EVRC-WIDEBAND: THE NEW 3GPP2 WIDEBAND VOCODER STANDARD

Venkatesh Krishnan, Vivek Rajendran, Ananthapadmanabhan Kandhadai, Sharath Manjunath

QUALCOMM Incorporated,
5775 Morehouse Drive, San Diego, CA 92121

ABSTRACT

In this paper, the latest wideband vocoder standard adopted by the cdma2000 standardization body, 3GPP2, is described. Christened Enhanced Variable Rate Codec- Wideband (EVRC-WB), the proposed codec encodes wideband speech (16 KHz sampling frequency) at a maximum bit-rate of 8.55 kbit/s. EVRC-WB is based on a split band coding paradigm in which two different coding models are used for the signal components in the low frequency (LF) (0-4 KHz) and the high frequency (HF) (3.5-7 KHz) bands. The coding model used for the former is based on the EVRC-B narrowband (0-4 KHz) codec, modified to encode the LF band signal at a maximum bitrate of 7.75 kbit/s. The HF band coding model is a LPC based coding scheme where the excitation is derived from the coded LF band excitation using non-linear processing. Mean opinion scores from 3GPP2 characterization tests are provided to demonstrate that the EVRC-WB codec (8.55 kbit/s, max.) performs statistically significantly better than the Adaptive Multirate Wideband (12.65 kbit/s, max.).

Index Terms— Wideband, Speech Coder, Non-linear Processing .

1. INTRODUCTION

In traditional wired and wireless communication systems, vocoders are designed to encode speech signals that are bandlimited to the narrow-band (NB) frequency range of 300 Hz to 3400 Hz. In CDMA wireless communication systems, the Enhanced Variable Rate Codec (EVRC) [1], standardized by the cdma2000 standardization body, 3GPP2, in 1996 is the most commonly used speech coding scheme. The EVRC vocoder encodes narrowband speech on a frame-by-frame basis at a bit-rate of either 8.55, 4.0, or 0.8 kbit/s depending on the characteristics (voiced and unvoiced, transients, or silence) of the current speech frame (20ms in duration). In May 2006, a new source-controlled narrowband variable rate vocoder called EVRC-B [2] that encodes a 20 ms speech frame at one of 8.55, 4.0, 2.0 or 0.8 kbit/s bit-rate (designated full, half, quarter and eighth rate modes, respectively) modes was standardized. While EVRC was designed to yield high quality narrowband coded speech at low bit-rates, EVRC-B was designed to have a multitude of operating points to enable wireless operators to achieve a desired capacity gain with a small, associated trade-off in the coded speech quality.

The bandwidth limitation in traditional telephony networks is largely due to the presence of PSTN sub-systems that are inherently narrowband. As wireless voice communication systems continue to evolve, networks are moving towards transcoder-free operation where no PSTN subsystem is involved. Additionally, IP Core networks and Voice over IP on wireless networks are becoming increasingly popular. These, coupled with advances in the electro-acoustic capabilities on wireless devices, have broadened the scope for employing speech coding techniques that encode wideband speech signals, covering the larger frequency range from 50 Hz to 7 or 8 kHz. A significant improvement in perceived speech quality and intelligibility can be obtained by encoding the wideband speech compared to the narrowband speech signal.

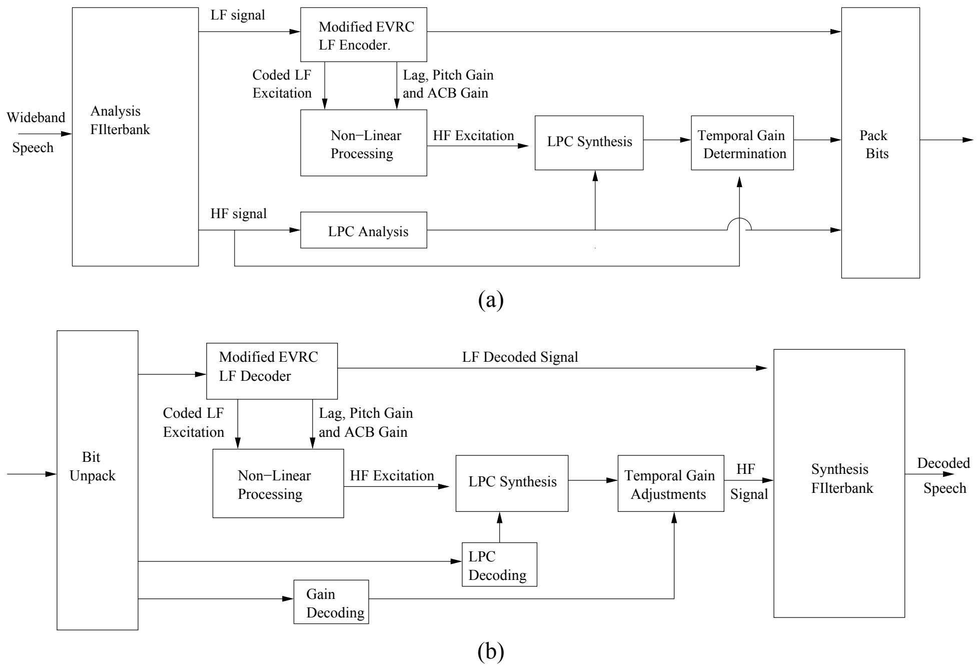
To enable wideband voice communication services in cdma2000 systems, the 3GPP2 standardization organization recently approved the EVRC-WB algorithm [3]. EVRC-WB met the 3GPP2 quality and average data-rate requirements viz: (1) encode wideband speech at a maximum bit-rate of 8.55 kbit/s (imposed by physical layer constraints in CDMA systems) while achieving quality equivalent to Adaptive Multi-Rate Wideband (AMR-WB) mode 2 [4] which operates at higher bit-rates of 12.65 kbit/s and (2) facilitate (reduced complexity/ memory) interoperability with network components that are bandlimited to the narrowband frequency range of 0-4 KHz (eg: interoperability during mobile to land line calls).

The EVRC-WB is based on a split-band coding approach in which the wideband input speech signal is separated into a low frequency (LF) band (0-4 KHz) signal and the high frequency (HF) band (3.5-7 KHz) signal using an analysis filterbank. The LF signal is encoded using an appropriate coding mode from the EVRC-B coding modes, modified to free up bits for the HF band signal coding. The HF signal is encoded using a LPC based coding scheme where the excitation is derived from the coded LF band excitation using non-linear processing. The parameters transmitted corresponding to the HF signal include a set of LSP coefficients, and a set of gain parameters that are obtained by comparing the input HF signal and the HF excitation (derived by non-linear processing) filtered by the LPC synthesis filter. Unlike AMR-WB where a single coding model is used to encode the entire wideband speech bandwidth, the HF signal coding scheme in EVRC-WB exploits the correlation between the LPC residuals of the LF and HF signals to enable encoding the HF signal with very few bits. Further the split band approach facilitates (reduced complexity/ memory) interoperability between wideband and narrowband components of a network.

The architecture of EVRC-WB is described in detail in Sec. 2. The mean opinion score (MOS) results from the 3GPP2 characterization test that compare EVRC-WB with vocoders such as AMR-WB (at bit-rate of 12.65 kbit/s) are provided in Sec. 3. The conclusions are provided in Sec. 4.

2. EVRC-WB ARCHITECTURE

The EVRC-WB encoder and decoder architectures are shown in Fig. 1. The input signal to the encoder is wideband speech sampled at 16 kHz. An analysis filterbank generates a LF band signal sampled at 8 kHz, and a HF band signal which is sampled at 7 kHz. This HF band signal represents the band from 3.5 to 7 kHz in the input signal, and the final decoded wideband signal is limited in bandwidth to 7



(a)

(b)

Fig. 1. The EVRC-WB architecture (a) Encoder and (b) Decoder

kHz. Note that there is a 500 Hz overlap between the low and high bands, allowing for a more gradual transition between the bands.

EVRC-WB encodes voiced frames using 171 bits (designated full-rate), unvoiced frames using 80 bits (half-rate) and silence frames using 16 bits (eighth-rate). The distribution of bits between the LF band signal and the HF band signal codecs is shown in Table 1.

2.1. LF (0-4 KHz) band signal coding

In EVRC-WB, the LF band signal is encoded using either the code-excited linear predictor (CELP), noise-excited linear predictor (NELP), or silence coding modes of the EVRC-B codec with minor modifications to free up bits for encoding the HF band signal. Details of EVRC-B can be found in [2]. In this section, a brief description of the EVRC-B coding modes is provided and the modifications to the coding modes in the context of EVRC-WB are highlighted.

The EVRC-B narrowband vocoder encodes voiced, unvoiced and silence frames of a speech signals using the following coding modes:

2.1.1. Coding Voiced Frames

Voiced frames are encoded using the Relaxed CELP algorithm [5]. The short term prediction residuals obtained from LPC analysis on a given frame of voiced speech are time warped to conform to a simplified pitch contour. The modified prediction residuals are then encoded every 6.67 ms (thrice in a frame) using contributions from the adaptive codebook (ACB) and a fixed codebook (FCB) [2]. The changes incorporated into the EVRC-B full-rate algorithm in the context of wideband coding include: (a) the number of fixed codebook pulses per subframe has been reduced from 7 to 6 pulses, or 5 pulses and (b) the delta lag (the difference between current and previous pitch lags, used after a frame erasure) is no longer transmitted. With these modifications, every voiced frame (20 ms duration) of LF band speech is encoded using 155 bits. For a detailed description, interested readers are referred to [2].

2.1.2. Coding Unvoiced Frames

LF signal frames classified as “unvoiced” are encoded using the Noise-Excited Linear prediction (NELP) model. A pseudo-random noise is colored using a set of filters, and scaled appropriately by gains in 10 subframes of 2 ms each to model the excitation. The use of a 28-bit line spectral frequency (LSF) vector quantization (VQ) (same as in voiced frames) and some minor filtering changes were made to the EVRC-B NELP encoding scheme in the context of wideband coding.

2.1.3. Coding Silence Frames

Silence encoding is similar to that of EVRC-B eighth-rate coding mode except for a new 10-bit LSF VQ and an energy quantization.

2.2. HF (3.5-7 kHz) band signal coding

In this section the high-band coding method employed by EVRC-WB is described. The coder provides high-quality reconstruction of the HF band signal at bit-rate approx. 1 kbit/s. The 7 KHz-sampled high-band input signal is analyzed to produce parameters that, together with signals from the narrowband coder, are used to reconstruct the high band. The parameters transmitted corresponding to the HF band signal represent the spectral and temporal envelope of the HF band signal. The components of the high-band encoder used in encoding voiced and unvoiced frames are described in detail below. The wideband band silence coding scheme follows the same split-band approach used for voiced/ unvoiced frames, except that no bits are spent on encoding the high band for silence frames (Table 1). Instead, temporal extrapolation is done at the decoder to synthesize the high band signal using smoothed gain and LSF parameters from some unvoiced frames which are selected as “good” background noise descriptors.

2.2.1. LPC analysis and coding

Six spectral parameters are obtained by the traditional LPC analysis. For efficient coding, these prediction coefficients are converted to LSFs and quantized using VQ. The quantization index is transmitted to the decoder. The LSF quantizer is improved by using temporal noise shaping on the encoder side, thus reducing spectral fluctuations without additional delay. The noise shaping works as follows. For each frame, the LSF quantization error vector is computed and multiplied by a scale factor less than 1.0. The following frame, this scaled quantization error is added to the LSF vector before quantization. The scale factor is adjusted dynamically depending on the amount of fluctuations already present in the unquantized LSF vectors: when the difference between the current and previous LSF vectors is large, the scale factor is close to zero and almost no noise shaping results. When the current LSF vector differs little from the previous one, the scale factor is close to 1.0. The resulting LSF quantization minimizes spectral fluctuations when the speech signal is relatively constant from one frame to the next. The number of bits used to encode the HF band LSFs for voiced and unvoiced modes of operation is provided in Table(1).

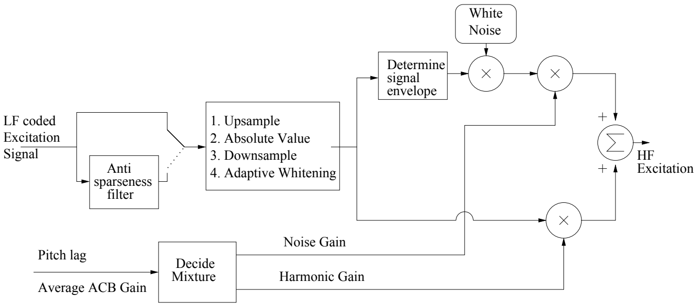


Fig. 2. Generation of HF Band Excitation

2.2.2. High-Band excitation generation using non-linear processing

The high-band excitation signal is derived from the coded low-band excitation as provided by the narrowband coder. To preserve the harmonic structure of the LF excitation signal in the HF excitation signal, a non-linear function (absolute value) is used [6]. The nonlinear function is applied after sufficiently over-sampling the narrowband signal in order to minimize aliasing. Fixed or adaptive whitening can be applied to the output of the nonlinear function to flatten the spectrum. The main drawback of the nonlinear function is that for many voiced speech signals, the lower frequencies exhibit a stronger harmonic structure than higher frequencies. As a result, the output of the nonlinear function can lead to a HF excitation signal that is too harmonic, leading to objectionable, ‘buzzy’-sounding artifacts. As a solution, a combination of a nonlinear function and noise modulation is used to produce a pleasantly-sounding high-band signal.

Fig. 2 depicts the process that generates the high-band excitation from the narrowband excitation signal. The excitation signal may be first run through an all-pass filter. This filter reduces the sparseness that results from encoding the low-band signal with a sparse fixed codebook, and is intended to be used during unvoiced speech. The decision to use the output of the filter or the original input signal may be derived from the quantized narrowband adaptive codebook gain and first reflection coefficient: unvoiced signals are characterized by a low adaptive codebook gain and a first reflection coefficient that is close to zero or positive (indicating a flat or upwards-tilted spectrum).

	LF signal coder	HF signal coder	Total
Voiced	155	16	171
Unvoiced	53	27	80
Silence	16	0	16

Table 1. Bit allocation between the LF and HF band coders in EVRC-WB

Parameters	Voiced	Unvoiced
LSP	8	12
Gain	4	8
Gain Frame	4	7
Total	16	27

Table 2. Bit allocation for the high-band encoder

Next is the nonlinear function. This module up-samples the signal to 64 kHz, takes the absolute value, and then down-samples it to 16 kHz. From here, a 7 kHz-sampled signal is produced using the same high-band analysis filter that was used to split the input signal in a low and a high band. The result is spectrally flattened with an adaptive 4th order linear prediction filter, to create the harmonically-extended excitation signal.

A modulated noise signal is generated by multiplying a unit-variance white noise signal with the envelope of the harmonically-extended excitation signal. This envelope is obtained by taking the squared value of each sample, smoothing with a first order IIR low-pass filter and taking the square-root of each smoothed sample.

The modulated noise and harmonically-extended excitation signals are now mixed together to create a signal with the right amount of harmonic and noise contents. The noise gain G_{Noise} is determined from the coded pitch-lag P_L and average ACB gain (P_G) parameters obtained from the LF signal encoder according to

$$G_{Noise} = 1 - \frac{P_G \times P_L}{60 + P_L}. \qquad (1)$$

P_G is set to 0 for unvoiced frames. The harmonics gain, G_{Harm} is derived from the noise gain as $G_{Harm} = \sqrt{1 - G_{Noise}^2}$, so that the energy of the output signal is not affected by changing the gains.

2.2.3. Gain parameter determination

To effectively reconstruct the temporal envelope of the HF band signal, a gain-frame and five temporal gain parameters are encoded. The gain-frame parameter is obtained by matching the energy of the input high-band signal with the unscaled, spectrally shaped (using LPC synthesis filter) high-band excitation signal. The five temporal gain parameters are found by comparing the energies in each 4 ms subframe of the input high-band signal with the spectrally shaped high-band excitation signal, scaled by the gain-frame parameter. For voiced and unvoiced frames, the gain-frame parameter is scalar quantized. The gain-frame parameters are vector quantized by a single stage VQ in voiced frames and a multistage (4+4 bits) for unvoiced frames. The number of bits used to encode the HF band gain parameters for voiced and unvoiced modes of operation is provided in Table (2).

It must be noted that the HF band excitation is generated in both encoder and decoder. The HF band encoder and the decoder may be

Codec	ASR	NL		LL		HL		1%		2%		3%		6%		1%DB		**Mean**
(kbit/s, max.)	(kbit/s)	MOS	SD	MOS	SD	MOS	SD	MOS	SD	MOS	SD	MOS	SD	MOS	SD	MOS	SD	**of MOS**
EVRC-WB (8.55)	7.42	4.08	0.78	4.01	0.76	3.86	0.74	3.87	0.86	3.73	0.93	3.54	0.88	3.15	0.92	3.97	0.80	**3.77**
AMR-WB (12.65)	12.65	4.13	0.76	4.09	0.79	3.87	0.81	3.88	0.87	3.65	0.95	3.33	0.91	2.91	0.98	3.81	0.83	**3.71**

Table 3. Extract from 3GPP2 characterization MOS test for clean input signal. SD refers to standard deviation and ASR to active speech bit-rate. The test conditions include NL (Nominal level: signal level at -22 dB); LL (Low level: signal level at -32 dB); HL (High level: signal level at -12 dB); 1,2,3, and 6% frame erasure rates and; 1%DB where the system experiences 1% dim-and-burst signaling and 1% packet-level dimming.

Codec	ASR	CN-10		CN-20Er		SN-15		OB-10Er		**Mean**
(kbit/s, max.)	(kbit/s)	MOS	SD	MOS	SD	MOS	SD	MOS	SD	**of MOS**
EVRC-WB (8.55)	7.42	3.15	0.73	3.74	0.86	3.29	0.78	3.23	0.83	**3.35**
AMR-WB (12.65)	12.65	3.22	0.65	3.53	0.83	3.13	0.75	2.91	0.81	**3.20**

Table 4. Extract from 3GPP2 characterization MOS test for noisy input signals. SD refers to standard deviation and ASR to active speech bit-rate. The test conditions including CN-10 (car noise at 10 dB wrt to the signal), SN-15 (street noise at 15 dB), and two condition (CN-20ER and OB-10Er) where the input signal is corrupted by 20 dB car noise and 10 dB office babble noise respectively in addition to the system experiencing 2% frame erasures and 2% dim and burst signaling.

designed to generate the identical HF band excitation by (1) using the quantized narrowband excitation, which is available to both encoder and decoder, and (2) having the same state for the random white noise generator in encoder and decoder at all times. The latter can be assured, for instance, by making the state of the random generator a deterministic function of information coded earlier within the same frame, such as the narrowband bit stream.

3. 3GPP2 CHARACTERIZATION MOS RESULTS

In this section, the results from the 3GPP2 characterization testing of EVRC-WB based on the Mean Opinion Score (MOS) are provided. The purpose of this test was to verify the quality of EVRC-WB (8.55 kbit/s, max.) and compare it to AMR-WB (12.65 kbit/s), in clean, noisy, and frame error conditions.

An extract form the 3GPP2 characterization test report for clean and noisy input signal conditions is provided in Table 3 and 4, respectively. The test coders, EVRC-WB(8.55 kbit/s max) and AMR-WB (12.65 kbit/s) were evaluated for their performance given clean input signal at three different signal levels and various noisy input signals. In AMR-WB, discontinuous transmission (DTX) [4] was enabled. The ability of these coders in handling frame error conditions, dim-and-burst signaling and packet-level dimming is also evaluated. Frame erasures occur when the coded bit-stream corresponding to a frame of speech fails to arrive at the decoder due to losses in the wireless communication system. Dim-and-burst signaling refers to a multiplexing technique used on traffic channels to send overhead signaling in which a less-than- 8.55 kbit/s bit-rate frame of speech data and overhead are combined and transmitted at a 8.55 kbit/s rate.

The 3GPP2 standards body required the new wideband codec (8.55 kbit/s, max.) to match the AMR-WB (12.65 kbit/s) in performance. For the 8 clean test conditions (three input signal levels: nominal, low, and high; 1%, 2%, 3% and 6 % frame erasure rates and 1% dim and burst signalling condition), ANOVA (analysis of variance)-based Global ToR analysis [7] using the MOS scores presented in Table 3 indicated that EVRC-WB at max. bit-rate of 8.55 kbit/s was significantly (statistically) better than AMR-WB at a bit-rate of 12.65 kbit/s . This satisfied the 3GPP2 requirement. A similar ANOVA test performed for the noisy input signal conditions (car noise at 10 dB SNR, car noise 20 dB SNR with 2% FER and 2% dim-and-burst, street noise at 15 dB SNR, office babble @ 20 dB SNR with 2% FER and 2% dim-and-burst) using MOS scores in Table 4 concluded that that EVRC-WB (8.55 kbit/s, max.) was significantly (statistically) better than AMR-WB (12.65 kbit/s.).

4. CONCLUSIONS

In this paper, the architecture of the new 3GPP2 wideband vocoder standard, EVRC-WB is described. EVRC-WB encodes wideband input speech signal using a split band approach where the low frequency band (0-4000 Hz) signal is independently encoded and the high frequency band (3500-7000 Hz) is encoded using information from the LF band encoder. 3GPP2 characterization test results are provided to demonstrate that EVRC-WB at a maximum bit-rate of 8.55 kbit/s out-performs the 3GPP standard codec AMR-WB at a maximum bit-rate of 12.65 kbit/s.

5. REFERENCES

[1] 3GPP2 C.S0014-A, "Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems," http://www.3gpp2.org/public_html/specs/C.S0014-A_v1.0_040426.pdf, May 2004.

[2] 3GPP2 C.S0014-B, "Enhanced Variable Rate Codec, Speech Service Option 3 and 68 for Wideband Spread Spectrum Digital Systems," http://www.3gpp2.org/public_html/specs/C.S0014-B_v1.0_060501.pdf, May 2006.

[3] 3GPP2 C.S0014-C v1.0, "Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems," http://www.3gpp2.org/public_html/specs.

[4] 3GPP TS 126.190, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions," http://www.3gpp.org/ftp/Specs/html-info/26190.htm, Dec 2004.

[5] W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP speech coding algorithm," *European Transactions on Telecommuncations*, vol. 4, no. 5, pp. 573–582, 1994.

[6] J. Makhoul and M. Berouti, "High frequency regeneration in speech coding systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, pp. 428–431.

[7] H .R .Lindman, *Analysis of variance in complex experimental designs*, W. H. Freeman & Co, San Francisco, CA, 1974.