band coder

			Bits/Sample Allocation		
				Dynamic	
Band	Center Frequency (Hz)	Bandwidth (Hz)	Fixed	Voiced	Unvoiced/ Silence
1	356	213	3	3	1½
2	640	356	2	2	2
3	F2	320	1½	1½	2
4	F3	320	1¼	1¼	2

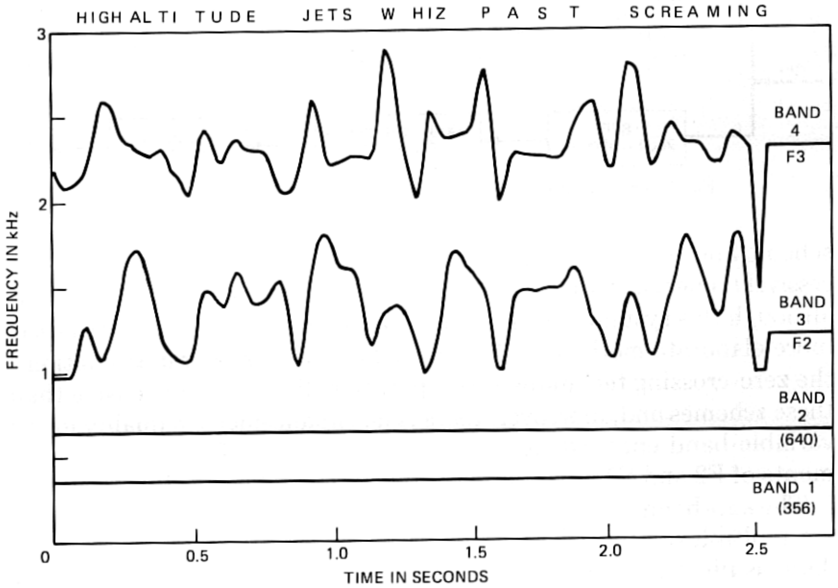

Fig. 4—Center frequencies of bands as a function of time for the sentence "High altitude jets whiz past screaming."

"High altitude jets whiz past screaming." A comparison of this plot to the spectrogram in Fig. 5a shows that these center frequencies do in fact track the major F2 and F3 resonances in the speech signal.

Two different bits/sample allocation schemes were tried in the simulations, a fixed allocation and a simple dynamic allocation scheme. In the fixed allocation 3, 2, 1½, and 1¼ bits/sample were used for encoding sub-bands 1 to 4, respectively. In the dynamic allocation scheme, 3, 2, 1½, and 1¼ bits/sample were used for encoding the voiced regions of the speech signal for bands 1 to 4, respectively. For unvoiced/silence regions, an allocation of 1½, 2, 2, and 2 bits/sample for bands 1 to 4 was used to encode more accurately the stronger energy in the higher frequencies during these intervals. A simple voiced/unvoiced decision was made by observing the variable step size of the APCM coder in the lowest band. If this step size was greater than five times its minimum allowed size, then the speech was assumed to be voiced and the 3, 2, 1½, and 1¼ bits/sample allocation was used. If it was less than five times the minimum step size, then the unvoiced/silence condition was assumed and the bits/sample allocation of 1½, 2, 2, and 2 was used.

Figure 5 shows spectrograms of the resulting computer simulations. The original sentence is represented by the upper spectrogram of Fig. 5a. Figure 5b corresponds to a sentence that was sub-band filtered (without encoding) with a fixed-band scheme (the two upper bands had center frequencies of 1200 Hz and 2300 Hz). In contrast, Fig. 5c shows

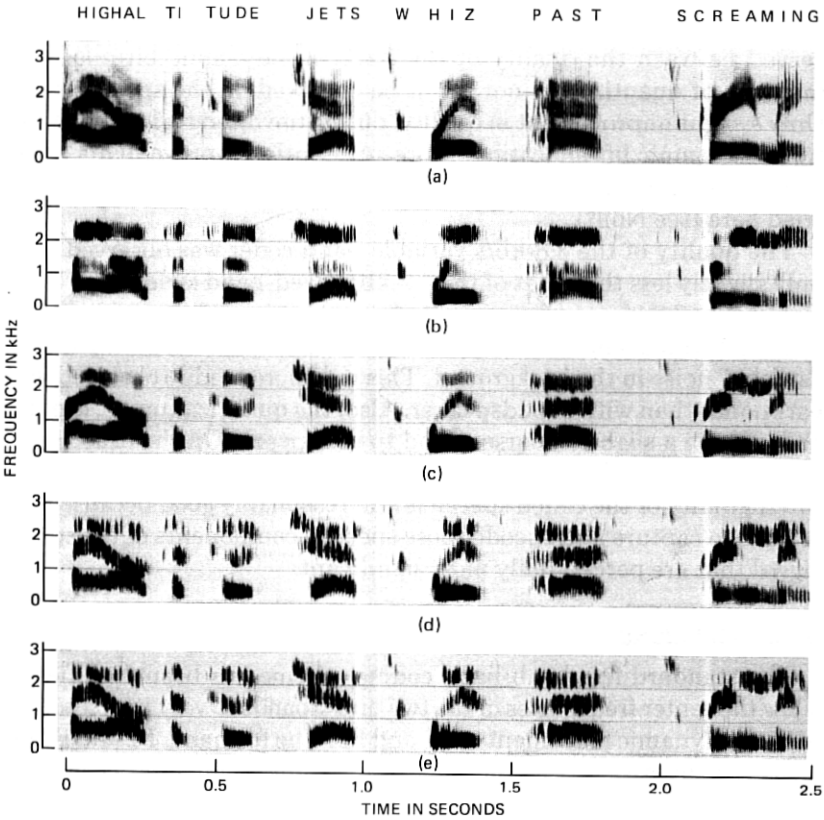

Fig. 5—Spectrograms of computer simulations. (a) Original. (b) Fixed-band filtered speech (uncoded). (c) Variable-band filtered speech (uncoded). (d) Variable-band coded with fixed bits/sample allocation. (e) Variable-band coded with dynamic bits/sample allocation.

the same sub-band arrangement except that the center frequencies of the two upper bands were allowed to vary according to Fig. 4. Again, the sub-bands were not encoded but simply filtered. A comparison of these two spectrograms (Figs. 5b and 5c) shows that the variable-band scheme gives a better representation of the important spectral features of the speech signal than the fixed-band scheme for the same total bandwidths. For example, in the words "high" and "whiz," the F2 resonance is lost in the gap between bands 3 and 4 of the fixed-band scheme; however, it is clearly present in the variable-band scheme.

Figures 5d and 5e show the results of the variable-band coder with the fixed and dynamic bit allocations discussed earlier. By comparison of these sentences with the unquantized sentence of Fig. 5c, the effects of the quantization can be observed. Typically, the quantized sentences have spectrograms that are more "ragged" in appearance due to the presence of the quantization errors and noise. Little difference was observed between the quality of the fixed and dynamic bit-allocation methods of quantization nor can it be observed on the spectrograms. Only a slight improvement in quality, during unvoiced regions, is gained by the dynamic bit allocation. More substantial improvements might be possible through more sophisticated allocation schemes than the one tried here (see Noll[6]).

The quality of the 4.8-kb/s variable-band coder was observed to be only slightly less than that of the 7.2-kb/s fixed-band coder reported in Ref. 2 (the 7.2 kbps coder was rated equal to that of an 18-kbps ADM coder). The movement of the two upper bands produced a noticeable "swishy" noise in the background. This was more readily observed with earphones than with a loudspeaker. Also, the quantization noise of the coders gave a slightly hoarse sound to the speech. On the other hand, although the quality of the variable-band coder is only moderate, the intelligibility of the coded speech is still reasonably good because it attempts to capture and encode those spectral components of the speech signal that are perceptually most significant.

IV. CONCLUSIONS

The standard fixed sub-band coding scheme has been modified to allow the center frequencies of the two upper bands to vary in accordance with the dynamic movement of F2 and F3. The formants F2 and F3 are measured by a relatively simple zero-crossing technique. Using this variable-band system, it is possible to produce moderate-quality, intelligible speech at 4.8 kb/s.

The variable-band system can be viewed as a hybrid type coder that combines the simplicity of a sub-band coder with the low-bit-rate potential of a vocoder type system. The ability of the variable-band coder to achieve narrowband transmission is directly associated with its vocoder-like utilization of the perceptually significant regions around the formants F2 and F3. But, unlike the vocoder, it is a true waveform coder that does not attempt merely to *model* the signal in terms of such features as pitch and vocal tract resonances.[4] It directly codes the entire 250-Hz to 818-Hz region of the spectrum and two 320-Hz bands centered about the crudely estimated values of F2 and F3. The variable-band coder can thus avoid the computationally expensive analysis-synthesis systems required of a vocoder, and can produce moderate-quality, intelligible speech in a relatively inexpensive manner.

REFERENCES

1. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Subbands," B.S.T.J., *55*, No. 8 (October 1975), pp. 1069–1085. See also Proc. of the 1976 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 233–236.

2. R. E. Crochiere, "On the Design of Sub-band Coders for Low Bit Rate Speech Communication," B.S.T.J., this issue, pp. 747–770.
3. L. L. Beranek, "The Design of Communication Systems," Proc. IRE, *35* (Sept. 1947), pp. 880–890.
4. J. L. Flanagan, *Speech Analysis Synthesis and Perception,* New York: Springer-Verlag, 1972.
5. N. S. Jayant, "Delta Modulation of Pitch, Formant and Amplitude Signals for the Synthesis of Voiced Speech," IEEE Trans. Audio Electroacoust., *AU-21,* No. 3 (June 1973), pp. 135–140.
6. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Acoust. Speech Sig. Process., *ASSP-23,* No. 6 (December 1975), pp 552–557.
7. B. S. Atal and S. L. Hanaver, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer., *50,* (1971) pp. 637–655
8. R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," IEEE Trans. Acoust. Speech Sig. Process., *25,* No. 4 (August 1977).