THE MULTIMODAL MULTIPULSE EXCITATION VOCODER

Takahiro Unno[1] *Thomas P. Barnwell III*[2] *Mark A. Clements*[2]

[1]LSI laboratories, Asahi Chemical Industry, Atsugi-shi, Kanagawa 243-02, Japan
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250
unno@ljk.atsugi.asahi-kasei.co.jp, tom,clements@ee.gatech.edu

ABSTRACT

This paper presents a new high-quality, variable-rate vocoder in which the average bit-rate is parametrically controllable. The new vocoder is intended for use with data-voice simultaneous channel (DVSC) applications, in which the speech data is transmitted simultaneously with video and other types of data. The vocoder presented in this paper achieves state-of-the-art quality at several different bit-rates between 5.5 Kbps and 10 Kbps. Further, it achieves this performance at acceptable levels of complexity and delay.

1. INTRODUCTION

The recent past has seen a rapid growth in many classes of networks along with an associated dramatic increase both the types and amounts of network services being offered. This includes a dramatic increase in the use of data-voice simultaneous channel (DVSC) applications such as video-phones. In such channel, It is best if the bit-rates of speech coders can be parametrically controlled so that speech coders can dynamically utilize excess channel capacity to improve quality and can also release channel capacity to other data streams when necessary.

A primary goal of this research was to design a controllable-rate speech coder which uses the same basic structure to achieve all of the bit-rates within its range. The speech coder presented in this paper uses a single, parametrically controllable structure which can operate at many different bit-rates between 5.5 Kbps and 10 Kbps. A second major goal of the research was to design a speech coder which produces state-of-the-art quality when operating at an individual bit-rate. The final goal was to design the speech coder to have acceptable computational properties (implementable on a single DSP microprocessor) and delay properties (less than 120 ms of additional delay) to operate in a DVSC environment. The coder presented in this paper achieves these goals, with transparent quality (for a speech signal sampled at 8000 samples per second) being achieved at a bit-rate of 9 Kbps.

The vocoder presented in this paper can be thought of as a variable-rate multipulse coder or as a self- excited vocoder [1] augmented by a multipulse excitation. Historically, fixed-rate multipulse vocoders have been capable of providing high quality speech, but they generally require relatively high bit-rates. Self-excited vocoders generally require lower bit-rates, but they do not perform well in transition regions. To reduce the bit-rate for multipulse coders, methods that constrain the positions and amplitudes of pulses [2] [3] [4] have been developed. As compared with unconstrained multipulse coders, constrained coders perform relatively well, but they do sometimes introduce additional distortion. For a variable bit-rate speech coder, the same number of pulses is not required for all excitation analysis frames. In stationary voiced regions, a long term predictor with a fast update rate works well, and fewer pulses are required. Conversely, in nonstationary regions, more pulses are usually required. For these reasons, the vocoder presented in this paper controls the number of pulses dynamically rather than constraining their positions and amplitudes. In general, this reduces both the bit- rate and the distortion. Also, in our system then LPC analysis frame-rate is also variable. Even for unconstrained multipulse vocoders, distortion can occur in rapidly changing transition regions of the speech signal. In order to remove the distortion in such regions, our system can apply a fast LPC analysis frame-rate in transition regions.

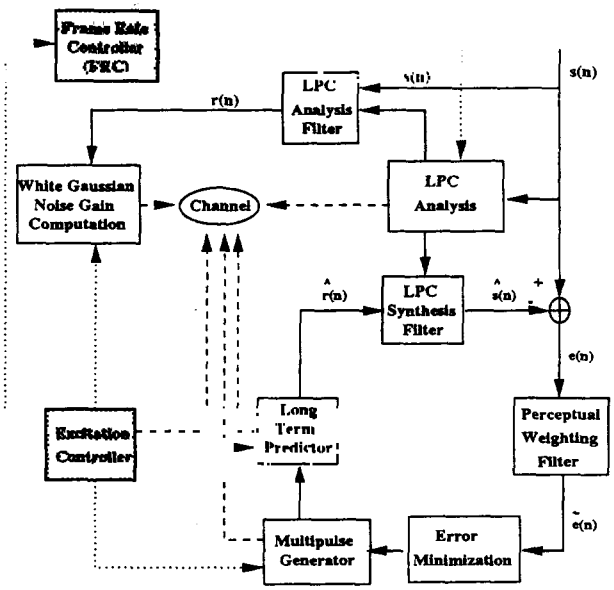

Figure 1: Encoder block diagram.

Table 1: Multimodal excitation model.

mode	Excitation Generator			Speech segment
	White noise	Long term predictor	Number of multipulse	
1	ON	OFF	0	Unvoice
2	OFF	ON	1	Stationary vowel
3	OFF	ON	4	Quasi-stationary vowel
4	OFF	ON	6	Less stationary vowel
5	OFF	OFF	3	Slow transition
6	OFF	OFF	5	Fast transition

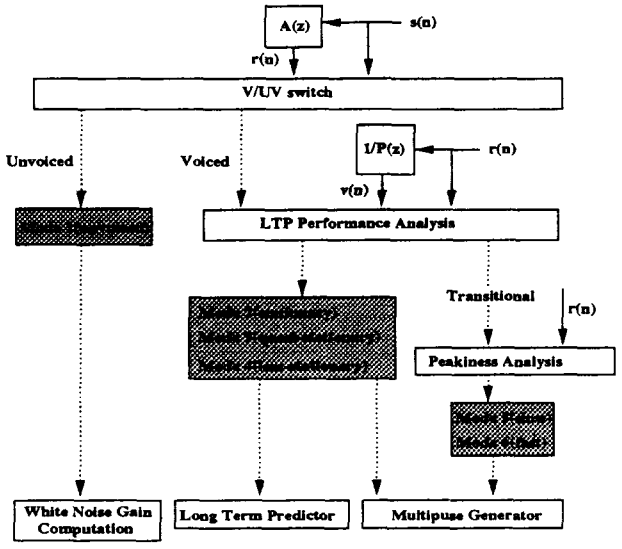

Figure 2: Excitation Controller

2. MULTIMODAL MULTIPULSE EXCITION VOCODER (MMEV)

The vocoder described here is called a *multimodal multipulse excitation vocoder* or *MMEV*. Like a self-excited vocoder, it uses a long term predictor with a relatively fast (5 ms) update rate. However, it has a variety of additional modes operating at different bit-rates which it can use to control the quality/bit-rate tradeoff. The operation of the MMEV is illustrated in Figure 1. Like a conventional multipulse coder, the MMEV divides the speech signal into frames for LPC analysis and then further divides the frames into subframes for excitation analysis. The basic coder uses a classical analysis-by-synthesis engine in which an LPC Synthesis Filter is excited by an excitation signal which is chosen from a set of alternatives by performing error minimization between perceptually weighted versions of the original and coded speech signals.

2.1. Multi-Excitation Model

The excitation generator has three components: a white noise generator, a long term predictor (LTP), and a multipulse generator. As shown in Figure 1, the Excitation Control function controls which combination of components is used for each subframe. As shown in Table 1, the excitation generator has six separate modes operating at six separate bit-rates. One of the modes, mode 1, involves only a white noise excitation. When using this mode, the MMEV functions exactly like the unvoiced mode of a pitch-excited vocoder. All of the other modes involve either multiple pulses or combinations of multiple pulses with the long term predictor. Excitation signals for mode 2 are generated by only fast update rate LTP and one multipulse, so mode 2 can be considered as self-excited-like model. On the other hand, mode 4 can be considered as traditional multipulse excitation model.

2.2. Excitation Controller

The operation of the Excitation Control function is illustrated in Figure 2. First, based on a comparison of the original speech signal s(n) and the residual signal r(n), a voiced/unvoiced decision is made. For subframes that are marked as *unvoiced*, only a white noise excitation (mode 1) is used. For *voiced* subframes, an analysis of the performance of the long term predictor is used to mark the subframes as stationary (mode 2), quasi-stationary (mode 3), less stationary (mode 4) or transitional (modes 5 and 6). The performance of LTP is analyzed by comparing the lowpass filtered input and output of the LTP inverse filter. Since the LTP inverse filter removes harmonics in lower frequency regions from LPC residual signals, the energy ratio of the lowpass filtered output of the filter to the input of the filter will decrease as the LTP inverse filter does a better job of removing the pitch redundancy. For transitional modes, an analysis of the peakiness of the residual signal is used to further classify transitional modes as slow (mode 5) or fast (mode 6).

2.3. Frame Rate Controller (FRC)

Another variable-rate control function is the Frame Rate Controller (FRC) (see Figure 1). In stationary segments, the LPC spectrum changes slowly, but in a transitional frame it changes rapidly and irregularly. The Frame Rate Controller finds onset transition segments by using the V/UV information and the multipulse/long-term predictor controller, and it assigns higher LPC frame-rates to such segments. In stationary segments, the loss in the LPC spectrum resolution caused by the linear interpolation of the LSPs is compensated by the multipulse excitation signals to some degree since a multipulse excitation is capable of generating a non-flat LPC spectrum. However, in rapid transitional segments, the loss of the LPC spectrum resolution is too significant to be recovered by the multipulse excitation signal. Thus, a higher LPC frame-rate is used in order to avoid significant LPC spectral distortions.

Table 2: Bit assignment of MMEV.

Mode	1	2	3	4	5	6
MP position	0	6	17	22	14	20
MP amplitude	0	3	6	10	9	16
MP sign	0	1	4	6	3	5
LTP delay	0	8	8	8	0	0
LTP gain	0	3	3	3	0	0
White noise gain	4	0	0	0	0	0
Total/5 ms	4	21	38	49	26	41
LSP/10,30 ms	18 (64 *levels*-3 *stages*)					
V/UV /15 ms	1					
Mode info/5 ms	0	2	2	2	3	3

3. COMPLEXITY AND PARAMETER CODING

In order to make the final speech coder practical in terms of both bit-rate and computational complexity, a number of techniques for bit-rate reduction and computational efficiency were used. For computational reasons, the pulse positions and amplitudes were computed using the sequential method and then the final amplitudes for the pulses were determined using the reoptimization method [5].

For parameter coding, multistage vector quantization (MSVQ) was employed for both the LPC parameter quantization and multipulse amplitude quantization. For the best performance, all MSVQs were designed such that every stage was jointly optimized [7], and an M-search algorithm (M=8) is employed for codebook design and encoding. For spectral quantization, the LSPs were quantized using an 18 bits MSVQ (64 *levels*-3 *stages*) with a weighted Euclidean distance measure [8]. With frequency-weighted spectral distortion [6] and informal subjective testing, it was observed that the 18 bits MSVQ performed better than a conventional 34 bits scalar quantizer. For multipulse amplitude quantization, the vector components are ordered so that the amplitude of the first pulse is the first component and the amplitude of the last pulse the last. This ordering reduces the combination of pulse position from $N!/(N-m)!$ to $N!/m!(N-m)!$ (where N is the frame length and m is the number of multipulse in the frame) and also reduces the number of bits used for pulse positions,. In order to reduce the number of entries in the vector codebooks for the multipulse amplitude quantization, a shape-gain MSVQ is also employed. All vector components were normalized by the maximum pulse amplitude, and the maximum log amplitude and normalized amplitudes are quantized separately. The lowest bit-rate MSVQ for multipulse amplitude that was judged to introduce no perceivable quantization distortion was selected for each mode. Table 2 shows bit assignments for all parameters.

4. BIT-RATE CONTROL AND DELAY FEEDBACK FUNCTION (DFF)

For the overall bit-rate control, the occurrence rate of each mode can be controlled by adjusting the thresholds of LTP performance analysis and peakiness analysis in Figure 2. In general, for higher bit-rates, and higher quality systems, the thresholds are set to make mode 4 occur more often. Conversely, for lower bit-rate, lower quality systems, the thresholds are set to make mode 2 more likely. Similarly, the threshold for peakiness analysis is set to make mode 6 occur more often for higher bit-rates.

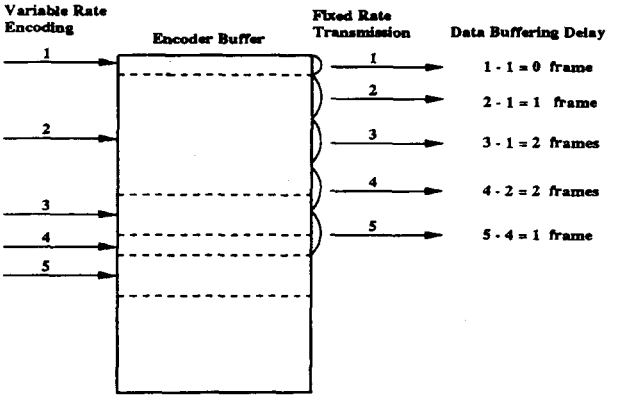

Figure 3: Data buffering delay model

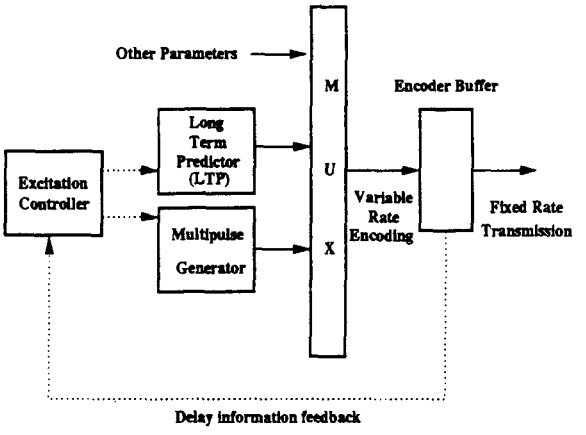

Figure 4: Delay feedback function

In order to provide a conversion between our variable-rate vocoder and a fixed rate channel, a buffer is provided between the encoder and the transmission channel. Figure 3 shows the buffer and data buffering delay model. Encoded data comes from left side with a variable bit-rate and transmitted data goes out the right side with a fixed bit-rate. Data is transmitted with fixed bit-rate, but only remaining data in the buffer will be transmitted if the amount of remaining data is less than a full data frame. The numbers on arrows show frame number of encoded and transmitted frame. Data buffering delay is defined the difference between transmitted frame number and encoded frame number as shown on the right side of Figure 3. Unacceptable data buffering delay will occur if expensive modes like mode 4 often occur successively. In order to avoid an unacceptable delay, a delay feedback function (DFF) is provided between the buffer and Excitation Controller as shown in Figure 4. If the data buffering delay is getting too long, the DFF informs the Excitation Controller of the current data buffering delay, and Excitation Controller suppresses the occurrence

Table 3: Occurrence of modes.

Occurrence of modes (%)						Target
1	2	3	4	5	6	bit-rate(bps)
26	0	0	46	0	28	10,000
26	0	12	34	6	22	9,000
26	4	22	20	16	12	8,000
26	18	17	10	20	9	7,000
26	35	8	2	20	9	6,000
26	42	2	0	26	4	5,500

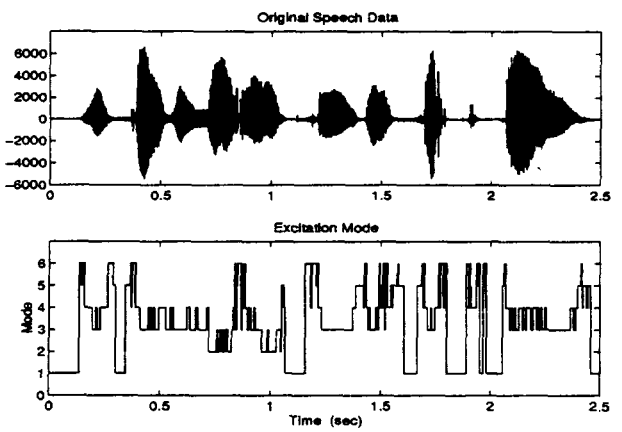

Figure 5: Original Speech and Excitation mode

of expensive mode by further adjusting the thresholds of LTP performance analysis and peakiness analysis.

5. RESULTS

For MSVQ codebook training, 230 sentences were selected at random from the TIMIT database, and another 40 sentences were selected for the evaluation of speech coder. The speech data was lowpass filtered (cut-off frequency of 3800 Hz) and downsampled to 8 kHz. The LPC analysis was performed every 30 ms for regular frame-rate segments and every 10 ms for high frame-rate segments. The V/UV switch frame interval was 15 ms. The subframe size for excitation analysis is 5ms. Table 3 shows the occurrence of each mode and the target bit-rate as controlled by the Excitation Controller, and Figure 5 shows input speech data and excitation mode. The number of each mode in Figure 5 corresponds to one in Table 1. It is observed that mode 1 occurs mostly in silence or fricative segments, mode 2-4 occurs mostly in voiced segments, and mode 5 and 6 occur mostly in transitional segments. From our experiments, we found that the bit-rate can be smoothly controlled from 5.5kbps to 10kbps in the MMEV and the maximum data buffering delay controlled by DFF is 120 ms. In informal listening tests, transparent quality was provided at 9 Kbps or higher, and the speech quality gradually degraded as the bit-rate decreased.

6. SUMMARY

The paper has presented a new type of vocoder for a DVSC application. *Multimodal Multipulse Excitation Vocoder (MMEV)* that can be parametrically controlled to have a range of different bit-rates using a single structure and can also provide state-of-the-art quality at individual bit-rates. The vocoder also achieves acceptable computational properties and delay properties by employing MSVQ and delay feedback function (DFF).

ACKNOWLEDGEMENT

The authors wish to thank to F. Kossentini at University of British Columbia and K. Truong at Atlanta Signal Processors, Inc. for valuable advice and discussions.

REFERENCES

[1] R. C. Rose and T. P. Barnwell III, "The self-excited vocoder - an alternate approach to toll quality at 4800 bps," *Proc. Inter. Conf. on Acoustic, Speech, and Signal Proc.*, pp. 453-456, April 1986

[2] "Draft Recommendation A V.25Y - Dual rate speech coder for multimedia telecommunication transmitting at 5.3 & 6.4 kbits/s," 1995.

[3] R. Salami, C. Laflamme, and J-P. Adoul, "ACELP speech coding at 8kbps/s with a 10ms frame: a candidate for CCITT standardization," *IEEE Workshop on Speech Coding for Telecommunications*, pp. 23- 24, 1993.

[4] S. Taumi, K. Ozawa, T. Nomura, M. Serizawa "Low-Delay CELP with Multi-Pulse VQ and Fast Search for GSM EFR," *Proc. Inter. Conf. on Acoustic, Speech, and Signal Proc.*, pp. 562-565, May 1996

[5] S. Singhal and B. S. Atal, "Amplitude Optimization and Pitch Prediction in Multipulse Coders," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, March 1989, pp. 317-327.

[6] J. S. Collura, A. McCree, T. E. Tremain "Perceptually Based Distortion Measure for Spectrum Quantization," *IEEE Workshop on Speech Coding for Telecommunications*, pp. 49-50, 1995.

[7] F. Kossentini, "Multistage Residual Vector Quantization with Application to Image Coding," *PhD thesis*, Georgia Institute of Technology, 1994.

[8] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 1, January 1993, pp. 3-14.