ON THE ARCHITECTURE OF THE CDMA2000® VARIABLE-RATE MULTIMODE WIDEBAND (VMR-WB) SPEECH CODING STANDARD

Milan Jelinek[†], *Redwan Salami*[‡], *Sassan Ahmadi*[*], *Bruno Bessette*[†], *Philippe Gournay*[†], *Claude Laflamme*[†]

[†]University of Sherbrooke, Sherbrooke, Quebec, Canada
[‡]VoiceAge Corporation, Montreal, Quebec, Canada
[*]Nokia Inc., San Diego, CA, USA

ABSTRACT

Description and design of the source-controlled variable-rate multimode wideband (VMR-WB) codec recently selected by the 3rd Generation Partnership Project 2 (3GPP2) for the cdma2000® system in Rate-Set II are presented. This paper gives an overview of the codec and the methodologies that enabled high quality wideband coding at average data rates ranging from TIA/EIA/IS-733 ADR to that of TIA/EIA/IS-127. The codec has three modes of operation at different average data rates and a fourth mode that is interoperable with 3GPP/AMR-WB (ITU-T/G.722.2). Despite the interoperability constraint, the codec was capable of meeting the aggressive performance requirements through the use of novel techniques such as noise suppression, efficient signal classification, new coding types optimized for stable voiced and unvoiced frames, novel post-processing technique for periodicity enhancement in the lower frequency band, and improved frame erasure concealment mechanisms.

1. INTRODUCTION

The variable-bit-rate (VBR) speech coding concept is crucial for optimal operation of the CDMA systems and interference control. In source-controlled VBR coding, the codec operates at several bit rates, and a rate selection mechanism is used to determine the bit rate suitable for encoding each speech frame based on the characteristics of the speech frame (e.g. voiced, unvoiced, transient, background noise). The goal is to attain the best speech quality at a given average data rate (ADR). The codec can operate in different modes by tuning the rate selection algorithm to obtain different ADRs where the codec performance is improved by increasing ADR. The mode of operation is imposed by the system depending on network capacity and the desired quality of service. This enables the codec with a mechanism of trade-off between speech quality and system capacity. In CDMA systems (e.g. cdmaOne and cdma2000), 4 bit rates are used and they are referred to as full-rate (FR), half-rate (HR), quarter-rate (QR), and eighth-rate (ER). In this system two rate sets are supported referred to as Rate-Set I and Rate-Set II. In Rate Set II, a variable-rate codec with rate selection mechanism operates at source-coding bit rates of 13.3 (FR), 6.2 (HR), 2.7 (QR), and 1.0 (ER) kbit/s, corresponding to channel bit rates of 14.4, 7.2, 3.6, and 1.8 kbit/s inclusive of error protection bits.

In January 2002, a process started in 3GPP2 for standardizing a variable-rate multi-mode wideband codec compliant with CDMA2000® Rate-Set II requirements. High-level specification for the CDMA2000® wideband codec required three modes of operation: a Premium mode with ADR equivalent to that of QCELP13 (IS-733), an Economy mode with ADR equivalent to that of EVRC (IS-127), and a standard mode with an ADR at the average of EVRC and QCELP13 ADRs. The quality requirements were such that the Premium mode is equivalent to AMR-WB at 14.85 kbit/s, the Standard mode to AMR-WB at 12.65 kbit/s, and the Economy mode to AMR-WB at 8.85 kbit/s. An optional mode Interoperable with AMR-WB was strongly recommended [4]. Following a highly competitive and stringent selection process including five candidate codecs in early 2003, the codec from Nokia/VoiceAge was selected.

This paper presents the description of the variable-rate multimode wideband (VMR-WB) codec selected by 3GPP2 for wideband speech services in the cdma2000® system. The codec has three modes of operations as described above with an additional mode that is interoperable with AMR-WB. The paper is organized as follows. Section 2 describes the VMR-WB codec design and structure and Section 3 gives the details of different coding types. Section 4 describes the VMR-WB decoder. Finally, Section 5 summarizes the codec performance.

2. VMR-WB CODEC DESIGN AND STRUCTURE

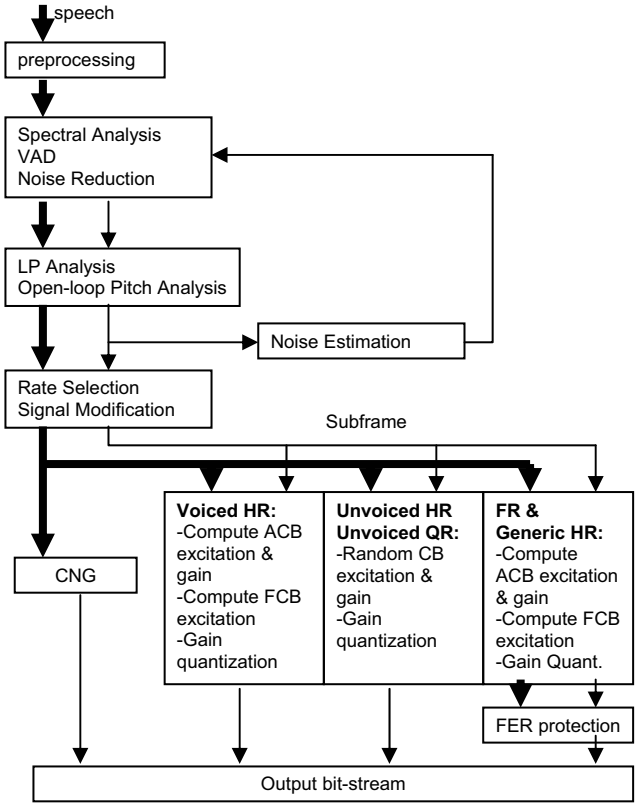
The full-rate (FR) coding types are based on the AMR-WB standard codec at 12.65 kbit/s [1,2,3]. This enables the design of a variable bit rate codec for the CDMA system capable of interoperating with other systems using the AMR-WB codec standard. Extra 13 bits per frame are added to fit in the 13.3 kbit/s full-rate of CDMA Rate-Set II. These bits are used to improve the codec robustness in case of erased frames and make essentially the difference between Generic FR and Interoperable FR coding types.

The codec operates on 20 ms speech frames. While its design has been focused on WB input, WB output signals, VMR WB accepts narrow-band (NB) input signals sampled at 8 kHz and it can also synthesize NB speech. The codec look-ahead varies depending on the sampling frequency of the input and the output being 13.75 ms for WB input, WB output and 15.0625 ms for NB input, NB output. The higher delay for NB processing is necessary for a sharper low-pass filter to avoid aliasing when down-sampling the synthesis signal from 12.8 kHz, VMR internal sampling rate, to 8 kHz at the output. This is important in particular for situations when processing a WB input, but synthesizing a NB output.

The flow diagram of the VMR-WB codec is shown in Figure 1. The input signal is preprocessed as in AMR-WB (down-sampling, high-pass filtering, and pre-emphasis). Spectral analysis is then performed on the preprocessed signal for use in noise reduction and voice activity detection (VAD). LP analysis

and open-loop pitch analysis is performed on a frame basis on the denoised signal. Then signal classification is performed to detect unvoiced frames (to be coded with Unvoiced HR or Unvoiced QR). A signal modification scheme that incorporates stable voiced signal classification is then applied to determine if the signal can be encoded using Voiced HR coding type. Then the frame is divided into 4 subframes and the signal is encoded on a subframe basis according to the selected coding type to find the adaptive codebook and fixed codebook indices and gains.

Figure 1: Flow diagram of the VMR-WB codec.



The spectral analysis is done twice per frame using a sine window. Its position is not centred but offset to take advantage of the available look-ahead. The signal energy is computed for each critical band. The VAD is based on the SNR per critical band. The VAD detection threshold is made adaptive with an estimation of a long-term SNR. An adaptive hangover is added at the end of a speech activity. The VAD also features a hysteresis favouring the previous state of VAD (i.e. active or inactive period).

The noise suppression (NS) is based on spectral subtraction technique using an overlap-add method with 50% overlap. For active speech frames, the attenuation gain varies with frequency and is dependent on the SNR. For inactive speech frames, the attenuation gain is constant.

The LP filter parameters are estimated similar to AMR-WB specifications [1,2]. The open-loop pitch is computed on the down-sampled weighted speech signal similar to AMR-WB. However, a new pitch-tracker is used in VMR-WB to improve the smoothness of the pitch contour by considering adjacent values. Three open loop pitch lags are computed every frame, two for every half-frame and one for the look-ahead.

Background noise estimation is then updated during inactive speech frames to be used in the VAD and the NS modules of the next frame. The decision about the speech activity for the noise update is based on several speech characteristic parameters and is independent of the VAD decision; i.e., independent of the frame SNR. This is an important feature as the noise update is thus very robust to the varying background noise level.

Rate determination

The rate determination is done implicitly during the selection of a particular encoding technique to encode the current frame. The rate selection is dependent on the mode of operation and the class of input speech. This classification is done in a number of stages. First, the VAD discriminates between active and inactive speech frames as described above.

Inactive speech frames are encoded using ER comfort noise generation (CNG) if the VMR-WB codec is operating in mode 0, 1 or 2. If the frame is not encoded with ER, then unvoiced signal classification is performed. The parameters used in unvoiced frame classification are the normalized pitch correlation, the spectral tilt, the energy variation within the frame, and the relative frame energy. The decision thresholds are function of the operation mode, and they are relaxed for modes with lower ADRs. Frames classified as unvoiced are encoded with Unvoiced HR coding type, and in the economy mode, if the unvoiced frame is not in a voiced/unvoiced transition it is encoded with Unvoiced QR.

If the frame is not classified as inactive or unvoiced frame, it is subjected to a stable voiced classification. The stable voiced classification is incorporated within the signal modification procedure used in encoding stable voiced frames. If the signal modification is successful, the frame is classified as voiced and encoded with the Voiced HR encoder. The signal modification is done prior to the sub-frame loop, so that conventional CELP encoding can follow. Further, the signal modification technique is frame synchronous so that the modified signal is always synchronous with the original signal at the end of each frame [5]. This allows for a transparent switching between Voiced HR frames and frames where conventional CELP coding is used. The signal modification uses a piecewise linear pitch contour to prevent pitch period fluctuations/discontinuities due to forced alignment at the end of each frame.

If the signal is not classified as inactive, unvoiced, or stable voiced then the frame is likely to contain an onset or a voiced transition and is encoded with FR coding type. Note that if the frame energy is lower than a certain threshold the frame is encoded using Generic HR to save the bandwidth on perceptually unimportant frames.

The bit allocation of the different coding types is given in Table 1. More details on each coding type are given in the next section.

3. VMR-WB CODING TYPES

Generic FR coder

The Generic FR coder is used for the CDMA specific modes (i.e., modes 0, 1 and 2). The core or the sub-frame loop of the Generic FR coder follows the specifications of AMR-WB [1,2].

Since AMR-WB frame at 12.65 kbps uses only 253 bits and the VAD bit is not used in the Generic FR coder, there are 14 bits left. These bits have been used to encode supplementary

information, which includes signal energy, frame classification, and phase information for the purpose of better frame erasure (FER) concealment and enhanced convergence to the normal operation at the end of frame erasure interval.

Table 1: Bit allocation of different coding types.

Parameter	Generic FR	Inter. FR	Signal. HR	Inter. HR	Generic HR
Class Info	-	-	3	3	1
VAD bit	-	1	-	1	-
LP Parameters	46	46	46	46	36
Pitch Delay	30	30	30	30	13
Pitch Filtering	4	4	4	4	-
Gains	28	28	28	28	26
Algebraic CB	144	144	-	-	48
FER	14	-	8	-	-
Unused bits	-	13	5	12	-
Total	266	266	124	124	124
Parameter	**Voiced HR**	**UV HR**	**UV QR**	**CNG QR**	**CNG ER**
Class Info	3	2	1	1	-
VAD bit	-	-	-	-	-
LP Parameters	36	46	32	28	14
Pitch Delay	9	-	-	-	-
Pitch Filtering	2	-	-	-	-
Gains	26	24	20	6	6
Algebraic CB	48	52	-	-	-
FER	-	-	-	-	-
Unused bits	-	-	1	19	-
Total	124	124	54	54	20

Interoperable FR

The interoperable FR coder is used only in the Interoperable mode and its bit allocation follows exactly the one of AMR-WB at 12.65 kbps. Given that the extra remaining bits cannot be transmitted over a gateway to 3GPP system, they are not used and are set to certain pattern that differentiates Generic FR and Interoperable FR coding types.

Signalling HR

The signaling HR coder is differentiated only in the VMR-WB decoder when a packet-level signaling has forced a rate reduction from the Generic FR to HR at the bit-stream level. In this case, the indices of the algebraic codebook are discarded at the gateway. At the decoder, the algebraic codebook indices are randomly generated. The supplementary energy and classification information bits of the FER protection are used in the Signaling HR decoder.

Interoperable HR

The Interoperable HR coder is similar in operation to the Signaling HR coder, but it is used only in the Interoperable mode. When considering the CDMA forward link, the difference between both techniques at the decoder side is essentially the fact that here no supplementary FER protection information can be used because the bit-stream has been generated at the 3GPP side.

The Interoperable HR technique is also used in the case of the Dim-and-Burst signaling in the CDMA reverse link; i.e., when the signaling command is received at the VMR-WB encoder. The reason is that in the Interoperable mode, the bit-stream must be transparent for the AMR-WB decoder and hence no specific HR encoder can be used (the algebraic codebook indices must be generated randomly or pseudo-randomly at the intermediate gateway).

Generic HR

The Generic HR coder is specific for CDMA modes. It has been designed for frames not classified as voiced or unvoiced. It is used in frames with low perceptual importance to reduce the ADR or during the dim-and-burst signaling (when classification requires the use of Generic FR and the system requests the use of HR). Similar to the previous encoding techniques, it is based on ACELP®. Considering the lower bit budget, the close-loop pitch is evaluated only twice per frame. All other parameters are computed and quantized for each sub-frame. This is also the only VMR-WB encoding technique that exploits the AMR-WB phase dispersion algorithm.

Voiced HR

The Voiced HR encoder is used when the frame successfully passes through the signal modification process. This module comprises an inherent powerful classifier of voiced frames suitable for voiced HR encoding. The result of the signal modification is a modified signal following a pre-defined pitch contour. The modified signal is then the input to the sub-frame loop that can use conventional ACELP® techniques with the exception that the close-loop pitch determination is skipped. The pitch contour is defined so that it can be encoded with only 9 bits achieving simultaneously a synchrony with the original signal at the end of the frame [5].

The Voiced HR uses an optimized Immitance Spectral Frequencies (ISF) quantizer. Despite a relatively low bit allocation for ISF quantization, an approximately transparent quantization is obtained by using an autoregressive (AR) prediction. The AR prediction has been long time avoided in speech coding quantizers due to the long-term propagation of the spectral error after frame erasures. This does not cause any problem here because the Voiced HR coder is used only during stable voiced frames and because all other encoders use the conventional moving average (MA) prediction quantizers. Hence the propagation of an error is stopped whenever another encoder is used; i.e., whenever the spectral characteristics tend to vary.

Unvoiced HR

The Unvoiced HR coder does not use the closed-loop pitch analysis and no pitch lag is transmitted. Consequently, a scalar quantizer quantizes the fixed codebook gain. The Unvoiced HR coder uses random Gaussian codebook for the excitation. A 13 bit random codebook is used based on a table of only 64 random vectors. The excitation vector is given as the sum of two signed vectors from the table. That is,

$$\mathbf{c} = s_1 \mathbf{v}_{p_1} + s_2 \mathbf{v}_{p_2}$$

where s_1 and s_2 are signs equal to -1 or 1, and p_1 and p_2 are the indices of the random vectors from the random table. Each index is encoded with 6 bits and the signs with only one bit giving a total of 13 bits (similar to encoding two pluses per track in AMR-WB [1]). The codebook is searched using a very efficient procedure in which only 8 vectors with their corresponding signs are first pre-selected out of the 64 vectors, then the search is reduced to finding the addition of 2 vectors out of the 8 vectors giving a total of 36 tests. The pre-selection is performed by determining the 8 vectors that give absolute maximum cross correlation with the backward filtered target vector (the correlation between the target vector and the impulse response of the weighted synthesis filter [3]) and the corresponding signs are given by the signs of the cross-correlations.

Unvoiced QR
The Unvoiced QR coder is similar to the Unvoiced HR, but the indices and signs of the two excitation vectors are randomly generated.
CNG QR
The CNG QR is used only in the Interoperable mode using the same quantization tables as in AMR-WB. The estimation of the quantized parameters is however done in such a way that the necessary active speech hangover and consequently the ADR are significantly reduced.
CNG ER
The CNG ER is used to encode inactive speech frames in CDMA specific modes and non-SID frames in the Interoperable mode. Similar to CNG QR, the CNG ER consists of a random excitation signal with quantized energy filtered through the LP synthesis filter. Both filter coefficients and the excitation energy are smoothed over time.

4. VMR-WB DECODER

The decoder operates similar to AMR-WB. For HR and QR modes the excitation is reconstructed as described in the previous section. The same post-processing of the excitation used in AMR-WB is also used here (pitch enhancement, background noise enhancement, phase dispersion in Generic HR). Further, a new post-processing procedure is applied on the synthesis signal before resampling back to 16 or 8 kHz to enhance the periodicity in low frequency region. This new post-processing consists of splitting the synthesis signal in two bands and applying a pitch enhancement filter to the lower band having the form

$$y(n) = \left(1-\frac{\alpha}{2}\right) x(n) + \frac{\alpha}{4}\{x(n-T) + x(n+T)\}$$

where α $(0 \le \alpha \le 1)$ is a coefficient that controls the inter-harmonic attenuation, T is the pitch period, $x(n)$ is the lower band of the reconstructed signal and $y(n)$ is the post-processed lower band signal. The factor α is derived from the normalized correlation between the full band signal $x'(n)$ and its delayed version $0.5\{(x'(n-T) + x'(n+T)\}$. The band splitting filter is combined with the upsampling filter to simplify the implementation.

In the case of a WB output, the high frequencies are then regenerated for active speech frames in the same way as in AMR-WB and added to the upsampled post-processed synthesis signal.

4.1 Frame erasure concealment

The erased frames are processed differently depending on the last correctly received frame. If this frame happens to be a CNG frame, no special processing is needed as the CNG parameters are already interpolated. The only major difference is that the decoding of these parameters is skipped and previous parameters are used instead.
Processing of erased frames following an active speech frame is done separately. The processing is however the same independent of the coding type used in the preceding active speech frame. The concealment is done by estimating the excitation signal for the whole frame and filtering it with an estimated LP synthesis filter, interpolated 4 times per frame as in normal processing. The FER concealment can be summarized such that the energy of the excitation signal and the spectral envelope represented by the LP filter coefficients are gradually moved to the corresponding estimated parameters of the background noise. The excitation periodicity converges to zero. The rate of the convergence depends on the last good frame classification.

The second stage of the FER processing involves the recovery of the normal processing after an erasure interval is over. To improve the convergence to the normal operation when a voiced onset has been lost, the onset is reconstructed artificially to enable rapid synthesis convergence of the voiced speech in the first good frame after the erasure and before continuing with the ACELP® sub-frame decoding. This is done however only if the first good frame after the erasure is a Generic FR frame, because this is the only frame where all the supplementary information for FER protection is transmitted. Limiting the artificial onset reconstruction to the Generic FR coding type only is not a constraint because voiced onsets are typically encoded using Generic FR type. The convergence to the normal operation is further improved by a careful control of the synthesized speech energy in the first good frame after an erasure.

5. PERFORMANCE

The performance of the VMR-WB codec was evaluated during the selection test. The VMR-WB codec clearly outperformed other candidates in different test conditions. The test results can be found in [6]. In clean conditions, the performance of the premium mode was equivalent to AMR-WB at 14.85 kbit/s and the standard mode equivalent to AMR-WB at 12.65 kbit/s. The economy mode was better than AMR-WB at 8.85 kbit/s. The quality improvement compared to AMR-WB is primarily due to the post-processing described in the previous section. In frame erasure conditions, the quality improvement is due to the supplementary information and careful FER concealment. In background noise conditions, the reference AMR-WB codec used the reference noise suppression based on EVRC noise suppression. The superiority of the VMR-WB noise suppression was evident as the results were significantly better than AMR-WB with the reference noise suppression.

ACKNOWLEDGMENTS

The authors wish to thank Tommy Vaillancourt, Mikko Tammi, Roch Lefebvre and Joachim Thiemann for their valuable contributions to the codec development.

REFERENCES

[1] 3GPP TS 26.190, "AMR Wideband Speech Codec: Transcoding Functions," *3GPP Technical Specification,* March 2002.

[2] ITU-T Recommendation G.722.2 "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)", Geneva, January 2002.

[3] B. Bessette, et al, "The Adaptive Multi-Rate Wideband Speech Codec (AMR-WB)," *IEEE Trans, on Speech and Audio Processing*. Vol 10, No 8, Nov. 2002, pp. 620-636.

[4] 3GPP2 C11-20030714-003, "Stage Two Requirements (Test Plan) for CDMA2000® Wideband Speech Codec", *3GPP2 Technical Specification,* July 2003.

[5] *M. Tammi and M. Jelínek,* "Signal Modification for Voiced Wideband Speech Coding and Its Application for IS-95", 2002 IEEE Workshop on Speech Coding, Japan, Oct. 2002.

[6] 3GPP2 C11-20030915-010, "cdma2000® Wideband Speech Codec Selection Test Report", *3GPP2 Technical Specification,* September2003.