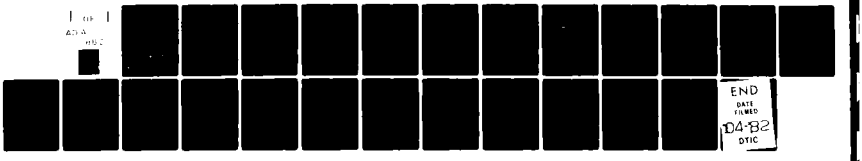
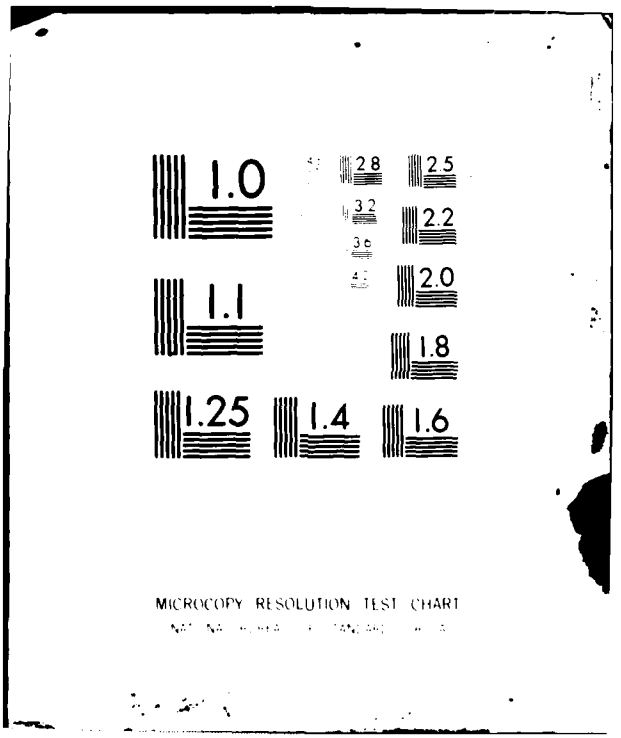
AD-A111 862 NAVAL RESEARCH LAB WASHINGTON DC F/G 9/4
ON PITCH-SYNCHRONOUS ABRIDGMENT OF THE LINEAR PREDICTION RESIDU--ETC(U)
MAR 82 N D SMITH
UNCLASSIFIED NRL-8570 NL



END
DATE FILMED
04-82
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART

NRL Report 8570

On Pitch-Synchronous Abridgment of the Linear Prediction Residual

N. D. SMITH

Communications Systems Engineering
Information Technology Division

March 12, 1982

NAVAL RESEARCH LABORATORY
Washington, D.C.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8570	2. GOVT ACCESSION NO. AD-A111	3. RECIPIENT'S CATALOG NUMBER 862
4. TITLE (and Subtitle) ON PITCH-SYNCHRONOUS ABRIDGMENT OF THE LINEAR PREDICTION RESIDUAL		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem. 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) N. D. Smith		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 33401N, X0734CC, 75-0114-0-1
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Electronic Systems Command Washington, DC 20360		12. REPORT DATE March 12, 1982 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear predictive coding (LPC)
Residual-excited linear prediction (RELP)
Residual processing
Speech compression
Multirate processors
Medium-band speech coding

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We developed an approach to abridging the short-term LPC residual and used the abridged information to generate speech at 9.6 to 16 kilobits per second (kb/s). This is an alternative to the baseband residual-excited coder now being used in NRL's Multirate Processor (MRP). The abridgment logic requires less than one multiply per input sample, is independent of voiced/unvoiced decisions, restricts the influence of errors to a single frame, strongly resists brief input anomalies, and provides *anchor periods* of excellent speech reproduction once per frame. This approach appears promising for practical application in real-time systems where simplicity, stability, and quality are important. It scores somewhat lower overall on the Diagnostic Rhyme Test than the more complicated MRP approach.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS

ON PITCH-SYNCHRONOUS ABRIDGMENT OF THE LINEAR PREDICTION RESIDUAL*

INTRODUCTION

We have developed an approach to abridging the short-term LPC residual* and used the abridged information to generate speech at 9.6 to 16 kilobits per second (kb/s). The abridgment logic requires very little computation — an average of less than one multiply per speech input sample. It offers a simple way to add a high-rate, improved quality capability to existing 2.4 kb/s LPC voice communication systems. Our simulated 16 kb/s RELP† transmission system using abridgment generates very good quality speech output. In addition, our abridgment approach is quite stable; it is independent of voiced/unvoiced decisions, restricts the influence of errors to a single frame, strongly resists brief input anomalies, and provides *anchor periods* of excellent speech reproduction once per frame. Our approach appears promising for practical application in real-time systems where simplicity, stability and quality are important.

We designed this medium-rate system as an alternative to the baseband residual-excited coder now used in the Naval Research Laboratory (NRL) Multirate Processor (MRP). The MRP is a design option for upgraded DOD digital secure voice communications, where noisy acoustic input is common. Under such conditions, voicing detectors tend to be very fallible. For this reason, the baseband residual-excited coder in the 16 kb/s MRP system was specifically designed to be independent of voicing decisions. Our emphasis on resistance to voicing errors and other anomalies also derives from the practical objectives of the MRP program. (See Appendix A for background and description of the MRP).

Our approach scores somewhat lower overall on the Diagnostic Rhyme Test (DRT) [2] than the MRP baseband approach at 16 kb/s‡. The MRP coder scored 93.9 on the DRT, while our approach scored 87.4. Listening tests on longer utterances suggest that the quality of speech output from our approach is a bit closer to that from the MRP system for continuous speech than for isolated DRT words. For some applications, the computational efficiency of our approach could outweigh the quality advantage of the MRP. We have no experimental results to indicate which of the systems is most robust.

---

*Linear prediction of speech ("linear predictive coding (LPC)") models speech samples as weighted sums of previous samples; weights are calculated by minimizing the mean-squared error over small (e.g., 22.5 milliseconds (ms)) intervals. (For further details, see Markel and Gray [1] Chapter 1, Section 4, pp. 10-16.) Markel and Gray define the *prediction error* on page 10 in the same way we define the residual (r_n):

$$r_n = s_n - \sum_{k=1}^{10} a_k \cdot s_{n-k} ,$$

where s_n is the speech at time n.

†"RELP" means residual-excited linear prediction. The excitation for the synthesis filter is derived from the residual (defined in the previous footnote). RELP approaches are intuitively appealing because using the (fully resolved) residual as an excitation would guarantee "perfect" reproduction of the speech input. The method of deriving the excitation varies; our technique is one possibility.

‡The DRT is a standardized intelligibility test in which the listener is asked to distinguish between rhyming words whose initial consonants differ by a single phonemic attribute (e.g., 'foal' vs 'vole', or 'rest' vs 'nest'). DRT scores are not a measure of overall speech quality or of listener preferences.

Manuscript submitted 7 December 1981.

The following description of our abridgment approach assumes the residual processing is part of a RELP system using a single (short-term) prediction filter. We choose a pitch-period-length segment of the residual once per three or four periods, and transmit a quantized version of the segment, along with prediction filter coefficients and other parameters once per analysis frame. At the receiver, we assemble the excitation signal by substituting the transmitted residual segment for its missing neighbors, as shown in Fig. 1. We use this excitation signal to drive the speech synthesis filter and to generate speech.

Ignoring quantization for now, we know the excitation exactly matches the residual one-third to one-quarter of the time.* The rest of the time, in quasi-periodic regions, our selection criteria ensure that the form of the excitation resembles the residual. As shown in Fig. 2, the excitation function approximates the quasi-periodic portions of the residual by making them locally periodic (for the duration of an analysis frame). Sample-by-sample differences between excitation and residual waveforms can be large, due to phase shifts in the glottal excitation, sampling/pitch frequency mismatches, and natural variation. Nonetheless, as Fig. 3 illustrates, the output speech waveform is smooth, with few visible distortions. In sustained vowel regions, and other near periodic portions of the waveform, the small phase and frequency shifts are not audible. The differences between speech input and output can be seen mainly in the frequency domain, and can be heard primarily in sharp onsets and rapid transitions. (This is one reason why we should expect our approach to score somewhat low on the DRT.) Spectrograms of speech input and output in Fig. 4 show some of the frequency-domain differences. (We discuss the relationship of input and output at length in the Performance section of this report.)

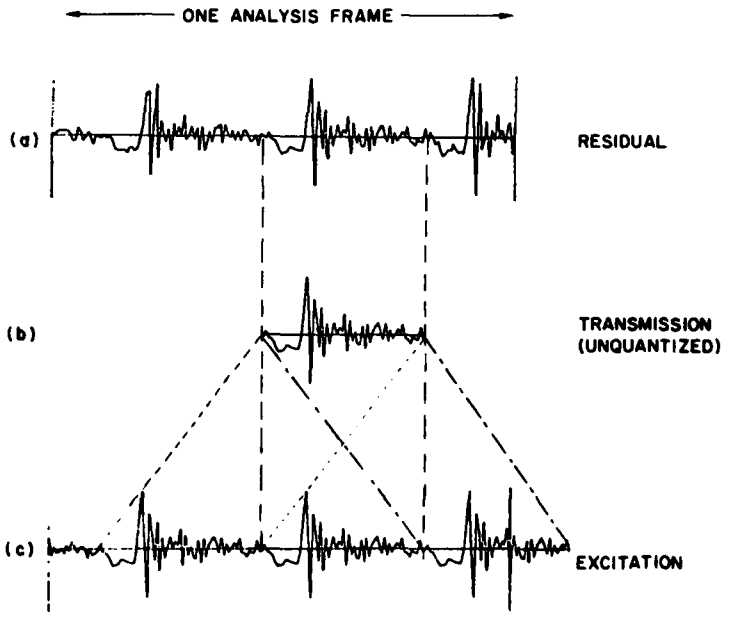


Fig. 1 — Abridging the residual and assembling the excitation. In (b), we choose one pitch-period length segment of the residual from three in the analysis frame (not necessarily the middle one). Transmission (b) excludes all but the chosen segment. We assemble the excitation signal (c) by substituting the transmitted segment for the segments we excluded from the transmission.

*We discuss reduction ratios further in the Refinements section of this report.

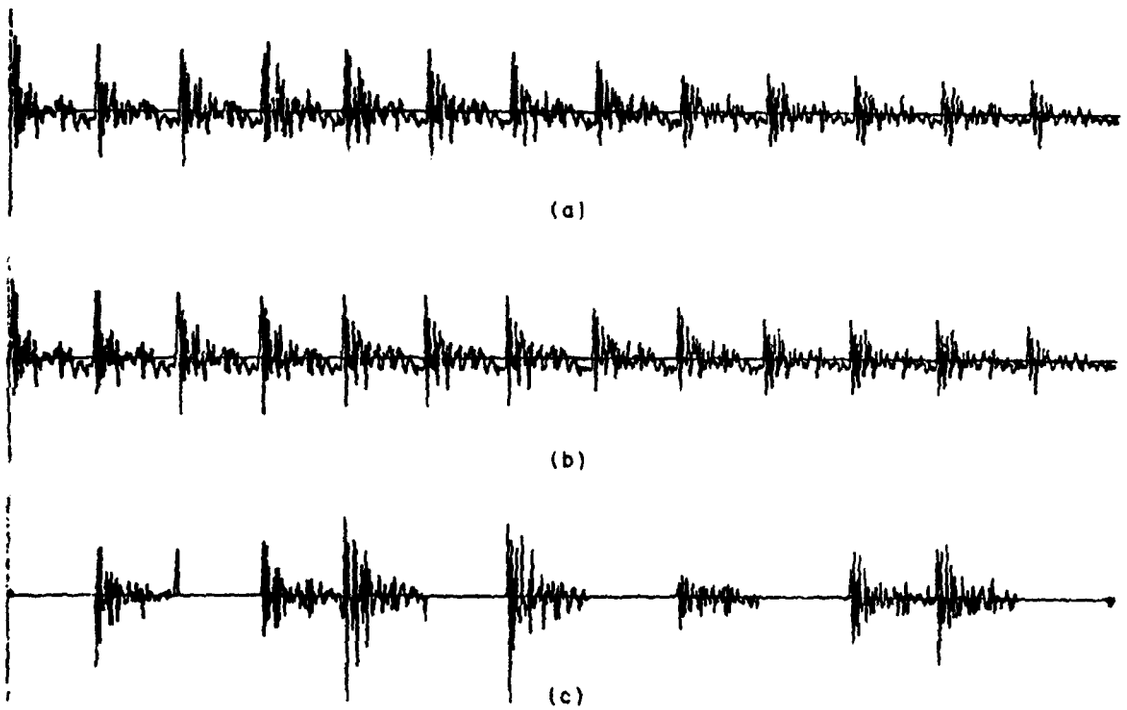


Fig. 2 — Comparison of the residual waveform (a), the excitation waveform (b), and the sample-by-sample difference of the residual and the excitation (c)

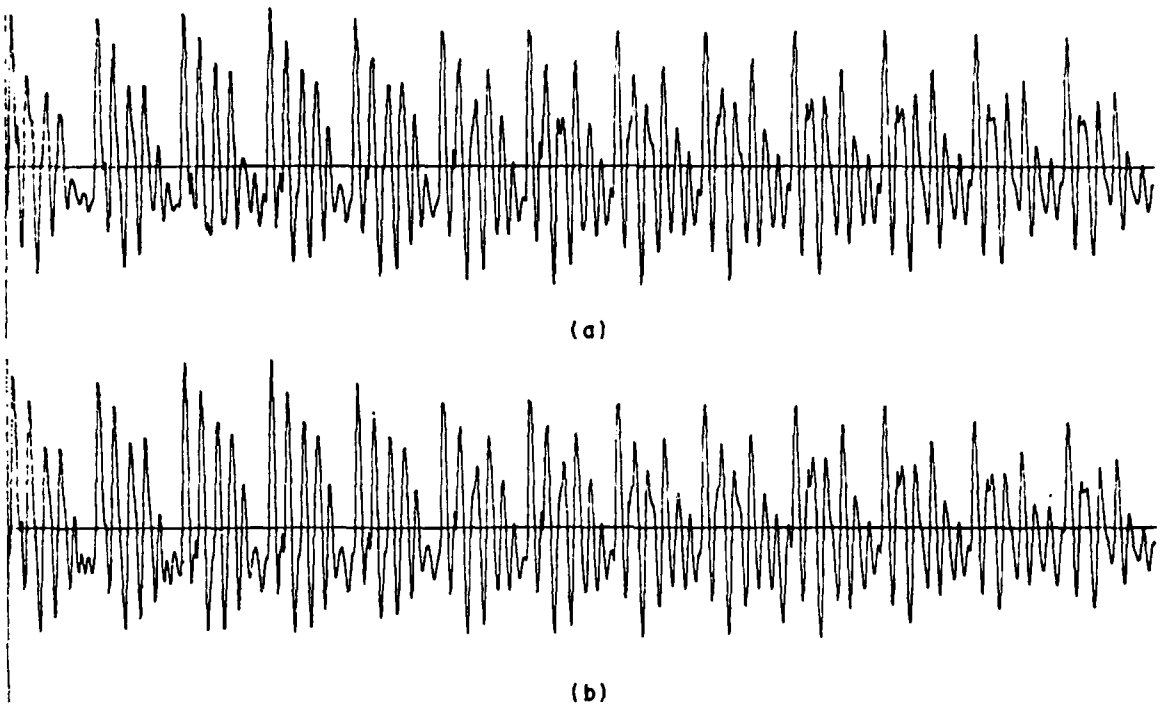


Fig. 3 — Comparison of input speech waveform (a), and output speech waveform (b) for portion of vowel in "dogs," from "Cats and dogs each hate the other." (Male speaker)

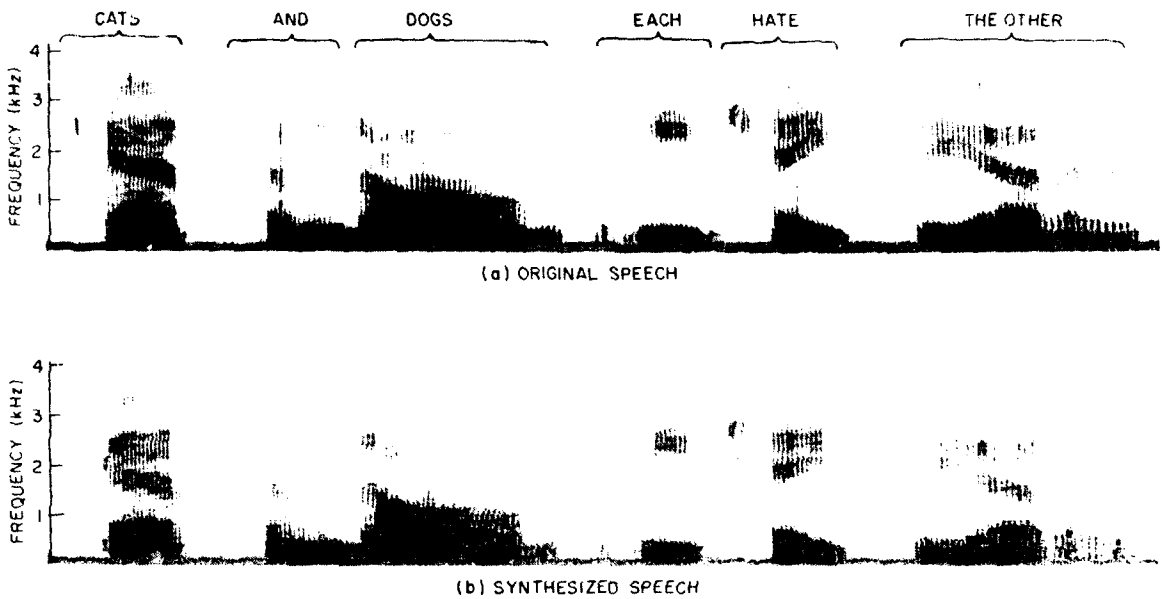


Fig. 4 — Spectrograms of original speech and synthesized speech output

The quality of the output speech from the unquantized system is the upper bound for the potential of our approach. The DRT score for the speech output from an unquantized excitation is 89.7 — surprisingly close to the score for the quantized excitation.

Our abridgment approach is related to previous data rate compression proposals, such as David and McDonald's (1956) [3] implementation of Dudley's (1938) [4] pitch-synchronous speech abridgment approach, Malah's (1979) [5] time domain/harmonic scaling of speech, Maitra's *improved LPC* model (1980) [6], and Abzug's (1981) 9.6 kb/s RELP system [7]. The novel features of our approach mitigate some problems reported earlier, such as generation of subharmonic distortion, sensitivity to voicing errors, and propagation of errors.*

RELATIONSHIPS WITH PREVIOUS SYSTEMS

The earliest precedent we know of is Dudley's analog speech processing technique [4] tested by David and McDonald [3], in which (say) every third period of the speech was retained and substituted for discarded parts. The authors reported "*minor distortion, which appears as submultiples of the pitch harmonics,*" as the major problem with the technique. We do not observe this problem in our results, even when we abridge the speech directly (rather than the residual). Our approach may not generate audible subharmonic distortion because the distance, measured in pitch periods, between the segments we transmit varies with time in a nonconstant (random?) way.

Malah described a related approach [4], in which he compressed the speech waveform by averaging neighboring pitch cycles. Our approach differs from Malah's, not only because it operates on the residual waveform (rather than the speech); but also because:

- No averaging is performed on the transmitted segment.

The absence of averaging helps to eliminate the propagation of errors and reduces the overall distortion. We discuss this important design decision at length in the Why Not Average? section of this report.

Superficially, our approach most closely resembles one described by Abzug [7], (based partially on Maitra [6]). Our approach differs from that RELP system in several significants ways.

- Voiced and unvoiced regions do not require separate processing.

- The selection criteria for the segment to be transmitted emphasize the similarity of the candidate segment to other segments in its frame, rather than its correlation with the past frame excitation.

- The updating of synthesis filter weights occurs at the onset of new residual information in the excitation, rather than at frame boundaries.

- No correlation information (other than that implicit in the pitch estimate) is required to generate the excitation signal.

---

*The two types of errors which our approach handles particularly well are its own selection errors and acoustic input errors or anomalies. By selection errors, we mean those cases in which the selection algorithm chooses a portion of the residual with some undesirable characteristic (e.g., anomalous spike). (Abzug [6] reports problems with this type of error.) By acoustic input errors or anomalies, we mean those (we hope) rare instances of transient non-voice inputs of high power and brief duration.

- The lengths of transmitted segments vary.

- The excitation signal we generate comprises entire periods of the residual spliced together without intervening zeros.

SYSTEM DESCRIPTION

Overview

We perform an LPC analysis and synthesis of the input speech, using our abridged residual information to create the excitation function. The input speech is sampled at 8 kHz. We calculate root-mean-square (rms) values for each half-frame of input speech (90 samples). The rms values are used for amplitude calibration of the speech synthesizer output. We next count the number of zero crossings in each half-frame of the speech. (These counts are used by the pitch smoother. They are unnecessary otherwise.) We then preemphasize the input speech with a factor of 0.5. We calculate 10 reflection coefficients using the covariance method, quantize them, and convert them to predictor coefficients. We can perform an optional second pass, as recommended by Atal and Schroeder [8], to obtain coefficients for our noise-shaping filter.

We inverse filter the preemphasized speech samples to generate the residual signal, and estimate the pitch period by peak-picking the autocorrelation function of the residual at lags 20 to 160 samples; (i.e., pitch frequencies between 400 Hz and 50 Hz). We also check the correlation at three harmonics of the detected pitch frequency, and change the estimate if the correlation at any shorter lag is at least 80% of the correlation at the original lag. (See Appendix B.) When appropriate, we check the pitch estimate further with a timewise pitch smoothing technique. (We describe this technique in the Pitch Estimation section of this report.)

We then choose, quantize, code, and transmit a *typical* period of the residual, of length equal to one pitch period. We transmit the quantized reflection coefficients, residual values, pitch, rms, and the next frame starting point. (See definition in the following section of this report and bit allocation scheme in Fig. 5.)

No. Bits	*Information*
1	synchronization
7	pitch/voicing decision
5	root-mean-square (gain)
41	10 reflection coefficients
54	LPC parameter set (2.4 kb/s)
6	residual maximum value
6	residual minimum value
6	$IS_{(n+1)}$ (next frame starting point)
280	residual segment samples
5	half-frame rms
3	synchronization
360	TOTAL (16 kb/s)

Fig. 5 — Bit allocation for 16 kb/s RELP system using embedded 2.4 kb/s LPC bit stream (54 bits per frame)

At the receiver, we verify the pitch value received, using the current and next frame starting points. We splice the transmitted segment to the previous excitation and to itself until we reach the next starting point, and use the resultant signal to excite the synthesis filter. We update the coefficients of the synthesis filter at the onset of new residual information in the excitation. (See details in the following section of this report and Fig. 6.) The speech output is deemphasized and smoothed, and half-frames are gain-adjusted using the transmitted rms values, before d-to-a conversion. The residual and excitation processing are the novel features of our system. We will describe them in detail.

SELECTION OF RESIDUAL SEGMENTS

We first describe our basic approach for abridging the residual. We discuss some possible refinements in the next section.

Defining Terms

Our abridging and splicing operations require three numbers per frame: pitch (IP), number of repeats (NR), and starting point (IS). NR and IS are determined from IP and the framelength, 180 samples. For the first frame,

$$IS_1 = 0 ,$$

and NR is the smallest integer such that

$$(NR_1 \cdot IP_1) + IS_1 \geq 180 .$$

For the following frames,

$$IS_n = (NR_{n-1} \cdot IP_{n-1}) + IS_{n-1} - 180 ,$$

and NR_n is the smallest integer such that

$$(NR_n \cdot IP_n) + IS_n \geq 180 .$$

Identifying Candidates

The candidate segments for each frame are determined by the pitch period and the starting point. We designate as candidates the NR-1 pitch-length segments beginning with the sample after the starting point. In addition, if there are more than IP/2 samples left over after the last full candidate segment, we create an additional candidate. We obtain enough samples to complete the NRth candidate from the end of the previous segment. As illustrated in Fig. 6, if the end of the last repeat in frame 1 is exactly at the frame boundary (IS = 0), and the pitch value in frame 2 is 64, samples 1-64 and 65-128 will be the first two candidates. The third candidate will be samples 129-180, followed by samples 117-128. The first candidate in frame 3 will begin at sample 13. To illustrate a case in which we have only NR-1 candidates, suppose the pitch in frame 2 is 41. The candidates will be 1-41, 42-82, 83-123, and 124-164. Because there are only 16 samples remaining — less than

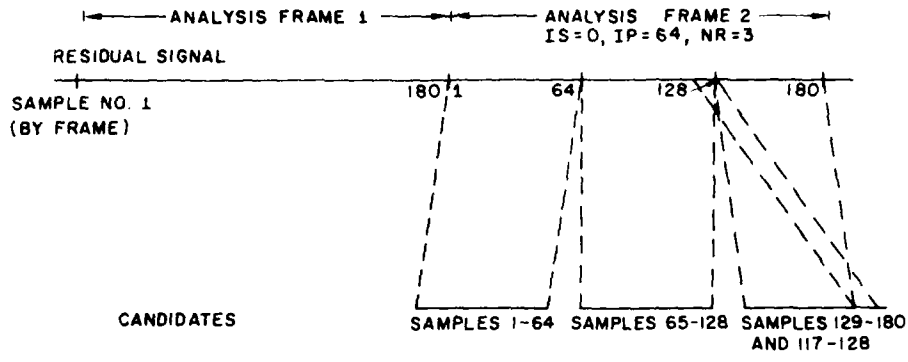


Fig. 6 — Identifying candidates. (Refer to text for definitions of IS, IP, and NR.) First two candidates are taken directly from residual signal, beginning at IS + 1. Third candidate is formed by wrapping around to previous period to pick up 12 samples. (We do not wrap around if the number of samples needed to make up a full candidate is greater than IP/2.)

half a pitch period — we will not construct a fifth candidate. We will however, repeat the chosen segment five times in the generation of the excitation function, so the first candidate of frame 3 will begin at sample 26.

Choosing Between Candidates

The objective of our selection process is to find the candidate which can best substitute for the other segments in the frame. Our highest priority for a substitute is that it match the high-amplitude portions of the waveform, because any errors in these portions sound louder. Candidates with small extreme values cannot match their larger neighbors' extremes very well. Furthermore, we can use our speech outlier* smoother to compensate for effects of high-amplitude portions of the excitation. Accordingly, our selection process always chooses the segment containing the absolute maximum value for the frame (the *largest candidate*), unless we decide the extreme value is an artifact of some sort. We consider the extreme value of an artifact if the candidate segment is a *different shape* from the others in the frame.

Deciding, by eye, if segments are nearly the same shape is easy. Our algorithm emulates this human visual process with two simple rules. First, the ratio of the segment maximum to the segment minimum should be close to the corresponding ratio for the other segments. (We consider this ratio a crude measure of *symmetry*.) Second, the distance in number of samples, between the segment maximum and the segment minimum should be near the distance in the other segments.

We calculate symmetry and distance statistics for all candidates in the frame and rank them according to size. We then examine the rankings for the two largest candidates. We assign two demerits to a candidate if it has either the largest or the smallest symmetry value. We add one or two more demerits if the symmetry (or its reciprocal) exceeds one or both of our threshold values: 2.0 and 3.0 (empirically determined). We assign one demerit if the candidate has either the largest or the smallest distance statistic. We select the candidate with fewest demerits (or the largest candidate, if there is a tie). Note that when we have only two candidates, each candidate gets three demerits based on rankings, so we automatically choose the largest candidate unless it is highly unsymmetric.

*An outlier is a segment whose extreme points are both inside or both outside of both its neighbors' extreme points.

Choosing the largest candidate works well most of the time, and the demerit system eliminates the few anomalies we do observe. The sorting and ranking process is very fast because we rarely have more than five candidates.

Why Not Average?

One reason we prefer not to average waveforms from adjacent pitch periods is that corresponding samples (e.g., the third sample following the beginning of each period) often come from slightly different points in the (continuous) pitch cycle. A slight phase difference between pitch cycles or a pitch frequency which is not an integer fraction of the sampling rate, can shift the sampling function relative to a single period of the waveform. As a result, even in a relatively stationary region, averaging across pitch periods can broaden and flatten peaks (and valleys), and can smooth out high-frequency features. The consequences are likely to be most serious for high-pitched speakers because the waveform completes more of its cycle in the same amount of time.

A second reason we avoid averaging is that in transitional regions, we may find adjacent periods quite dissimilar. Substituting either period for its neighbor will create significant errors in the other. Averaging the two, on the other hand, usually creates significant errors in both, and sounds worse. A third reason we do not use averaging is that it provides a mechanism for spreading waveform anomalies across several periods.

Listening suggests that the synthesized speech becomes crisper as a result of not averaging. The sharp, high-energy peaks of the chosen segment generate excellent dynamic range in the speech output, and for one period per frame, the excitation *is* the residual. We do not gain-adjust the excitation, either, because the adjustment would rob us of the one-period-per-frame match of the excitation to the residual.

An intuitive argument for averaging can be based on the *conventional wisdom* which states that transient effects are more audible than persistent effects of comparable severity. The one-period-per-frame match of the excitation and the residual results in an exceptionally good match of speech output to speech input for that period and a lower quality fit in the remaining periods in the frame. One can argue that we thereby create an undesirable variation in the quality of the speech output. Several colleagues agreed that such a variable-quality fit could sound worse than a more constant fit (even) of slightly lower quality [9]. Further perceptual testing would be required to determine if such an effect exists. But based on informal listening, we believe that the fit in our *best* period may not be dramatically better than in neighboring periods from a perceptual standpoint. A slight phase shift or pitch frequency change lowers our objective measures of goodness of fit greatly, but may not make the speech sound much worse. We suspect the differences are subtle and nondistracting most of the time. The perceptual effect of the quality variation also may be minimized because the frequency of the better periods is not constant.

QUANTIZATION AND USE OF THE ABRIDGED RESIDUAL

Quantization

We quantize the maximum and minimum values for the chosen segment, and normalize the segment to the range (−1.0, +1.0). To quantize the normalized residual samples, we have tried both fixed and variable (based on segment length) bit assignments. Based on its better performance, we use a quantizer with a variable bit assignment. We also use Atal and Schroeder's proposed noise shaping filter [8], which provides a very noticeable perceptual improvement. Given the narrow gap

between our 16 kb/s quality and our upper bound (unquantized) results, we would like to investigate quantization to 9.6 kb/s. In addition to the time-domain schemes, we tried quantizing the Fourier components of harmonic frequencies, and performed an inverse Fourier transform at the receiver. The results did not seem to justify the added complexity. We would like to experiment with a center-clipper and a finer quantizer for the largest residual values, but have not as yet. We believe that some further research on quantization could be fruitful despite the fact that the DRT scores indicate a small proportion of the total degradation is due to quantization; listening suggests that the quantization effect is more noticeable for longer utterances.

Assembling the Excitation

At the receiver, we splice together strings of repeated segments to create an excitation signal. We calculate the number of repeats, using the received values for IP and IS. This is the only logic required unless the pitch value is lost in transmission. In that event, we can use the past frame pitch value and the past and present frame IS values to recalculate the pitch value. We divide the number of samples between past and current starting points by the previous frame's pitch value and round to the nearest integer to get the new NR. Then we divide the distance between starting points by the new NR to get the new IP. Good error protection for the pitch value should make this latter process unnecessary most of the time.

Synthesizing the Speech

We filter the excitation signal using the short-term synthesis filter defined by our ten predictor coefficients. We update the coefficients at each starting point (IS), to prevent the old residual values from over- or under-driving the new filter at frame boundaries. The old excitation and the new filter do not match whenever the residual peaks change scale at frame boundaries (especially in transitional regions, and when coefficients change significantly).

We deemphasize the output speech and then perform some postprocessing. (The deemphasis complements the transmitter preemphasis.) To smooth the output speech before gain adjusting, we check for apparent outlier segments within frames and adjust them, when necessary.* If we find an outlier segment, we calculate the factors for interpolating its extreme points between its neighbors' extremes. We use the interpolating factor for the segment maximum to adjust all positive samples of the segment, and the factor for the minimum to adjust all negative samples. Finally, we adjust the output amplitude of each half-frame to match the original speech amplitude, and then convert to an analog signal.

REFINEMENTS

Frame Splitting

Choosing one segment per frame, we reduce the number of residual samples to be transmitted by 2:1 up to 6:1 per frame. We would like to narrow that range because it makes quantization simpler and because listening suggests that a 3:1 or 4:1 reduction ratio is preferable to higher ratios. For high-pitched speakers (above 220 Hz), we modify our candidate identification procedure by splitting the residual frame (NOT the analysis frame) roughly in half and choose two segments

*Smoothing is necessary when the equal-power periods in the excitation generate uneven speech output periods. This uneveness occurs because we do not scale the excitation periods individually to match the corresponding residual periods.

for transmission. For instance, if NR = 6, we choose one segment from the first three candidates and one from the fourth through sixth. This strategy gives better quality output than simply increasing the number of bits per sample in the quantized residual. There is some risk involved in frame splitting, since we can create more waveform discontinuities. We have not, however, identified any audible effects of such discontinuities. We have not attempted to raise the reduction ratio for very low-pitched speakers (e.g., from 2:1 to 4:1), either. To do so probably would require altering the frame rate, either implicitly — by discarding every other set of coefficients — or explicitly — by doubling the number of samples in the covariance calculation. We doubt that the benefits of raising the ratio would justify the costs of altering the frame rate adaptively.

Truncating Segments

Because frame boundaries can fall anywhere in a pitch period, the phase of the transmitted segment is unconstrained (where a phase is arbitrarily referenced to the time of the zero crossing preceding the segment absolute maximum). The high-power portion of the period sometimes falls near the end of the transmitted segment, and can extend well into the next frame of the excitation, after splicing. In this case, we might prefer to use the high-amplitude portion of the current frame's segment. We could move IS closer to the beginning of the frame by truncating the repeated segment in a low-amplitude region. (Truncation would slightly lower the reduction ratio for the given pitch.)

We might want to move IS when there is too much overlap into the next frame, or if there is too much power in the overlapping portion of the segment. Reasonable quantitative definitions of *too much overlap* and *too much power in the overlap* might be: 1) more than half a pitch period of overlap, or 64 samples, whichever is smaller (64 because IS is encoded in 6 bits), and 2) the segment maximum and minimum are in the overlapping portion. We could test for either of these conditions, and then search forward from the beginning of the frame. If we found a long enough region (e.g., 20 percent of a pitch period) in which the amplitude remains below the residual rms, we would truncate near its middle, and change IS.

Adding Reference Segments

As previously stated, our selection criteria are degenerate when there are only two candidates. We have considered, but not tested, the idea of ranking the immediately preceding segment along with the two current frame candidates. For this purpose we would create a new segment from samples straddling the frame boundary. This new segment would not be considered as a candidate, but only used as a reference point in the rankings.

PITCH ESTIMATION

Accurate pitch estimates are critical to the performance of the residual abridgment approach. We use a peak-picking algorithm to estimate the pitch from the autocorrelation function of the residual signal, and check the first three harmonic frequencies to guard against frequency division. (This pitch detector is part of the MRP system. See Appendix B.) In addition to this within-frame precaution, we developed an across-frame pitch smoother to correct occasional errors. Based on the recent past, the pitch smoother establishes an expected range for the pitch of each frame; it forces the pitch estimate for that frame into the range when we are confident that it belongs there.

The range is centered at the running average pitch value for the utterance, and its width is equal to that average. If the pitch estimate is in the range, we update the running average, and proceed. If not, we use the ratio of the running average to the current frame estimate to calculate an

appropriate adjustment factor. We allow five factors: 3.0, 2.0, 1.0, 0.5, and 0.333. We require that the correlation at the new lag exceeds a threshold value (= 0.7 · (maximum for frame)). As an additional precaution, we will not change the estimates in more than three successive frames.

We use a criterion based on voicing-related statistics (e.g., number of zero crossings in half-frames) to decide whether to include the current frame's pitch in the running average. (We could just as well use a power threshold.) We include the pitch in the average only when we are in a strongly voiced frame (or a frame with rms greater than 500). If the smoother changes the pitch estimate, we do not include it in the average.

Our smoother is effective for eliminating the occasional multiplication or division of pitch. To apply the concept more broadly (e.g., to unrestricted conversation) would require additional constraints to prevent smoothing across real pitch (or speaker) transitions. Before developing the smoother logic further, we would determine how often it is required. Depending on the accuracy of the pitch estimation, the benefits of the smoother might not justify the added processing requirements.

PERFORMANCE

What kind of an excitation signal do we create with our approach? The excitation signal captures some characteristics of the residual extremely well: the voiced and unvoiced regions, the distribution of energy within periods, the frames' peak residual energies, the amplitude values at harmonic frequencies, and the regularity of the harmonic structure. (Baseband approaches tend to disrupt the harmonic structure.) The excitation represents transitions fairly well. Parts of transitions are dropped by the abridgment process, but the listener tends to fill in these parts, since the remaining parts indicate the correct trajectory for the transition. Transitions involving stop consonants are not handled particularly well. Depending on the duration of the burst, the selection criteria can erroneously determine it is an anomaly, and avoid it. The detailed DRT scores indicate that this error can degrade the system's performance. We tested an approach to differentiating stop consonants from undesirable artifacts, with some success. (There was no *stop detector* included in the system that we submitted for DRT testing.)

The excitation lacks the residual's intra-(analysis) frame variation. The pitch structure is absolutely regular throughout each string of repeated segments in the excitation. This pitch regularity is passed on to the output speech. Spectral comparisons between the residual and the excitation, and between the original speech and the output speech, reveal small frequency mismatches in some frames. (See Figs. 7 and 8.) The mismatches are due to the regularity of the pitch in the excitation, and to the limited resolution of the pitch detector. They are more frequent for high-pitched speakers because the resolution of the pitch detector is inversely proportional to pitch. Amplitude values at harmonic frequencies generally match well, even when the frequencies themselves do not. Both amplitudes and frequencies match well for many frames for some speakers. (See Fig. 9.)

In addition to a regular pitch, our excitation signal's sample values are the same in neighboring periods except across starting points. Our excitation signal has roughly four times the resolution at each sample point as a one-bit quantized version of the entire residual signal, however. The intra-period variation is thus higher than for a nonabridged residual quantized at the same bit rate.

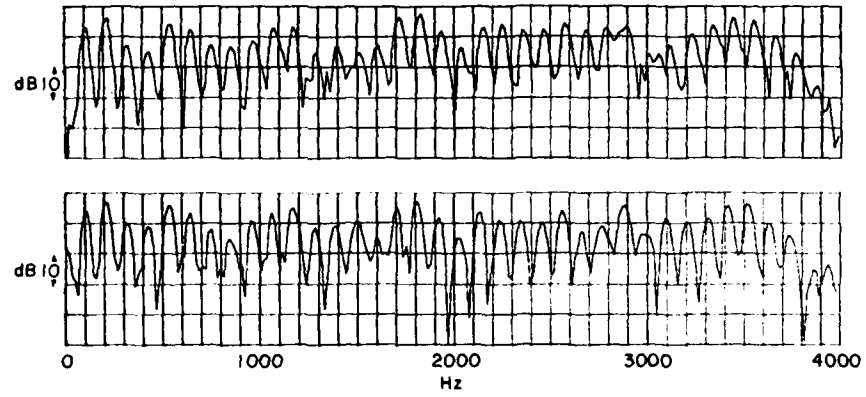


Fig. 7 — Residual spectrum (top) and excitation spectrum (bottom) shows slight frequency mismatch. Mismatch is most evident near 2 kHz. Amplitudes match well at harmonics.

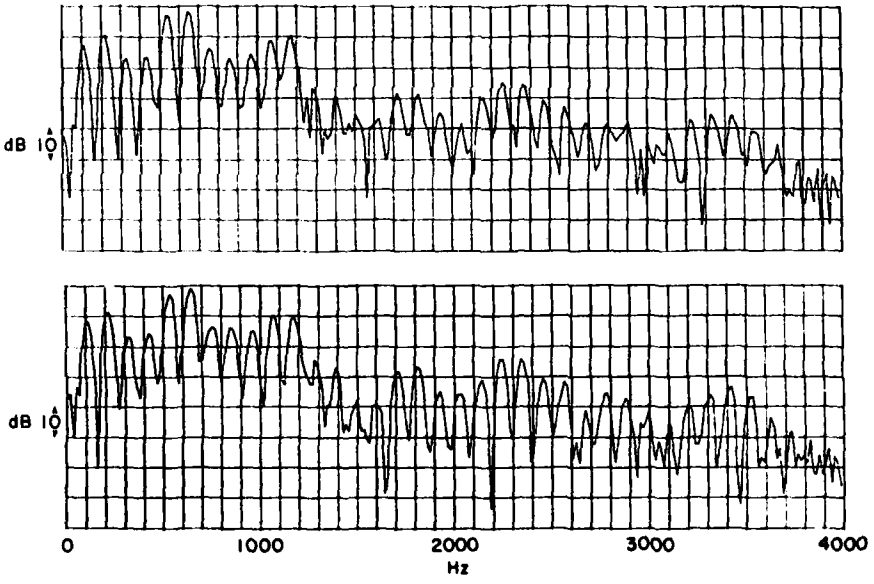


Fig. 8 — Input speech spectrum (top) and output speech spectrum (bottom) shows frequency mismatch. Although amplitudes at null frequencies are sometimes quite different, amplitudes match well at peaks.

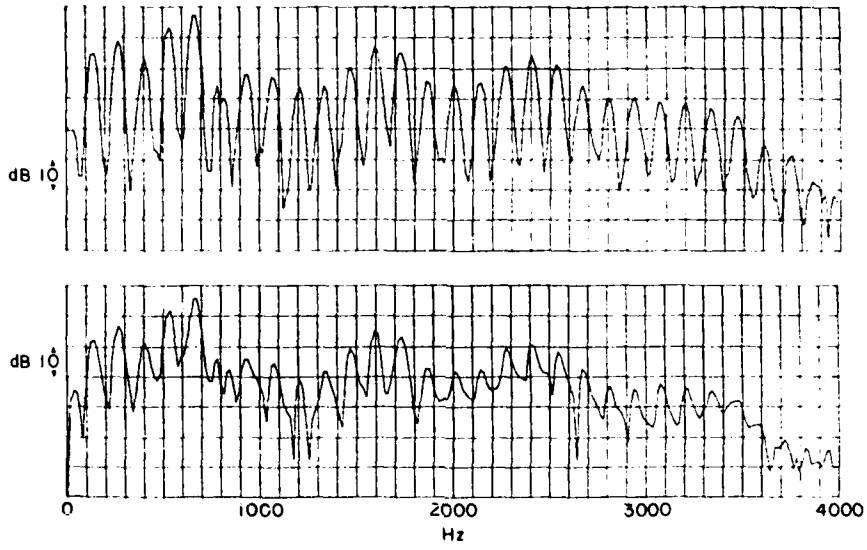


Fig. 9 — Input speech spectrum (top) and output speech spectrum (bottom) shows good match of harmonic frequencies and peak amplitudes

CONCLUSIONS

As in all waveform abridgment approaches, our system is affected by pitch errors. The implications of pitch detection errors are somewhat less than for a pitch-excited LPC system (or Abzug's approach) [7], however, since pitch frequency halving causes no audible problems. Frequency doubling causes the same sort of pitch jumps in our approach as in pitch-excited LPC.

The greatest strengths of our approach are its independence from voicing decisions, its reluctance to introduce distortions or to propagate errors, and its computational simplicity. Our splicing technique frees us from the voicing decision by eliminating the use of intervening zeros between residual segments in voiced regions. The splicing works well in unvoiced regions because repeated segments of white (or nearly white) noise are nearly white (assuming the segments are relatively long).

Our selection criteria strongly resist brief input anomalies and, combined with the absence of averaging and frame-to-frame correlation, virtually eliminate the error propagation [7]. Our approach addresses another potential source of error or distortion not mentioned in previous discussions, i.e., excitation pulses at frame boundaries over- or under-driving the synthesis filter. By updating the filter weights at each starting point, our approach avoids this problem without doing any normalization. We believe that recognizing and solving this problem is important, because it reduces the system's sensitivity to the placement of frame boundaries relative to pitch epochs. We tested our approach using several different frame alignments for each utterance, and heard no differences in the results.

Our selection criteria and splicing technique are computationally simple. They require no correlation calculations and only about ten multiplies per frame. The most complicated portion of the simulated system (after the LPC analysis) is the outlier processor, which can require 90-100 multiplies per frame.

Pitch-synchronous residual abridgment is a computationally simple, resistant approach for adding good quality, high-rate speech to narrowband LPC systems. Our approach seems to strike an appropriate balance between output quality and stable performance, particularly considering its low computational overhead. Our results are incomplete; further testing of pitch-synchronous residual abridgment is required to verify its actual quality and stability. Additional analysis of our results could indicate which features we should try hardest to improve.

RECOMMENDATIONS

Further work on pitch-synchronous abridgment of the residual should focus briefly on finding a better quantization scheme. As mentioned earlier, listening suggests that the quantization effect is more noticeable for longer utterances. The work should concentrate intensely on conducting thorough perceptual tests and on analyzing the results. The analysis of test results should attempt to identify those features of the abridgment approach which cause the largest degradations in speech quality.

Among other things, the search for a better quantization method should examine center-clipping with finer quantization of high-amplitude portions, and perhaps some frequency-domain schemes. The inquiry should also examine the possibility of a 9.6 kb/s implementation. The merits of variable and fixed bit assignments should be compared for each quantization scheme. Comparing variable and fixed bit assignments requires testing the quantizers on speakers with a variety of pitches because the number of samples to quantize changes with pitch. The final analysis of the quantization results should weigh the complexity of the various schemes explicitly, and justify added complexity in terms of improved performance. Our results indicate that the performance benefits of a noise-shaping filter in a loop around the quantizer easily justify the additional complexity introduced by such a loop.

The perceptual testing should establish how well a RELP system using our residual abridgment approach performs under a variety of acoustic input conditions, relative to other 16 kb/s systems. Tests comparing a system using our abridgment approach with other systems should be coordinated with the quantizer search, so that the best possible system is used for the tests. At a minimum, we would expect to use some binary comparison preference tests, some intelligibility tests (both isolated and in-context words), and some multiple comparison tests (perhaps using presentation, recording and analysis techniques like those in Atal and David [10]). Such tests would be performed for a variety of acoustic input conditions and for a broad range of speakers, and speaking situations (e.g., conversation and isolated sentences). Assessments of the relative performances of the systems tested should mention any significant differences in complexity between the systems.

In addition, some attempt should be made to test those properties of the output speech directly attributable to the residual abridgment and excitation splicing processes. Ideally, we would like the testing to show precisely how these properties affect the perceived quality of the output speech under a variety of acoustic input conditions. One approach would be to isolate the effects by comparing our output with contrived output lacking specific properties of our system's output. Developing a method of generating meaningful test materials is a worthwhile area for future work. The analysis of the results from such tests should indicate what properties of the abridged residual and spliced excitation contribute most to the nonquantizer degradation.

Understanding more about the effects of certain properties of our approach should provide the basis for a future work agenda. Limited additional development work should establish the performance potential of pitch-synchronous residual abridgment. The principal task should be to determine the appropriate range of applications for this simple, resistant approach to transmitting good quality speech at high rates.

ACKNOWLEDGMENT

We gratefully acknowledge the contributions of Mr. George S. Kang to this project. His kind criticism, thoughtful suggestions, and keen listening greatly enhanced our results. His enthusiasm and excellence provided both support and inspiration.

REFERENCES

1. J. D. Markel, and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.
2. William D. Voiers, "Diagnostic Evaluation of Speech Intelligibility", in *Speech Intelligibility and Recognition*, M. E. Hawley (ed.), Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1977.
3. E. E. David, Jr., and H. S. McDonald, "A Note on Pitch-Synchronous Processing of Speech," Journal of the Acoustical Society of America, V28, p. 159, 1956.
4. H. W. Dudley, U.S. Patent No. 2,115,803, May 3, 1938.
5. D. Malah, "Time Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 2, April 1979.
6. Sidhartha Maitra, and Charles R. Davis, "Improvements in the Classical Model for Better Speech Quality," Record of the 1980 IEEE ICASSP, p. 23.
7. B. M. Abzug, "Using the Prediction Residual to Improve LPC Synthesis for 9600 BPS Applications," Record of the 1981 IEEE ICASSP, p. 812.
8. B. S. Atal, and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, June 1979.
9. E. E. David, Jr., and M. R. Schroeder, personal communication.
10. B. S. Atal, and N. David, "On Synthesizing Natural-Sounding Speech by Linear Prediction," Record of the 1979 IEEE ICASSP, p. 44.

Appendix A

A DESCRIPTION OF THE NRL MULTIRATE PROCESSOR

The following figures and description of the NRL Multirate Processor (MRP) are excerpts from the original NRL Report 8295 proposing the system [A1]. Our comments and additions are enclosed in curly brackets { }.

MRP CONCEPT

The unique characteristic of the MRP is that the bit stream representing voice encoded at 9.6 kb/s also contains the bit stream representing voice encoded at 2.4 kb/s. Figure A1 depicts the data structure of the MRP for each frame of speech data (22.5 ms, 180 speech samples). The 54 bits represent the set of data required for the generation of 2.4 kb/s speech: one bit for synchronization, 41 bits for synthesis filter weights, and 12 bits for excitation signal. {These 12 bits include the pitch estimate, the gain, and the voicing decision.} The 162 bits represent the supplementary data required for the generation of 9.6 kb/s speech: three bits for synchronization and 159 bits for improved excitation signal.

The data are structured in a building-block form, in which the wideband mode shares the data belonging to the narrowband mode. The use of a single voice-processing principle, linear predictive coding (LPC), makes it possible to form the embedded data structure. Because of this embedded data structure, 9.6 kb/s can be converted to 2.4 kb/s by simply deleting 162 bits. On the other hand 2.4 kb/s can be converted to 9.6 kb/s by inserting 162 bits (the message indicator which tells the receiver that speech is originally encoded at 2.4 kb/s) in each frame. {Thus, there is no analog tandeming (and associated speech degradation) in the operation between a 2.4 kb/s and a 9.6 kb/s system.} The rate conversion can be made anywhere within the link without user intervention while synchronization and encryption are maintained.

DESCRIPTION OF THE MRP ALGORITHM

Based on the preceding discussion the general approach to the MRP algorithm is:

- Ten prediction coefficients are extracted from the speech waveform by the same procedures employed in 2.4 kb/s LPC {covariance method}.

- The same filter is used for the speech synthesis at 2.4 and 9.6 kb/s.

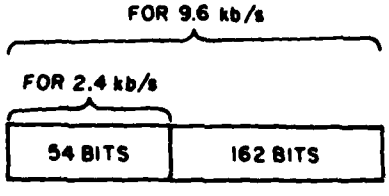


Fig. A1 — MRP data structure of each frame

- The excitation signal for the narrowband mode is either a quasiperiodic broadband signal (if speech is voiced) or random noise (if speech is unvoiced) as is used with the current 2.4 kb/s LPC.

- The excitation signal for the wideband mode is derived from the prediction residual.

- The residual is transformed into a set of amplitude and phase spectrum components and transmitted for the baseband only from {160 to 1160 Hz.}.

- The residual information is quantized on an open-loop basis.

- The phase information is quantized with a finer resolution than is the amplitude information.

- The lower-frequency components are quantized with a finer resolution than are the high-frequency components.

- The upperband excitation signal is generated by frequency-shifting of the baseband signal.

In essence the MRP is a 2.4 kb/s LPC device with an add-on processor for a 9.6 kb/s capability. The add-on processing is nothing more than a means to transmit the residual information. Figure A2 is a block diagram of the MRP. The two blocks that are hatched imply the add-on processing to 2.4 kb/s LPC to produce the MRP. The equipment listed in these two blocks performs the following operations:

- Residual transformation.

- Baseband residual encoding/decoding.

- Excitation signal generation.

{Extending the MRP system to 16 kb/s allows us to transmit twice as many amplitude and phase components, covering the range 160 to 2120 Hz.}

REFERENCE

A1. G. S. Kang, L. J. Fransen, and E. L. Kline, "Multirate Processor (MRP) for Digital Voice Communications", Naval Research Laboratory Report 8295, March 1979.

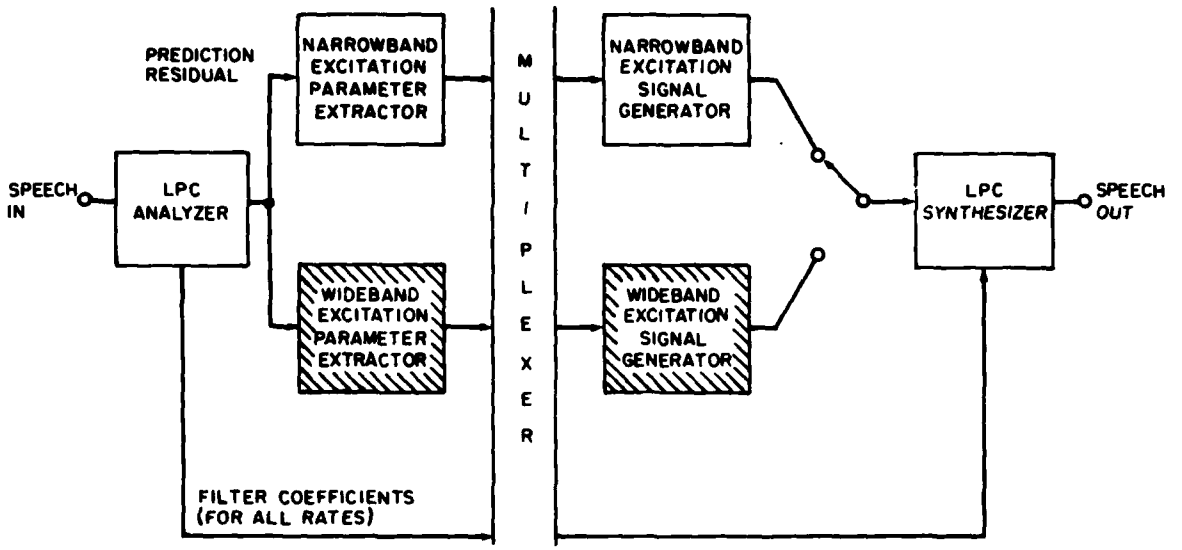


Fig. A2 — MRP voice processor. The hatched blocks are additions to the current 2.4-kb/s LPC device.

Appendix B

A DESCRIPTION OF MRP PITCH DETECTOR

The pitch detector we use was described by Kang (1974) [B1], and implemented as part of the MRP. The following description is from Kang's 1974 paper.

Pitch period may be estimated either directly from the input speech waveform, as in virtually all previous vocoders, or from the prediction residual. Because the prediction residual is relatively less contaminated by the formant structure, the use of the prediction residual is preferred, particularly since it is available from the analysis filter. Despite the inverse filtering or formant-reducing effect of the analysis filter, the prediction residual is not always clean enough for a simple waveform peak-picking process to determine the glottal excitation period. A recommended procedure is to derive a statistical descriptor from the prediction residual that is reliably dependent on the pitch period. One of the simplest descriptors is a regression coefficient between the prediction residual and the time-delayed prediction residual. The usefulness of this type of regression coefficient lies in the fact that it represents a best-fit slope in the least-square error sense of a scatter diagram, in which the prediction residual is on one axis and the time-delayed prediction residual is on the other axis. If the magnitude of the time delay equals the pitch period, the points on the scatter diagram tend to cluster along at 45° line, signifying that the regression coefficient is near unity.

The first order regression analysis assumes a linear dependency between the prediction residual and the time-delayed prediction residual. Thus,

$$\epsilon_t = \theta \epsilon_{t+\tau} + \delta_t \,, \qquad \{B1\}$$

where θ is a constant, ϵ_t is the prediction residual, $\epsilon_{t+\tau}$ is the time-delayed prediction residual, and δ_t is an error whose mean-square value is expressed by

$$S(\vartheta,t) = E[(\epsilon_t - \theta \epsilon_{t+\tau})^2] \,. \qquad \{B2\}$$

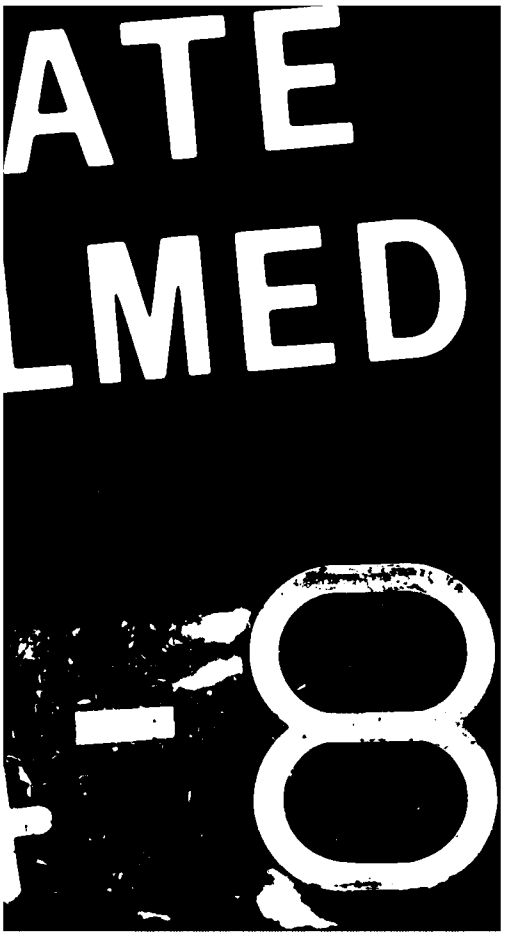
From the theory of least squares, the best estimate of θ is

$$E\{\hat{\theta}\} = \frac{E[\epsilon_t \epsilon_{t+\tau}]}{E[\epsilon_t^2]} \,. \qquad \{B3\}$$

The numerator and the denominator of Eq. {B3} are the autocorrelation coefficients at a delay time τ and at no delay, respectively. In essence, the pitch period is the time separation between the main peak at zero delay and the first side peak.

REFERENCE

B1. George S. Kang, "Application of Linear Prediction Encoding to a Narrowband Voice Digitizer," Naval Research Laboratory Report 7774, October 1974.


ATE
LMED
8