MULTIMODE VARIABLE BIT RATE SPEECH CODING: AN EFFICIENT PARADIGM FOR HIGH-QUALITY LOW-RATE REPRESENTATION OF SPEECH SIGNAL

Amitava Das, Andy DeJaco, Sharath Manjunath, Ananth Ananthapadmanabhan, Jeff Huang, Eddie Choy

Qualcomm Inc.
6455 Lusk Boulevard
San Diego, CA 92121, USA

ABSTRACT

The speech signal consists of a time-varying ensemble of different types of segments with distinct characteristics, which require different degrees of coding resolution in order to retain an overall high voice quality. A fixed-rate coder can capture such time-varying characteristics only if it operates at a high enough bit rate. At low bit rate, a fixed-rate coder will not be able to capture all of these various segments well and will fail to render high voice quality. A multimode variable bit rate (VBR) coder uses an arsenal of modes, operating at different bit rates. These modes are designed to represent these different speech segments optimally with the right amount of coding resolution. Thus, a multimode VBR codec adapts the coding mechanism to the input speech and delivers high quality at low (average) rates. This paper presents the essential framework and the unique advantages of a multimode VBR codec and suggests algorithms for the different modes.

1. INTRODUCTION

Low bit rate speech coding, particularly at rates of 4kbps and below, is motivated by the expanding needs of wireless and satellite-based communication systems as well as various network applications such as Internet telephony, multimedia communications, voice mail and other voice storage systems. The main driving force is the need for higher capacity. A number of recent standard activities, such as the ITU 4kbps codec standardization [1] are motivating research and development in low-rate speech coding. The challenge is to retain toll-quality at lower bit rate, unlike the lower "communication quality" earlier low-rate codec developments were shooting for.

Since speech signal is essentially a time-varying ensemble of different types of segments with distinct characteristics, it is important to study these segments closely to determine the most optimal way to represent them at low bit rate. It is well known that these different speech segments, such as voiced, unvoiced, transitions, do require different degrees of coding resolution in order to retain an overall high voice quality. A fixed-rate coder can capture such time-varying characteristics only if it operates at a high enough bit rate. At low bit rate, a fixed-rate coder will not be able to capture all of these various segments well and fail to render high voice quality.

A multimode variable bit rate (VBR) coder [2] uses an arsenal of modes (coding algorithms), operating at different bit rates. These modes are designed to represent these different speech segments optimally with the right amount of coding resolution. Thus, a multimode VBR codec can dynamically adapt the overall coding scheme to the input speech signal and deliver very high voice quality at low (average) rate.

In this paper, we focus on the unique advantages of the multimode VBR coding paradigm and illustrate how it is a very promising direction towards retaining toll quality at our target low rates. Finally, a number of techniques are suggested to design the modes of such a multimode VBR codec.

2. VBR SPEECH CODING FRAMEWORK

2.1 An Example VBR Speech Codec

To exemplify the concepts presented in this paper, we introduce a hypothetical VBR codec as a framework of our discussion. This codec applies the conventional Linear Prediction (LP) based coding scheme [3] shown in Figure 1. The LP Parameters as well as the LP residual are coded with one of the four modes: Full-Rate[FR], Half-Rate[HR], Quarter-Rate[QR] and the Eighth-Rate[ER] modes.

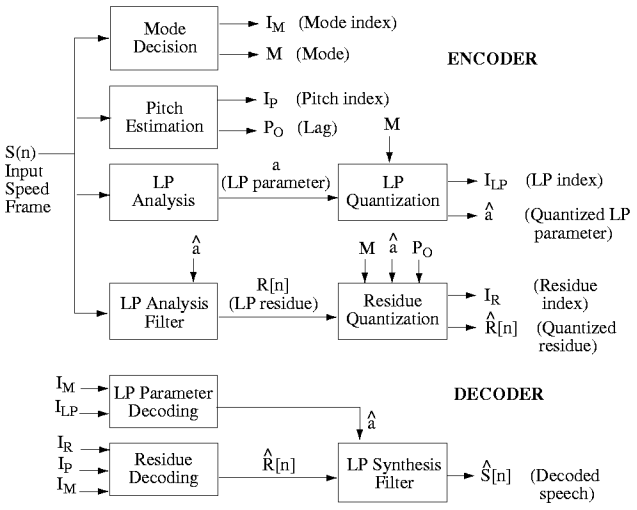

Figure 1: A linear prediction based multimode VBR codec.

2.2 Open-Loop Mode Selection

An open-loop mode decision mechanism analyzes the input speech and picks the best mode suitable to represent the characteristics of the incoming speech frame. Thus, the mode-selection mechanism can be completely source-controlled, adapting the codec to the incoming speech signal. The open-loop mode-selection can also be tailored to deliver a pre-selected rate-mixture suitable for a target rate-quality point. Table 1 shows a possible rate allocation to our modes and how different rate-mixtures will deliver various average-rates.

Mode	Rate (kbps)	Mix1	Mix2	Mix3	Mix4
Full-Rate	8	0.4	0.25	0.2	0.1
Half-Rate	4	0	0.5	0.1	0.2
QuarterRate	2	0	0.1	0.1	0.1
Eighth-Rate	1	0.6	0.6	0.6	0.6
Codec-Avg-Rate (kbps)		3.8	3.0	2.8	2.4

Table 1: Rates and rate-mixtures of the example VBR codec.

2.3 Closed-Loop Mode Selection

The performance of a VBR codec can be further enhanced and a guaranteed quality of service can be delivered by a closed-loop mode decision mechanism. In closed-loop mode decision [Figure 2] a suitable error measure is used to decide whether the selected lower-rate mode produced good quality speech or not, and if the performance is not satisfactory, a higher-rate mode is applied.

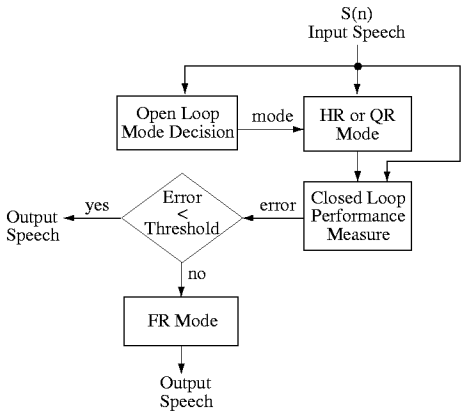

Figure 2: Closed-loop mode decision.

3. ADVANTAGES OF VARIABLE BIT RATE SPEECH CODING

3.1 Optimal Capture of Different Speech Segments with Different Characteristics

A close look at different types of speech segments (Figures 3-4) and the corresponding LP residual reveals that a multimode VBR coding is the natural and most optimal way to encode them. These different types of speech segments shown here have different amount of information content and therefore, to attain a target voice quality, they do require varying degree of encoding precision or different bit-rates. Therefore, customized modes with the appropriate bit rates, can be designed to represent these segments optimally.

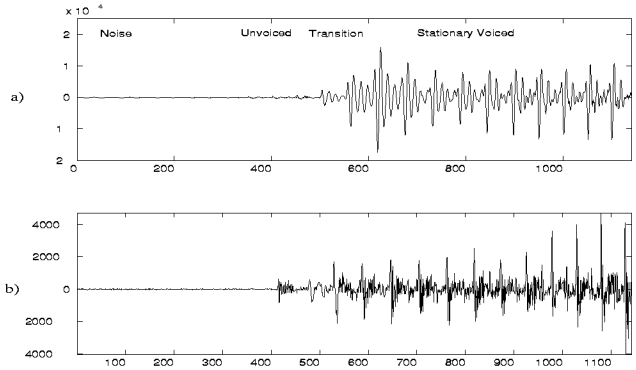

Figure 3: a) Speech segment; b) LP residue.

For example, the background noise and silence can be preserved well with our Eighth-Rate mode. The stationary voiced segments can be captured with a well-designed Half-Rate mode, tactfully exploiting the repetition of the pitch periods. Transitional segments, such as the unvoiced to voiced segment transition shown here, will lack the strong correlation between past and present frame, and need a Full-Rate mode with high amount of bits offering higher coding resolution. Unvoiced speech segments exhibit both a transient and stationary characteristics, and can be captured by a well-designed Quarter-Rate mode

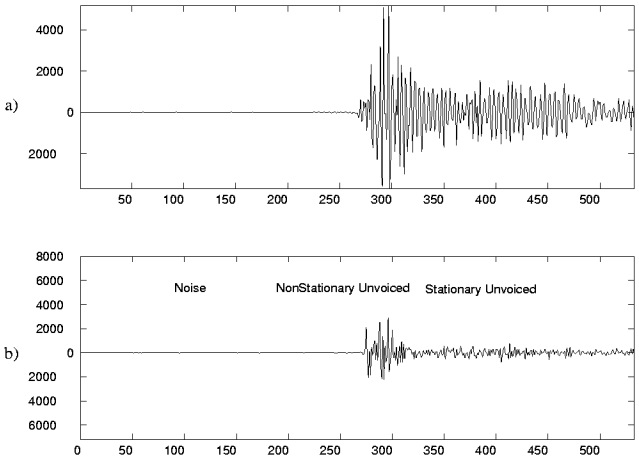

Figure 4: a) Unvoiced speech; b) LP residue.

Therefore, we see that VBR multimode coding can capture the dynamics of the speech signal in an efficient manner, employing only the required amount of coding resolution.

Figure 5 shows the variation of the rates of a source-controlled VBR codec employing the rate-set of Table 1. Note how sparsely the Full-Rate mode is used (only when it is needed). As a result, the average rate will be significantly lower than the full rate, while high voice-quality, similar to the quality of a fixed-rate coder operating at the full-rate can be attained.

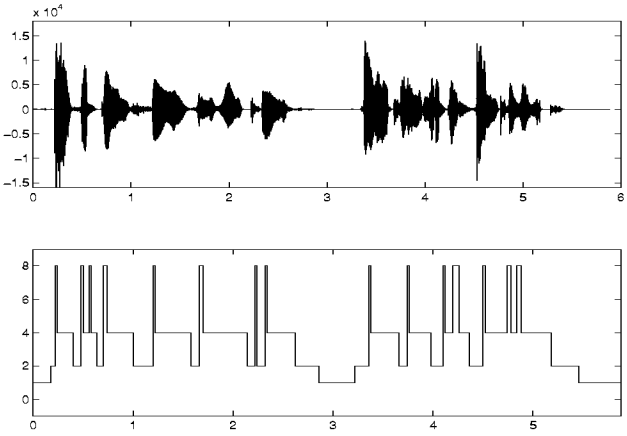

Figure 5: a) Original speech signal b) Corresponding rate variation of a source-controlled VBR coder.

3.2 Equivalent Voice Quality of a Fixed Full Rate Codec at Lower Average Rate

The rate-mixture 1 of Table 1 shows that the separation of incoming signal into active speech and non-speech segments (background noise or silence), and the subsequent application of the full-rate mode to active speech and the eighth-rate mode to noise, create a VBR codec with significantly less average rate. This 3.8 kbps VBR codec will deliver the same voice quality as the 8 kbps full-rate codec. Judicious application of Half-Rate and Quarter-Rate modes in active speech will reduce the percent of use of the Full-Rate, and reduce the average rate further without any significant loss of quality.

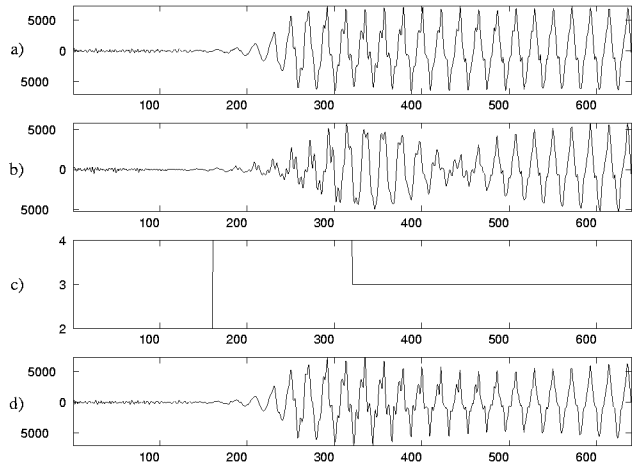

Figure 6: a) Original speech; b) Output of a fixed 4 kbps codec, c) and d) Rate variations and output speech of a VBR codec.

3.3 Improved Voice Quality at Equivalent Average Low Rate

Due to the varying nature of speech and the need for the varying amount of coding precision, a fixed low rate (say, 4 kbps) speech coder will have extreme difficulty to offer a robust and high-quality performance. Such a codec will fail to capture the transition regions well (Figure 6a) which will degrade voice quality. Being a low-rate implementation, the coding mechanism will have to rely on some kind of prediction scheme, employing some form of memory (past quantized frame) and differential coding. Therefore, poor capturing of any high-information-content frame will result in the performance degradation of subsequent frames (Figure 6b). The prediction mechanism will suffer from poor memory and the accumulative effect will cause further degradation.

The injection of Full-Rate modes during such high information content segments will not only improve the codec performance during those segments, but also improve the performance of future prediction-based Half-Rate modes (Figure 6d). Such injections of full-rate modes are also extremely helpful in situations when a prolonged application of the prediction-based half-rate mode lowers the performance below some minimum level of acceptability. Note that such injections of Full-Rate mode, at a first glance, may seem to increase the average rate a bit. However, judicious application of the Quarter-Rate and Eighth-Rate modes can compensate for the rate-increase imposed by the Full-Rate injections. Therefore, it is possible to design a multimode VBR codec, which will operate on the same low 4 kbps (average) rate, but will offer voice quality significantly higher than the 4 kbps fixed-rate coder.

3.4 The Quality-Rate Dial

The different rate-mixtures of a VBR codec (Table 1) effectively produce a Quality-Rate Dial, or a virtual set of codecs packed in a single codec implementation, operating at different (average) rates with different voice quality. Figure 7 shows such a Quality-Rate Dial, offered by the TIA-standard IS733 VBR codec [4].

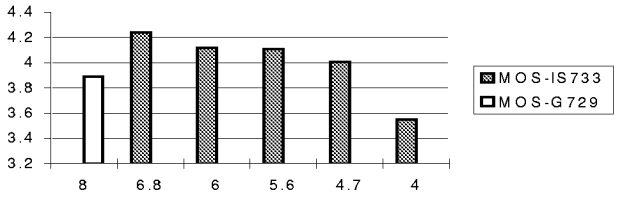

Figure 7: The quality-rate dial: MOS scores of the different rate mixtures (different average-rates) of the IS733 VBR Codec.

Such a Quality-Rate dial, a useful byproduct of the VBR coding paradigm, enables applications (such as video-conferencing) to dynamically reallocate bandwidth to the audio channel, while using a single codec.

4. SUGGESTED ALGORITHMS FOR THE MODES OF THE EXAMPLE VBR CODEC

In this section, we look at possible algorithms to design the modes of our hypothetical VBR codec. For the Full-Rate mode, conventional CELP algorithm, such as the Full-Rate mode of the IS-127 VBR codec [5] can be used.

For the Quarter-Rate mode, a high-time-resolution, coarsely-quantized, gain-shape representation of the unvoiced residual can be used. In such a scheme, the decoded LP residue is generated

by shaping a random noise segment with a gain contour formed by interpolating a set of quantized gain parameters. The large number [10-20] of gain parameters are extracted from a similar number of sub-frames of the original residue. These gains are very coarsely quantized with the low amount of bits available. The resulting high time-resolution and coarse preservation of the energy dynamics of the LP residue seems sufficient to retain the perceptual quality of such segments. The Eighth-Rate can be designed by a similar but scaled down version of the Quarter-Rate mode.

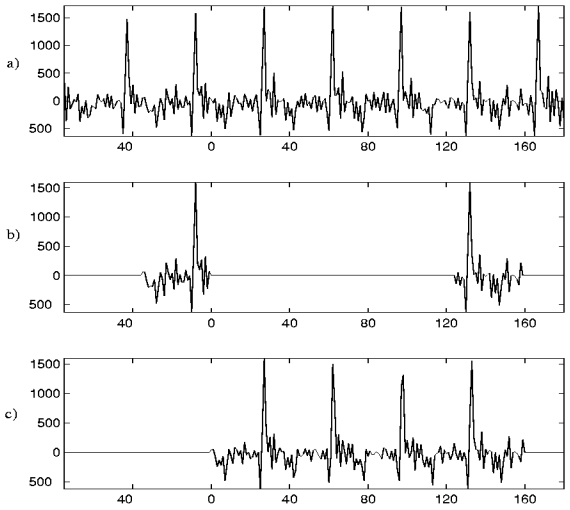

Figure 8: a) Original LP residue; b) Current and past prototypes; c) LP residue reconstructed by TSWI.

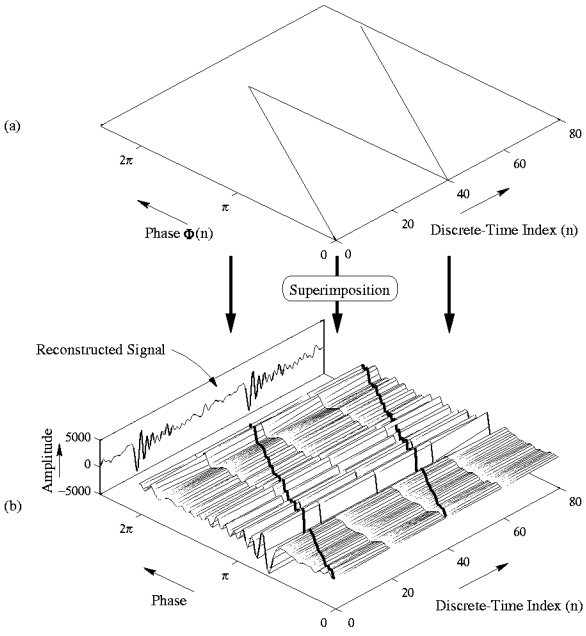

Figure 9: a) Cubic phase contour (wrapped); b) TSWI synthesis: re-sampling of the interpolated 2-D CW surface with the proposed phase track to create the 1-D reconstructed residue.

The most challenging task is to design a suitable Half-Rate mode that captures stationary voiced speech segments well at 4kbps. One technique we are currently investigating is the Time-Synchronous Waveform Interpolation [TSWI]. TSWI is a variation of the Waveform Interpolation [WI] scheme [6], in which instead of a number of prototypes (as in WI) only one waveform is extracted for each frame (Figure 8a). Unlike, conventional WI scheme, TSWI attempts to preserve the time-synchrony between the original and the reconstructed signals by applying a novel cubic-based phase track during the conversion from the 2-D characteristic waveform surface to the 1-D signal as shown in Figure 9b. Such a phase contour can be written as:

$$\Phi(n) = \alpha n^3 + \beta n^2 + \chi n \ + \delta$$

where n is the discrete-time index and the coefficients $\{\alpha, \beta, \chi, \delta\}$ can be derived from the pitch values, the initial phase offset and also the alignment shifts introduced to the extracted pitch prototypes.

5. CONCLUSIONS

We presented the essential framework of a multimode variable bit rate speech coder, highlighted the important features, and presented the unique advantages that only the multimode VBR coding paradigm can offer. One attractive feature of multimode VBR coding is the Quality-Rate Dial which effectively delivers a virtual set of codecs operating at different quality-rate points. A multimode VBR codec can deliver equivalent voice quality of a full-rate codec at significantly lower (average) rate. It can also deliver significantly higher voice quality than a fixed low rate codec at equivalent average rate. Multimode VBR coding is, therefore, the most promising direction, perhaps the only way, to retain toll quality at low (4kbps and below) rates.

6. REFERENCES

[1] ITU SG16/Q21: Terms of reference for the 4 kbps codec, September 19, 1998

[2] A. Das, E. Paksoy, and A. Gersho, "Multimode and Variable-Rate Coding of Speech", in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds., Chapter 7, pp 257-288, Elsevier, 1995

[3] W. B. Kleijn and K. K. Paliwal, "An Introduction to Speech Coding", in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds., Chapter 1, pp 1-47, Elsevier, 1995

[4] TIA/EIA/IS-733, High Rate Speech Service Option for Wideband Spread Spectrum Communication Systems

[5] TIA/EIA/IS-127, Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, as amended by IS-127-1, Addendum to IS-127

[6] W.B. Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis", in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds., Chapter 5, pp 175-208, Elsevier, 1995