MULTI-RATE SPEECH CODING FOR WIRELESS AND INTERNET APPLICATIONS

J. E. Kleider and R. J. Pattison

Motorola, Systems Solutions Group
Scottsdale, Arizona, USA

ABSTRACT*

Fixed-rate speech codecs are unable to provide synthesized speech with fixed delay when the channel capacity changes, and can not dedicate additional forward error correction bits for protection against noisy channels. We propose a multi-rate method for variable bandwidth applications, such as the Internet, and severely degraded wireless channels, such as mobile cellular. The technique uses a multi-rate version of the sinusoidal transform coder (MRSTC), operates at 9.6/4.8/2.4/1.2 kilobits/sec (kb/s), and is switchable "on-the-fly." The algorithm produces high quality speech, even when transitioning between rates. We compare two switching techniques, one method uses a "frame-deletion" (FD) technique, and a second method which utilizes "parameter-history" (PH) information. PH produces the best speech quality. FD is attractive because it requires no additional speech memory. Experimental results show greater than a 9 dB gain in receiver C/N_o operating range using the MRSTC over a fixed-rate system with STC operating at 9.6 kb/s.

1. INTRODUCTION

Good quality voice services are in high demand, primarily due to the emergence of global communication capabilities, such as those provided by cellular systems, the Iridium system, and other interconnected satellite, landline, and wireless systems. Digital speech coders are often designed assuming that the voice and channel coders operate at fixed bit rates. In actuality, however, speech communication is a nonstationary process with random-like intervals of silence, while for internet/network applications the user access statistics are a variable process. For wireless applications, the channel capacity can change dramatically, and thus impose a variable limit on the maximum bit rate that can be passed through the channel. Variable-rate speech coding has been shown to provide significant gains in communication system capacity while sustaining adequate levels of voice quality [1].

Historically, variable-rate speech coding has been separated into three categories: source-controlled, network-controlled, and channel-controlled variable-rate speech coding. Our objective is to propose a variable-rate voice coding method with a wide range of speech encoding rates, while providing optimal voice quality under variable-bandwidth demands and noisy channel environments. Our multi-rate method is designed using a multimode coder because multimode coders are found to produce superior speech quality compared to embedded speech coders [2].

One proposed multi-rate method, based on multimode CELP coding at 16 kb/s, 8 kb/s, and 4kb/s [3], produced good speech quality at those rates; however, CELP quality suffers below 4 kb/s. We believe that a coder exhibiting good speech quality below 4 kb/s will be necessary to satisfy our objective. Multi-rate Multiband Excitation speech coding was developed for 4.8/3.6/2.4/1.2 kb/s, however the coder does not provide switching between rates on demand [4]. Our MRSTC technique is switchable between any of the rates, can be requested at any time, and produces smooth transitions at the switch locations without producing annoying artifacts in the speech. Good speech quality is achieved by utilizing an optimized coder design for each source-encoding rate. Computational complexity is minimized using a modular multi-rate architecture.

MRSTC can be used in network-controlled and channel-controlled modes. In network-controlled mode, the MRSTC switches between any rate at the request of a network rate control signal. In channel-controlled mode, similar switching is provided, but in response to changing channel conditions, as determined by the communication system. For example, in a wireless communication system, long-term fading can cause very low received signal to noise ratios resulting in excessively high bit error rates over long time duration. To minimize the speech distortion due to uncorrected bit errors, the system requests a reduction in the MRSTC speech coding bit rate, and either an increase in forward error correction bit rate or a reduction in the modem symbol rate. Both methods have been shown to produce significant improvements in speech quality for severely degraded wireless channels when compared to fixed-rate systems [5].

We utilize a variable-size/rate buffer (VSRB) to arrange and format the speech bits before data transmission. The VSRB system is shown in Figure 1, and supports fixed or variable block delivery bit rates. This buffering approach is useful for adaptive-rate wireless voice systems [5], but also minimizes end-to-end delay when transmitting voice data over the Internet. Two switching techniques, FD and PH, are evaluated, where PH provides the best overall speech quality and FD provides good speech quality with the lowest complexity. Spectral distortion is used as an objective measure of speech quality, while informal listening tests are utilized to support the objective findings. We also show a wireless system application, where high quality speech is produced using the MRSTC in a severely degraded long-term fading channel.

Section 2 describes the variable buffering system of Figure 1, and in addition provides details of the multi-rate STC and switching algorithms. Experimental results are provided in Section 3, with conclusions given in Section 4.

* Prepared through collaborative participation in the Advanced Telecommunications & Information Distribution Research Program (ATIRP) Consortium sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0002.

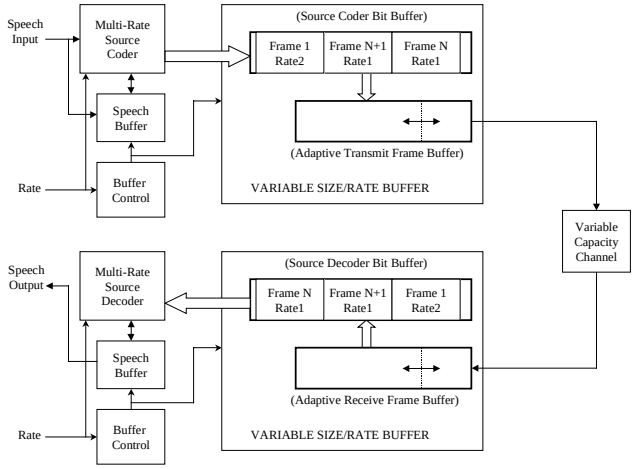

Figure 1: Block diagram of multi-rate coder using VSRB.

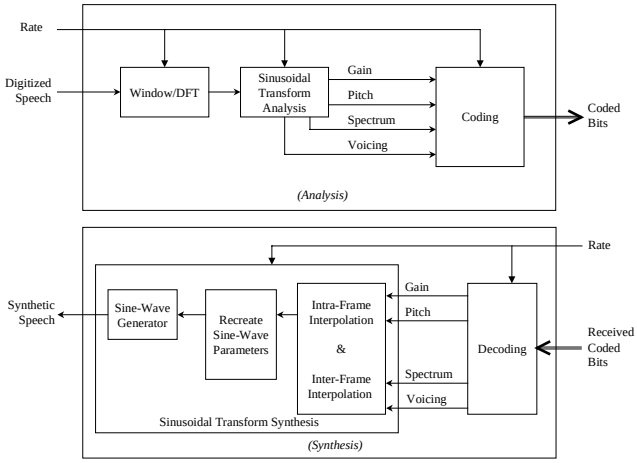

Figure 2: Multi-rate STC architecture.

2. SYSTEM DESCRIPTION

This section provides a detailed description of the VSRB system. The dynamic buffering approach is motivated by the work done in [6], but is unique because of its application to variable-bit-rate channels.

2.1 Buffer Control. The function of the buffer control is to process the rate request from the communication system, and pass the appropriate control parameters to the multi-rate source (de)coder (MRSTC), speech buffer, and the VSRB. Note that there will be an inherent delay in the response of the MRSTC to the control parameter sent from buffer control. This is due to the difference between the rate request time and the time left to finish coding the current speech frame from the MRSTC. In many instances, however, this delay can be eliminated by ignoring the current speech frame, and backing up the appropriate amount in the speech buffer to begin coding at the new rate. The "rate" change is sent to the receiver via a low-bandwidth side information channel. Alternatively, a separate field could be created to carry this rate information as part of the transmitted frame structure.

2.2 Speech Buffer. The speech buffer's purpose at the transmitter is different from that at the receiver. At the transmitter, the speech buffer stores past samples of the digitized speech. This is done to ensure smooth rate changes when using the PH switching method. It can also be used to minimize the time delay between a switch request and when the rate switch is actually executed. If this is not a critical requirement, or if the FD switching method is used, the speech buffer can be eliminated. The buffer's purpose at the receiver is to remove any jitter due to variance in data delivery rates.

2.3 Variable Size/Rate Buffer (VSRB). The VSRB consists of two main components, the source coder bit buffer (SCBB), and the adaptive transmit frame buffer (ATFB). The main function of the ATFB is to allow variable block sizes of bits to be transmitted. This greatly aids in minimizing end-to-end delay of digital voice data transmitted over the Internet, and supports adaptive-rate modulation for transmission of digital voice data over wireless channels. The operation of the ATFB is very straightforward. The block size that is transmitted is set to be directly proportional to the rate of the source coder. The ATFB frame delivery rate is then proportional to the time taken to fill the SCBB with the integer number of speech frames, I_f, at the current source coding rate. (Note: The frame delivery rate is not restricted to a fixed value, provided correct timing coordination is selected between the ATFB and SCCB.) When I_f is reached, the buffer control block sends out a control signal indicating it is time to transfer I_f frames to the ATFB and to output the block of bits from the VSRB. The data flow at the receiver is in general the reverse of that at the transmitter. The buffer control operation is less complex at the receiver since the rate request at the receiver coincides with the actual switch time at the multi-rate vocoder frame boundary.

The advantage in using the VSRB buffering method can be seen by showing the end-to-end delay compared to a fixed size/rate buffering (FSRB) approach, which generally would be required for fixed-symbol rate wireless systems. For the FSRB, we assume the transmit frame buffer holds a fixed number of bits, the frame transmit rate is fixed, but allows storage variance in the number and size of source coder frames (same as the VSRB). For the VSRB, we assume that the output bit rate is equal to the source coder bit rate. The total delay, t_{dv}, for the VSRB method can be written as

$$t_{dv} = t_{sw} + t_{fsr} = (t_{fsc} - t_{req}) + t_{fsr} \text{ (msec)}, \quad (1)$$

where t_{fsc} is the vocoder frame size at the current rate, t_{fsr} is the vocoder frame size at the new rate, t_{req} is the time of the "rate request," relative to the end of the current frame boundary, and t_{sw} is the time difference between t_{fsc} and t_{req}. We assume that t_{req} occurs such that it is uniformly distributed within a frame of length equal to t_{fsc}. The total delay, t_{df}, for the FSRB method is

$$t_{df} = t_{sw} + t_{buf} \text{ (msec)}, \quad (2)$$

where t_{sw} is as defined above, and t_{buf} is the time required to fill the transmit frame buffer. t_{buf} can be expressed as

$$t_{buf} = (B_t/B_v)t_{fsr} \text{ (msec)}, \quad (3)$$

where B_t is the transmit frame buffer size (bits per frame), B_v is the vocoder frame size at the new rate (also in bits per frame).

2.4 Multi-Rate Speech Coder Description. Figure 2 shows a block diagram of the multi-mode coder used for this work and is based on a modularized STC architecture. The objective of the original vocoder architecture from [7] was to provide good speech

quality at a wide range of bit rates. The four vocoder bit rates used are 9.6 kb/s, 4.8 kb/s, 2.4 kb/s, and 1.2 kb/s. Our objective in utilizing this modular approach is to maximize speech quality at each desired bit rate via the proper design interface to the multi-mode coder. The interface provides the ability to switch between any of the 4 rates, at any time, without producing annoying artifacts at the switch locations. MRSTC produces a graceful degradation in speech quality as the rate decreases.

The sinusoidal transform analysis and synthesis blocks can be used at any of the desired rates with only minor differences in the algorithm at each rate. The primary differences are in the parameters used to perform the signal processing; for example, LP analysis order changes with the rate. The coding/decoding blocks are rate specific, as unique quantization codebooks are necessary at each rate to produce reasonable speech quality at the lower encoding rates. Table 1 provides a summary of the MRSTC algorithmic details at each of the four rates.

Bit Rate (kb/s)	Frame Size (msec)	Bits/Frame	LPC Order
1.2	40	48	10
2.4	30	72	14
4.8	30	144	16
9.6	25	240	16

Table 1: Multi-rate voice coder parameters.

3. EXPERIMENTAL RESULTS

An important consideration of a multi-rate vocoder is the algorithmic delay incurred when switching through a wide range of bit rates. To show the effectiveness of the VSRB method compared to a FSRB approach, we ran a simulation to model the delay probabilities, $P(t_{df})$ and $P(t_{dv})$. The simulation tested the switching algorithm over 50k independent switch requests. For FSRB, B_t is fixed, with a size which is limited to an integer multiple of the vocoder frame size in bits. This means that no data is available at the output until the integer number is reached. For example, if B_t is 400 bits and B_v is 100 bits, then it does not output data until 4 frames have been stored in the SCBB. Figure 3 shows the delay probability simulation results. The VSRB system clearly has much less delay, by a factor of 1/6 the delay of FSRB.

As mentioned previously, we utilized two methods, PH and FD, for switching the MRSTC. Within each switching technique, it was also critical to evaluate switched speech quality based on the type of speech activity. We found that both switching techniques worked well in low-energy, un-voiced, and silence regions, so results are given for switching during voiced speech segments.

The FD technique is used to minimize algorithmic complexity. It is performed by removing any delay frames at the requested rate. (Note: For MRSTC there is a two-frame delay before any output of useful coded data is seen.) The FD method is then characterized by ignoring (removing) the delay frames and performing a short-time discrete amplitude filter on each of the merged speech sections. The FD method is illustrated in Figure 4 a) for the speech sequence of "mabel" (mebxl). The switch occurs near the middle of the voiced portion of the vowel /e/. Even though, visually, the speech sequence appears to change dramatically, listening

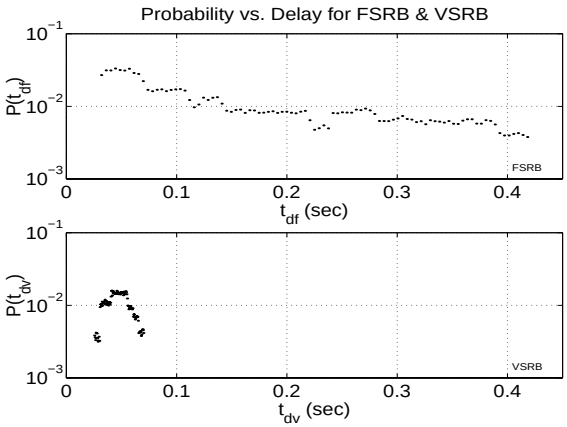

Figure 3: Delay probability plots for FSRB and VSRB.

tests revealed very little loss in intelligibility and quality. A slight "thud" was perceptible, however, due to the variance in energy near the switch location.

PH is the alternate switching method which is proposed for our multi-rate coder. Given that no interpolation is performed between frame boundaries at two different rates, the PH method requires speech buffering to enable the MRSTC to produce the necessary parametric information to provide a seamless transition to the new rate. Two different techniques can then be utilized so no overlap or gaps are present in the synthesized speech at the receiver. One method is to utilize another speech buffer at the receiver to remove jitter due to inconsistent frame rates between the various rates of the MRSTC. This would add additional end-to-end delay, however. An alternative would be to run both the old and new coders in parallel during the switching interval so that no overlap or gaps occur due to frame rate differences between the two rates. The alternative requires a short period of additional processing during the switching interval, but this additional complexity is necessary for low delay applications. Figure 4 b) shows the synthesized speech using the PH method. The degradation in speech quality was nearly unperceptible, producing a smooth transition between rates.

To show the improvement in system operating range for a wireless application, the MRSTC was incorporated into an adaptive-rate communication system. BPSK modulation was used along with rate ½ convolutional channel coding (constraint length, K = 7) at the transmitter, while BPSK demodulation and hard-decision Viterbi decoding were used in the receiver. A long-term fading channel model was utilized with complex additive white Gaussian noise. Specifically, we assume that the i^{th} received symbol, $y(i)$, is related to the i^{th} transmitted symbol, $x(i)$, by the equation

$$y(i) = A(i)x(i) + n(i). \quad (4)$$

Here, $x(i)$ is one of 2 possible symbols in the BPSK signal set and $n(i)$ is a zero-mean complex Gaussian random variable with variance $N_o/2$. $A(i)$ is the multiplicative fading coefficient for the log-normally distributed fading channel.

The performance metric of speech is shown by measuring the speech spectral distortion (*SD*) at the receiver [5]. The improvement in quality is shown by comparing the *SD* of the adaptive-rate system (ARS) to the *SD* of a fixed-rate system (FRS). The FRS operates at a symbol rate of 19.2 ksymbols/sec (ks/s), a MRSTC

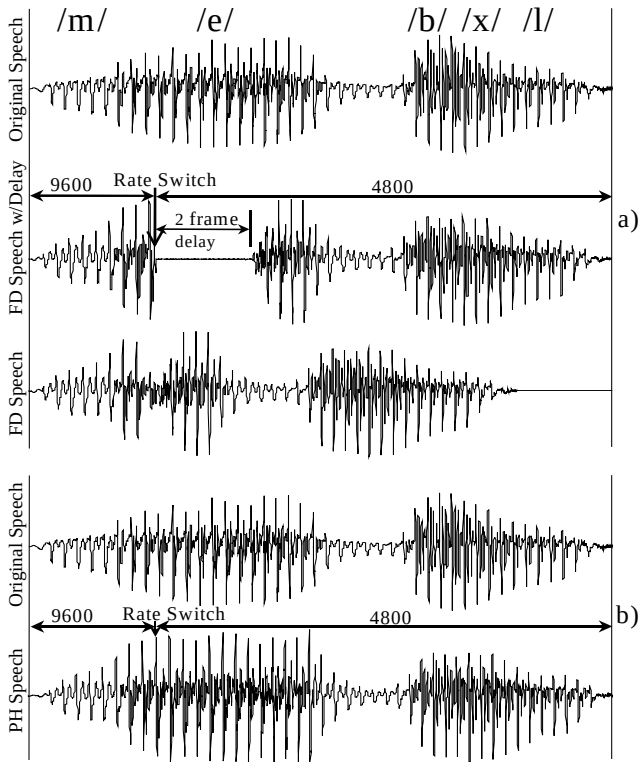

Figure 4: Resulting speech, a) FD and b) PH switching methods.

vocoder rate of 9.6 kb/s, with the same channel coding as utilized in the adaptive-rate system. The received SNR versus time characteristic of the channel is shown in Figure 5 for a fixed-rate transmitter. Note that for the adaptive-rate system, the received SNR can be written as $E_s/N_o = (C/N_o)(1/R_s)$, where C is the average received power, N_o is the noise spectral density, R_s is the modem symbol rate, and the transmitter power is fixed.

For both systems the short-time *SD*, SD_{st}, is averaged over a 3-to-5 frame window of a 30 second speech sequence. The ARS details can be found in [5]. We implemented the adaptive-rate system, which utilizes rate ½ channel coding, modem rates of 19.2/9.6/4.8/2.4 ks/s, and MRSTC rates of 9.6/4.8/2.4/1.2 kb/s. The SD_{st} results are plotted in Figure 6. It is clear that ARS's speech quality is superior to that of the FRS. Informal listening tests confirmed a large improvement in ARS speech quality compared to the FRS. The results in Figure 6 show that the increase in *SD* at a low values of E_s/N_o is due to an excessively large number of bit errors entering the vocoder, and is much greater than the increase in *SD* due to lower source encoding rates in the MRSTC. An increase in C/N_o system operating range, of greater than 9 dB, can then be achieved with the ARS, given that the FRS degrades rapidly below 0 dB E_s/N_o. *SD* for both systems, averaged over the 30-second sequence, was 9.5 dB and 0.6 dB, respectively.

4. CONCLUSIONS[#]

We have presented a multi-rate speech coding method providing improved performance over fixed-rate systems, allowing good quality speech in wireless systems operating in severely degraded channels and other systems experiencing variation in bandwidth demand. We maximize speech quality by jointly minimizing the distortion due to two parameters: (1) source-encoding rate and

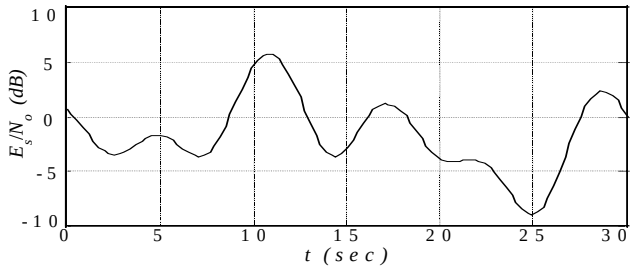

Figure 5: Channel model SNR characteristic versus time.

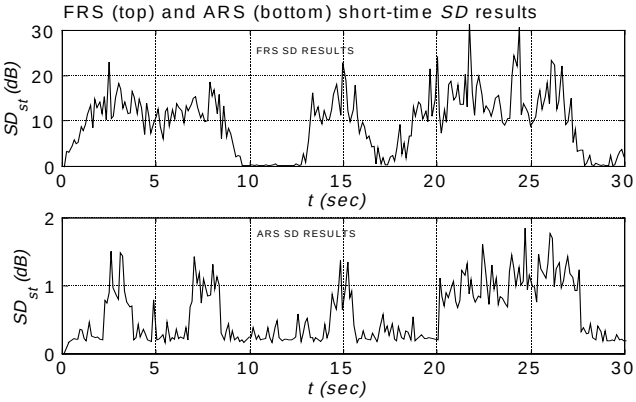

Figure 6: Short-time *SD* comparison between FRS and ARS.

(2) channel bit errors. Experimental results show a reduction of nearly 9 dB in average spectral distortion when the proposed adaptive-rate system is used in lieu of a fixed-rate system. We designed a technique to minimize end-to-end delay using a variable size/rate buffer, providing optimal performance in both Internet and wireless applications.

5. REFERENCES

[1] Ramachandran R. and Mammone R. *Modern Methods Of Speech Processing*, KAP, 1995.

[2] Yuen E., Ho P. and Cuperman V. "Variable rate speech and channel coding for mobile communication," *43rd IEEE/VTS Tech. Conf.*, pp. 1709 - 1713, 1994.

[3] Lupini P., Cox N. and Cuperman V. "A multi-mode variable rate CELP coder based on frame classification," in *Proceedings of ICC*, pp. 406 - 409, 1993.

[4] Xionwei Z., Lixin C. and Xianzhi C. "Real-time implementation of 4.8/3.6/2.4/1.2 kb/s high quality multi-rate speech coding on TMS320C31," in *Proceedings of ICCT*, Vol. 1, pp. 449 - 452, 1996.

[5] Kleider J.E. and Campbell W.M. "An adaptive-rate digital communication system for speech," in *Proceedings of ICASSP*, Vol. III, pp. 1695 - 1698, 1997.

[6] Dubnowski J.J. and Crochiere R.E. "Variable rate coding of speech," BSTJ, Vol. 58, No. 3, pp. 577 - 600, 1979.

[7] McAulay R.J. and Quatieri T.F., "Low-rate speech coding based on the sinusoidal model," Advances in Speech Processing, S. Furui and M.M. Sondi (Eds.), Marcel Dekker, Chapter 1.6, pp. 165 – 208, 1992.

[#] The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.