This paper presents novel techniques for source controlled variable rate wideband speech coding. These techniques have been used in the variable-rate multimode wideband (VMR-WB) speech codec recently selected by 3GPP2 for wideband (WB) speech telephony, streaming, and multimedia messaging services in the cdma2000 third generation wireless system. The coding algorithm contains several innovations that enable very good performance at average bit rates as low as 4.0 kbit/s in typical conversational operating conditions. These innovations include: Efficient noise suppression algorithm, signal classification and rate selection algorithm that enables high quality operation at low average bit rates, efficient post-processing techniques tailored for wideband signals, and novel frame erasure concealment techniques including supplementary information for reconstruction of lost onsets and improving decoder convergence. Further, the coder utilizes efficient coding types optimized for different classes of speech signal including a generic coding type based on AMR-WB for transients and onsets, voiced coding type optimized for stable voiced signals, and novel frame erasure concealment techniques including supplementary information for reconstruction of lost onsets and improving decoder convergence. Further, the article describes in detail some of the codec novel features. An overview of the codec design and structure is given. Section 2 describes the source-controlled VBR mechanism and details some of the codec novel features. The codec has three modes of operation as described above with additional modes for encoding/decoding wideband speech (50-7000 Hz) sampled at 16 kHz using 20 ms frames. The operation of VMR-WB is controlled by instantaneous speech signal characteristics (i.e., source-controlled) and by traffic condition of the network (i.e., network-controlled mode switching).

2. OVERVIEW OF VMR-WB CODEC DESIGN AND STRUCTURE

2.1 VMR-WB Features
VMR-WB is a variable-rate multimode speech codec designed for encoding/decoding wideband speech (50-7000 Hz) sampled at 16 kHz using 20 ms frames. The operation of VMR-WB is controlled by instantaneous speech signal characteristics (i.e., source-controlled) and by traffic condition of the network (i.e., network-controlled mode switching).

VMR-WB is based on 3GPP/AMR-WB (ITU-T/G.722.2) codec [2,3,4], recently standardized by 3GPP and ITU-T for 3G high-quality multimedia applications, and is interoperable with this codec in one of its operational modes. VMR-WB is fully compliant with cdma2000 rate-set II (Radio Configuration 4 (reverse link)/5 (forward link)) [1].

Depending on the traffic conditions, one of 4 operational modes is selected. The transition from one mode to another is seamless, memoryless (i.e. no transition period to achieve the

---

**ADVANCES IN SOURCE-CONTROLLED VARIABLE BIT RATE WIDEBAND SPEECH CODING**

*Milan Jelinek*, Redwan Salami†, Sassan Ahmadi‡, Bruno Bessette†, Philippe Gourlay§,* Claude Laflamme†, Roch Lefebvre†

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

**ABSTRACT**

This paper presents novel techniques for source controlled variable rate wideband speech coding. These techniques have been used in the variable-rate multimode wideband (VMR-WB) speech codec recently selected by 3GPP2 for wideband (WB) speech telephony, streaming, and multimedia messaging services in the cdma2000 third generation wireless system. The coding algorithm contains several innovations that enable very good performance at average bit rates as low as 4.0 kbit/s in typical conversational operating conditions. These innovations include: Efficient noise suppression algorithm, signal classification and rate selection algorithm that enables high quality operation at low average bit rates, efficient post-processing techniques tailored for wideband signals, and novel frame erasure concealment techniques including supplementary information for reconstruction of lost onsets and improving decoder convergence. Further, the coder utilizes efficient coding types optimized for different classes of speech signal including a generic coding type based on AMR-WB for transients and onsets, voiced coding type optimized for stable voiced signals, and novel frame erasure concealment techniques including supplementary information for reconstruction of lost onsets and improving decoder convergence. Further, the article describes in detail some of the codec novel features. The article describes in detail some of the codec novel features.
ABR of the new mode) and there is no need to transmit the mode information to the decoder. Modes 0, 1, and 2 are specific to CDMA systems (i.e., TIA/EIA/IS-95, cdma2000) with mode 0 providing the highest quality and mode 2 the lowest ABR. Mode 3 is interoperable with AMR-WB at 12.65 kbit/s. Its ABR is slightly higher than the ABR of mode 0. The active speech ABRs are approximately 12.8, 10.5, 8.1 and 13.3 kbit/s for modes 0, 1, 2 and 3 respectively which translates into 5.7, 4.8, 3.8 and 6.1 kbit/s for a typical conversational speech of about 40% of speech activity. Additionally, the system can also force maximum and minimum rates.

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.

Finally, VMR-WB is equipped with an integrated novel noise reduction (NR) algorithm with adjustable maximum allowed reduction and with a sophisticated frame error concealment technique making it very robust for background noise and channel impairments.

2.2 VMR-WB Encoder Overview

The flow diagram of the VMR-WB encoder 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 twice per frame on the preprocessed signal for use in noise reduction and voice activity detection (VAD). The signal energy is computed for each critical band [6].

The Spectral analysis is done using a sine window. The sine window (or square root of Hanning window) has been chosen because it is well suited for overlap-add methods and because the VMR-WB noise reduction is based on spectral subtraction technique using an overlap-add with 50% overlap. The position of the sine window is not centred but offset to take advantage of the whole available look-ahead.

Linear Prediction (LP) analysis and open-loop (OL) pitch analysis are performed on a frame basis on the denoised signal. The LP filter parameters are estimated similar to AMR-WB specifications. A new OL 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 jointly estimated 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 NR modules of the next frame. The parameters used for the noise update decision are: pitch stability, signal non-stationarity, voicing, and ratio between 2nd order and 16th order LP residual error energies. These parameters have generally low sensitivity to the noise level variations. The reason for not using the encoder VAD decision for noise update is to make the noise estimation robust to rapidly changing noise levels [7]. If the encoder VAD decision were used for the noise update, a sudden increase in noise level would cause an increase of SNR in inactive speech frames, preventing the noise estimator to update, which in turn would maintain the SNR high in following frames, and so on. Consequently, the noise update would be blocked and some other logic would be needed to resume the noise adaptation.

Then signal classification is performed on frame basis to detect unvoiced frames, voiced frames and frames with low perceptual significance. Finally, 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. Supplementary information is estimated and send in FR coding type to enhance the codec robustness to frame erasures (FER).

3. SOURCE AND CHANNEL CONTROLED OPERATION

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.

Figure 2 shows a simplified high-level description of the signal classification procedure. If the signal is classified as a non-speech signal by the VAD algorithm, then it is usually encoded as an ER frame. Otherwise, unvoiced signal classification is applied. If the frame is classified as unvoiced, it is encoded with either HR Unvoiced or QR Unvoiced coding. If the frame is not classified as unvoiced, then stable voiced classification is applied. If the frame is classified as stable voiced, it can be encoded using HR Voiced coding. Otherwise, the frame is likely to contain a non-stationary speech segment such as a voiced onset or rapidly evolving voiced speech signal. These frames typically require a general purpose coding model at high bit rate for sustaining good speech quality. Thus in this case
FR coding is mainly used. Frames with very low energy and not detected as non-speech, unvoiced or stable voiced can be encoded using Generic HR coding in order to reduce the average data rate. Generic HR is also used if the system mandates a HR maximum rate.

The above is a simplified description of the rate determination procedure. The actual choice of coding type in a certain frame is based both on the frame classification and the required mode of operation and the classification thresholds depend on the mode of operation. Further, the rate determination procedure has to cope also with maximum and minimum rate signaling information.

Except for VAD decision, the signal classification described in the following sections does not concern the AMR-WB interoperable mode. As AMR-WB is not a source-controlled codec, only FR can be used to encode active speech frames. However, a special HR coding type has been implemented to cope with the CDMA signaling when requiring to use a maximum HR frame. In this case, Interoperable HR coding mode is used. It is basically the Generic FR (AMR-WB at 12.65 kbit/s) without indices describing the algebraic excitation. These indices must then be generated randomly or in a predetermined way at the system gateway.

### 3.1 Voice Activity detection

VAD is determined based on the signal-to-noise ratio (SNR) per critical band. Let $E_{CB}^{(1)}(i)$ and $E_{CB}^{(2)}(i)$, $i=0,...,19$, denote the energy per critical band information for the first and second spectral analysis, respectively. The average energy per critical band for the whole frame and part of the previous frame is computed as

$$E_{av}(i) = 0.2E_{CB}^{(0)}(i) + 0.4E_{CB}^{(1)}(i) + 0.4E_{CB}^{(2)}(i)$$

where $E_{CB}^{(0)}(i)$ denote the energy per critical band information from the second analysis of the previous frame. The SNR per critical band is then computed as

$$SNR_{CB}(i) = E_{av}(i) / N_{CB}(i), \text{ bounded by } SNR_{CB} \geq 1,$$

where $N_{CB}(i)$ is the estimated noise energy per critical band. The average SNR per frame is estimated as

$$SNR_{av} = 10 \log \left( \sum_{i=0}^{19} SNR_{CB}(i) \right).$$

The voice activity is detected by comparing the average SNR per frame to a certain threshold which is a function of the long-term SNR. The long-term SNR is given by

$$SNR_{LT} = \bar{E}_f - \bar{N}_f$$

where $\bar{E}_f$ and $\bar{N}_f$ are the long-term average signal energy and noise energy, respectively. The threshold is a piece-wise linear function of the long-term SNR. Two functions are used, one for clean speech and one for noisy speech.

If $SNR_{LT} < 35$ (noisy speech) then

$$th_{VAD} = 0.4346SNR_{LT} + 13.9575$$

else (clean speech)

$$th_{VAD} = 1.0333SNR_{LT} - 7$$

Further, a hysteresis in the VAD decision is added to prevent frequent switching at the end of an active speech period. It is applied in case the frame is in a soft hangover period or if the last frame is an active speech frame. The soft hangover period consists of the first 10 frames after each active speech burst longer than 2 consecutive frames. In case of noisy speech ($SNR_{LT} < 35$), the hysteresis decreases the VAD decision threshold by $th_{VAD} = 0.95th_{VAD}$. In case of clean speech it decreases the VAD decision threshold by $th_{VAD} = th_{VAD} - 11$.

If the average SNR per frame is larger than the VAD decision threshold, that is, if $SNR_{av} > th_{VAD}$, then the frame is declared as an active speech frame and the VAD flag and a local VAD flag are set to 1. Otherwise the VAD flag and the local VAD flag are set to 0. However, in case of noisy speech, the VAD flag is forced to 1 in hard hangover frames, i.e. one or two inactive frames following a speech period longer than 2 consecutive frames (the local VAD flag is then equal to 0 but the VAD flag is forced to 1).

![Figure 2: Overview of Signal Classification Module.](image)

### 3.2 Unvoiced frame classification

The unvoiced parts of the signal are characterized by missing the periodic component and can be further divided into unstable frames, where the energy and the spectrum changes rapidly, and stable frames where these characteristics remain relatively stable. The classification of unvoiced frames uses the following parameters:

- A voicing measure, computed as an averaged normalized correlation,
- spectral tilt measures, $e_{tilt}(0)$ and $e_{tilt}(1)$, for both spectral analysis per frame,
- a signal energy ratio ($dE$) used to assess the frame energy variation within a frame and thus the frame stability, and
- relative frame energy.

A description of these parameters is given below.

**Voicing Measure**

The normalized correlation $r_n$, used to determine the voicing measure, is computed on the weighted speech at the OL pitch.
delay as part of the OL pitch search. The voicing measure is given by the average correlation \( \overline{r}_s \) which is defined as

\[
\overline{r}_s = (r_s(0) + r_s(1) + r_s(2))/3
\]

where \( r_s(0) \), \( r_s(1) \) and \( r_s(2) \) are respectively the normalized correlation of the first half of the current frame, the normalized correlation of the second half of the current frame, and the normalized correlation of the look-ahead (beginning of next frame). The correlations \( r_s(k) \) are computed on the weighted speech signal decimated by 2 for lower complexity. The length of the autocorrelation computation \( L_k \) is dependant on the pitch period (decimated by 2) and given by

- \( L_4 = 40 \) samples for \( p_k \leq 31 \) samples
- \( L_4 = 62 \) samples for \( 62 < p_k \leq 61 \) samples
- \( L_4 = 115 \) samples for \( p_k > 61 \) samples

These lengths assure that the correlated vector length comprises at least one pitch period, which helps for a robust open loop pitch detection.

**Spectral Tilt**

The spectral tilt parameter contains the information about the frequency distribution of energy. The spectral tilt is estimated in the frequency domain as a ratio between the energy concentrated in low frequencies and the energy concentrated in high frequencies.

The energy in high frequencies \( E_h \) is computed as the average of the energies of the last two critical bands. The energy in low frequencies \( E_l \) is computed as the average of the energies in the first 10 critical bands. The middle critical bands have been excluded from the computation to improve the discrimination between frames with high energy concentration in low frequencies (generally voiced) and with high energy concentration in high frequencies (generally unvoiced). In between, the energy content is not characteristic for any of the classes and increases the decision confusion.

The energy in low frequencies is computed differently for long pitch periods and short pitch periods. For voiced female speech segments, the harmonic structure of the spectrum is exploited to increase the voiced-unvoiced discrimination. Thus for short pitch periods, \( E_l \) is computed bin-wise and only frequency bins sufficiently close to the speech harmonics are taken into account in the calculation.

The resulting low and high frequency energies are obtained by subtracting estimated noise energies \( N_l \) and \( N_h \) from the values \( E_l \) and \( E_h \) respectively (to account for the presence of background noise), where \( N_l \) and \( N_h \) are the averaged noise energies in the last 2 critical bands and first 10 critical bands respectively.

The spectral tilt is given by

\[
e_{\text{tilt}}(i) = E_l / E_h.
\]

Note that the spectral tilt computation is performed twice per frame to obtain \( e_{\text{tilt}}(0) \) and \( e_{\text{tilt}}(1) \) corresponding to both spectral analysis per frame. The average spectral tilt used in unvoiced frame classification is given by

\[
e_{\text{tilt}} = \frac{1}{3}(e_{\text{tilt}}(0) + e_{\text{tilt}}(1))
\]

where \( e_{\text{tilt}} \) is the tilt in the second half of the previous frame.

**Energy Variation**

The energy variation \( dE \) is evaluated on the denoised speech signal \( s(n) \), where \( n=0 \) corresponds to the current frame beginning. As the unvoiced frames are encoded using Gaussian excitation, the performance of unvoiced coding decreases if the energy varies rapidly inside a frame, as it is the case of stop consonants for example. To track energy spikes, the maximum energy is evaluated based on short-time segments of 32 samples. The short-time maximum energies are computed as

\[
E^{(1)}_s(j) = \max_{i=0}^{31}(s^2(i + 32j)), \quad j = -1,...,8,
\]

where \( j = -1 \) and \( j = 8 \) correspond to the end of previous frame and the beginning of next frame. Another set of 9 maximum energies is computed by shifting the origin by 16 samples. That is

\[
E^{(2)}_s(j) = \max_{i=0}^{31}(s^2(i + 32j - 16)), \quad j = 0,...,8,
\]

The maximum energy variation \( dE \) is computed as the maximum of the ratios of maximum energies between adjacent segments, that is the maximum of the following:

\[
\begin{align*}
& E^{(1)}_s(0)/E^{(1)}_s(-1), \\
& E^{(1)}_s(7)/E^{(1)}_s(8), \\
& \max(E^{(1)}_s(j), E^{(1)}_s(j-1))/\min(E^{(1)}_s(j), E^{(1)}_s(j-1)) \quad \text{for } j=1 \text{ to } 7 \\
& \max(E^{(2)}_s(j), E^{(2)}_s(j-1))/\min(E^{(2)}_s(j), E^{(2)}_s(j-1)) \quad \text{for } j=1 \text{ to } 8
\end{align*}
\]

**Relative Frame Energy \( E_{rel} \)**

Low energy frames are of little perceptual importance and need not be encoded in FR. The relative frame energy is computed as the difference between the current frame total energy (computed as the addition of the average energies per critical band) and the long-term average of active speech energy.

**Unvoiced Speech Classification**

The classification of unvoiced speech frames is based on the parameters described above, and decision thresholds are set based on the operating mode. Basically for operating modes with lower allowable average bit rates, the thresholds are set to favor more unvoiced classification (since a half-rate or a quarter rate coding will be used to encode the frame). Unvoiced frames are usually encoded with Unvoiced HR encoder. However, in case of the Economy mode, Unvoiced QR is also used in order to further reduce the ABR.

In both the Standard and Economy modes, if the local VAD is zero, the frame is classified directly as an Unvoiced frame. Otherwise, for a frame to be classified as unvoiced, both \( R \) and \( e \) must be under a certain threshold, independently of mode. Further, for Premium mode, \( dE \) must be also under a certain threshold. For Standard and Economy modes, \( dE \) or \( E_{rel} \) must be also under a certain threshold. If the system mandates a HR maximum frame, the unvoiced decision is similar to the Economy mode.

In Premium and Standard modes, unvoiced frames are encoded using Unvoiced HR coding type. In Economy mode, unvoiced frames can also be encoded with Unvoiced QR if the following further conditions are also satisfied: If the last frame is either unvoiced or background noise frame, and if at the end of the frame no potential voiced onset is detected \((r_s(2) \text{ and } e_{\text{tilt}}(1) \text{ are under certain thresholds})\). Note that \( r_s(2) \) is the normalized correlation in the lookahead and \( e_{\text{tilt}}(1) \) is the tilt in the second
spectral analysis which spans the end of the frame and the lookahead.

3.3 Classification of Stable Voiced Frames:
If the frame is not classified as inactive speech frame or unvoiced frame, it is subjected to stable voiced classification, which is an integral part of the signal modification procedure [8]. Several operations in the signal modification procedure yield indicators quantifying the attainable performance of long-term prediction in the current frame. If any of these indicators is outside its allowed limits, the signal modification procedure is terminated. In this case, the original denoised signal is preserved intact and encoded with a generic coding model. Otherwise, the frame is classified as stable voiced and the original signal is modified to simplify the encoding of long-term prediction parameters. The indicators include the evolution of pitch period, the fitness of the selected delay contour for describing this evolution, and the pitch prediction gain attainable with signal modification. If these indicators enable signal modification, long-term prediction is able to model the modified speech frame efficiently at a low bit rate without degrading subjective quality. In this case, the adaptive codebook excitation has a dominant contribution in describing the excitation signal, and the bit rate allocated for the fixed-codebook excitation can be reduced. The frame is classified as stable voiced frame and encoded using Voiced HR coding type. When these indicators disable signal modification, the frame is likely to contain a non-stationary speech segment such as a voiced onset or rapidly evolving voiced speech signal. These frames typically require a high bit rate for sustaining good subjective quality and are usually encoded with generic FR coding type.

3.4 Classification of FR and Generic HR frames:
If the rate selection reaches this stage, then the frame has not been declared as inactive speech frame, nor declared as unvoiced frame (not encoded with Unvoiced HR or Unvoiced QR), nor declared as stable voiced frame (not encoded with Voiced HR). Thus the frame is likely to contain a non-stationary speech segment such as a voiced onset or rapidly evolving voiced speech signal. These frames typically require a general purpose coding model at high bit rate for sustaining good speech quality. Thus in this case FR coding is mainly used. However, frames with very low relative energy ($E_n$ below a certain threshold) can be encoded using Generic HR coding in order to reduce the average bit rate. The Generic HR coding is used only in the Standard and Economy modes or if the system requires a maximum HR frame.

4. VMR-WB CODING TYPES

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

FR Coding Types
The Generic FR and the interoperable FR coding types are based on AMR-WB at 12.65 kbit/s [2,3,4]. The interoperable FR is used only in the Interoperable mode and its bit allocation follows exactly the one of AMR-WB. Given that the 13 remaining bits to 13.3 kbit/s of CDMA FR cannot be transmitted over a gateway to 3GPP system, they are not used and are set to certain pattern that differentiates both full-rates. The Generic FR is used for the CDMA specific modes (i.e., modes 0, 1 and 2). Since the Generic FR does not use the AMR-WB VAD bit, 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 FER concealment and enhanced convergence to the normal operation at the end of frame erasure interval.

Table 1: Bit Allocation of Different Coding Types.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Voiced HR</th>
<th>UV HR</th>
<th>UV QR</th>
<th>CNG HR</th>
<th>CNG ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Info</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>VAD bit</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LP Parameters</td>
<td>36</td>
<td>46</td>
<td>32</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Pitch Delay</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pitch Filtering</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gains</td>
<td>26</td>
<td>24</td>
<td>20</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Algebraic CB</td>
<td>48</td>
<td>52</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FER</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unused bits</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>124</td>
<td>124</td>
<td>54</td>
<td>54</td>
<td>20</td>
</tr>
</tbody>
</table>

Signalling HR coding types
The signalling HR and the Interoperable HR are derived respectively from the Generic FR and the Interoperable FR by dropping the indices describing the algebraic excitation. They differ, similarly as the FR coding types in the absence of FER additional information in case of Interoperable HR. For lack of available bits, the Signalling HR does not contain the phase information for FER concealment anymore. The Signalling HR is used in CDMA specific modes when the system mandates FR to HR conversion on bitstream level, which means that the conversion is unknown to the encoder. This coding mode is thus differentiated only at decoder side where the algebraic codebook indices are generated randomly. (Note that when the FR to HR conversion is known to the encoder, dedicated HR coding types can be used instead.)

The Interoperable HR is used in mode 3 similarly to Signalling HR in other modes. However, unlike the Signalling HR, the Interoperable HR is used whether the FR to HR conversion is known to the encoder or not. The reason is that in the Interoperable mode, the bitstream must be transparent for the AMR-WB decoder and hence no specific HR encoder can be used.

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 (FR to HR conversion known to the encoder 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. 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 [8].

The Voiced HR uses an optimized Inmittance Spectral Frequencies (ISF) quantizer. Despite a relatively low bit allocation for ISF quantization, a practically transparent prediction error is observed after frame erasures. This does not cause any problem here because the Voiced HR encoder is used only during stable voiced frames and because all other coding types use the conventional moving average (MA) prediction quantizers. Hence the propagation of an error is stopped whenever another coding type is used; i.e., whenever the spectral characteristics tend to vary.

Unvoiced Coding Types
The Unvoiced coding types do not use the closed-loop pitch analysis and no pitch lag is transmitted. A scalar quantizer is used to quantize the fixed codebook gain. The Unvoiced HR coding type uses random Gaussian codebook for encoding the excitation with 13 bits per subframe. The Unvoiced QR coding type is similar, but the indices of the Gaussian excitation vectors are randomly generated (i.e. no indices are transmitted) and fewer bits are allocated for gain quantization.

The 13 bit random codebook is 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 [3]).

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. The 64 random vectors are stored using a sequence of only 190 samples where a shift-by-2 is used to obtain each random vector.

Comfort Noise Generation (CNG)
CNG technique is used to encode inactive speech frames. It comprises a random excitation signal with quantized energy filtered through the LP synthesis filter. Both filter coefficients and the excitation energy are smoothed over time.

In general, CNG ER is used to encode inactive speech frames. CNG QR is used only in the Interoperable mode to update noise characteristics at regular intervals as the Silence Descriptor (SID) of AMR-WB does not fit in CDMA ER frame [9].

Usage of Different Coding Types
Table 2 summarizes for each operational mode the coding types used in that mode as a percentage of their usage in active speech coding. The data has been measured on a clean speech nominal level database.

Table 2: Coding Types Relative Usage in Active Speech.

<table>
<thead>
<tr>
<th>Coding Type</th>
<th>Mode 0</th>
<th>Mode 1</th>
<th>Mode 2</th>
<th>Mode 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic FR</td>
<td>93.4 %</td>
<td>60.4 %</td>
<td>34.1 %</td>
<td>-</td>
</tr>
<tr>
<td>Interoperable FR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100.0 %</td>
</tr>
<tr>
<td>Generic HR</td>
<td>-</td>
<td>7.1 %</td>
<td>13.1 %</td>
<td>-</td>
</tr>
<tr>
<td>Voiced HR</td>
<td>-</td>
<td>13.0 %</td>
<td>33.2 %</td>
<td>-</td>
</tr>
<tr>
<td>Unvoiced HR</td>
<td>6.6 %</td>
<td>19.5 %</td>
<td>5.6 %</td>
<td>-</td>
</tr>
<tr>
<td>Unvoiced QR</td>
<td>-</td>
<td>-</td>
<td>14.0 %</td>
<td>-</td>
</tr>
</tbody>
</table>

4. VMR-WB DECODER
The decoder uses similar post-processing of the excitation signal as AMR-WB (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. The post-processing of the lower band is given by

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

where \( \alpha \) 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 \( \alpha \) 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 FER concealment is different for erasures following active speech frames and CNG frames. If the last correctly received 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.

If a FER follows an active speech frame, the concealment technique is independent of the coding type of that frame. It 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. Probably the most difficult situation arises when a voiced onset is lost. To improve the convergence to the normal operation after a lost onset, 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 normal ACELP sub-frame decoding. The artificial onset reconstruction is done by placing a low-pass filtered impulse at an estimated position of the glottal pulse. 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 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

5.1 Performance of CDMA Specific Modes with WB Inputs

The performance of the CDMA specific modes of VMR-WB codec was evaluated during the selection test using Mean Opinion Score (MOS) subjective testing methodology and two-stage statistical analysis. The first stage consisted in statistical comparison by Dunnett’s Test of the MOS for the candidate codecs against the MOS for the reference codecs. The second stage consisted in head-to-head comparison of the candidate codecs that passed the first stage using pair-wise Analyses of Variance.

The VMR-WB codec clearly outperformed other candidates in different test conditions. The test results can be found in [7]. 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. In FER and background noise conditions, the performance of VMR-WB was systematically superior.

The high performance of modes 1 and 2 is primarily due to the sophisticated signal classification and rate selection scheme together with efficient coding of HR frames. The improvement of mode 0 compared to AMR-WB in clean speech conditions is primarily due to the new low frequency post-processing. 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.

![Figure 3: Comparison of MOS Scores in Clean Conditions.](image)

![Figure 4: Comparison of MOS Scores in FER Conditions.](image)

![Figure 5: Comparison of MOS Scores in Noisy Conditions.](image)

Figures 3 to 5 illustrate the performance of VMR-WB as compared to that of the reference codecs. Ref 0, 1, 2 correspond respectively to AMR-WB at 14.25, 12.65 and 8.85 kbit/s. Test 0, 1, 2 correspond respectively to VMR-WB mode 0, 1 and 2. Note also that the 1%, 3% and 6% FER conditions were tested in presence of 1% dim-and-burst signaling and the Car 20dB and Office 20dB conditions were tested in presence of 2% bitstream-level signaling. As AMR-WB codec does not support CDMA signaling, the signaling has been simulated in the reference codec by using the AMR-WB at 6.6 kbit/s mode in signaling frames.
5.2 Performance in Case of Narrowband Inputs

Although VMR-WB was optimized for wideband input signals, it can process signals with 8 kHz input/output sampling frequency and it exhibits good performance with narrow band inputs. The codec performance with narrowband inputs was tested in the first phase of the characterization tests and presented in December 2003 [11].

![Figure 6: Comparison of MOS Scores for NB processing.](image)

The NB performance of VMR-WB is illustrated in Figure 6 for clean channel nominal input speech signals. EVRC codec [12] and SMV codec [14] with modes 0, 1, and 2 were included in the test.

5.3 Performance of the Interoperable Mode (WB input)

As mentioned previously, mode 3 of VMR-WB is directly interoperable with AMR-WB at 12.65 kbit/s. The interoperability feature has been tested during Characterization phase. The performance for clean channel is given in Figure 7 for both links.

6. CONCLUSION

We have presented new techniques for variable-rate WB speech coding employed in variable-rate multi-mode VMR-WB codec, recently Selected as the new 3GPP2 standard for wideband speech services. While the paper gave an overall codec description, the main focus has been on a detailed description of signal classification and rate selection mechanism as it is the crucial part of the variable-rate multi-mode system.

The codec has been selected as a clear winner in a competitive selection process despite a supplementary constraint of interoperability with 3GPP/ITU AMR WB standard. The codec has been superior to other candidates in clean speech, channel error and background noise conditions. Apart from the rate selection algorithm, the performance on clean speech is mainly due to a novel low-frequency post-processing. The improvement in channel error conditions has been achieved by a new FER concealment method, focused not only on erased frames reconstruction, but also on rapid convergence after an erased segment. Finally, the high performance in noisy conditions is mainly due to a new noise reduction algorithm.

VMR-WB has been optimized for an operation with WB input/output signals. However, its performance with traditional telephony bandwidth signals is comparable to the state-of-the-art NB codecs.

![Figure 7: Comparison of MOS Scores for Interworking with AMR-WB.](image)

ACKNOWLEDGMENTS

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

REFERENCES