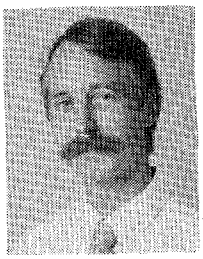
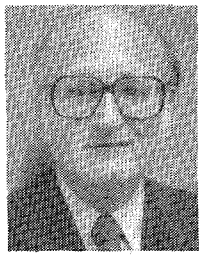


Dale E. Veeneman (S'81-M'84) received the B.S.E. degree in aerospace engineering in 1972, M.S. degrees in bioengineering and electrical science in 1980, and the Ph.D. degree in bioengineering in 1984, all from the University of Michigan, Ann Arbor.

He served as a United States Peace Corps Volunteer in Liberia, West Africa, from 1973 to 1976. From 1981 to 1984 he taught a signal processing laboratory course at the University of Michigan and received a Distinguished Achievement Award in Teaching from the Department of Electrical and Computer Engineering in 1983. In 1984 he joined the Signal Processing Department of GTE Laboratories Inc., Waltham, MA, as a member of Technical Staff. His primary research interests include digital signal processing and the application of physiological modeling to speech processing.

Dr. Veeneman is a member of Sigma Xi, Phi Eta Sigma, and the IEEE Acoustics, Speech, and Signal Processing and Engineering in Medicine and Biology Societies.



Spencer L. BeMent (S'61-M'67-SM'77) received B.S.E. and M.S.E. degrees in electrical engineering and the Ph.D. degree in bioengineering from the University of Michigan, Ann Arbor, in 1960, 1962, and 1967, respectively.

He participated in psychophysical research in the Sensory Intelligence Laboratory of the University of Michigan from 1960 to 1967. Since then he has performed research in sensory neurophysiology in the Bioelectrical Sciences Laboratory at the University of Michigan and at the University of Vermont in 1976. His research interests include acoustical signal processing and recognition, applications of solid-state electrodes for prosthetic devices, biopotential waveform analysis, and electrical stimulation of neural components. He is presently Professor of Electrical Engineering and Computer Science at the University of Michigan, and is also associated with the University's Bioengineering Program.

Dr. BeMent is a member of Sigma Xi, Tau Beta Pi, Eta Kappa Nu, and the IEEE Acoustics, Speech, and Signal Processing, the Biomedical Engineering, and the Education Societies.

Improvement of the Excitation Source in the Narrow-Band Linear Prediction Vocoder

GEORGE S. KANG, MEMBER, IEEE, AND STEPHANIE S. EVERETT

Abstract—**The major weakness of the current narrow-band LPC synthesizer lies in the use of a "canned" invariant excitation signal. The use of such an excitation signal is based on three primary assumptions, namely, 1) that the amplitude spectrum of the excitation signal is flat and time invariant, 2) that the phase spectrum of the voiced excitation signal is a time-invariant function of frequency, and 3) that the probability density function of the phase spectrum of the unvoiced excitation signal is also time invariant. This paper critically examines these assumptions and presents modifications which improve the quality of the synthesized speech without requiring the transmission of additional data. Diagnostic acceptability measure (DAM) tests show an increase of up to five points in overall speech quality with the implementation of each of these improvements. These modifications can also improve the speech quality of LPC-based speech synthesizers.**

INTRODUCTION

THE narrow-band LPC operating at 2.4 kbits/s [1] is becoming a vital part of military and civilian communication systems because it is capable of providing adequate communicability under less than ideal operating conditions such as the limited transmission bandwidth of high-frequency (HF) channels or telephone lines. In general, the intelligibility of narrow-band LPC speech compares favorably to that of voice processors operating at higher data rates. However, the speech quality of the narrow-band LPC is still relatively poor. The objective of this investigation, a continuation of the authors' previous work on LPC analysis improvements [2], is to enhance the speech quality of the narrow-band LPC by modifying the synthesizer without altering the data rate, the speech sampling rate, the frame rate, or the parameter coding formats.

The weakest part of the narrow-band LPC is the excitation signal source. Like most other narrow-band speech encoders, the LPC relies on a strictly binary voicing decision, using either a repetitive pulse (voiced) or random noise (unvoiced) as the excitation source in the synthesizer. Very little information related to the desired characteristics of the excitation signal is transmitted. It is this oversimplification of the speech excitation that allows the transmission of speech at the rate of 2.4 kbits/s. Although some people prefer LPC speech over the raspy sound of many high-rate processors [3], the use of such a simplified excitation signal causes the speech to sound indistinct or fuzzy and tense. It also tends to smear abrupt changes present in the original speech.

We can reduce some of these undesirable characteristics through modifications of the assumptions used for the generation of the conventional excitation signal. To do this, it is necessary to express the narrow-band LPC excitation signal in a more general form:

$$e(i) = \sum_{k=0}^{K} a(k) \left[\cos\left(\frac{2\pi k}{I}\right) i + \phi(k) \right] \qquad 1 \leqslant i \leqslant N \qquad (1)$$

where $a(k)$ and $\phi(k)$ are the kth amplitude and phase spectral

Manuscript received March 9, 1984; revised September 17, 1984.

The authors are with the Naval Research Laboratory, Washington, DC 20375.

TABLE I
SUMMARY OF PARAMETERS FOR THE CONVENTIONAL AND MODIFIED LPC EXCITATION SIGNALS

Parameters	Conventional Narrow-Band LPC Excitation Signal	Our Modified Narrow-Band LPC Excitation Signal
Amplitude Spectrum $a(k)$	Frequency independent and time invariant	With weak resonant frequencies updated pitch synchronously
Phase Spectrum $\phi(k)$		
Voiced Speech	A nonlinear function of frequency and time invariant	A quadratic function of frequency, with frequency-dependent phase jitter
Unvoiced Speech	N/A[a]	A stationary random process with a uniform distribution between $-\pi$ and π rad, superimposed by amplitude-weighted, randomly spaced pulses for plosives

[a]Most commonly, the conventional unvoiced excitation signal is read out randomly from a table containing uniformly distributed random numbers.

components, respectively, I is the number of excitation signal samples, K is the total number of spectral components within the passband, and N is the number of samples in a frame. In the narrow-band LPC, the phase spectrum is dependent on the voicing decision (i.e., deterministic if voiced or random if unvoiced) and I is equal to the pitch period. For unvoiced speech, the pitch period is arbitrary, usually a quarter of a frame. As stated earlier, characteristics of both the amplitude and phase spectra are assumed to be time invariant for the narrow-band LPC excitation signal. This paper examines the effects of these assumptions on the synthetic speech, and modifies them so that their general characteristics are closer to those of the prediction residual, which represents the ideal excitation signal for an LPC analysis/synthesis system. While the current narrow-band LPC transmits very little information related to the excitation source, some of the other data (filter coefficients, change of speech rms values, etc.) contain implicit information related to the desired excitation signal characteristics. These have been exploited in the improvement of the narrow-band LPC excitation signal. Table I summarizes the differences between the conventional and modified excitation signals.

AMPLITUDE SPECTRUM OF THE VOICED EXCITATION SIGNAL

The amplitude spectral envelope of the conventional voiced excitation signal is flat (i.e., $a(k)$ in (1) is the same for all k's). The use of such an excitation would be logical if the LPC analysis filter were capable of removing speech resonant and anti-resonant frequency components completely so that the prediction residual had a flat amplitude spectral envelope. In actuality, because of limitations inherent in the linear predictive analysis (all-pole modeling of the speech, the use of a limited number of filter weights, quantization of filter coefficients, etc.), the prediction residual retains a considerable amount of speech resonant and anti-resonant components (see Fig. 1). The presence of these resonant frequencies makes the prediction residual itself highly intelligible—using only the prediction residual, the average diagnostic rhyme test (DRT) score for a set of three male speakers was 83.5.

There are two major causes of resonant frequencies in the prediction residual. First, the magnitudes of the resonant peaks of an all-pole filter, such as the LPC synthesis filter, are dependent on the pole locations; they cannot be independently

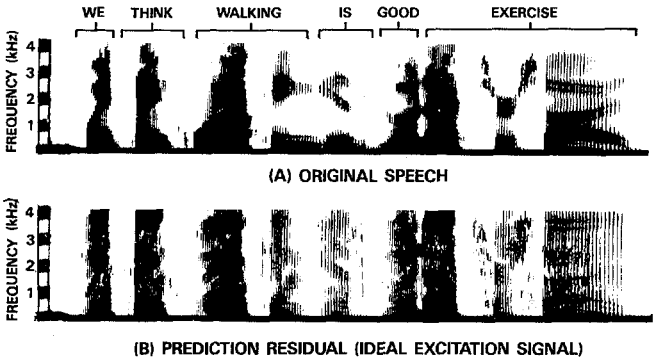


Fig. 1. Spectra of original speech and LPC excitation signals. The prediction residual contains a considerable number of resonant frequency components unfiltered by the LPC analysis filter.

controlled as they can in a parallel formant synthesizer. In other words, for a given set of pole locations, the magnitudes of the resonant peaks are predetermined. It has been our observation that the formant amplitudes in the LPC synthesizer are often lower than those of actual speech. The greater the magnitude of the original formants, the stronger the resonant frequency components in the prediction residual. Therefore, a voice with unusually intense formant frequencies will not be reproduced well by the narrow-band LPC unless the excitation signal is augmented with formant frequencies similar to those in the prediction residual.

Resonant frequencies in the prediction residual also result from the quantization of the filter coefficients which tends to reduce the spectral peaks attained by an all-pole filter, as illustrated in Fig. 2. This reduction is partly due to the clipping of LPC coefficients by the LPC quantizer. Again, the differentials in the spectral peaks will appear as formant frequencies in the prediction residual. (Fig. 2 is based on the coefficient quantization rule for the narrow-band LPC specified by Federal Standard 1015, but other parameter quantization rules designed for 2.4 kbit/s LPC's produce similar results.)

Amplitude Spectrum Modification of the Voiced Excitation Signal

An exact shaping of the voiced excitation spectrum is impossible because no direct information is transmitted by the narrow-band LPC. However, an approximation is possible through the exploitation of the following observations.

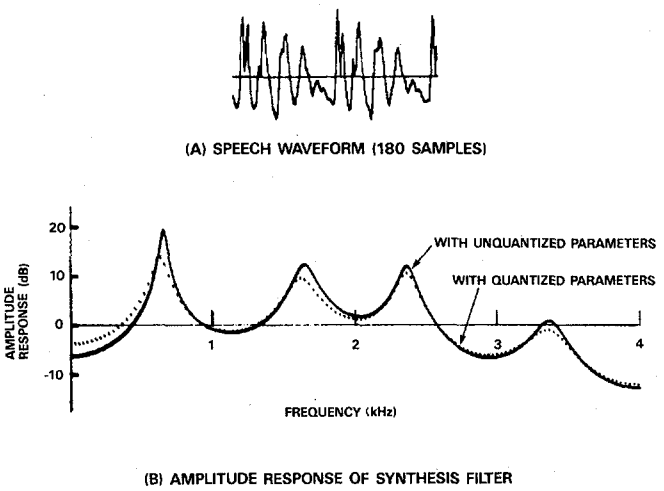


Fig. 2. Effect of LPC coefficient quantization on the amplitude response of the synthesis filter. Quantization of LPC coefficients results in a reduction of resonant peaks in the synthesis filter.

The first observation is that the predominant resonant frequencies of the prediction residual track closely with those of the original speech, as illustrated in Fig. 1. This is why the prediction residual is so intelligible. While the prediction residual has extraneous resonant frequencies not found in the original, omission of these does not seem to have a significant impact on the output speech. However, the resonant peaks in the prediction residual are nearly equalized, unlike those of the original speech. Thus, the all-pole spectrum of the prediction residual may be approximated by the all-pole spectrum of the speech with a reduced feedback gain:

$$A(z) = \frac{1}{1 - G \sum_{n=1}^{N} \alpha_n z^{-n}} \qquad G < 1 \tag{2}$$

where α_n is the nth prediction coefficient of the speech available at the LPC synthesizer. The factor G is related to the overall reduction of the pole moduli. Since the root loci of $A(z)$ do not lie along the radial direction, there will be a slight but insignificant shift in the frequency of the resonant peaks.

The second observation used in shaping the excitation signal is that the resonant peaks in the residual become smaller when the efficiency of prediction is fairly high (i.e., the residual rms is much smaller in comparison to the speech rms). This usually occurs with front vowels, murmurs, and nasals, which are well suited to the all-pole modeling of the LPC. Hence, it is reasonable to assume that the modulus reduction factor G is proportional to the ratio of the residual rms to the speech rms, namely,

$$G = G' \sqrt{\prod_{n=1}^{N} (1 - k_n^2)} \tag{3}$$

where k_n is the nth reflection coefficient available at the LPC synthesizer and the proportionality constant G' is yet to be determined.

The constant G' is dependent on several factors, including the preemphasis factor, the nature of filter coefficient quantization, the voice characteristics, and the speech itself. G' minimizes the mean-square difference between the all-pole spectrum of the prediction residual and the all-pole spectrum of the spectral shaper expressed by (2) with (3).

We analyzed approximately 1200 frames (180 samples/frame) of male and female voiced speech samples to obtain a preferred value for G', using a frequency-domain computational approach which enabled us to exclude the effect of frequency components below 150 Hz since they are not audible at the narrow-band LPC output anyway. Not surprisingly, G' varies from speaker to speaker. According to our analyses, a reasonable choice for G' would be somewhere around 0.25, although from listening to processed speech while varying G' from 0 to 1.0, it appears that there is a broad range of acceptable values for G'.

The excitation spectrum defined by (2) may be incorporated in the narrow-band LPC in two ways: one is a direct method in which the amplitude spectral components in the excitation signal model in (1) are made equal to the amplitude spectrum of (2); the other is an indirect method in which the amplitude spectrum is modified by passing the flat-spectrum excitation signal through an all-pole filter whose transfer function is described by (2). We tried both methods and noted virtually no difference in the sound quality.

Test and Evaluation

We incorporated the amplitude spectral modification of the voiced excitation signal in NRL's programmable real-time narrow-band voice processor and in another narrow-band LPC currently under development. The diagnostic acceptability measure (DAM) was used to evaluate the speech quality of these two systems. Both tests yielded virtually identical results, with a 4.8-point improvement for male speakers (from 48.6 to 53.4) and a 5.6-point improvement for female speakers (from 44.9 to 50.5). The scores for the modified LPC compare favorably to those for a 9.6 kbit/s voice processor (54.8 for males and 53.5 for females).

Although we did not expect the amplitude spectrum modification of the voiced excitation signal to noticeably affect consonant intelligibility, we nevertheless conducted diagnostic rhyme tests (DRT's) to ensure that it did not hurt the speech intelligibility. The DRT scores for three male and three female speakers in a quiet environment were 87 both with and without the amplitude spectrum modification. Likewise, the DRT scores for three male speakers in a shipboard environment were virtually unchanged—78 with modification and 77 without modification. These results confirm that the amplitude spectral modification of the voiced excitation signal significantly improves the quality of the narrow-band LPC without degrading the intelligibility.

Phase Spectrum of the Voiced Excitation Signal

The amplitude and phase spectra of the conventional voiced excitation signal are time-invariant and repeat exactly from one pitch cycle to the next. In contrast, the actual prediction residual varies substantially from cycle to cycle. This is due to irregularities in the vocal cord movement and the turbulent airflow from the lungs during the glottis-open period of each pitch cycle. The amount of waveform jitter varies with the

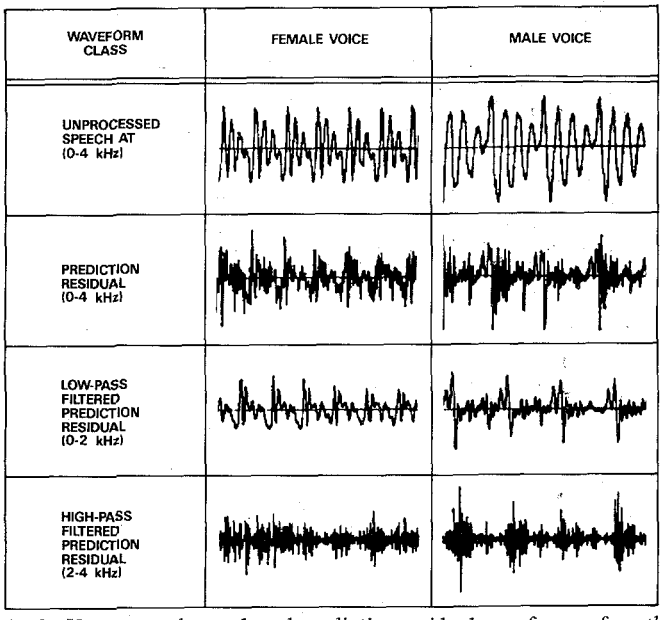


Fig. 3. Unprocessed speech and prediction residual waveforms of *soothing, mellow* voices. Note the randomness of the prediction residual, particularly the high-passed prediction residual, and compare this waveform to the conventional narrow-band LPC voiced excitation signal shown in Fig. 6. Some amount of randomness in the excitation signal is essential for the production of natural-sounding speech. Note also the highly oscillatory speech waveform characteristic of mellow voices. The prediction residual waveforms illustrated in this figure (and Figs. 4–6) have been amplified four times for clarity.

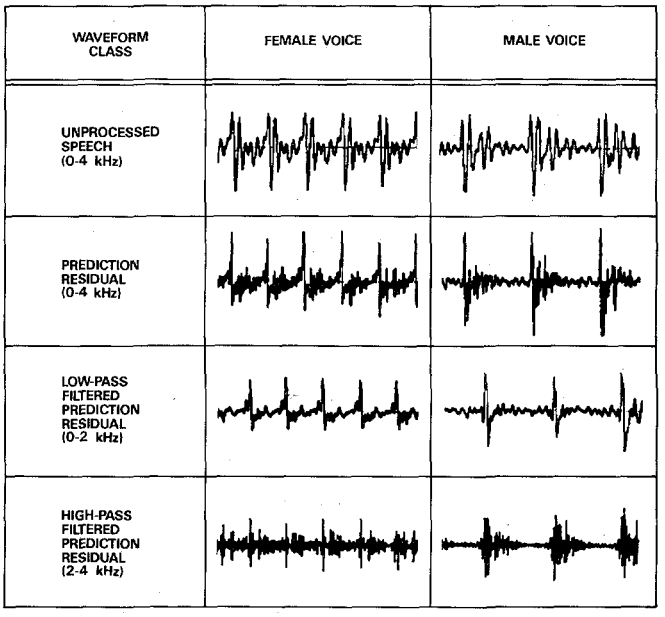


Fig. 4. Unprocessed speech and prediction residual waveforms of *normal voices.* Note that the periodicity of the prediction residual is better defined than in the preceding figure, but less than for the tense voices in the following figure. The modified voiced excitation has a similar amount of randomness, as illustrated in Fig. 6.

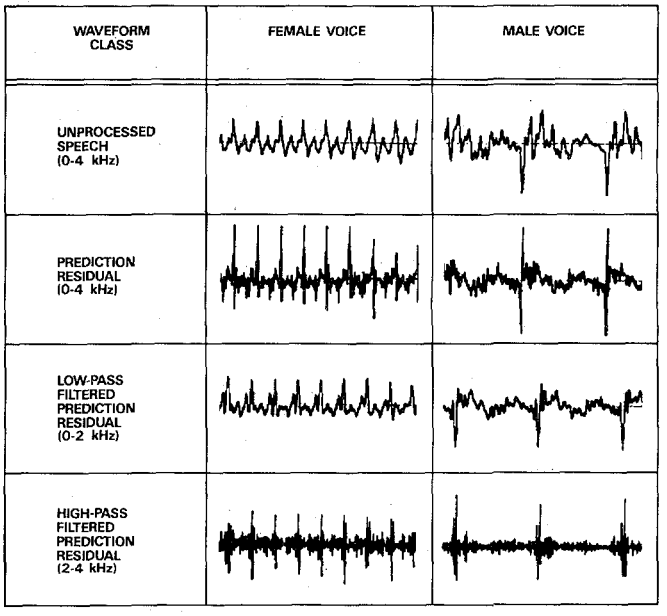


Fig. 5. Unprocessed speech and prediction residual waveforms of *tense voices.* Note that the well-defined periodicity of the prediction residual, even the high-pass filtered prediction residual, is very similar to that of the conventional narrow-band LPC voiced excitation signal shown in Fig. 6. Note also the highly damped speech waveform.

age of the speaker, the speaker's nervous condition and degree of muscular elasticity, and with the nature of the speech sound. Without an appropriate amount of waveform jitter in the excitation signal, the synthetic speech sounds flat, machine-like, and usually buzzy; it may also sound edgy, tense, or angry.

This last effect deserves special attention because of its particularly insidious nature. When we look at the waveform structure of a smooth, mellow voice, we immediately notice that it lacks the strong, regular pitch harmonics so prevalent in the synthetic LPC speech. This is due to the presence of a certain amount of breath air during the glottis-open period, which introduces flutter in the pitch harmonics. On the other hand, strong, regular pitch harmonics similar to those of the LPC synthesized speech are characteristic of sharp, clear voices and of speakers who are tense or angry.

Figs. 3–5 are vivid illustrations of how the speech and prediction residual waveforms differ in three selected types of voices—unusually mellow, normal, and tense—for both male and female speakers. Note that the periodicity of the prediction residual, particularly that of the high-passed prediction residual, is progressively better defined as the tenseness of the voice increases. In very tense voices, the prediction residual looks much like the conventional voiced excitation signal used in the narrow-band LPC (compare Figs. 5 and 6). This is one of the reasons LPC speech sounds unnecessarily tense regardless of the quality of the speaker's own voice.

All of these observations lead us to the conclusion that a small amount of irregularity in the narrow-band LPC speech is highly desirable. A similar conclusion was reached by Makhoul *et al.* [4], who introduced irregularity in LPC synthesized speech by using a mixed excitation source in which the periodic pulse train was low-pass filtered while the noise was high-pass filtered at the same cutoff frequency. The cutoff frequency was variable, and was estimated to be the highest frequency at which the speech spectrum was considered periodic. This cutoff frequency was quantized into 2 or 3 bits and transmitted to the receiver. The frequency quantization step was as coarse as 500 Hz, and low-order Butterworth filters were used. Mixed excitation sources have also been applied to channel vocoders

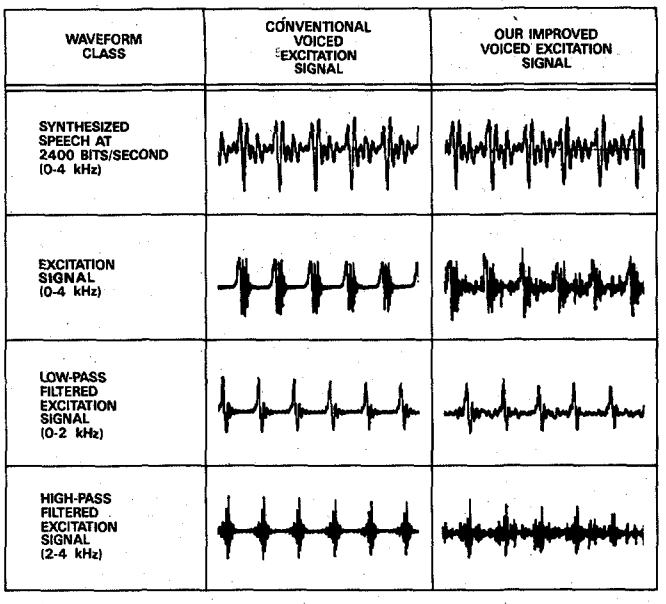


Fig. 6. Synthesized speech and excitation signal waveforms for the narrow-band LPC. These waveforms are generated by the use of LPC parameters extracted from the normal female speech waveform shown in Fig. 4. The absence of randomness in the conventional voiced excitation signal is in part responsible for the tense and unnatural speech quality of the narrow-band LPC. (Compare the left column of this figure to Fig. 5.) The presence of randomness in the modified voiced excitation signal (right column) adds naturalness to the synthesized speech. This voiced excitation signal is an approximation of the actual prediction residual of the normal female voice shown in Fig. 4.

[5] and to the formant synthesizer [6] in efforts to improve voice quality.

In our approach, the mixed excitation source is simply a special case of the excitation signal generator described in (1). The phase spectrum consists of two parts, reflecting both the stationary (time-invariant) and random (time-variant) characteristics of voiced excitation:

$$\phi(k) = \phi_0(k) + \Delta\phi(k) \qquad k = 0, 1, \cdots, K \tag{4}$$

where $\phi_0(k)$ and $\Delta\phi(k)$ are the kth stationary and random phase components, respectively. These phase spectral components are discussed in the following section.

Stationary Part of the Phase Spectrum

The stationary part of the phase spectrum of the voiced excitation signal is important because it has a direct bearing on the peakiness and dispersiveness of the excitation signal. For example, if the phase spectrum is a linear function of frequency or the differential delay is zero, the corresponding time function is an impulse. The use of an impulse for the voiced excitation is undesirable for two reasons. First, a spiky excitation signal produces a spiky narrow-band LPC output which does not operate well in tandem with high-rate voice processors that encode the difference between two consecutive speech samples, such as continuously variable slope delta (CVSD) systems. Because CVSD cannot accurately follow the steep changes in the input amplitude produced by the impulse excitation, the output speech is distorted. Over the years, the narrow-band LPC has improved its tandem performance with the CVSD. At one time, the DRT score for a 16 kbit/s CVSD operating from the narrow-band LPC output was 78 for three male and three female voices; it is now 82. One reason for this improvement is the use in the LPC of a time-dispersed voiced excitation signal in lieu of an impulse excitation.

Second, a spiky excitation signal requires a greater dynamic range in the LPC signal processor, so the output amplitude often has to be lowered in order to avoid clipping. We can reduce the required dynamic range by as much as 10 dB through the use of a time-dispersed voiced excitation signal such as that discussed below.

In addition, it has been reported that the use of a nonimpulse excitation source produces synthetic speech preferable to that produced by an impulse excitation [7], [8]. In the past, a number of different approaches have been investigated in an effort to design a family of signals with flat amplitude spectra and low peak amplitudes [9], [10]. One option is a signal having a flat amplitude spectrum and a phase spectrum which is a quadratic function of frequency [10]. Thus,

$$\phi_0(k) = (2\pi)\,\xi\left(\frac{k}{K}\right)^2 \qquad k = 0, 1, \cdots, K \tag{5}$$

where $\phi_0(k)$ is the kth stationary phase component defined in (1) and K is the total number of spectral components. The quantity ξ is an integer number; the larger the ξ, the greater the dispersion of the excitation signal. The differential delay, as obtained from (5), is

$$D_0(k) = \frac{\Delta\phi_0(k)}{\Delta w} = \frac{(2\pi)\,(2\xi)}{K(\Delta w)}\left(\frac{k}{K}\right) \qquad k = 0, 1, \cdots, K \tag{6}$$

in which Δw is a uniform frequency spacing between two adjacent spectral components. In our narrow-band LPC, $K(\Delta w)$ is (2π) 4000 rad/s. Thus, (6) may be written as

$$D_0(k) = \frac{\xi}{2}\left(\frac{k}{K}\right) \text{ ms} \qquad k = 0, 1, \cdots, K. \tag{7}$$

Equation (7) states that if the phase angle is a multiple of 2π rad at 4000 Hz, the differential delay at the same frequency is a multiple of 0.5 ms.

For purposes of illustration, we generated four different voiced excitation signals using $\xi = 3, 4, 5$, and 6 in (5) and (7). The spectral and temporal characteristics of these signals are listed in Table II. In Example 1 ($\xi = 3$), the differential delay increases linearly from 0 ms at 0 Hz to 1.5 ms at 4000 Hz. The excitation signal samples are dispersed over 25 sampling time intervals as given in Table III. The peak amplitude reduction factor—defined as the maximum signal magnitude when the signal is normalized to have a unity power—is 8.98 dB. In the second example ($\xi = 4$), the differential delay at 4000 Hz is increased to 2 ms, and the excitation signal samples are dispersed over 31 sampling time intervals. The resulting peak amplitude reduction factor is increased to 9.51 dB, and so on.

Any of these excitation signals is satisfactory if the pitch period is greater than the number of excitation signal samples. If the pitch period is smaller than the number of excitation sig-

TABLE II

FOUR EXAMPLES OF TIME-INVARIANT VOICED EXCITATION SIGNAL CHARACTERISTICS. IN THESE EXAMPLES, THE PHASE SPECTRUM IS A QUADRATIC FUNCTION OF FREQUENCY (i.e., THE DIFFERENTIAL DELAY IS A LINEAR FUNCTION OF FREQUENCY). FOR PURPOSES OF COMPARISON, THE DISPERSION WIDTH IS DEFINED AS THE TIME INTERVAL IN WHICH EVERY SAMPLE OUTSIDE THIS TIME INTERVAL HAS A MAGNITUDE LESS THAN OR EQUAL TO 1/256 WHEN THE SIGNAL AMPLITUDE IS NORMALIZED TO HAVE A UNIT VARIANCE.

Example	Amplitude Spectrum	Phase Shift @ 4000 Hz $(2\pi)\,\xi$	Diff. Delay @ 4000 Hz $0.5\,\xi$	Absolute Maximum Amplitude when $\Sigma\, e^2(n) = 1$		Dispersion Width
1	Flat	$3(2\pi)$ rad	1.5 ms	0.3555	−8.98 dB	25 samples
2	Flat	$4(2\pi)$	2.0	0.3344	−9.51	31
3	Flat	$5(2\pi)$	2.5	0.3194	−9.91	35
4	Flat	$6(2\pi)$	3.0	0.2835	−10.95	41

TABLE III

FOUR EXAMPLES OF TIME-INVARIANT VOICED EXCITATION SIGNALS. THESE ARE COMPARABLE TO THE CONVENTIONAL VOICED EXCITATION SIGNAL, AND MAY BE OBTAINED THROUGH AN INVERSE FOURIER TRANSFORM OF THE SPECTRA SPECIFIED IN TABLE II. EACH EXCITATION SIGNAL IS NORMALIZED TO HAVE A VARIANCE OF $(1024)^2$.

Time Index	Example 1[a]	Example 2	Example 3	Example 4
1				4
2				−6
3				8
4			−5	−12
5			7	19
6		−4	−11	−29
7		6	17	44
8		−10	−28	−69
9	5	16	44	104
10	−8	−26	−72	−154
11	13	44	114	212
12	−24	−76	−175	−262
13	43	128	244	267
14	−81	−204	−295	−183
15	147	289	271	−9
16	−252	−335	−113	228
17	359	239	−155	−290
18	−364	44	327	60
19	92	−342	−152	253
20	336	231	−245	−194
21 (center)	−306	250	225	−212
22	−336	−231	245	194
23	92	−342	−152	253
24	364	−44	−327	−60
25	359	239	−155	−290
26	252	335	113	−228
27	147	289	271	−9
28	81	204	296	183
29	43	128	244	267
30	24	76	174	262
31	13	44	114	212
32	8	26	72	154
33	5	16	44	104
34		10	28	69
35		6	17	44
36		4	11	29
37			7	19
38			5	12
39				8
40				6
41				4

[a]Our choice.

nal samples, the customary procedure is to superimpose the trailing samples of the excitation signal onto the next pitch cycle. Such a superposition is not completely valid, however, because the LPC synthesizer is a dynamic system in which the filter coefficients are updated pitch synchronously and the output amplitude is normalized after synthesis. Our choice for the modified voiced excitation signal is Example 1 ($\xi = 3$); if the pitch period is less than 25 samples, the trailing samples are simply discarded. Our computations show that even if only the first 20 samples are used (i.e., a fundamental pitch frequency of 400 Hz at a 8 kHz sampling rate), the excitation signal spectrum is flat within 0.5 dB for frequencies greater than the fundamental pitch frequency.

Random Part of the Phase Spectrum

There are two types of randomness present in the natural voiced speech waveform. One is pitch-epoch variation or jitter caused by irregularities in vocal cord movement; the other is period-to-period waveform variation caused by the turbulent air flow from the lungs. The magnitude of pitch-epoch variation is small, having a standard deviation somewhere between 10 and 60 μs for adult male speakers [11]; it may therefore be ignored in the narrow-band LPC.

The period-to-period waveform variations caused by breath air are very complex. On the one hand, they are random because the air coming from the lungs is turbulent. On the other hand, they are pitch modulated because the air passes through the glottis as it opens and closes at the pitch rate. These waveform variations constitute a substantial portion of the prediction residual, and are disproportionally strong in the high-frequency regions because the LPC analysis filter boosts the treble to flatten the spectral envelope of the voiced speech. The amount of period-to-period waveform variation in the prediction residual differs substantially from speaker to speaker, as can be seen in Figs. 3-5. In addition, there is evidence to indicate that the amount of waveform variation is dependent on the speech sound—for example, there is more randomness in back vowels than in front vowels.

The presence of these variations in the excitation signal is essential to the synthesis of natural-sounding speech. Unfortunately, period-to-period waveform variations cannot be reproduced exactly in the narrow-band LPC because relevant information is not available at the receiver. However, since there is a many-to-one transformation between random noise and its perception by the human ear, the nature of any artificially introduced randomness in the voiced excitation signal need not be exactly identical to that of the prediction residual.

We listened to a large number of speech samples processed by our real-time narrow-band LPC as we varied the nature of the random components in the voiced excitation signal. While there seemed to be a wide range of acceptable characteristics, we noted that the overall intensity and the frequency distribution of the random components appeared to be more signifi-

cant than other parameters. The overall intensity is important because the speech quality suffers both if it is too low or too high; the frequency distribution characteristics are important because the speech will sound warbly if there is too much low-frequency jitter. It is interesting to note that these are the only two parameters utilized by the narrow-band LPC to synthesize unvoiced speech.

We are interested in extracting average values for these two parameters from the actual prediction residual so that they may be used as constants in the LPC receiver. This analysis is by no means straightforward; the selection of the proper prediction residual samples and the choice of the analysis method are both critical. The prediction residual samples must be selected carefully because period-to-period waveform variations in the prediction residual are caused not only by breath noise and the instability of the excitation source (i.e., the glottis), but also by the changes in the vocal tract during speech transitions. We would like to exclude the effects of the speech transitions in the estimated parameters so we must select prediction residual samples from voiced frames where the LPC coefficients (i.e., the vocal tract filtering characteristics) do not vary significantly from one frame to the next. In other words, we must select the prediction residuals for analysis from sustained vowels.

Once the residual samples are selected, the choice of the analysis method is critical for obtaining reliable results. The most direct way of estimating the intensity and frequency distribution parameters is through a variance analysis of the phase spectra derived from the prediction residual using a pitch-synchronous analysis window. However, this approach is insurmountably difficult and risky since even visual inspection cannot reliably determine the pitch epoch in a highly noise-like prediction residual (for example, see Fig. 3). The phase spectrum is sensitive to the location of the window with respect to the waveform under analysis, and frequent window placement errors will degrade the estimated parameters beyond any usefulness. Since we are primarily interested in the gross characteristics of the frequency dependency and the overall intensity, we choose to use an indirect analysis method involving the spectral analysis of the pitch-filtered prediction residual defined by

$$r'(i) = r(i) - \beta r(i - T) \tag{8}$$

where $r(i)$ is a prediction residual sample, $r'(i)$ is a pitch-filtered prediction residual sample, T is the pitch period, and β is a first-order prediction coefficient of $r(i)$ T samples apart. As usual, β is obtained by minimizing the mean-square value of the right-hand member of (8). Thus,

$$\beta = \frac{\sum_i r(i)\, r(i - T)}{\sum_i r^2(i - T)}. \tag{9}$$

Since we select only stationary prediction residuals for the analysis, β may be expressed by

$$\beta = \frac{\sum_i r(i)\, r(i - T)}{\frac{1}{2}\left[\sum_i r^2(i) + \sum_i r^2(i - T)\right]} \tag{10}$$

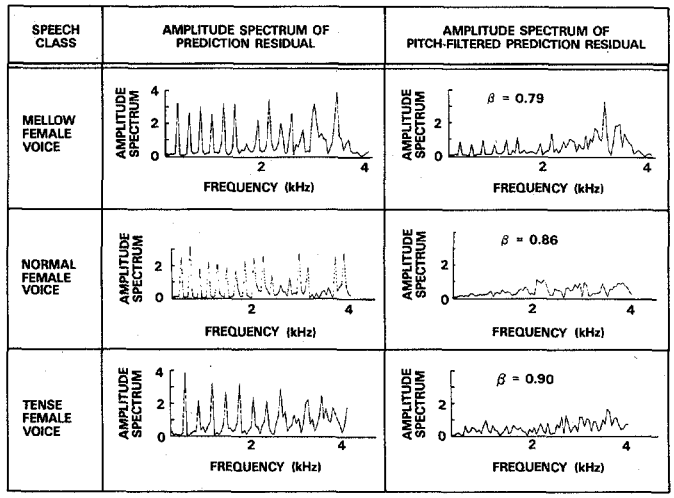


Fig. 7. Amplitude spectra of prediction residuals and pitch-filtered prediction residuals from the three female voices shown in Figs. 4–6. As noted, the amplitude spectrum of the pitch-filtered prediction residual generally increases with frequency.

where the magnitude is bounded between 1 and -1. Equation (8) represents the input-output relationship of a notch filter which suppresses harmonically related frequencies (in this case, the fundamental pitch frequency and its harmonics). The quantity β is related to the notch filter bandwidth and is dependent on the randomness of the input. For example, in the absence of randomness, as in the conventional voiced excitation signal, β is unity. For actual prediction residuals from steady vowels, β lies somewhere between 0.7 and 0.9.

With a steady vowel as the input, the pitch-filtered prediction residual contains mainly period-to-period waveform variations of the prediction residual. Thus, the spectral analysis of the pitch-filtered prediction residual indicates both the nature of the frequency dependency and the overall intensity of the random parts of the prediction residual. Fig. 7 shows the amplitude spectra of pitch-filtered prediction residuals generated from the three types of female voice waveforms previously illustrated in Figs. 3-5. For reference, the amplitude spectra of the corresponding prediction residuals are also shown in Fig. 7.

The spectral distribution of the pitch-filtered prediction residual is significant because it represents the spectrum of the period-to-period waveform variations in the prediction residual. We introduce random components in the voiced excitation signal such that the amplitude spectrum of the pitch-filtered excitation signal has a spectral distribution similar to that of normal voices as shown in Fig. 7. As can be seen in this figure, and in similar plots of other voices, the amplitude spectrum of the pitch-filtered prediction residual is an approximately linear function of frequency, and the pitch prediction coefficient β is somewhere around 0.85. Thus, the random part of the phase spectrum $\Delta\phi(k)$, as obtained numerically through the use of (1), (8), and (10), is approximately

$$\Delta\phi(k) = \frac{\pi}{2}\,\sigma(k)\left(\frac{k}{K}\right) \quad \text{rad} \tag{11}$$

where $\sigma(k)$ is a uniformly distributed random variable between -1 and 1, k is the frequency index, and K is the total number of components within the 0-4 kHz passband. Fig. 8 is similar

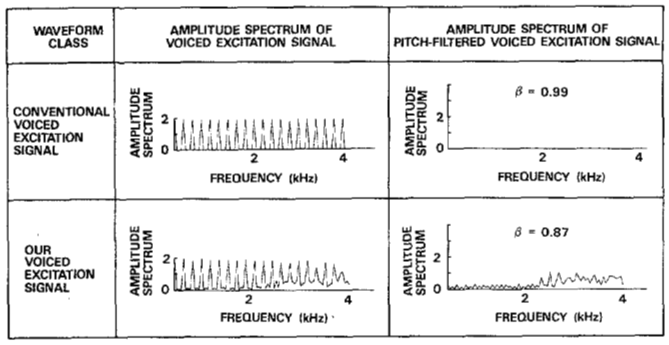


Fig. 8. Amplitude spectra of the voiced excitation signal and the pitch-filtered voiced excitation signal for the conventional excitation (upper illustrations) and our modified excitation (lower illustrations). Both are derived from LPC parameters generated using the speech waveform of the normal female voice shown in Fig. 5. (The prediction residual spectrum and pitch-filtered residual spectrum of this voice are shown in Fig. 7.) The conventional voiced excitation signal has a small amount of randomness because we carefully introduced the actual LPC parameter quantization and interpolation effects in the excitation signal, but the amount of randomness is negligible. On the other hand, our voiced excitation signal has randomness in which the frequency dependency and magnitude (in terms of the β value) are similar to those of the pitch-filtered prediction residual of the actual speech as shown in Fig. 7.

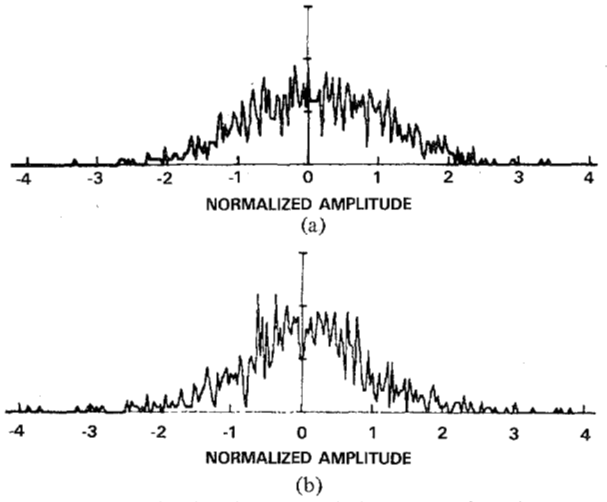


Fig. 9. Probability density function of the unvoiced excitation signal used to synthesize /s/. The first illustration (a) shows the unvoiced excitation signal generated by (1) with random phases between π and $-\pi$. The second (b) shows the prediction residual of a natural /s/. The normalized amplitude is the excitation signal amplitude divided by its root-mean-square value.

to the plot in Fig. 7, and compares the conventional voiced excitation signal and our modified voiced excitation signal. Note that our pitch-filtered excitation signal has characteristics more similar to those of the prediction residual of the normal voice. (The time samples of both excitation signals are shown in Fig. 6.)

Test and Evaluation

When our voiced excitation signal is used in the narrow-band LPC, one can readily hear that the output speech has a quality of breathiness not unlike that of the unprocessed speech, and that the buzzy, twangy qualities often present in the conventional narrow-band LPC output are greatly reduced. DAM tests were conducted to ascertain the degree of quality improvement achieved. The test results show a 4.7-point improvement for male speakers (from 48.6 to 54.3) and a 4.8-point improvement for female speakers (from 44.9 to 49.7). A DRT was also conducted to ensure that the phase spectral modification did not produce such strong improvements in speech quality at the expense of the speech intelligibility. As expected, the DRT score of 85.8 for the modified LPC was only slightly better than the 85.3 for the conventional LPC.

Probability Density Function of the Unvoiced Excitation Signal

In the past, the unvoiced excitation signal has not received as much attention as the voiced excitation signal. The excitation signal traditionally used for generating all unvoiced sounds has been uniformly distributed random noise; no distinction is made between fricative sounds (/h/, /s/, /sh/, /f/, /th/) and burst or stop sounds (/p/, /t/, /k/). Usually the excitation signal is generated by randomly choosing numbers from a table of random numbers.

In natural speech, a fricative sound, generated by a turbulence in the airflow caused by a constriction somewhere in the vocal tract, is essentially stationary random noise. As it is whitened by the LPC analysis filter, the prediction residual becomes virtually bandlimited white noise. Based on an examination of various samples, the probability density function of the prediction residual of fricative sounds is closer to Gaussian than uniform in distribution, similar to that obtained from (1) when the phase spectral components are randomly distributed between π and $-\pi$ (see Fig. 9). The exact form of the probability density function, however, is not too critical in the reproduction of fricative sounds. As long as the output spectrum and loudness are similar to those of the original speech, the ear tends to perceive them as fricatives. The conventional unvoiced excitation signal is therefore satisfactory for the generation of fricative sounds.

However, the conventional unvoiced excitation signal is not satisfactory for reproducing plosives. A plosive is a sequence of events: a rapid closure of the oral cavity, a build up of air pressure, and the subsequent release of a short burst of broadband energy. Since this sudden burst cannot be predicted well by the past speech samples, the prediction residual is particularly large at the onsets of these sounds, producing large spikes in the residual, as can be seen in Fig. 10. Since the conventional unvoiced excitation signal lacks similar spikes, the reproduced plosives often sound more like fricatives—CAT is often heard as HAT, and TICK may sound like THICK or SICK. As a result, DRT scores may suffer, particularly for those attributes important to the distinction of plosives (i.e., "Graveness").

The improvement of unvoiced plosives is accomplished by the introduction of randomly spaced spikes into the conventional unvoiced excitation signal. For the relatively gentle onsets of fricatives, the randomly spaced pulses are negligible; thus, the excitation is equivalent to the conventional unvoiced excitation signal. For the sudden onsets of plosives, however, the train of randomly spaced pulses becomes a significant portion of the excitation signal. The addition of random pulses to the conventional unvoiced excitation signal does not affect the loudness of the synthetic speech if the system calibrates the output speech rms after synthesis, as do most current narrow-band LPC's.

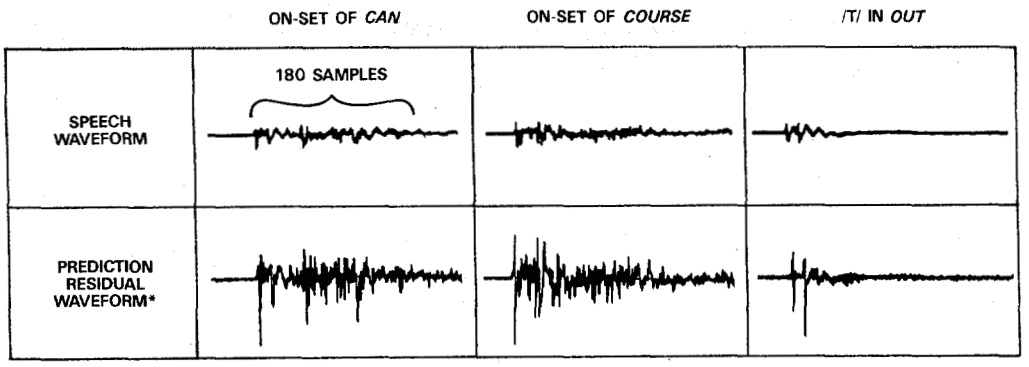


Fig. 10. Three examples of unvoiced plosives and their prediction residuals. Note large spikes in the prediction residual at the onset. Without those spikes, the plosives often sound more like fricatives.

TABLE IV
SPEECH rms RATIOS FROM TWO CONSECUTIVE UNVOICED FRAMES. UNDERLINE INDICATES WHERE THE RATIO WAS COMPUTED.

Test Word		rms Ratio
Abrupt Unvoiced Plosives	out	14
	stop	17
	to	32
	blunt	34
	can	19
	take	20
	course	25
	took[a]	26
	town[a]	19
	at your	22
	pipe	11
Nonabrupt Unvoiced Fricatives	stop	2
	self	5
	he	4
	his	3
	sharp	2
	Fred	2

[a]With shipboard background noise.

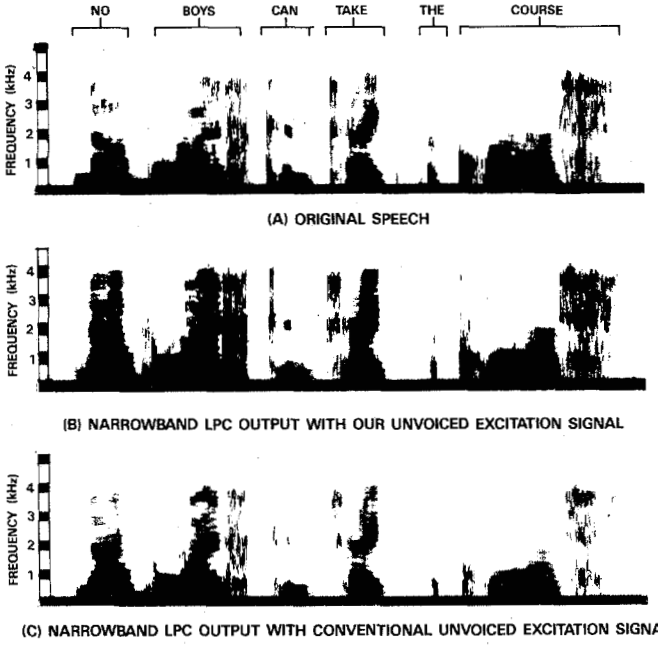


Fig. 11. Spectrograms of narrow-band LPC input and output. When our modified unvoiced excitation is used, the onsets of CAN, TAKE, and COURSE are reproduced better at the narrow-band LPC output. Note the sudden bursts of speech energy at these onsets in Fig. 11(b) and compare them to those in Fig. 11(c).

The exact form of each individual random pulse is not too critical because the ear cannot accurately analyze a broad-band random signal. We chose to use a doublet composed of a positive-going pulse followed by a negative-going pulse because such pairs do occur frequently in the prediction residuals of plosives. The spacing of the pulses is random, with the average rate fixed at four per frame. Each pulse amplitude is made proportional to the abruptness of the speech.

Measure of Abruptness

The abruptness of the speech is related to the amount of change in the speech energy over a short period of time. Thus, the ratio of the speech rms values from two consecutive frames should indicate the degree of abruptness. To test this hypothesis, we randomly selected words containing abrupt and nonabrupt unvoiced consonants and computed the speech rms ratios at the consonant onsets. The test words were excerpted from casually spoken sentences, so they were not articulated any more carefully than would be expected in normal conversational speech. The computed speech rms ratios, listed in Table IV, are consistently larger for the stops and smaller for the fricatives. This is also true for the two words (TOOK and TOWN) contaminated by helicopter carrier noise.

In general, the presence of background noise decreases the magnitude of the speech rms ratio, so unvoiced stops tend to sound like fricatives unless the noise interference is reduced somehow. For this reason, we recommend the use of a noise-cancellation microphone and noise-suppression preprocessing in noisy platforms. If this is done, then our investigations indicate that the effect of the noise floor on the rms ratio is not significant. However, we fixed the minimum rms at four in order to reduce the contrast between noise-free and noisy cases when computing the rms ratio. The values in Table IV were obtained on this basis.

Test and Evaluation

Our modified unvoiced excitation signal was developed to improve the reproduction of unvoiced speech, particularly unvoiced plosives (see Fig. 11). The DRT is an excellent

TABLE V
DRT SCORES OF NARROW-BAND LPC PROCESSED SPEECH FOR THREE FEMALES. THE FIRST SET OF SCORES WAS OBTAINED USING THE CONVENTIONAL UNVOICED EXCITATION SIGNAL; THE SECOND SET WAS OBTAINED USING OUR MODIFIED UNVOICED EXCITATION SIGNAL. NOTE PARTICULARLY THE INCREASE IN THE SCORE FOR "GRAVENESS" WHICH TESTS /p/ VERSUS /t/, /f/ VERSUS /t/, ETC.

Sound Class	With Conventional Unvoiced Excitation Signal	With Our Unvoiced Excitation Signal	Change
Voicing	88.0	83.6	-4.4
Nasality	94.5	99.2	+4.7
Sustention	74.0	77.1	+3.1
Sibilation	80.2	84.9	+4.7
Graveness	63.5	77.9	+14.4
Compactness	88.5	87.8	-0.7
Overall	81.5	85.1	+3.6

means of evaluating this improvement because it specifically tests the intelligibility of initial consonants including unvoiced plosives. We selected female speakers for the testing because the performance of the narrow-band LPC is notoriously poorer with female voices than with male voices (average DRT scores are about 5.5 points lower).

Table V lists DRT scores for three female speakers using the narrow-band LPC with the conventional unvoiced excitation signal and with our modified unvoiced excitation signal. The improvement for the attribute "Graveness" is highly significant. A look at the score changes for the features within "Graveness" reveals that this improvement is due primarily to better reproduction of unvoiced sounds, particularly plosives (/p/ versus /t/).

The slight drop in the "Voicing" score can be attributed largely to faithful reproduction of the overly strong voiced plosives in the original speech caused by articulation directly into the microphone. Since the bursts of voiced stops are normally weak, this led listeners to mistakenly identify these sounds as unvoiced. This tendency is consistent with the improvements produced by our modified unvoiced excitation signal.

CONCLUSIONS

The objective of this effort was to improve the excitation signal in the conventional narrow-band LPC synthesizer without altering the data rate, the speech sampling rate, the frame rate, or the parameter coding formats. Since the narrow-band LPC transmits speech at such a low bit rate, some of the parameters (specifically those of the excitation source) are not transmitted, but are introduced at the receiver. The major weakness of the narrow-band LPC synthesizer lies in the use of fixed excitation signal parameters which do not reflect the changing characteristics of natural speech.

We have modified the amplitude and phase spectra of the voiced excitation signal to simulate the natural irregularities found in the prediction residual of a normal voice. These modifications each improve DAM quality scores by almost five points for both male and female speakers. Through modification of the temporal characteristics of the unvoiced excitation signal, we have improved the reproduction of unvoiced plosive onsets and raised the DRT score for female speakers nearly four points. Quality and intelligibility scores with each of these LPC synthesis improvements indicate that the 2.4 kbit/s LPC is only slightly worse than a 9.6 kbit/s vocoder.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the NRL Research Advisory Committee, Dr. B. Wald and Dr. J. Davis.

REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655, 1971.
[2] G. S. Kang and S. S. Everett, "Improvement of the narrowband linear predictive coder, Part 1-Analysis improvements," NRL Rep. 8645, Dec. 1982.
[3] G. S. Kang and L. J. Fransen, "Second report of the multirate processor (MRP) for digital voice communications," NRL Rep. 8614, Sept. 1982.
[4] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A mixed-source model for speech compression and synthesis," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1577-1581, Dec. 1978.
[5] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 68-72, Mar. 1968.
[6] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel format synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 298-305, June 1973.
[7] M. R. Sambur, A. E. Rosenberg, L. R. Rabiner, and C. A. McGonegal, "On reducing the buzz in LPC synthesis," *J. Acoust. Soc. Amer.*, vol. 63, pp. 918-924, Mar. 1978.
[8] B. S. Atal and N. David, "On synthesizing natural-sounding speech by linear prediction," in *Conf. Rec., 1979 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979, pp. 44-47.
[9] L. R. Rabiner and R. E. Crochiere, "On the design of all-pass signals with peak amplitude constraints," *Bell Syst. Tech. J.*, vol. 55, no. 4, pp. 395-407, 1975.
[10] M. R. Schroeder, "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 85-89, Jan. 1970.
[11] D. C. Coulter, C. L. Ludlow, and F. H. Gentges, "Limits of frequency perturbation in normal phonation," *J. Acoust. Soc. Amer.*, to be published.

George S. Kang (M'60) received the M.S. degree in electrical engineering from the University of Wisconsin, Madison, in 1960.

From 1960 to 1971 he was on the Research Staff at General Dynamics/Electronics and Bunker-Ramo Corporations. He joined the Naval Research Laboratory, Washington, DC, in 1971, where he has designed several experimental voice processors at various data rates. Some of these processors are currently deployed by the U.S. Navy.

Stephanie S. Everett was born in Cooperstown, NY, on October 4, 1957. She received the B.A. degree in linguistics from Cornell University, Ithaca, NY, in 1979.

Since that time, she has been with the Voice Systems Section at the U.S. Naval Research Laboratory, Washington, DC. Her primary interests include acoustic phonetics and automatic speech recognition.

Ms. Everett is a member of the International Society of Phonetic Sciences and the Linguistic Society of America.