Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/5531--93-7371

Template Based Low Data Rate Speech Encoder

LAWRENCE FRANSEN

Human-Computer Interaction Lab
Information Technology Division

September 30, 1993

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 30, 1993	Interim

4. TITLE AND SUBTITLE

Template Based Low Data Rate Speech Encoder

5. FUNDING NUMBERS

33904N

6. AUTHOR(S)

Lawrence Fransen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/5531--93-7371

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AC NCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

The 2400-b/s linear predictive coder (LPC) is currently being widely deployed to support tactical voice communication over narrowband channels. However, there is a need for lower-data-rate voice encoders for special applications: improved performance in high bit-error conditions, low-probability-of-intercept (LPI) voice communication, and narrowband integrated voice/data systems.

An 800-b/s voice encoding algorithm is presented which is an extension of the 2400-b/s LPC. To construct template tables, speech samples of 420 speakers uttering 8 sentences each were excerpted from the Texas Instrument - Massachusetts Institute of Technology (TIMIT) Acoustic-Phonetic Speech Data Base.

Speech intelligibility of the 800-b/s voice encoding algorithm measured by the diagnostic rhyme test (DRT) is 91.5 for three male speakers. This score compares favorably with the 2400-b/s LPC of a few years ago.

14. SUBJECT TERMS

Speech analysis/synthesis
Line-spectrum frequencies
Vector quantization

15. NUMBER OF PAGES

15

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std 239-18
298-102

Contents

Template Based Low Data Rate Speech Encoder

INTRODUCTION

The 2400-b/s linear predictive coder (LPC) is currently being widely deployed to support tactical voice communication over narrowband channels. However, there is a need for lower-data-rate voice encoders for the following special applications.

Increased tolerance to channel bit errors: The intelligibility of the 2400-b/s LPC degrades rapidly in the presence of transmission bit errors. With 3% random errors, the intelligibility decreases to a level often described as having "poor intelligibility." To increase the tolerance to bit errors, error protection code is added to the 800-b/s speech data for transmission at 2400 b/s.

Voice/Data Integration: Recently, voice/data integration has drawn much attention. The use of the 800-b/s voice encoding algorithm allows integration of voice and data over a single 2400-b/s channel. For example, a visual aid (written text, hand-drawn scribbles, etc.) can be transmitted with voice to enhance communicability.

Voice Multiplexing (Voice/Voice Integration): Currently, a single voice net can be transmitted over a 3-kHz narrowband channel. If the 800-b/s voice processor is used, however, three independent voice nets can be multiplexed and transmitted over a single narrowband channel. This multiplexing capability permits secure conferencing. Current secure conferencing requires a conference director to moderate the traffic flow by designating who can talk. This is not a satisfactory solution to conferencing. With voice multiplexing available, however, it is possible to transmit three individual voices independently over a single channel. As a result, all the participants can hear each other, even if two people accidentally talk at the same time. In addition, voice multiplexing can achieve a more effective utilization of RF assets because one radio can be shared by three independent voice circuits.

We present an 800-b/s voice encoding algorithm which is an extension of the 2400-b/s LPC. In essence, the 800-b/s voice algorithm is a 2400-b/s LPC with modified parameter encoders. Speech intelligibility of the 800-b/s voice encoding algorithm measured by the diagnostic rhyme test (DRT) is 91.5 for three male speakers evaluated by impartial listeners not associated with our R&D effort. This score compares favorably with the 2400-b/s LPC of a few years ago. This paper is an improvement of our recent report (Ref. 1).

TECHNICAL APPROACH

The 800-b/s voice encoder is an extension of the 2400-b/s LPC. In essence, the 800-b/s encoder is the 2400-b/s LPC with an 800-b/s parameter encoder and decoder (Fig. 1). Significant features of the 800-b/s voice encoder are:

(1) *Joint parameter encoding over two consecutive frames:* Two sets of parameters for two frames are encoded as a unit, except for the pitch period. By transmitting two frames of data as a unit, the parameter correlation existing in two adjacent frames can be exploited. For example, a person cannot change speaking volume from a maximum to a minimum over one frame of time (20 milliseconds). Hence such a transition can be eliminated from the coding of amplitude information. A similar argument holds for filter coefficients.

Manuscript approved May 10, 1993.

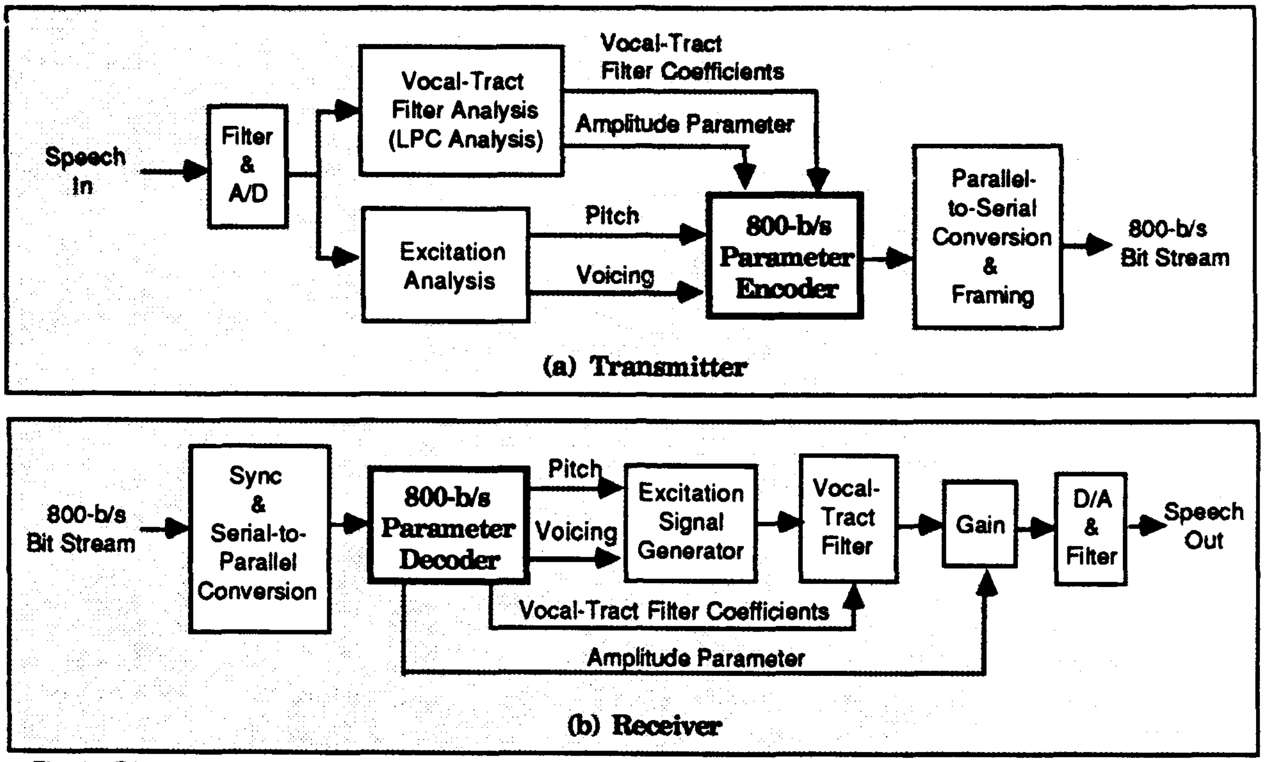


Fig. 1 - Block diagram of 800-b/s voice encoder. The general layout of computational blocks are identical to that of the 2400-b/s LPC. The only blocks unique to the 800-b/s voice encoder are the parameter encoder and parameter decoder identified by heavy-lined blocks. Since the other blocks are well-known, we will not elaborate further on them.

(2) Speech-spectrum-dependent voicing decision:

No separate voicing information is transmitted; rather, the voicing information is implicitly specified by the filter coefficients. We exploit the fact that filter coefficients from voiced speech are substantially different from those from unvoiced speech. Thus, each filter coefficient set has an associated voicing decision.

(3) Reduction of Frame Size:

Frame size is the time interval between parameter updates. In the past, frame size was often determined after considering the number of bits required to encode all the parameters per frame. This is not a good design approach because there is a preferred value for frame size in terms of speech intelligibility for voice processors that use an artificial excitation signal (i.e., pitch-excited vocoders such as the 2400 LPC and the 800-b/s voice encoder). In these voice encoders, rapid speech changes can be reproduced only by rapid filter and amplitude parameter updates. Intelligibility is adversely affected by slow speech onsets. There are many ways to encode speech parameters efficiently, but speech degradation resulting from improper frame size is irreversible.

Some years ago, a study was conducted to investigate the relationship between frame size and speech intelligibility (Ref. 2). According to this study, a marked speech degradation occurs as the frame size increases from 20 to 30 ms. Recently, we also examined the effect of frame size on speech intelligibility as measured by the DRT (Ref. 1). By using a 10-tap LPC without parameter quantization, we obtained DRT scores for three frame sizes: 17.5 ms, 20 ms, and 22.5 ms. As indicated in Fig. 2, a frame of 20 ms is the preferred choice.

Accordingly, we used a frame size of 20 ms in the 800-b/s voice encoder. It is significant that a pitch-excited LPC can achieve a DRT score of 95 with unquantized parameters.

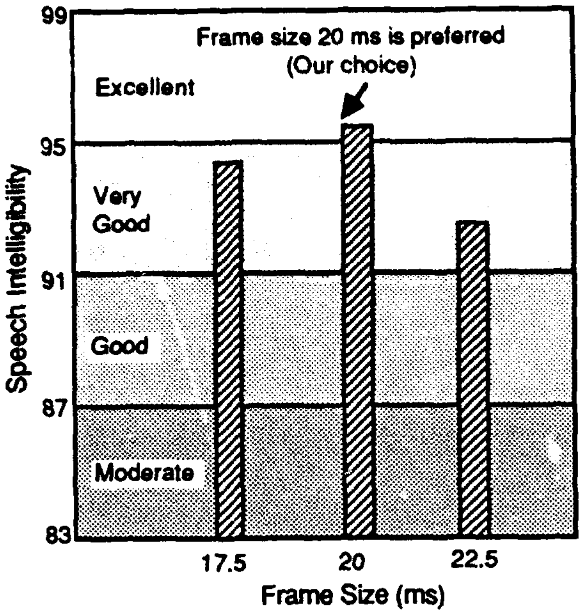


Fig. 2 - Frame size vs. speech intelligibility. This figure shows DRT scores for a 10-tap LPC with three different frame sizes. Most 2400-b/s voice processors have a frame size of 22.5 ms, but the preferred size is 20 ms.

(4) LSPs as Vocal Tract Filter Coefficients

We observed that the intelligibility of an 800-b/s voice encoder improves significantly after LSPs are used as filter parameters. LSPs have been gaining interest because their intrinsic properties permit more efficient encoding than the better-known reflection coefficients:

- *Frequency-selective spectral error:* An error in one member of the LSPs affects the spectrum only near that frequency (i.e., frequency selective). Thus, LSPs can be quantized in accordance with properties of auditory perception (i.e., coarser representation of the higher-frequency components of the speech-spectral envelope).

- *Unequal spectral-error sensitivity:* For a given LSP set, spectral-error sensitivity of each line spectrum can be determined easily (as will be shown). Thus, fewer bits are needed to encode spectrally less sensitive LSPs.

The LPC analysis filter, A(z), that transforms speech samples to residual samples is expressed by

$$A(z) = 1 - \sum_{k=1}^{10} \alpha(k)\, z^{-k} \qquad (1)$$

where z^{-1} is a one-sample delay operator. A(z) may be decomposed to a set of two transfer functions, one having an even symmetry, and the other having an odd symmetry. This can be accomplished by taking a difference and sum between A(z) and its conjugate function

$A^*(z)$ (i.e., the transfer function of the filter whose impulse response is a mirror image of A(z)). Thus,

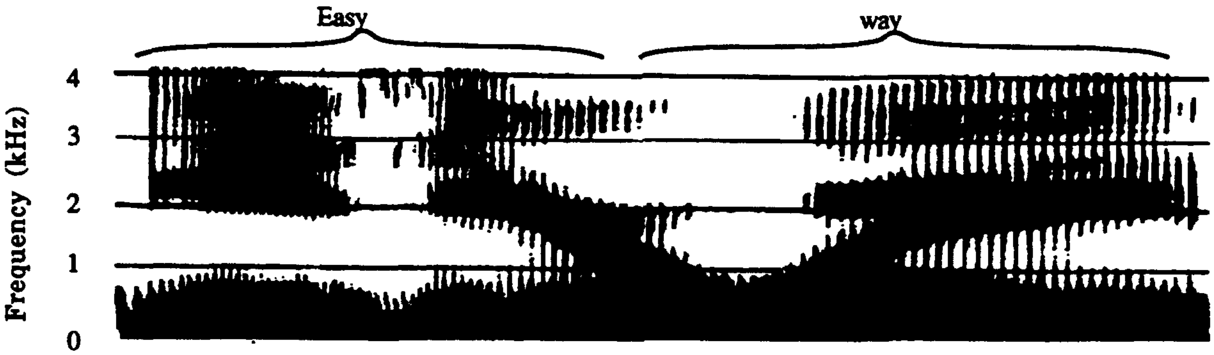
$$P(z) = A(z) + z^{-11} A^*(z) \tag{2}$$

and

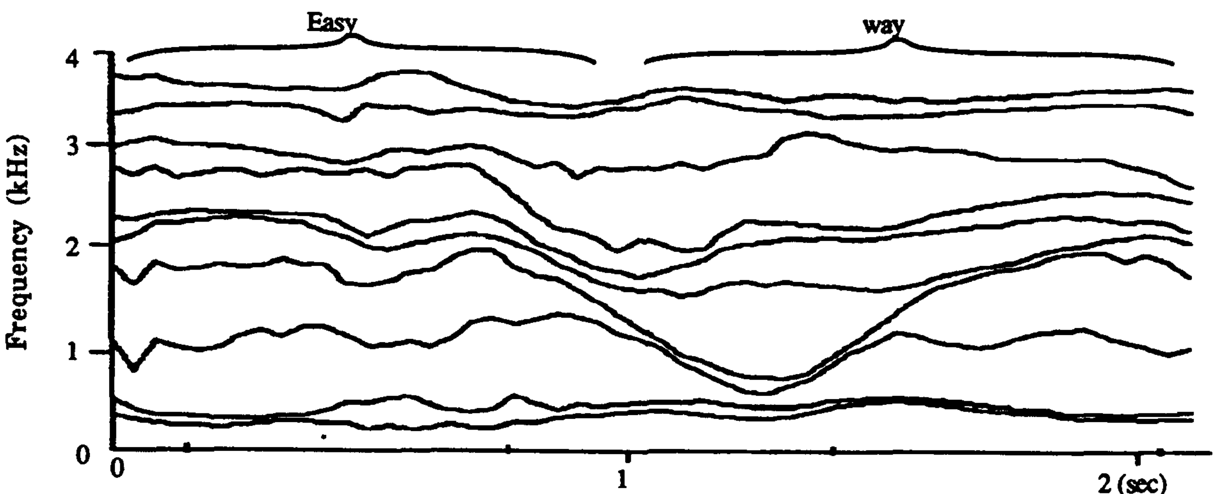
$$Q(z) = A(z) - z^{-11} A^*(z) \tag{3}$$

where $z = EXP(j2\pi ft_s)$ in which f is frequency in Hz and t_s is the sampling-time interval.

The roots of P(z) and Q(z) in Eqs. (2) and (3) are LSPs. LSPs may be computed using Chebyshev polynomials [3]. We obtain LSPs from null frequencies of P(z) and Q(z) computed at a 20-Hz interval. A parabolic approximation using three consecutive frequencies around each null frequency produces LSPs having an accuracy of a few Hz (Ref. 1). Figure 3 shows typical LSP trajectories.



(a) Spectrogram



(b) LSP Trajectories

Fig. 3 - Comparison of spectrogram and LSP trajectories derived from the same speech. As noted, line-spectrum frequencies are close together where formant frequencies are located.

5) Bit Assignment

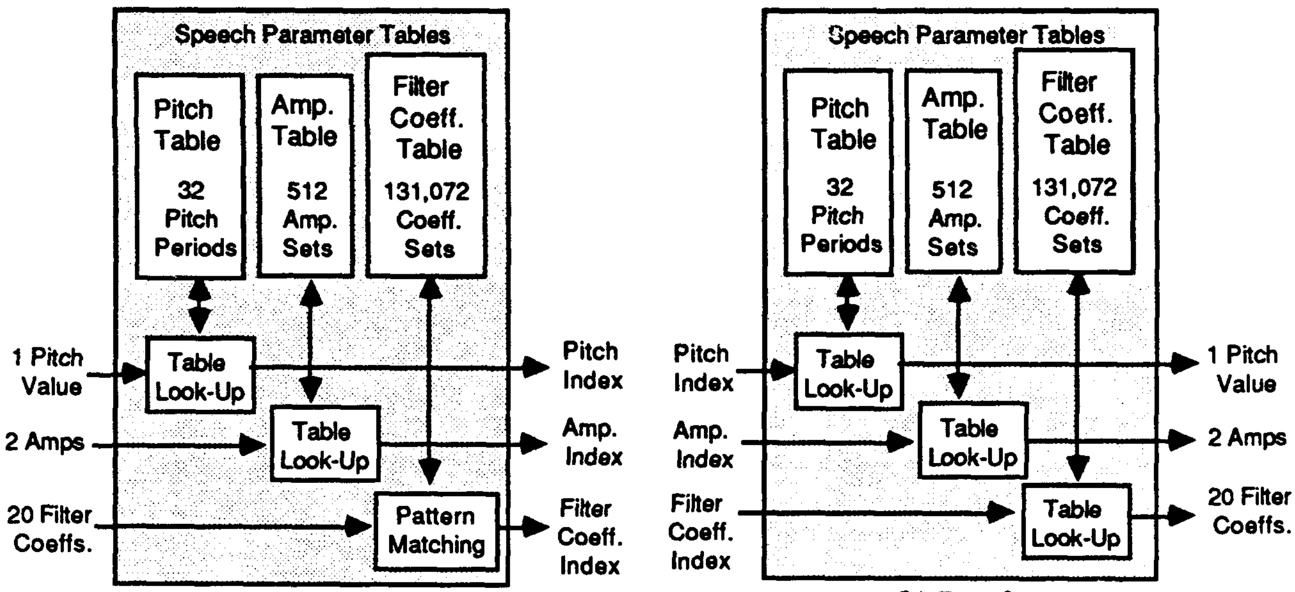
The 800-b/s voice encoder transmits the following speech parameters for two frames (Table 1). For comparison, bit assignments for a current 2400-b/s LPC are also listed.

Table 1 - Bit Assignments for 800-b/s Voice Encoder. Note that the frame rate of 2400-b/s LPC is 44.44 Hz, whereas the frame rate for 800-b/s voice encoder is 50 Hz.

	2400 b/s LPC	800 b/s Encoder
Pitch Period	6 bits/frame	5 bits/2 frames
Amplitude	5	9
Filter Coeffs	41	17
Voicing Decision	1	None
Frame Sync	1	1
TOTAL	54 bits/frame	32 bits/2 frames

PARAMETER QUANTIZATION

Speech parameters are encoded by table-look up. Figure 4 is a block diagram of the 800-b/s parameter encoder and decoder identified in the overall block diagram previously shown in Fig. 1.



(a) Encoder (b) Decoder

Fig. 4 - Block diagrams of 800-b/s parameter encoder and decoder. As noted, with an exception of filter coefficient encoding, encoding and decoding are performed by table look-up.

1) Pitch Quantization (Scalar Quantization)

The pitch period does not change as rapidly as other parameters in normal conversation. Therefore, only one pitch period (pitch period of the first frame) is encoded, and it is also used for the second frame. Pitch period is encoded from 20 to 120 sampling-time intervals (which correspond to the fundamental pitch frequencies from 400 to 66.6667 Hz). The pitch resolution is 12 steps per octave, and the number of bits required to transmit pitch period is only 5 bits for two frames. Pitch encoding is a table look-up operation where, for a given pitch value, the pitch code is read directly from Table 2. Pitch decoding is the reverse operation.

Table 2 - Pitch Encoding/Decoding Table. The pitch periods listed are those allowed by the 2400-b/s LPC.

Pitch Period	Pitch Code	Pitch Period	Pitch Code	Pitch Period	Pitch Code
20	0	40	12	80	24
21	1	42	13	84	25
22	2	44	14	88	26
23	3	46	15	92	26
24	4	48	15	96	27
25	5	50	16	100	28
26	5	52	17	104	28
27	6	54	17	108	29
28	6	56	18	112	30
29	7	58	18	116	30
30	7	60	19	120	31
31	8	62	20	124	31
32	8	64	20	128	31
33	9	66	21	132	31
34	9	68	21	136	31
35	10	70	22	140	31
36	10	72	22	144	31
37	11	74	23	148	31
38	11	76	23	152	31
39	12	78	24	156	31

(2) <u>Amplitude Quantization (Vector Quantization)</u>

The amplitude parameter is the root mean-square value of the speech waveform computed for each frame. Initially, each amplitude parameter is logarithmically quantized into one of 26 values over the entire dynamic range of the speech signal. Then, two amplitude parameters over two consecutive frames are jointly encoded. According to extensive analyses of various speech samples, only 512 are significant among 676 (= 26 x 26) possible amplitude transitions. Each of the allowable amplitude transitions is assigned a code, as tabulated in Table 3.

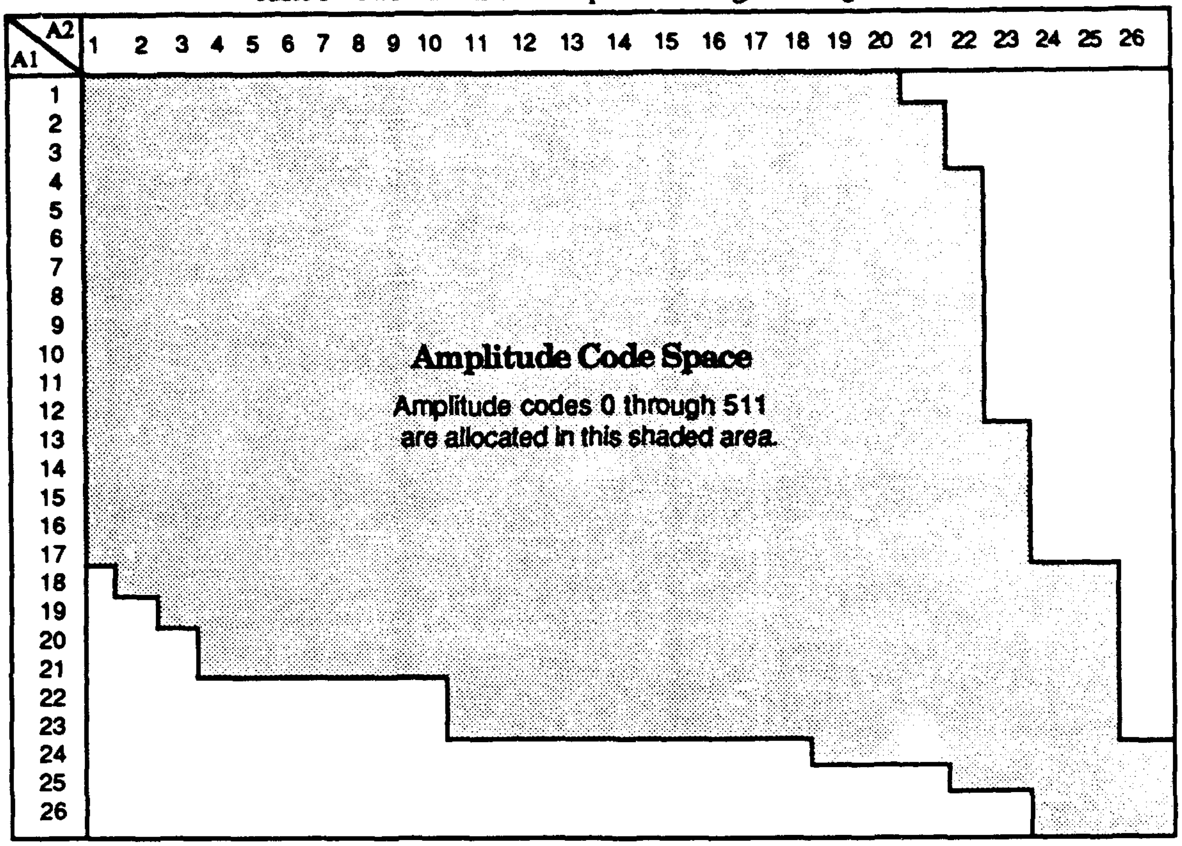
Amplitude encoding is achieved by a table look-up process. For two logarithmically quantized amplitudes (A1 and A2), the corresponding code is read directly from the 26-by-26 matrix. Unallowable amplitude transitions (unshaded areas) are excluded from the coding space. Decoding is the reverse operation which converts an amplitude code to two amplitudes (A1 and A2) by look up Table 3.

(3) <u>Filter Coefficient Quantization (Matrix Quantization)</u>

Previously, template matching (often called vector quantization) of filter coefficients has shown remarkable results (Refs. 4 through 7). In this approach, speech is synthesized from the filter coefficients selected from the reference templates that are free from nonspeech sounds. We again use a similar technique but take it one step further. We apply a pattern matching technique for jointly encoding filter coefficients from two adjacent frames. In this way, we not only eliminate nonspeech sounds from encoding, but we also eliminate improbable filter coefficient transitions across two adjacent frames.

The filter coefficient coding/decoding table consists of LSP templates, each containing 20 frequencies. The number of LSP sets, as stated in Table 2 is 131,072 (= 2^{17} or a 17-bit quantity). LSP templates are collected through the procedures outlined next.

Table 3 - Two Dimensional Amplitude Coding/Decoding Table



LSP Template Collection

We collect a representative number of LSP templates by analyzing 420 speakers uttering 8 sentences each. LSP templates are collected by the following steps:

<u>Step 1</u>: The first incoming LSP set (20 frequencies from two consecutive frames) is the first LSP template, and it is stored in memory.

<u>Step 2</u>: The second incoming LSP set is compared with all the stored templates. If the average spectral difference between the incoming LSP set and one of the templates is less than 2 dB, both LSP sets are regarded as being the same family, and therefore the incoming LSP sets is discarded. Otherwise, it will be stored as a new template.

<u>Step 3</u>: Step 2 is repeated until the maximum allowable template size (i.e., 2^{17} = 131,072) is reached. Actually we collect more than the maximum number, pending elimination of least-frequently-used templates later on to meet the required maximum template size.

A similar approach was also successfully used by Gold (Ref. 6) for the channel vocoder, and Paul (Ref. 7) for the spectral-envelope-estimation vocoder. A difficulty of designing a satisfactory vector quantizer is that there are always speakers whose speech parameters are far outside the hyperspace defined by the templates. Therefore, it is desirable to collect LSP templates from vastly different voice characteristics.

LSP Template Storage in Tree Arrangement

An exhaustive search of 131,072 LSP templates in two frames cannot be performed in real time with present-day hardware. Thus, the templates must be partitioned in such a way that only a fraction of the total templates are searched. We present a method of LSP template partitioning where the maximum number of templates in any one group is only 2048. Since each filter-coefficient template has two voicing decisions associated with it, filter-coefficient templates are initially partitioned in the following four ways.

Case 1: Both frames are unvoiced: This case includes fricatives, plosives, and silence. For this case, the number of templates is on the order of 1000. The best-matched template can be found by exhaustive search.

Case 2: The first frame is voiced, and the second frame is unvoiced: This case includes trailing ends of words and phrases. For this case, the number of filter-coefficient templates is on the order of 2000. The best-matched template can be found by exhaustive search.

Case 3: The first frame is unvoiced, and the second frame is voiced: This case is for speech onsets, and it is critical to speech intelligibility. The number of templates for this case is on the order of 16,000. To facilitate the search for the best-matched template, templates are partitioned based on the indices of seven closely spaced line-spectral frequencies (Fig. 5).

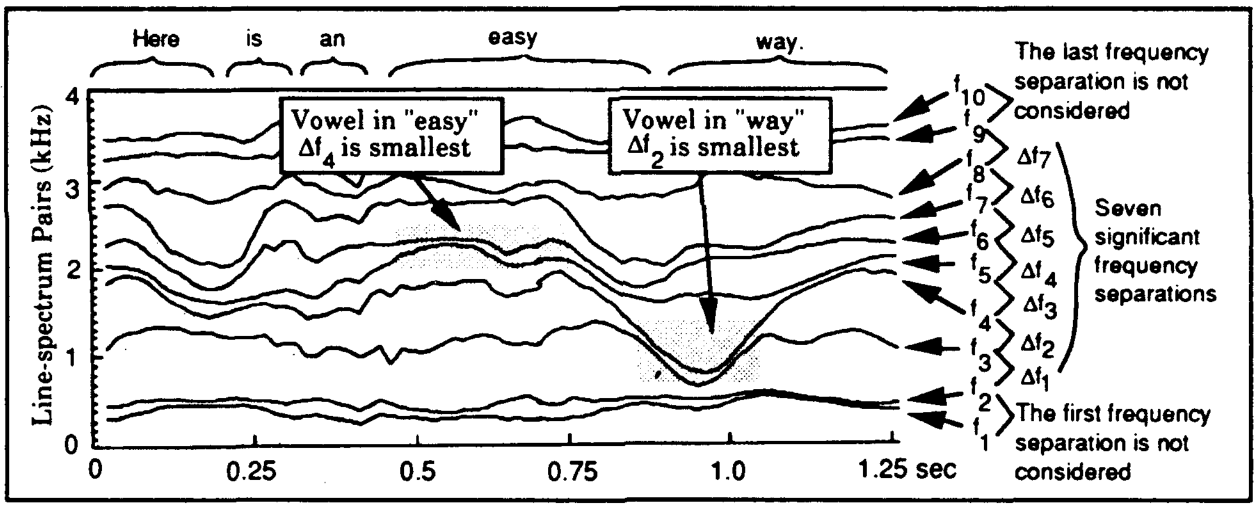


Fig. 5 - Seven significant frequency separations in LSP trajectories. The first and last frequency separations are not considered because they are more or less stationary, therefore, they not too useful for LSP partition.

As illustrated in Fig. 5, closely-spaced line-spectral frequencies vary from phoneme to phoneme. By clustering filter-coefficient templates in terms of indices of closely-spaced line-spectral frequencies, templates are grouped in terms of similar speech sounds. Figure 6 is a tree search of filter-coefficients templates for Case 3.

Case 4: Both frames are voiced: This case is for vowels. The number of filter coefficient templates is on the order of 110,000. Templates are partitioned on the stationarity of line-spectral frequencies over two frames. If the speech is a sustained vowel over two frames, the indices of the closely spaced frequency separations will be identical. For transitional vowels, they are expected to be different. Figure 7 is a tree diagram of further partitioning of the filter-coefficient templates for Case 4.

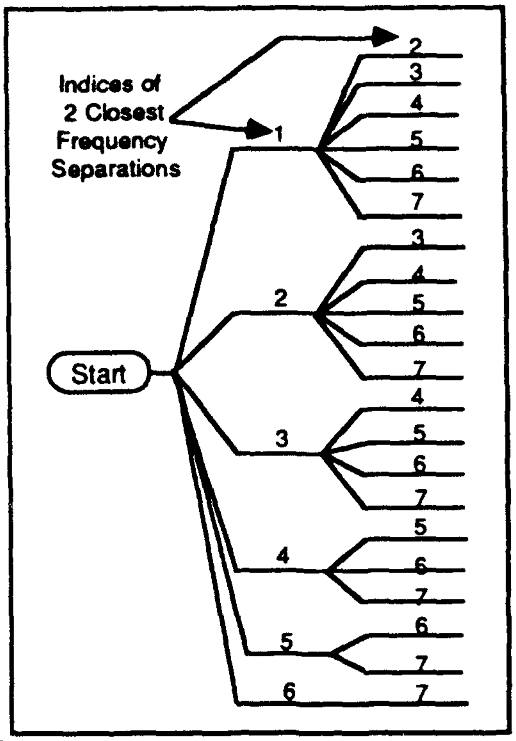


Fig. 6 - Filter coefficient partition for Case 3 (unvoiced-transition)

LSP Template Matching

The incoming LSP matrix (LSP sets from two adjacent frames) are compared with all of the LSP templates (each template is likewise made of two LSP sets). The index corresponding to the closest match is transmitted. We use the error criterion expressed as the sum of the absolute weighted differences between two sets of LSP matrices, $\{F_a\}$ and $\{F_b\}$, each comprised of 20 line-spectrum frequencies. Thus,

$$d(F_a, F_b) = \sum_{i=1}^{20} | w_a(i) [F_a (i) - F_b(i)] | \quad (4)$$

and

$$d(F_b, F_a) = \sum_{i=1}^{20} | w_b(i) [F_a (i) - F_b(i)] | \quad (5)$$

where $w_a(i)$ and $w_b(i)$ are the weights of the i^{th} line spectrum of $\{F_a\}$ and $\{F_b\}$, respectively.

The magnitude of the weighting factor is proportional to the spectral-error sensitivity (i.e., a larger magnitude for closely-spaced LSPs (Ref. 1)). For each comparison, we generate two-way errors based on both Eqs. (4) and (5); then we choose the largest error of the two. We compute the weighting factors beforehand and store them along with the LSP templates.

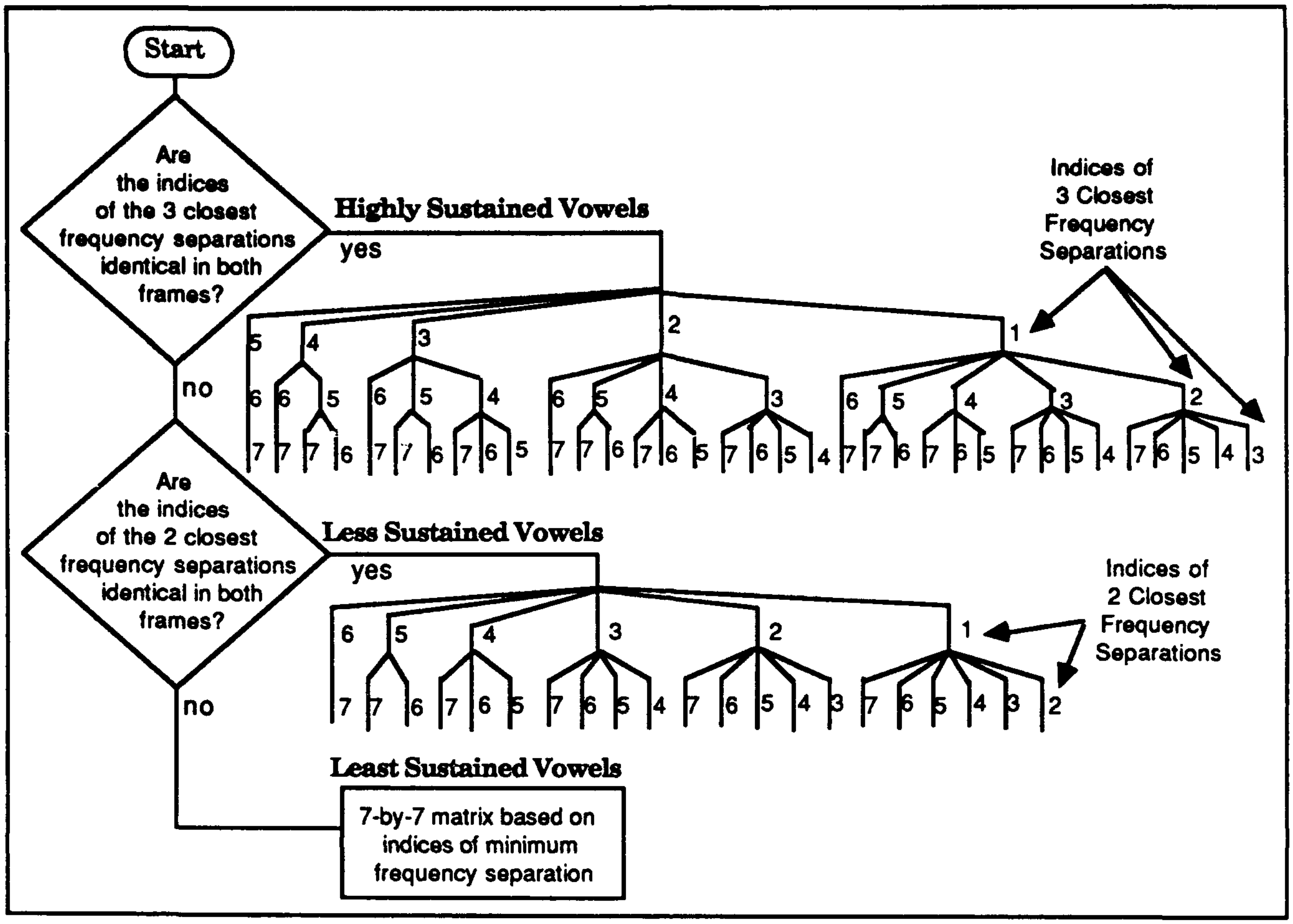


Fig. 7 - Filter coefficient partition for Case 4 (both frames are voiced)

INTELLIGIBILITY TEST SCORES

The DRT evaluates the discriminability of initial consonants of monosyllable rhyming word pairs. According to our experience, DRT scores are dependable (i.e., scores are repeatable under retesting), and they often reveal latent defects of synthetic speech that are not easily discernible through casual listening. As listed in Table 3, the average DRT score of the 800-b/s voice algorithm is 91.5. Three male speakers are used for this test. As far as we can determine, these are the highest DRT scores for any 800-b/s voice processor. For comparison, DRT scores for the latest 2400-b/s LPC are also entered in this table.

Table 3 - DRT Scores of the 800-b/s Voice Processor.

DRT Attribute		Data Rate (b/s)	
		800	2400
Voicing	Distinguishes /b/ from /p/, /d/ from /t/, /v/ from /f/, etc.	94.0	95.1
Nasality	Distinguishes /n/ from /d/, /m/ from /b/, etc.	95.6	96.9
Sustention	Distinguishes /f/ from /p/, /b/ from /v/, /t/ from /θ /, etc.	87.5	88.3
Sibilation	Distinguishes /s/ from /θ /, /ʃ / from /d/, etc.	95.8	93.8
Graveness	Distinguishes /p/ from /t/, /b/ from /d/, etc.	82.8	87.0
Compactness	Distinguishes /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc.	93.2	96.4
	TOTAL	91.5	92.9

REAL-TIME IMPLEMENTATION

The 800-b/s voice encoder has been implemented on commercially-available signal processors. Figure 8 is the block diagram. The INTEL i860 signal processor is the key element in the implementation of the invention. It is capable of performing 40 MIPS and 80 MFLOPS. The INTEL i860 processor can handle four independent 800-b/s voice channels. The analog I/O digitizes the speech waveform into a bit stream and vice versa. The VME bus allows the i860 (via i960) to access the analog I/O facilities.

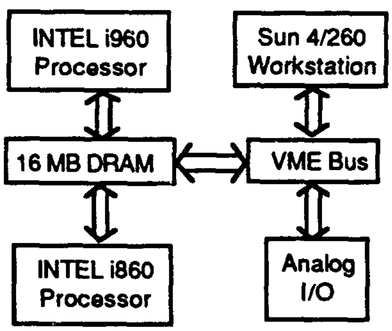


Fig. 8 - Real-time emulation of 800-b/s Voice Encoder

The INTEL i960 processor performs mainly input/output (I/O) operations. The dynamic random access memory (DRAM) has 16 million bytes of storage capacity. To execute the 800-b/s voice algorithm, the following amount of memory is needed: 5 MB for tables, 1.5 MB for program, and 30 KB for other miscellaneous operations.

A Sun 4/260 workstation hosts the software development environment, and it is not needed once the 800-b/s software is complete.

CONCLUSIONS

After nearly a decade of research and development, we were able to generate 800-b/s speech that can be classified as "very good" speech. The factors that most contributed to the high intelligibility are: choice of a 20-ms frame, vector quantization of amplitude parameters and matrix quantization of LSP coefficients, both over two consecutive frames. Speech intelligibility of the 800-b/s voice processor exceeds that of the 2400-b/s LPC of a few years ago. We expect that very-low-data-rate voice processors will be increasingly used to enhance bit-error performance, low-probability of intercept, and narrowband voice/data integration.

ACKNOWLEDGMENTS

We thank Timothy McChesney and Sharon James of SPAWAR PMW151 for support of this R&D effort.

REFERENCES

1. G.S. Kang and L.J. Fransen, "High-Quality 800-b/s Voice Processing Algorithms," NRL Report 9301. (1991)

2. A.W.F. Huggins, R. Viswanathan, and J. Makhoul, "Quality Rating of LPC Vocoders; Effects of Number of Poles, Quantization and Frame Rate," 1977 IEEE ICASSP Record, 413-416. (1977)

3. P. Kabal and R.P. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials," *IEEE Trans. Acoustics, Speech and Signal Proc.* **ASSP-34**(6), 1419-1426. (1986)

4. D.Y. Wong, B.H. Jung, and A.H. Gray, Jr., "An 800 bits/s Vector Quantization LPC Vocoder," *IEEE Trans. Acoustics, Speech and Signal Proc.* **ASSP--30**(5), 770-780. (1982)

5. G.S. Kang and L.J. Fransen, "Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)," NRL Report 8857. (1985)

6. B. Gold, "Experiments with a Pattern-Matching Channel Vocoder," IEEE ICASSP Record, 32-34. (1981)

7. D.B. Paul, "An 800-b/s Adaptive Vector Quantization Vocoder Using Perceptual Distance Measures," IEEE ICASSP Record, 73-76. (1983)