The Adaptive Multirate Wideband Speech Codec (AMR-WB)

Bruno Bessette, Redwan Salami, *Senior Member, IEEE*, Roch Lefebvre, *Member, IEEE*, Milan Jelínek, *Member, IEEE*, Jani Rotola-Pukkila, Janne Vainio, Hannu Mikkola, and Kari Järvinen

***Abstract*—This paper describes the Adaptive Multirate Wideband (AMR-WB) speech codec recently selected by the Third Generation Partnership Project (3GPP) for GSM and the third generation mobile communication WCDMA system for providing wideband speech services. The AMR-WB speech codec algorithm was selected in December 2000 and the corresponding specifications were approved in March 2001. The AMR-WB codec was also selected by the International Telecommunication Union—Telecommunication Sector (ITU-T) in July 2001 in the standardization activity for wideband speech coding around 16 kb/s and was approved in January 2002 as Recommendation G.722.2. The adoption of AMR-WB by ITU-T is of significant importance since for the first time the same codec is adopted for wireless as well as wireline services. AMR-WB uses an extended audio bandwidth from 50 Hz to 7 kHz and gives superior speech quality and voice naturalness compared to existing second- and third-generation mobile communication systems. The wideband speech service provided by the AMR-WB codec will give mobile communication speech quality that also substantially exceeds (narrowband) wireline quality. The paper details AMR-WB standardization history, algorithmic description including novel techniques for efficient ACELP wideband speech coding and subjective quality performance of the codec.**

***Index Terms*—Analysis-by-synthesis, CELP, speech coding, wideband speech coding.**

I. Introduction

THE CURRENT second-generation (2G) and third-generation (3G) mobile communication systems operate with narrow audio bandwidth limited to 200–3400 Hz. As wireless systems are evolving from voice-telephony dominated services to multimedia and high-speed data services, the introduction of a wider audio bandwidth of 50–7000 Hz provides substantially improved speech quality and naturalness. Compared to narrowband telephone speech, the low-frequency enhancement from 50 to 200 Hz contributes to increased naturalness, presence and comfort. The high-frequency extension from 3400 to 7000 Hz provides better fricative differentiation and therefore higher intelligibility. A bandwidth of 50 to 7000 Hz not only improves the intelligibility and naturalness of speech, but adds also a feeling of transparent communication and eases speaker recognition.

Recent advances in speech coding have made wideband coding feasible in the bit-rates applicable for mobile communication. In 1999 3GPP together with ETSI (European Telecommunication Standards Institute) started development and standardization of a wideband speech codec for the WCDMA 3G and GSM systems. A feasibility study phase of wideband coding preceded the launch of standardization. After almost two years of intense development and two competitive codec selection phases, the wideband codec algorithm was selected in December 2000. The speech codec specifications were finalised and approved in March 2001. The wideband codec is an adaptive codec capable of operating with a multitude of speech coding bit-rates from 6.6 to 23.85 kb/s. The codec is referred to as Adaptive Multirate Wideband (AMR-WB) codec [1].

The AMR-WB codec includes a set of fixed rate speech and channel codec modes, a Voice Activity Detector (VAD), Discontinuous Transmission (DTX) functionality in GSM and Source Controlled Rate (SCR) functionality in 3G, in-band signaling for codec mode transmission and link adaptation to control the mode selection. The AMR-WB codec adapts the bit-rate allocation between speech and channel coding, optimising speech quality to prevailing radio channel conditions. While providing superior voice quality over the existing narrowband standards, AMR-WB is also very robust against transmission errors due to the multirate operation and adaptation. The adaptation is based on similar principles as in the AMR codec (referred to also as the AMR narrow-band codec, AMR-NB) standardized previously for GSM and WCDMA 3G systems.

The AMR-WB codec has been developed for use in several applications: the GSM full-rate channel, the GSM EDGE Radio Access Network (GERAN) 8-Phase Shift Keying (8-PSK) Circuit Switched channels, the 3G Universal Terrestrial Radio Access Network (UTRAN) channel and also packet based voice over internet protocol (VoIP) applications.

During the standardization of AMR-WB in 3GPP, a parallel process was undertaken in ITU-T for standardizing a wideband speech codec at bit rates around 16 kb/s. A wide range of applications are envisioned for the ITU-T wideband coding standard including ISDN wideband telephony and audio/video teleconferencing, Voice over IP and Internet applications such as IP video conferencing, voice mail, voice chat, broadcast, and voice streaming.

Manuscript received September 18, 2001; revised August 7, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Harry Printz.

B. Bessette, R. Lefebvre, and M. Jelínek are with the Department of Electrical and Computer Engineering, University of Sherbrooke, QC J1K 2R1, Canada (e-mail: bruno.bessette@courrier.usherb.ca; roch.lefebvre@courrier.usherb.ca; milan.jelinek@courrier.usherb.ca).

R. Salami is with the VoiceAge Corporation, Montreal, QC H3R 2H6, Canada (e-mail: redwans@voiceage.com).

J. Rotola-Pukkila, J. Vainio, H. Mikkola and K. Järvinen are with the Nokia Research Center, 33721 Tampere, Finland (e-mail: jani.rotola-pukkila@nokia.com; janne.m.vainio@nokia.com; hannu.j.mikkola@nokia.com; kari.ju.jarvinen@nokia.com).

Digital Object Identifier 10.1109/TSA.2002.804299

Based on the test results of the AMR-WB codec in 3GPP, ITU-T approved the 3GPP AMR-WB codec to participate in the selection phase of ITU-T competition. Also interoperability with 3G wireless systems was considered important. In the July 2001 Rapporteur's meeting of ITU-T Q.7/16, the selection test results were presented and the AMR-WB codec was selected. This resulted in a single harmonized wideband codec for GSM, WCDMA 3G and ITU-T. The codec has been approved by the ITU in January 2002 and it is known as ITU-T Recommendation G.722.2 [2].

The adoption of AMR-WB by ITU-T is of significant importance since for the first time the same codec is adopted for wireless as well as wireline services. This will eliminate the need for transcoding and ease the implementation of wideband voice applications and services across a wide range of communication systems and platforms.

This article will give a detailed description of the AMR-WB codec. Section II gives a brief history of the speech coding standardization for GSM and WCDMA 3G. The encoding functions of the algorithm will be described in Section III including novel approaches such as efficient perceptual weighting and pitch analysis relevant for wideband signals. Section IV gives the algorithmic description of the AMR-WB decoder including new efficient postprocessing techniques such as gain smoothing, pitch enhancement and high frequency generation. Section V gives a brief description of the codec complexity. Section VI explains the adaptive operation of AMR-WB in GSM channels while Section VII briefly describes the VAD operation and comfort noise generation. The performance of the AMR-WB codec is described in Section VIII. Finally Section IX gives the conclusions.

II. Evolution of GSM and WCDMA 3G Speech Coding

The 13 kb/s Full-Rate (FR) codec was the first voice codec defined for GSM. The codec was standardized in 1989 and it is used in the GSM full-rate channel whose gross bit-rate is 22.8 kb/s. FR is the default codec to provide speech service in GSM. The 5.6 kb/s Half-Rate (HR) codec was standardized in 1995 to provide channel capacity savings through operation in the half-rate channel at a gross bit-rate of 11.4 kb/s. The HR codec provides the same level of speech quality as the FR codec, except in background noise and in tandem (two encodings in mobile-to-mobile calls) where the performance is somewhat lower.

The 12.2 kb/s Enhanced Full-Rate (EFR) codec was the first GSM codec to provide voice quality equivalent to that of a wireline telephony [3]. The EFR codec brought a substantial quality improvement over the two previous GSM codecs. EFR provides wireline speech quality across all typical radio conditions down to carrier-to-interference ratio (C/I) of approximately 10 dB [3]. Wireline quality was required since GSM had become increasingly used in communication environments where it started to compete directly with fixed line or cordless systems. To be competitive also with respect to speech quality, GSM needed to provide wireline speech quality which is robust to typical usage conditions such as background noise and transmission errors. The EFR was standardized first for the GSM-based PCS 1900 system in North America during 1995 and was adopted to GSM in 1996 through a competitive selection process. In addition to voice quality performance, the advantage of using the same voice codec in PCS 1900 and in GSM was one factor in favor of this particular codec. The EFR codec was jointly developed by Nokia and the University of Sherbrooke.

EFR still left some room for improvements. In particular, the performance in severe channel error conditions could be improved by employing a different bit-allocation between speech and channel coding. Also, the GSM half-rate channel was not yet able to provide high speech quality. A further development in GSM voice quality was the standardization of the AMR narrowband codec in 1999 [4]. The AMR codec offers major improvement over EFR in error robustness in the GSM full-rate channel by adapting speech and channel coding depending on prevailing channel conditions. In the full-rate channel, AMR extends the wireline quality operating region from about $C/I \geq 10$ dB in EFR to about $C/I \geq 4$–7 dB. By switching to operate in the half-rate channel during good channel conditions, AMR also gives channel capacity gain over EFR while sustaining high voice quality. AMR contains eight speech coding bit-rates between 4.75 and 12.2 kb/s. EFR source codec is included as one mode in AMR. The AMR codec was adopted in 1999 by 3GPP as the default speech codec for the WCDMA 3G system. The AMR codec was jointly developed by Ericsson, Nokia, and Siemens.

The AMR wideband codec, which was jointly developed by Nokia and VoiceAge, is the most recent voice codec standardized for GSM and WCDMA 3G systems. The codec was standardized in 2001. While all the previous codecs in mobile communication operate on narrow audio bandwidth limited to 200–3400 Hz, AMR-WB extends the audio bandwidth to 50–7000 Hz bringing substantial quality improvement. The AMR-WB codec operates on nine speech coding bit-rates between 6.6 and 23.85 kb/s [1]. Like the other GSM and WCDMA 3G codecs, AMR-WB has also a low bit-rate source dependent mode for coding background noise [5].

The standardization of AMR-WB codec was launched in mid-1999. Prior to that, a feasibility study phase had been carried out in ETSI during spring 1999 on the applicability of wideband coding for mobile communication. The results showed that the target is feasible and, consequently, the standardization was started. The work was carried out as a joint effort in ETSI and 3GPP targeting development of wideband coding for both second- and third-generation mobile communication systems. In the year 2000, all the GSM standardization work was transferred from ETSI to 3GPP, including the finalization of the AMR-WB codec.

After the launch of standardization, detailed speech quality performance requirements and codec design constraints (e.g., for implementation complexity and transmission delay) were defined. The AMR-WB codec selection was then carried out as a competitive process consisting of two phases: Qualification Phase in spring 2000 and Selection Phase in June–October 2000. From altogether nine codec candidates, seven codecs were submitted for the Qualification Phase. The five best codecs proceeded into the Selection Phase.

In the Selection Phase, the codec candidates were tested thoroughly in six independent test laboratories. Testing was coordinated internationally and was conducted with five languages. Each experiment in the tests was performed in two languages to avoid any bias due to a particular language. The tests covered speech with and without background noise, channel errors, mode adaptation and also source controlled rate operation. The candidate codecs were implemented in C-code with fixed-point arithmetic using the same set of basic operators used to define the previous GSM and WCDMA 3G codecs.

Based on the test results and technical details of the codec proposals, the Nokia/VoiceAge codec was selected as the 3GPP AMR-WB codec in December 2000. Since then the speech codec specifications have been finalized and they were approved in March 2001. Specifications of 3GPP related to AMR-WB are found in [1], [5]–[12].

After approval of the codec specifications, a Characterization Phase took place. During this phase, the AMR-WB codec was subjected to extensive testing in various operating conditions. The results are summarized in the 3GPP Technical Report in [13].

A floating-point arithmetic version of the AMR-WB codec was developed later and was standardized in March 2002 to be used in multimedia applications applying general purpose floating-point processors [14].

The AMR-WB codec has found applications in WCDMA 3G also beyond speech telephony. The codec has been standardized as the default codec for "speech" media type at 16 kHz sampling frequency for Packet Switched Streaming Service (PSS) [15] and Multimedia Messaging Service (MMS) [16].

III. Algorithmic Description of the Speech Encoder

The AMR-WB codec is based on the algebraic code excited linear prediction (ACELP) technology [17]. The ACELP technology has been very successful in encoding telephone-band speech signals and several ACELP-based standards are being deployed in a wide range of applications including digital cellular applications and VoIP (e.g., 3GPP AMR (TS 26.090) [4], ETSI EFR (TS 06.60) [3], NA-TDMA IS-641, NA-CDMA IS-127 and ITU-T G.729 and G.723.1 codecs).

Although ACELP technology gives good performance on narrow-band signals, some difficulties arise when applying the telephone-band optimized ACELP model to wideband speech, therefore additional features need to be added to the model for obtaining high quality on wideband signals. The ACELP model will often spend most of its encoding bits on the low-frequency region, which usually has higher energy contents, resulting in a low-pass output signal. To overcome this problem, the perceptual weighting filter has to be modified in order to suit wideband signals. Further, the pitch contents in the spectrum of voiced segments in wideband signals do not extend over the whole spectrum range and the amount of voicing shows more variation compared to narrow-band signals. Thus, it is important to improve the closed-loop pitch analysis to better accommodate the variations in the voicing level. For the same reasons, it is also important to improve periodicity enhancement techniques at the decoder. Another important issue that arises in coding wideband signals is the need to use very large excitation codebooks. Therefore, efficient codebook structures that require minimal storage and can be rapidly searched become essential.

In this section, the AMR-WB algorithm will be described with emphasis on novel methods that address the above mentioned issues and other features resulting in high-quality wideband ACELP coding.

A. Codec Overview

The AMR-WB speech codec consists of nine speech coding modes with bit-rates of 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.6 kb/s [6]. AMR-WB includes also a background noise mode which is designed to be used in discontinuous transmission (DTX) operation in GSM and as a low bit-rate source dependent mode for coding background noise in other systems. In GSM the bit-rate of this mode is 1.75 kb/s.

The 12.65 kb/s mode and the modes above it offer high quality wideband speech. The two lowest modes 8.85 and 6.6 kb/s are intended to be used only temporarily during severe radio channel conditions or during network congestion.

The block diagrams of the encoding and decoding algorithms are shown in Figs. 1 and 2, respectively. The bit allocation of the codec at different bit rates is shown in Table I. The AMR-WB codec operates at a 16 kHz sampling rate. Coding is performed in blocks of 20 ms. Two frequency bands, 50–6400 Hz and 6400–7000 Hz, are coded separately in order to decrease complexity and to focus the bit allocation into the subjectively most important frequency range. Note that already the lower frequency band (50–6400 Hz) goes far beyond narrowband telephony.

The input signal is down-sampled to 12.8 kHz and pre-processed using a high-pass filter and a pre-emphasis filter of the form $P(z) = 1 - \mu z^{-1}$ with $\mu = 0.68$. The ACELP algorithm is then applied to the down-sampled and pre-processed signal. Linear Prediction (LP) analysis is performed once per 20 ms frame. The set of LP parameters is converted to immittance spectrum pairs (ISP) [18] and vector quantized using split-multistage vector quantization with 46 bits. The speech frame is divided into four subframes of 5 ms each (64 samples). The adaptive and fixed codebook parameters are transmitted every subframe. The pitch lag is encoded with 9 bits in odd subframes and relatively encoded with 6 bits in even subframes. One bit per subframe is used to determine the low pass filter applied to the past excitation. The pitch and algebraic codebook gains are jointly quantized using 7 bits per subframe. Algebraic codebooks of size 36, 44, 52, 64, 72 and 88 bits are used at rates 12.65, 14.25, 15.85, 18.25, 19.85 and 23.05 kb/s, respectively. The 23.85 kb/s mode differs from the 23.05 kb/s mode in using 4 bits per subframes to encode the higher band gain. The 8.85 kb/s mode uses 46 bits for the LP parameters, 8 and 5 bits for the pitch lag in odd and even subframes, respectively, 20 bits for the algebraic codebook and 6 bits for the two gains in each subframe. The 6.6 kb/s mode uses 36 bits for the LP parameters, 8 bits for the pitch lag in the first subframe and 5 bits in the other subframes, 12 bits for the algebraic codebook, and 6 bits for the two gains in each subframe.

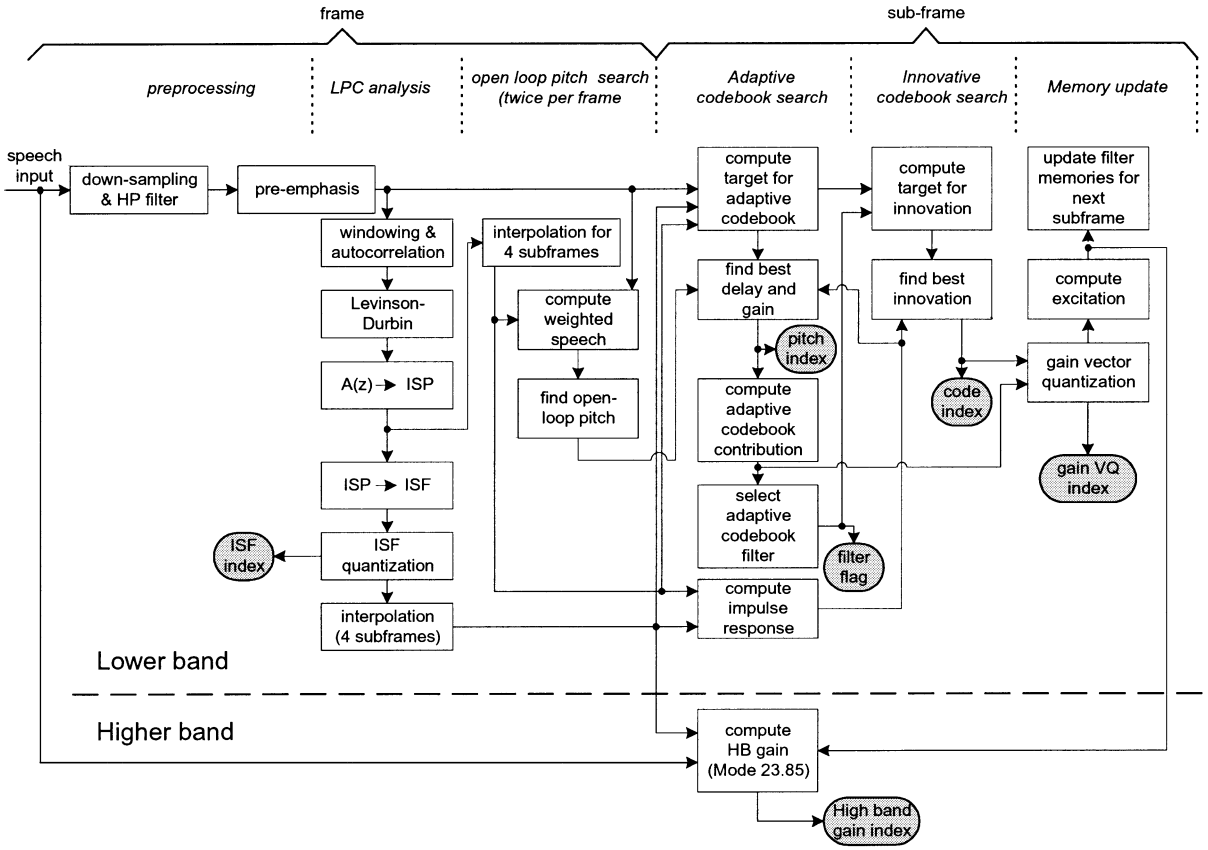

Fig. 1. Block diagram of the AMR-WB ACELP encoder.

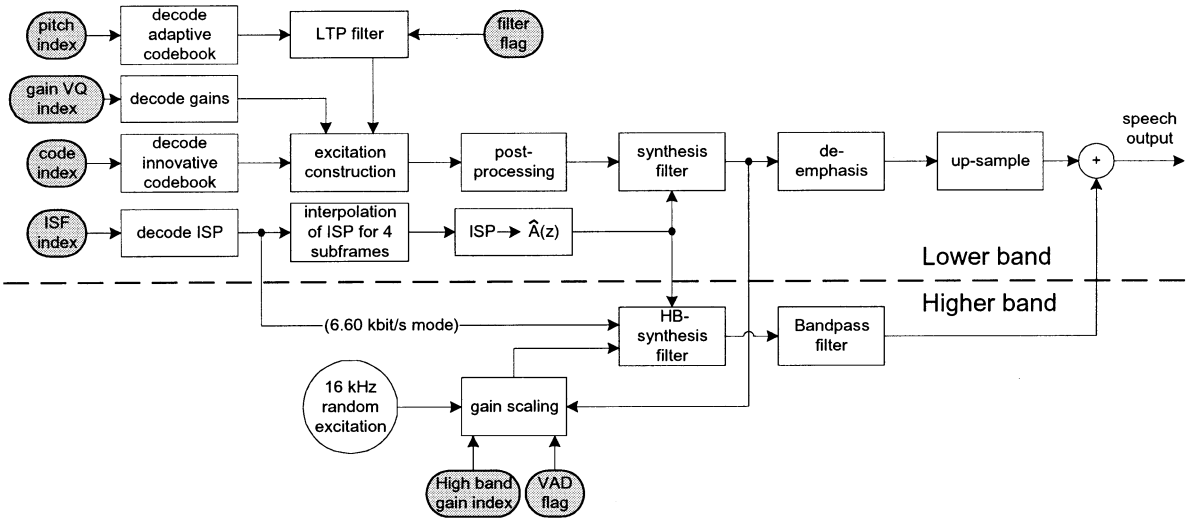

Fig. 2. Block diagram of the AMR-WB ACELP decoder.

The higher frequency band (6400–7000 Hz) is reconstructed in the decoder using the parameters of the lower band and a random excitation. No information about the higher band is transmitted, except in the 23.85 kb/s mode, where the higher band gain is transmitted using 4 bits per subframe. In other modes, the gain of the higher band is adjusted relative to the lower band using voicing information. The spectrum of the higher band is reconstructed by using a wideband LP filter generated from the lower band LP filter.

In the following subsections, several encoding functions will be discussed in more details.

B. Perceptual Weighting

In analysis-by-synthesis coders, the optimum pitch and innovation parameters are searched by minimizing the mean squared error between the input speech and synthesized speech in a perceptually weighted domain.

TABLE I
BIT ALLOCATION OF THE AMR-WB CODEC MODES

PARAMETER	CODEC MODE [kb/s] 6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
VAD flag	1	1	1	1	1	1	1	1	1
LTP filtering flag	0	0	4	4	4	4	4	4	4
ISP	36	46	46	46	46	46	46	46	46
Pitch delay	23	26	30	30	30	30	30	30	30
Algebraic code	48	80	144	176	208	256	288	352	352
Gains	24	24	28	28	28	28	28	28	28
High-band energy	0	0	0	0	0	0	0	0	16
Total per frame	**132**	**177**	**253**	**285**	**317**	**365**	**397**	**461**	**477**

Traditionally, perceptual weighting is performed through the use of a weighting filter

$$W'(z) = \frac{A'\left(\frac{z}{\gamma_1}\right)}{A'\left(\frac{z}{\gamma_2}\right)}, \quad 0 < \gamma_2 < \gamma_1 \leq 1 \tag{1}$$

where $A'(z)$ is the linear prediction filter and γ_1 and γ_2 are factors that control the amount of perceptual weighting bounded by $0 < \gamma_2 > \gamma_1 \leq 1$. It can be shown that the coding noise (which is assumed to have a white spectrum) is weighted by a transfer function $1/W'(z)$, which is the inverse of the transfer function of the perceptual weighting filter. Transfer function $1/W'(z)$ exhibits some of the formant structure of the input speech signal. Thus, the masking property of the human ear is exploited by shaping the quantization error so that it has more energy in the formant regions where it will be masked by the strong signal energy present in these regions.

This traditional perceptual weighting filter works well with telephone band signals where the transfer function $W(z)$ does not exhibit a strong spectral tilt. However, it is not suitable for efficient perceptual weighting of wideband signals due to the inherent limitations of $W(z)$ in modeling the formant structure and the required spectral tilt concurrently. The spectral tilt is more pronounced in wideband signals because of the wide dynamic range between low and high frequencies. It was suggested in the past to add a tilt filter into $W(z)$ in order to control the tilt and formant weighting of the wideband input signal separately [19].

A novel solution to this problem is to introduce a preemphasis filter of the form $P(z) = 1 - \mu z^{-1}$ at the input, compute the LP filter $A(z)$ based on the preemphasized speech $s(n)$ and use a modified filter $W(z)$ by fixing its denominator. For example, the perceptual weighting filter can be selected as

$$W(z) = \frac{A\left(\frac{z}{\gamma_1}\right)}{(1 - \gamma_2 z^{-1})}, \quad 0 < \gamma_2 < \gamma_1 \leq 1. \tag{2}$$

This structure substantially decouples the formant weighting from the tilt.

Note that because $A(z)$ is computed based on the preemphasized speech signal $s(n)$, the tilt of the filter $1/A(z/\gamma_1)$ is less pronounced compared to the case when $A(z)$ is computed based on the original speech. Since deemphasis is performed at the decoder end using a filter having the transfer function $1/P(z)$, the coding error is shaped by a filter $W^{-1}(z)P^{-1}(z)$. When γ_2 in (2) is set equal to μ, which is typically the case, the spectrum of the quantization error is shaped by a filter whose transfer function is $1/A(z/\gamma_1)$, with $A(z)$ computed based on the preemphasized speech signal. Subjective listening showed that this structure for achieving the error shaping by a combination of preemphasis and modified weighting filtering is very efficient for encoding wideband signals. Further, the preemphasis filter reduces the dynamic range of the input speech signal, which renders it more suitable for fixed-point implementation. Without preemphasis, LP analysis is difficult to implement in fixed-point using single-precision arithmetic.

Fig. 3 compares the noise shaping of the traditional and proposed weighting filters in case of a voiced speech segment. In the traditional case, the filter is given by $W'(z) = A'(z/0.9)/A'(z/0.6)$ where $A'(z)$ is computed using the original signal without pre-emphasis. The proposed filter is given by $W(z) = A(z/0.9)/(1 - 0.68z^{-1})$ where $A(z)$ is computed using the original signal after preemphasis with $P(z) = 1 - 0.68z^{-1}$. In the traditional case, the spectrum of the coding noise is shaped using the filter $1/W'(z) = A'(z/0.6)/A'(z/0.9)$ while in our case it is shaped using the filter $W^{-1}(z)P^{-1}(z) = 1/A(z/0.9)$. In Fig. 4, the same comparison is given in case of an unvoiced speech segment.

The proposed weighting filter shapes the coding noise in a way that it follows better the original speech spectrum. This is more evident on unvoiced segments where the traditional filter fails to shape the noise properly at low frequencies. In the example of Fig. 4, there is about 15 dB difference between the two filters at low frequencies.

C. Pitch Analysis

In the AMR-WB codec, the pitch search is composed of three stages. In the first stage, an open-loop pitch lag T_O is estimated every 10 ms based on the low-pass filtered decimated weighted speech signal. The details of the open-loop pitch analysis can be found in [6].

In the second stage, closed-loop pitch search is performed for integer pitch lags around the estimated open-loop pitch lag T_O at a range of ± 7 samples, which significantly simplifies the search procedure. Once an optimum integer pitch lag is found

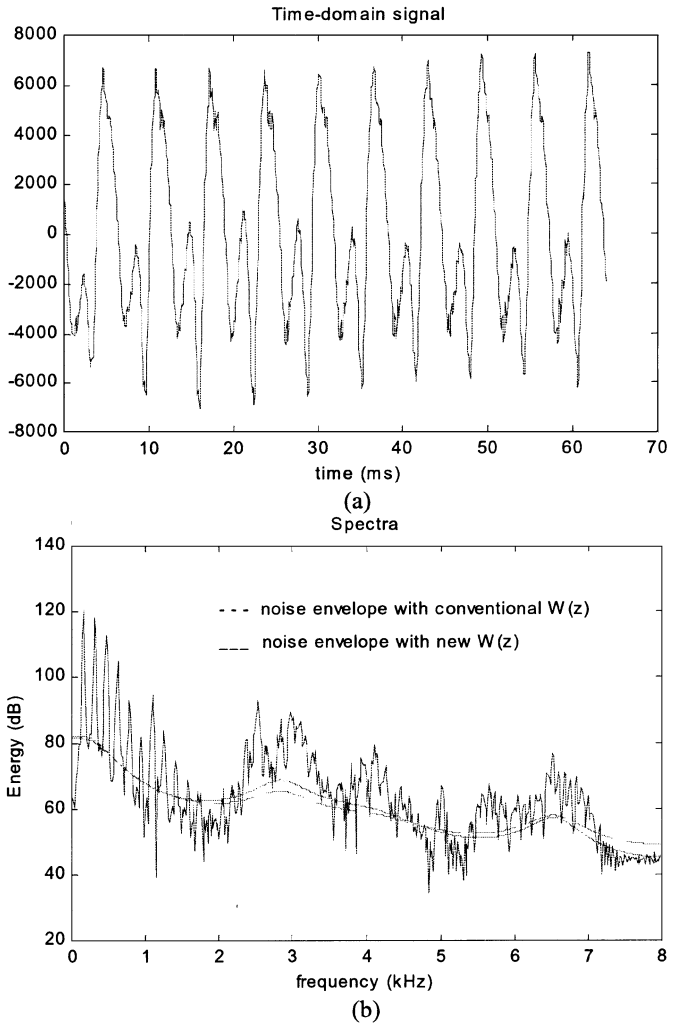

Fig. 3. (a) An example of a time domain voiced signal (b) Spectrum of the signal in (a) along with the quantization noise envelope after shaping by the conventional weighting filter and the proposed one.

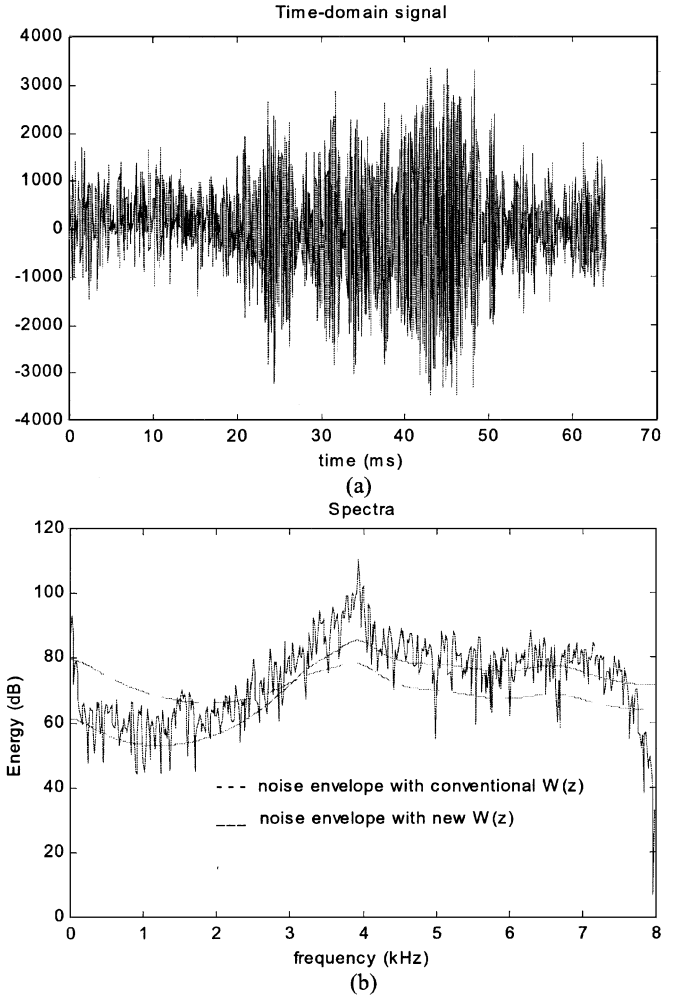

Fig. 4. (a) An example of a time domain unvoiced signal (b) Spectrum of the signal in (a) along with the quantization noise envelope after shaping by the conventional weighting filter and the proposed one.

in the second stage, a third stage of the search goes through the fractions around that optimum integer value. The pitch lag is bounded to the range [34, 231].

For the first and third subframe at bit rates of 12.65 kb/s and higher, a fractional pitch lag is used with resolutions of 1/4 and 1/2 in the ranges [34, 127(3/4)] and [128, 159(1/2)], respectively. Integer resolution is used in the range [160, 231]. For the second and fourth subframe, a resolution of 1/4 is always used in the search bounded to the range [−8, 7(3/4)] around the integer pitch lag found for the previous subframe.

The closed loop pitch search is performed by minimizing the mean-squared weighted error between the original and synthesized speech. This is achieved by maximizing

$$C_k = \frac{\sum_{n=0}^{N-1} x(n) y_k(n)}{\sqrt{\sum_{n=0}^{N-1} y_k(n) y_k(n)}} \tag{3}$$

where N is the subframe size, $x(n)$ is the target signal and $y_k(n)$ is the past excitation at lag k, filtered through the weighted synthesis filter $W(z)/\hat{A}(z)$ (note that $W(z)$ is the weighting filter described in Section III-B and $1/\hat{A}(z)$ is the synthesis filter with quantized LP coefficients). The signal $y_k(n)$ is computed by the convolution of past excitation at lag k, $u_k(n)$, with the impulse response of the weighted synthesis filter, $h(n)$. For pitch lags less than the subframe size, the past excitation is extended by the LP residual signal. Note that the search range is limited around the open-loop pitch estimate T_O as explained earlier.

Once the optimum integer pitch lag is determined, the fractions from −3/4 to 3/4 with a step of 1/4 around that integer are examined. The fractional pitch search is performed by interpolating the normalized correlation in (3) and searching for its maximum. Once the fractional pitch lag is determined, the interpolated past excitation $v'(n)$ is computed by interpolating the past excitation signal $u(n)$ at the given lag and fraction. The interpolation is performed using two FIR filters composed from a Hamming windowed sinc functions; one for interpolating the term in (3) with a filter order of 35 and the other for interpolating the past excitation with a filter order of 127. The filters have their cut-off frequency (−3 dB) at 6000 Hz, which means that the interpolation filters exhibit low-pass frequency response.

For enhancing the pitch prediction on wideband signals, a frequency-dependant pitch predictor is used. This is important

on wideband signals since the periodicity does not necessarily extend over the whole spectrum.

When the pitch predictor is represented by a filter $1/(1-bz^{-T})$, which is a valid assumption for pitch lags $T \geq N$, the pitch filter exhibits a harmonic structure over the entire frequency range, with a harmonic frequency related to $1/T$. On wideband signals, this structure is not efficient since the harmonic structure does not typically cover the entire spectrum up to 7 kHz. The harmonic structure exists only up to a certain frequency depending on the speech segment. Thus, in order to achieve efficient representation for the pitch contribution in the voiced segments of wideband speech, the pitch prediction filter needs to have the flexibility of varying the amount of periodicity at high frequency.

A new method which achieves efficient modeling of the harmonic structure of the speech spectrum on wideband signals is used, whereby several forms of low pass filters are applied to the past excitation and the low pass filter with the highest prediction gain is selected.

When fractional pitch resolution is used, the low pass filters can be incorporated into the interpolation filters used to obtain a higher pitch resolution. In this case, the third stage of the pitch search, in which the fractions around the chosen integer pitch lag are tested, is repeated for the several interpolation filters having different low-pass characteristics and the fraction and filter index which maximize the search criterion are selected.

A more pragmatic approach is to complete the search in the three stages described above to determine the optimum fractional pitch lag using only one interpolation filter (usually the filter with wider frequency response). The optimum low-pass filter shape can be determined at the end by applying the predetermined candidate filters to the chosen pitch codebook vector $\mathbf{v}_T$ and selecting the filter that minimizes the pitch prediction error.

In the AMR-WB codec, one bit per subframe is allocated for characterizing the low pass filter used for shaping the adaptive codebook excitation. When the low pass filter is disabled, the adaptive codebook excitation becomes simply $v(n) = v'(n)$. Otherwise, a second order FIR filter $B_L(z) = b_L(0)z + b_L(1) + b_L(2)z^{-1}$ is used for filtering the adaptive codebook resulting in

$$v(n) = \sum_{i=-1}^{1} b_L(i+1)v'(n+i). \tag{4}$$

The filter coefficients are chosen such that $b_L(0) = b_L(2) = 0.18$ and $b_L(1) = 0.64$. This second order filter gives a practical compromise between the complexity and desired high-frequency attenuation.

The adaptive codebook gain is then found by

$$g_p = \frac{\sum_{n=0}^{N-1} x(n)y(n)}{\sum_{n=0}^{N-1} y(n)y(n)} \quad \text{bounded by} \quad 0 \leq g_p \leq 1.2$$

where $y(n) = v(n)*h(n)$ is the filtered adaptive codebook vector (zero-state response of the weighted synthesis filter $W(z)/\hat{A}(z)$ to $v(n)$). To avoid potential instability at the decoder in case of channel errors, the adaptive codebook gain g_p is upper bounded by 0.95, if the adaptive codebook gains of the previous subframes have been high and the LP filters of the previous subframes have been close to being unstable.

TABLE II
PULSE POSITIONS FOR THE TRACKS OF THE ALGEBRAIC CODEBOOK

Track	Valid Pulse Positions in Subframe
T_0	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60
T_1	1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, 45, 49, 53, 57, 61
T_2	2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62
T_3	3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, 47, 51, 55, 59, 63

D. Structure and Search of the Algebraic Fixed Codebook

In wideband speech coding, very large codebooks are needed in order to guarantee a high subjective quality. In this section we describe an algebraic codebook structure where codebooks of size as high as 88 bits can be designed and efficiently searched.

In our implementation, the subframe size is 64 samples corresponding to 5 ms at a 12.8 kHz sampling rate. The codebook structure is based on *interleaved single-pulse permutation* (ISPP) design. As shown in Table II, the 64 positions in the codevector are divided into 4 tracks of interleaved positions, with 16 positions in each track. The different codebooks at the different rates are constructed by placing a certain number of signed pulses in the tracks, from 1 to 6 pulses per track. The codebook index, or codeword, represents the pulse positions and signs in each track. Thus, no codebook storage is needed, since the excitation vector at the decoder can be constructed through the information contained in the index itself without lookup tables.

If a single signed pulse is placed in each track, the pulse position is encoded with 4 bits and its sign is encoded with 1 bit, resulting in a 20-bit codebook. If two signed pulses are placed in each track, the two pulse positions are encoded with 8 bits and their corresponding signs can be encoded with only 1 bit by exploiting the pulse ordering. Therefore a total of $4\times(4+4+1) = 36$ bits are required to specify pulse positions and signs for this particular algebraic codebook structure. Other codebook structures can be designed by placing 3, 4, 5, or 6 pulses in each track. The encoding of pulses in each track is detailed in [6].

An important feature of the used codebook is that it is a dynamic codebook consisting of an algebraic codebook followed by an adaptive pre-filter $F(z)$ which enhances special spectral components for improving subjective speech quality. In the AMR-WB codec, $F(z)$ consists of a two-filter cascade, a periodicity enhancement filter $1/(1-0.85z^{-T})$ and a tilt filter $(1-\beta_T z^{-1})$. The coefficient β_T is related to the voicing of the previous subframe and is bounded into [0, 0.5] and T is the integer part of the pitch lag. The codebook search is performed in the algebraic domain by combining the filter $F(z)$ with the weighed synthesis filter prior to the codebook search. This is done by convolving the impulse response of the weighted synthesis filter with the impulse response of $F(z)$. Below the result of this convolution is denoted by $h(n)$.

The algebraic codebook is searched by minimizing the mean-squared error between the weighted input speech and

the weighted synthesized speech. The target signal used in the closed-loop pitch search is updated by subtracting the adaptive codebook contribution.

Let $h(n)$ denote the impulse response of the weighted synthesis filter convolved with the prefilter $F(z)$. Let $\mathbf{z}_k$ denote the filtered codevector at index k (of length N), corresponding to the signal $z_k(n) = c_k(n)^* h(n) = \sum_{i=0}^{n} c_k(i)h(n-i)$ (note that since the output of the convolution is of size N then only the values $h(n)$, $N = 0$, to $N-1$ are needed). It can be shown that the algebraic codebook is searched by maximizing the search criterion

$$Q_k = \frac{(\mathbf{x}_2^t \mathbf{z}_k)^2}{\mathbf{z}_k^t \mathbf{z}_k.}$$

Let the matrix $\mathbf{H}$ be defined as the lower triangular Toeplitz convolution matrix of size $N \times N$ with diagonal $h(0)$ and lower diagonals $h(1), \ldots, h(N-1)$, then $\mathbf{z}_k$ is given by $\mathbf{z}_k = \mathbf{H}\mathbf{c}_k$. Let also $\mathbf{d} = \mathbf{H}^t \mathbf{x}_2$ denote the correlation between the target signal $x_2(n)$ and the impulse response $h(n)$ and $\mathbf{\Phi} = \mathbf{H}'\mathbf{H}$ denote the matrix of correlations of $h(n)$. The ijth element of $\mathbf{\Phi}$ is denoted as $\phi(i,j)$. The correlation $\mathbf{d}$ is also referred to as a backward filtered target vector and the corresponding signal is denoted by $d(n)$. Therefore, the search criterion can be now written as

$$Q_k = \frac{(\mathbf{x}_2^t \mathbf{H}\mathbf{c}_k)^2}{\mathbf{c}_k^t \mathbf{H}^t \mathbf{H}\mathbf{c}_k} = \frac{(\mathbf{d}^t \mathbf{c}_k)^2}{\mathbf{c}_k^t \Phi \mathbf{c}_k} = \frac{(C_k)^2}{E_k} \tag{5}$$

The vector $\mathbf{d}$ and the matrix $\mathbf{\Phi}$ are usually computed prior to the codebook search.

The algebraic structure of the codebooks allows for very fast search procedures since the innovation vector $\mathbf{c}_k$ contains only a few nonzero pulses. The correlation in the numerator of (5) becomes

$$C = \sum_{i=0}^{N_p - 1} a_i d(p_i) \tag{6}$$

where p_i is the position of the ith pulse, a_i is its sign and N_p is the number of pulses in $\mathbf{c}_k$. The energy in the denominator of (5) is given by

$$E = \sum_{i=0}^{N_p - 1} \phi(p_i, p_i) + 2 \sum_{i=0}^{N_p - 2} \sum_{j=i+1}^{N_p - 1} a_i a_j \phi(p_i, p_j). \tag{7}$$

To simplify the search procedure, the pulse amplitudes are predetermined based on a reference signal $b(n)$. In this signal-selected pulse amplitude approach, the sign of a pulse at position i is set equal to the sign of the reference signal at that position. Here, the reference signal $b(n)$ is defined as

$$b(n) = \sqrt{\frac{E_d}{E_r}} r(n) + \alpha d(n) \tag{8}$$

where $E_D = \mathbf{d}^T \mathbf{d}$ is the energy of the signal $d(n)$. Similarly, $E_R = \mathbf{r}^T \mathbf{r}$ is the energy of the residual signal $r(n)$ resulting from pitch prediction. The scaling factor α controls the dependence of the reference signal on $d(n)$ and it is lowered as the number of excitation pulses is increased. As an example, the value of α is set to 1 at 12.65 kb/s and reduced to 0.5 at 23.05 kb/s.

To simplify the search the signal $d(n)$ and matrix $\mathbf{\Phi}$ are modified to incorporate the pre-selected signs. Let $s_b(n)$ denote the sign of $b(n)$. The modified signal $d'(n)$ becomes

$$d'(n) = s_b(n) d(n) \quad n = 0, \ldots, N-1 \tag{9}$$

and the modified autocorrelation matrix $\mathbf{\Phi}'$ becomes

$$\begin{aligned} \phi'(i,j) =& s_b(i) s_b(j) \phi(i,j), \quad i = 0, \ldots, N-1; \\ j =& i, \ldots, N-1. \end{aligned} \tag{10}$$

The numerator of the search criterion Q_k is now simply

$$C_k = \sum_{i=0}^{N_p - 1} d'(i) \tag{11}$$

and the denominator

$$E_k = \sum_{i=0}^{N_p - 1} \phi'(p_i, p_i) + 2 \sum_{i=0}^{N_p - 2} \sum_{j=i+1}^{N_p - 1} \phi'(p_i, p_j). \tag{12}$$

The goal of the search now is to determine the codevector with the best set of N_p pulse positions assuming amplitudes of the pulses have been selected as described above. The basic selection criterion is the maximization of Q_k.

For reducing the complexity, a fast procedure known as *depth-first tree search* is used, whereby the pulse positions are determined N_m pulses at a time. More precisely, the N_p available pulses are partitioned into M nonempty subsets of N_m pulses, respectively, such that $N_1 + N_2 \cdots + N_m \cdots + N_M = N_p$. A particular choice of positions for the first $J = N_1 + N_2 \cdots + N_{m-1}$ pulses considered is called a level-m path or a path of length J. The basic criterion for a path of J pulse positions is the ratio $Q_k(J)$ when only the J relevant pulses are considered.

The search begins with subset #1 and proceeds with subsequent subsets according to a tree structure such that subset $\#m$ is searched at the mth level of the tree. The purpose of the search at level 1 is to consider the N_1 pulses of subset #1 and their valid positions for determining one or a number of candidate paths of length N_1. These candidate paths are the tree nodes at level l. The path at each terminating node of level $m-1$ is extended to length $N_1 + N_2 \cdots + N_m$ at level m by considering N_m new pulses and their valid positions. One or a number of candidate extended paths are determined to constitute level-m nodes. The best codevector corresponds to that path of length N_p which maximizes the criterion $Q_k(N_p)$ with respect to all level-M nodes.

A special form of the depth-first tree search procedure is used here, in which two pulses are searched at a time, that is, $N_m = 2$ and these two pulses belong to two consecutive tracks. Further, the search is performed such that only a part of the correlation matrix $\mathbf{\Phi}$ is precomputed and stored. This reduces memory consumption as the full matrix $\mathbf{\Phi}$ requires a memory of $N \times N$ words ($64 \times 64 = 4096$ words). The prestored part represents

the correlations of the impulse response corresponding to potential pulse positions in consecutive tracks, as well as the correlations corresponding to $\phi(j,j)$ for $j=0,\ldots,N-1$, that is, the diagonal elements of matrix $\mathbf{\Phi}$.

To reduce complexity while testing possible combinations of two pulses, a limited number of potential positions of the first pulse are tested. Further, in case of large number of pulses, some pulses in the higher levels of the search tree are fixed. For making an intellegent guess on the potential positions to be considered for the first pulse, $b(n)$ defined in (8) is used a "pulse-position likelihood-estimate" based on the character of speech signals. The value of $b(n)$ characterizes the "probability" of a pulse occupying position n in the best codevector we are searching for.

The search procedures for all modes are similar. Two pulses are searched at a time and these two pulses always correspond to consecutive tracks. That is the two pulses to be searched are in tracks T_0 and T_1, T_1 and T_2, T_2 and T_3, or T_3 and T_0.

Before searching the positions, the sign of a pulse at potential position n is set the sign of $b(n)$ at that position. Then the modified signal $d'(n)$ and correlations $\phi'(n,n)$ are computed as described above by including the predetermined signs.

For the first two pulses (first tree level), the numerator of the search criterion Q_k is given by

$$C = d'(p_0) + d'(p_1) \tag{13}$$

and the denominator by

$$E = \phi'(p_0,p_0) + \phi'(p_1,p_1) + 2\phi'(p_0,p_1) \tag{14}$$

where the correlations $\phi'(p_i,p_j)$ has been modified to include the preselected signs at positions p_i and p_j.

For subsequent levels, the numerator and denominator are updated by adding the contribution of two new pulses. Assuming two new pulses at a certain tree level with positions p_k and p_{k+1} from two consecutive tracks are searched, the updated values of C and E is given by

$$C = C + d'(p_k) + d'(p_{k+1}) \tag{15}$$

$$E = E + \phi'(p_k,p_k) + \phi'(p_{k+1},p_{k+1}) + 2\phi'(p_k,p_{k+1}) + 2R_{hv}(p_k) + 2R_{hv}(p_{k+1}) \tag{16}$$

where $R_{hv}(m)$ denotes the correlation between the impulse response $h(n)$ and

$$v_h(n) = \sum_{i=0}^{k-1} h(n-p_i).$$

That is,

$$R_{hv}(m) = \sum_{n=m}^{N-1} h(n)v_h(n-m).$$

Note that $v_h(n)$ is the addition of delayed versions of the impulse response at the previously determined positions. At each tree level, the values of $R_{hv}(m)$ are computed online for all possible positions in each of the two tracks being tested. It can be seen from (16) that only the correlations $\phi'(m_k,m_{k+1})$ corresponding to pulse positions in two consecutive tracks need to be stored ($4\times16\times16$ words), along with the correlations $\phi'(m_k,m_k)$ corresponding to the diagonal of the matrix $\mathbf{\Phi}$ (64 words). Thus the memory requirement in the present algebraic structure is 1088 words instead of $64\times64=4096$ words.

The search of a 36-bit codebook employed in the 12.65 kb/s mode will be described as an example. In this codebook, two pulses are placed in each track giving a total of 8 pulses per subframe of length 64. Two pulses corresponding always to consecutive tracks are searched at a time. That is the two pulses to be searched are in tracks T_0 and T_1, T_1 and T_2, T_2 and T_3, or T_3 and T_0. The tree has four levels in this case. In the first level, pulse P_0 is assigned to track T_0 and pulse P_1 to track T_1. In this level, no search is performed and the two pulse positions are set to the two maxima of $b(n)$ in each track. In the second level, pulse P_2 is assigned to track T_2 and pulse P_3 to track T_3. Four positions for pulse P_2 are tested against all 16 positions of pulse P_3. The four tested positions of P_2 are determined based on the maxima of $b(n)$ in the track. In the third level, pulse P_4 is assigned to track T_1 and pulse P_5 to track T_2. Eight positions for pulse P_4 are tested against all 16 positions of pulse P_5. Similar to the previous level, the eight tested positions of P_4 are determined based on the maxima of $b(n)$ in the track. In the fourth level, pulse P_6 is assigned to track T_3 and pulse P_7 to track T_0. Eight positions for pulse P_6 are tested against all 16 positions of pulse P_7. Thus the total number of tested combination is $4\times16+8\times16+8\times16=320$. The whole process is repeated four times by assigning the pulses to different tracks. For example, in the second iteration, pulses P_0 to P_7 are assigned to tracks T_1, T_2, T_3, T_0, T_2, T_3, T_0 and T_1, respectively. Thus in total $4\times320=1280$ position combination are examined.

Another example is the 15.85 kb/s mode where three pulses are placed in each of four tracks for a total of 12 pulses. The three pulses in a track can be encoded with 13 bits (3 positions and 1 sign and the other two signs can be deduced from pulse ordering) giving a 52-bit codebook. There are 6 levels in the tree search whereby two pulses are searched in each level. In the first two levels, four pulses are set to the maxima of $b(n)$. In the subsequent four levels, the numbers of tested combinations are 4×16, 6×16, 8×16 and 8×16, respectively. Four iterations are used giving a total of $4\times26\times16=1664$ combinations.

E. Gain Quantization

The adaptive codebook gain and the fixed codebook gain are vector quantized using a 6-bit codebook for modes 8.85 and 6.60 kb/s and a 7-bit codebook for all the other modes. The fixed codebook gain quantization uses moving-average prediction with constant coefficients. The fourth order MA prediction is performed on the innovation energy in the logarithmic domain. The prediction and codebook search are similar to those used in the GSM EFR codec [3] or G.729 [20].

IV. Algorithmic Description of the Speech Decoder

The function of the decoder consists of decoding the transmitted parameters (VAD-flag, LP parameters, adaptive codebook indices, adaptive codebook gains, fixed codebook

vectors, fixed codebook gains and high-band gains) and performing synthesis to obtain the reconstructed speech. The reconstructed speech is then postprocessed and upsampled. Finally the high-band signal is generated to the frequency band from 6.4 to 7.0 kHz. The signal flow at the decoder is shown in Fig. 2. Several postprocessing features have been added to the decoder to improve the performance of wideband signals as will be explained below.

A. Decoding, Excitation Postprocessing, and Speech Synthesis

The received indices of ISP quantization are used to reconstruct the quantized ISP vector. Four interpolated ISP vectors (corresponding to 4 subframes) are computed and then converted to LP filter coefficient domain a_k, which is used for synthesizing speech in the subframe. Then in each subframe, the received adaptive codebook index is used to find the integer and fractional parts of the pitch lag. The adaptive codebook excitation $v'(n)$ is found by interpolating the past excitation $u(n)$ at the pitch lag. If the pitch filter flag is disabled, the adaptive codebook excitation becomes $v(n) = v'(n)$. Otherwise $v'(n)$ is filtered through a pitch filter $B_L(z)$ for obtaining $v(n)$ as described above in (4).

The received algebraic codebook index is used to extract the positions and signs of the excitation pulses and to find the algebraic codebook excitation $c(n)$. If the integer part of the pitch lag is less than the subframe size 64, the pitch sharpening procedure is applied which translates into filtering $c(n)$ through the prefilter $F(z)$ described above. Then the adaptive and fixed codebook gains are decoded and the total excitation is constructed by

$$u(n) = \hat{g}_p v(n) + \hat{g}_c c(n) \tag{17}$$

Before the speech synthesis, a post-processing of excitation elements is performed for improving the codec subjective performance. Post-processing procedures applied to the excitation signal are given below.

Anti-Sparseness Processing in the 6.60 and 8.85 kb/s Modes: Adaptive anti-sparseness post processing is applied to the fixed codebook excitation $c(n)$ for reducing perceptual artefacts arising from the sparseness of the algebraic fixed codebook with only a few nonzero samples per subframe [21]. The anti-sparseness processing consists of circular convolution of the fixed codebook excitation with an impulse response. Three pre-stored impulse responses are used corresponding to no modification, medium modification and strong modification. The selection of the impulse response is performed adaptively based on the adaptive and fixed codebook gains. More details about the impulse response selection are given in [6].

Noise Enhancement: A nonlinear gain smoothing technique is applied to the fixed codebook gain $\hat{g}_c$ to enhance excitation on noise. Based on the stability and voicing, the fixed codebook gain is smoothed for reducing fluctuation in the energy of the excitation on stationary signals. This improves the subjective speech quality on stationary background noise.

The voicing of the subframe is estimated by

$$\lambda = \frac{E_c}{(E_v + E_c)}$$

where E_v and E_c are the energies of the adaptive codebook and fixed-codebook excitation, respectively. The voicing factor λ is bounded into [0, 1] by definition. It attains a value close to zero on purely voiced subframes and one on unvoiced subframes. The stationarity of the frame is estimated based on LP filters using the difference of the ISP parameters between the current and previous frame as the stationarity measure. The ISP distance measure between the ISP's in the present frame n and the past frame $n-1$ is given by the relation

$$D_s = \sum_{i=1}^{p-1} \left(isp_i^{(n)} - isp_i^{(n-1)} \right)^2$$

where $p = 16$ is the order of the LP filter. Note that the first $p-1$ ISP's being used are frequencies in the range 0 to 6400 Hz. The ISP distance measure is mapped to a stability factor θ in the range 0 to 1 using the relation

$$\theta = 1.25 - \frac{D_s}{400\,000.0} \quad \text{bounded by } 0 \le \theta \le 1.$$

Note that the value of the stationarity factor θ tends to unity on stationary frames.

Finally, the gain smoothing factor σ is computed as

$$\sigma = \lambda\theta. \tag{18}$$

The value of σ tends to unity for unvoiced and stationary signals, such as stationary background noise. For purely voiced or unstationary signals, the value of σ tends to zero.

The nonlinear smoothing uses the gain of the preceding subframe as a reference $\hat{g}_c^r$ for computing the modified gain $\hat{g}_c^m$. If $\hat{g}_c \ge \hat{g}_c^r$, the modified gain $\hat{g}_c^m$ is set 1.5 dB below $\hat{g}_c$ but lower bounded to the reference. Otherwise, $\hat{g}_c^m$ is set 1.5 dB above $\hat{g}_c$, but upper bounded to the reference. Finally, the fixed codebook gain is smoothed using the smoothing factor and the modified gain

$$\hat{g}_c := \sigma\hat{g}_c^m + (1-\sigma)\hat{g}_c. \tag{19}$$

Thus, in stationary unvoiced subframes, the fixed codebook gain used in the synthesis approaches the modified gain $\hat{g}_c^m$. The smoothing algorithm uses the outcome of (19) in the next subframe as a reference gain.

Pitch Enhancement: A pitch enhancer improves the subjective speech quality by filtering the fixed codebook excitation $c(n)$ through an innovation filter whose frequency response emphasizes high frequencies and attenuates low frequencies. The coefficients of the pitch enhancement filter

$$F_E(z) = -f_E z + 1 - f_E z^{-1} \tag{20}$$

are related to the periodicity of the signal as $f_E = 0.25\lambda$ with the voicing factor $\lambda = E_c/(E_v + E_c)$ as described above. The filter coefficients are updated once a subframe. The filter (20) gives the maximum enhancement on unvoiced signals.

With the pitch enhancement filter, the fixed codebook excitation becomes

$$c_E(n) = c(n) - f_E(c(n+1) + c(n-1)) \tag{21}$$

and consequently the total excitation is given by

$$\hat{u}(n) = \hat{g}_p v(n) + \hat{g}_c c_E(n). \tag{22}$$

The above procedure can be done in one step by updating the excitation as follows:

$$\hat{u}(n) = u(n) - \hat{g}_c f_E(c(n+1) + c(n-1)). \tag{23}$$

Speech Synthesis: The reconstructed speech for the subframe of size 64 is obtained by filtering the postprocesed total excitation through the LP synthesis filter. That is

$$\hat{s}(n) = \hat{u}(n) - \sum_{i=1}^{16} \hat{a}_i \hat{s}(n-i) \quad \text{for} \quad n = 0, \ldots, 63 \tag{24}$$

where $\hat{a}_i$ is the ith coefficient of the interpolated LP filter. The synthesis speech $\hat{s}(n)$ is then passed through an adaptive postprocessing which is described in the following section.

Note that in order to keep synchronization with the encoder, the memory of the adaptive codebook is updated using the excitation $u(n)$ without postprocessing.

B. Upsampling and High-Frequency Generation

The synthesis signal is high-pass filtered as a precaution against undesired low frequency components. The signal is then de-emphasized using the filter $1/P(z) = 1/(1 - 0.68z^{-1})$. Finally, the signal is upsampled to obtain the lower band synthesis $\hat{s}_L(n)$ at a 16 kHz sampling rate.

A high-frequency generation procedure is used to fill the frequency band between 6.4 and 7 kHz. The high frequency content is generated by filling the upper part of the spectrum with a white noise properly scaled in the excitation domain, then converted to the speech domain by shaping it with a filter derived from the same LP synthesis filter used for synthesizing the down-sampled signal.

The high-band excitation $u_H(n)$ is obtained from a white noise signal $w(n)$ as

$$u_H(n) = \frac{\hat{g}_H w(n)}{\sqrt{\sum_{k=0}^{63} w^2(k)}} \tag{25}$$

where is $\hat{g}_H$ is the high-frequency gain. In the 23.85 kb/s mode, $\hat{g}_H$ is decoded from the received gain index. In all other modes, $\hat{g}_H$ is estimated for every subframe using voicing information. First, the tilt τ of synthesis signal is found as

$$\tau = \frac{\sum_{n=0}^{63} \hat{s}(n)\hat{s}(n-1)}{\sum_{n=1}^{63} \hat{s}^2(n)} \tag{26}$$

conditioned by $0 \le \tau \le 1 - 2\lambda$ with λ being the voicing factor defined above, but bounded into [0.1, 1.0]. The high-frequency gain $\hat{g}_H$ is then computed as

$$\hat{g}_H = w_S(1 - \tau) + 1.25(1 - w_S)(1 - \tau) \tag{27}$$

TABLE III
IMPLEMENTATION COMPLEXITY OF AMR-WB AND AMR-NB SPEECH CODECS

COMPLEXITY FIGURE	CODEC	
	AMR-WB	AMR-NB
Computational complexity [WMOPS]		
Speech Encoder	31.1	14.2
Speech Decoder	7.8	2.6
Total	38.9	16.8
Data RAM [kWords, 16-bit]	6.5	5.3
Data ROM [kWords, 16-bit]	9.9	14.6
Program ROM [num. of ETSI basicops]	3889	4851

where w_S is one when VAD is ON and zero otherwise. Hence on active speech the high frequency gain is $1-\tau$ and on background noise signals a higher gain $1.25(1 - \tau)$ is used.

The tilt factor is incorporated into the gain computation to take into account the high frequency contents of the synthesized signal. On voiced segments where less energy is present at high frequencies, τ approaches one resulting in a lower high-frequency gain. This reduces the energy of the generated noise on voiced segments.

The high-band LP synthesis filter $1/\hat{A}_H(z)$ is obtained for every subframe from a low-band LP synthesis filter using

$$\hat{A}_H(z) = \hat{A}\left(\frac{12.8z}{16}\right) \tag{28}$$

where $1/\hat{A}(z)$ is the interpolated LP synthesis filter. The filter $1/\hat{A}(z)$ has been estimated for a sampling rate of 12.8 kHz but it is now used for a 16 kHz signal. Effectively, this means the band 5.1–5.6 kHz in 12.8 kHz domain is mapped into the band 6.4–7.0 kHz in 16 kHz domain.

The high-band synthesis $\hat{s}_H(n)$ is computed by filtering $u_H(n)$ through $1/\hat{A}_H(z)$ and band-limiting the output with a bandpass FIR filter from 6.4 to 7 kHz. Finally, the high-band synthesis $\hat{s}_H(n)$ is added to the low-band synthesis $\hat{s}_L(z)$ to produce the synthesized speech signal.

V. CODEC COMPLEXITY

The AMR-WB codec is defined in fixed-point arithmetic using a set of basic operators defined by 3GPP/ETSI. Basic operators are a C-language implementation of commonly found fixed-point DSP assembly instructions. Describing an algorithm in terms of basic operators allows for easy mapping of the C-code to a certain DSP assembly language as well as a rough estimate of the algorithmic complexity. A certain weight is associated with each basic operator which reflects the number of instruction cycles. The computational complexity of the AMR-WB speech codec is 38.9 WMOPS (Weighted Million Operations Per Second). Note that this estimate includes the VAD/DTX/CNG functions. The complexity is the theoretical worst case complexity of the codec, in which the worst case complexity path through the codec is assumed. The complexity of the AMR-WB speech codec is show in Table III. For comparison, the complexity of AMR-NB speech codec is also shown in the table.

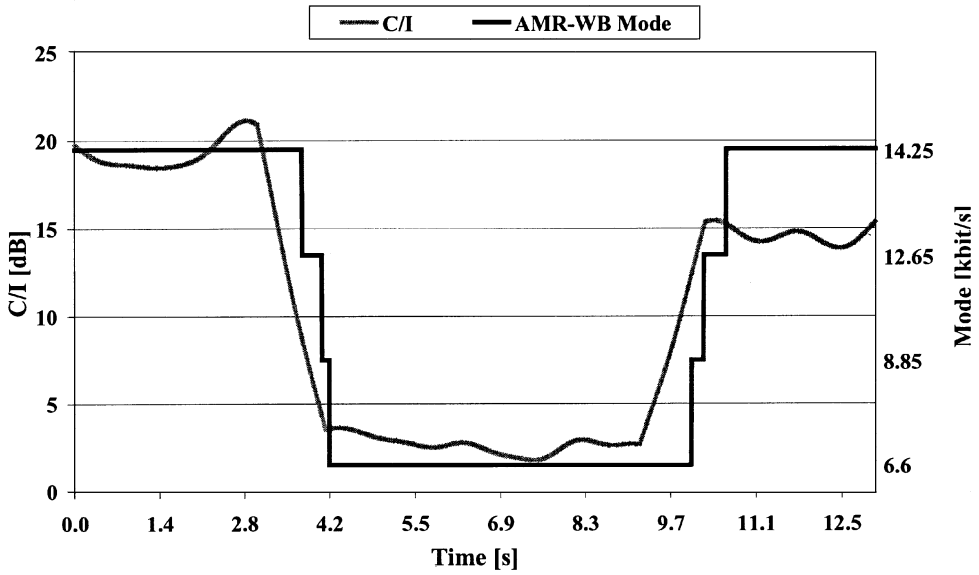

Fig. 5. Example of AMR-WB mode adaptation in GSM Full Rate channel.

VI. Adaptive Operation in GSM Channel

AMR-WB has high granularity of bit-rates between 6.6 and 23.85 kb/s. For GSM channels, this makes it possible to maximize speech quality by adapting the codec bit-rate to increase robustness against transmission errors. For nonadaptive 3G UTRAN channels using fast power control, operators can select the suitable bit-rates to make an optimal trade-off between the speech quality and network capacity.

The link adaptation process in GSM bears responsibility for measuring the channel quality and selecting the most appropriate mode according to prevailing channel conditions [22]. Link adaptation also takes into account constraints for available bit-rate set by the network (e.g., network load).

In-band signalling (400 b/s in full-rate channel) transmits both the requested mode for the reverse link (based on channel quality measurements) and active codec mode of the forward link over the air interface to the receiver side. For a more detailed description of the adaptation see [23].

Fig. 5 shows an example of how the codec mode adaptation works in the GSM full-rate channel when the Carrier to Interference ratio (C/I) varies between 22 and 2 dB. Based on estimated channel quality (e.g., C/I), one out of the activated codec modes (14.25, 12.65, 8.85, or 6.6 kb/s) is chosen, resulting in high speech quality throughout the call.

VII. Voice Activity Detection and Discontinuous Transmission

The AMR-WB codec includes a Voice Activity Detection (VAD). VAD allows the codec to switch to a lower-rate mode for coding of background noise. This feature saves power in the mobile station and also reduces the overall interference level over the air interface.

VAD computes a Boolean VAD decision for each 20 ms speech frame. The VAD decision is based on dividing the speech signal at the frequency band [0, 6.4 kHz] into 12 sub-bands and computing the level of the signal in each band. A tone detection function is used for indicating the presence of a signalling tone, voiced speech, or other strongly periodic signal. The function is based on the normalized open-loop pitch gains which are calculated by pitch analysis of the speech encoder. The tone detection function generates a tone-flag.

Background noise level is estimated in each frequency band based on the VAD decision, signal stationarity and the tone-flag. Intermediate VAD decision is then calculated by comparing input SNR (ratio of the signal level and the background noise level) to an adaptive threshold. The adaptation of the threshold is based on noise and long term speech level estimates. The final VAD decision is calculated by adding hangover period to the intermediate VAD decision. The details of the VAD algorithm can be found in [9].

In GSM, the VAD decision is used by the Discontinuous Transmission (DTX) mechanism, which allows the radio transmitter to be switched off most of the time during speech pauses. During speech pauses, synthetic noise similar to the transmit side background noise is generated on the receive side. This synthetic comfort noise is produced by transmitting parameters describing background noise at a regular rate during speech pauses. This allows the comfort noise to adapt to the changes of the background noise.

The comfort noise analysis is carried out using the 12.8 kHz input signal. The algorithm first determines the weighted average of the spectral parameters and the average of the logarithmic signal energy. These parameters are averages from the CN averaging period (eight most recent frames) and they give information on the level and the spectrum of the background noise. The encoder also determines how stationary the background noise is by using spectral distances between all the spectral parameter vectors in the averaging period as well as using the variation of the energy between frames. The comfort noise parameters are encoded into a special frame, called a Silence Descriptor (SID) frame for transmission to the receive side.

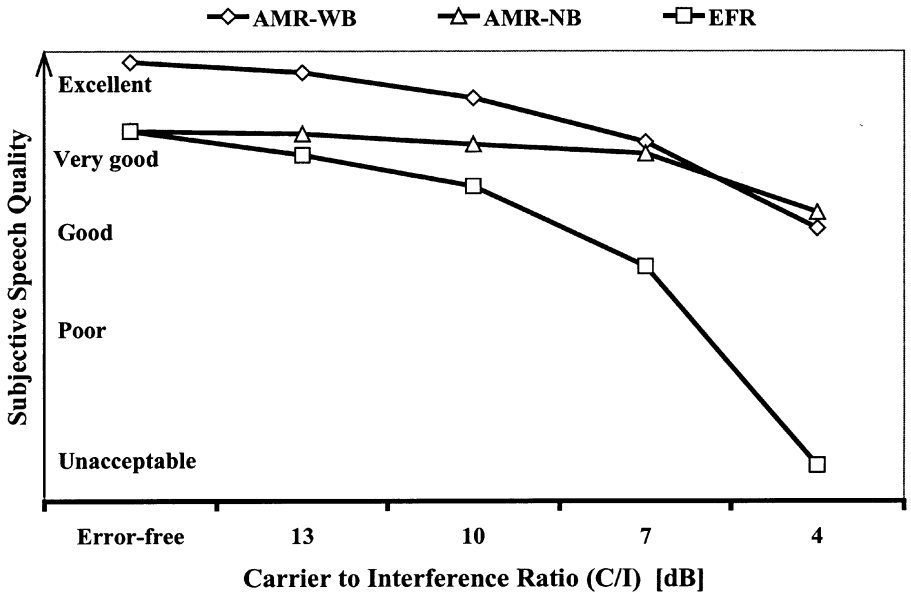

Fig. 6. Speech quality of AMR-WB in GSM FR channel compared to AMR-NB and EFR.

The comfort noise generation procedure generates a pseudo random noise vector that is scaled using an excitation gain computed from the logarithmic signal energy. The scaled pseudo random noise is filtered through a synthesis filter generated from the weighted average of the spectral parameters. For nonstationary background noise, dithering of both spectral parameters and excitation gain is employed. The synthesis is up-sampled back to 16 kHz and the higher band is generated like in the speech decoder. More detailed description of VAD, DTX and Comfort noise generation can be found in [5] and [10].

VIII. Speech Quality

The AMR-WB codec provides mobile communication with speech quality that substantially exceeds the quality provided by existing second– and third-generation systems.

Fig. 6 shows an illustrative graph comparing AMR-WB to narrowband codecs AMR-NB and EFR in the GSM full-rate channel. In typical operating conditions ($C/I > 10$ dB), AMR-WB gives superior quality over all other GSM codecs. Even in poor radio channel conditions ($C/I < 7$ dB), AMR-WB still offers comparable quality to AMR-NB and far exceeds the quality of the fixed rate GSM codecs.

AMR-WB provides high granularity of bit-rates making it suitable for many applications in 2G and 3G systems. The high speech quality makes the codec well suited also for wideband voice applications in wireline services as shown by its adoption by ITU-T.

A. Formal Testing of the AMR-WB Codec

During the AMR-WB standardization in 3GPP, the codec was extensively tested in several phases: Selection Phase, Verification Phase and Characterization phase. Later the codec participated also in the ITU-T Selection and Characterization Phases.

During the 3GPP selection phase, the AMR-WB codec was tested in six independent listening laboratories with five languages: Japanese, English, French, Mandarin Chinese, and Spanish. The testing covered different input levels, tandeming, background noise performance and performance of VAD/DTX. In addition, the codec was tested under different error conditions in mobile communication channels both in GSM and WCDMA 3G. The AMR-WB codec showed constant good performance. It met all performance requirements in all of the laboratories throughout the tests.

During the post-selection verification phase, the AMR-WB codec was tested in several additional conditions to verify its good performance. The tests included performance for DTMF tones and for other special input signals, overload performance, muting behavior, comfort noise generation and performance for music signals. The codec showed good performance throughout the tests. Transmission delay, frequency response, and implementation complexity were also analyzed in detail.

The characterization phase contained further testing to fully characterize all the nine codec modes of the chosen AMR-WB codec. A total of six different languages were used: English, Finnish, Spanish, French, German and Japanese. The experiments included tests for input levels and self-tandeming, interoperability in real world wideband and narrowband scenarios, VAD/DTX, clean speech and speech in background noise (four types), channel errors in GSM and WCDMA 3G channels (for both clean speech and speech under background noise).

Furthermore, the AMR-WB codec was tested during the ITU-T selection phase. The testing covered several input levels, tandeming, four types of background noise, frame erasure testing and testing with narrow-band speech signals. The tests were conducted in several languages including English, French, German, and Japanese. In the ITU-T testing only a subset of modes were tested: 12.65, 15.85, 19.85, and 23.85 kb/s.

Further characterization testing was carried out in ITU-T. These complement the earlier testing phases and cover quality of various sub-sets of modes with VAD/DTX on and off, additional background noise types and quality for music

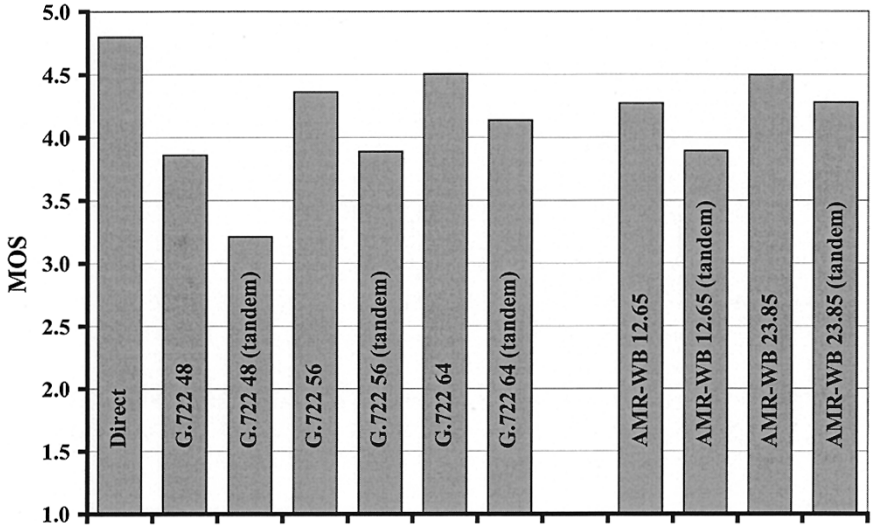

Fig. 7. AMR-WB quality with clean speech. From experiment 1a of ITU-T selection test performed in French language.

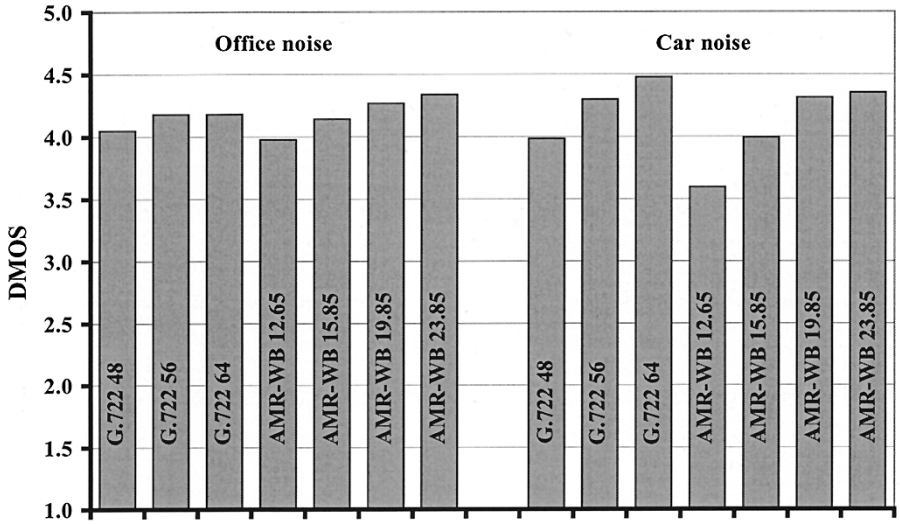

Fig. 8. AMR-WB quality in the presence of background noise. From experiment 3a of ITU-T selection test in American English.

signals. The quality of the AMR-WB codec is described in the following subsections based on the different phases of the testing.

B. Basic Quality

The clean speech quality provided by the six highest AMR-WB modes (23.85, 23.05, 19.85, 18.25, 15.85, 14.25) is equal to or better than ITU-T wideband codec G.722 at 64 kb/s. Results are consistent over all tested input levels and also in self-tandeming. The 12.65 kb/s mode is at least equal to G.722 at 56 kb/s. The 8.85 kb/s mode gives still quality equal to G.722 at 48 kb/s.

The clean speech quality of the AMR-WB codec is illustrated in Fig. 7. This is an extract from the ITU-T selection tests. The figure shows codec performance for nominal signal level (−26 dBov) with clean speech in single coding and in self-tandeming. The 12.65 and 23.85 kb/s modes were included in this experiment carried out in the French language. ITU-T wideband speech codec G.722 with three bit rates of 48, 56 and 64 kb/s was used as a reference codec. The results show that the performance of the 12.65 kb/s AMR-WB mode already exceeds the performance of G.722 at 48 kb/s and is about the same as the quality of G.722 at 56 kb/s. The highest AMR-WB mode 23.85 kb/s has performance equal to G.722 at 64 kb/s. The above observations are valid both in single coding and in self-tandeming.

The background noise performance of the AMR-WB codec is shown in Fig. 8. This is an extract from the ITU-T selection tests showing results for the English language. AMR-WB modes of 12.6, 15.85, 19.85, and 23.85 kb/s were included in the test. Office and car noise were used with SNR 15 dB for both types of noise. For office noise, the lowest tested AMR-WB mode of 12.65 kb/s has about equal performance to G.722 at 48 kb/s while the other modes perform better or equal to G.722 at 64 kb/s. For car noise, the performance of the two highest modes are about equal to G.722 at 56 kb/s and the 15.85 kb/s mode has about the same performance as G.722 at 48 kb/s.

The music performance of AMR-WB codec is illustrated in Fig. 10. This extract from the ITU-T characterization tests show that for music (classical and modern, both instrumental music and vocal music) the highest bit-rate mode of AMR-WB gives equivalent quality to G.722 at 56 kb/s.

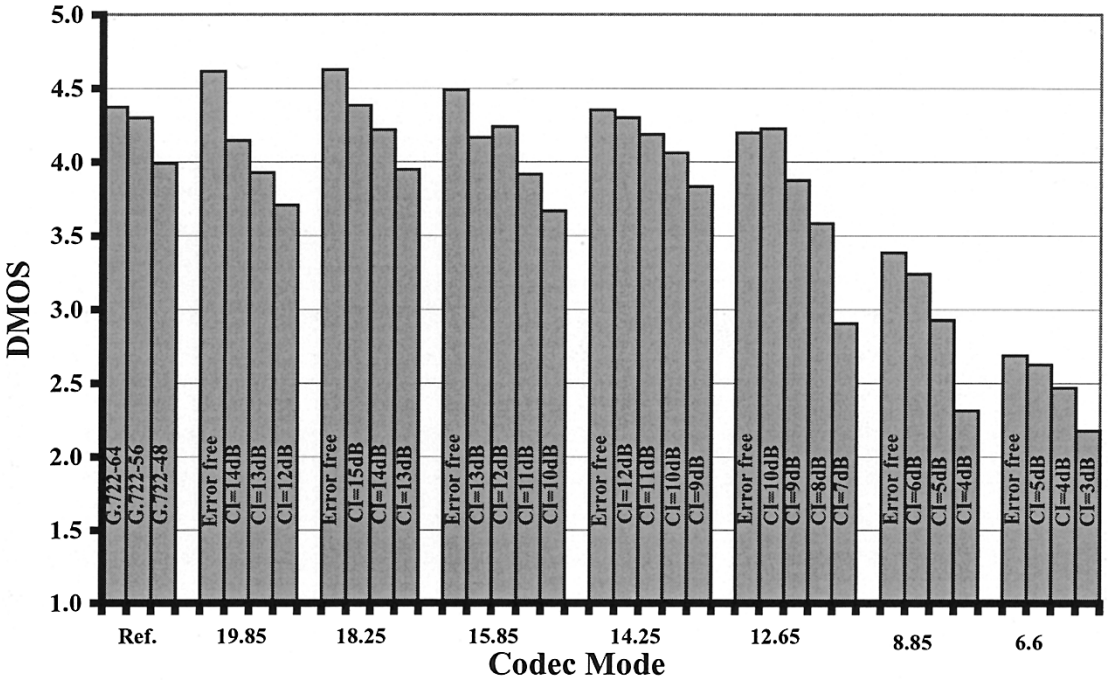

Fig. 9. AMR-WB performance in GSM full-rate channel under channel errors and with 15 dB Car background noise. From experiment 6a of 3GPP characterization phase with English language.

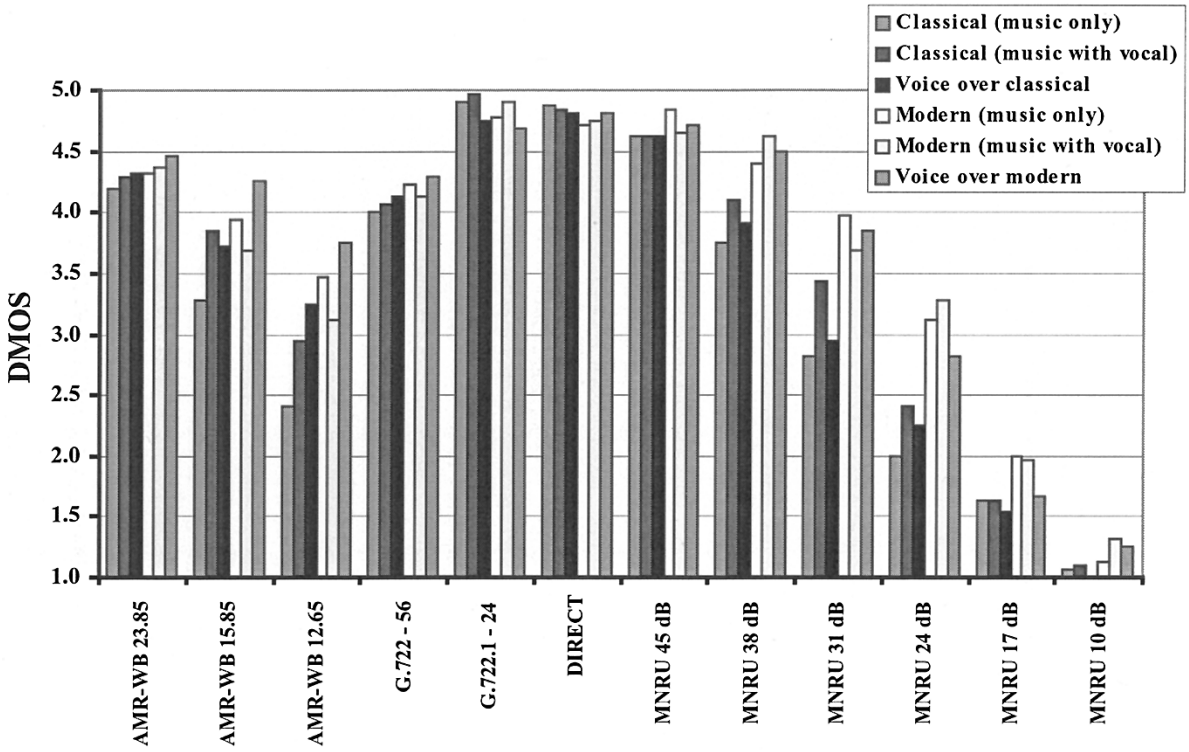

Fig. 10. AMR-WB performance with music. From experiment 3b of ITU-T characterization test.

C. Cellular Environment

The AMR-WB codec has been developed for use in mobile radio environment where typical usage conditions may include both channel errors and high level background noise. AMR-WB provides robust operation even in these conditions.

In the GSM full-rate channel (with clean speech), AMR-WB provides quality better or equal to G.722 at 64 kb/s at about 11 dB C/I and above. A quality at least equal to G.722 at 56 kb/s is obtained for error rates at about 10 dB C/I and above. Under background noise (15 dB Car Noise and 20 dB Office Noise), AMR-WB provides in the GSM full-rate channel quality equal to or better than G.722 at 64 kb/s at C/I-ratios about 12 dB and above. AMR-WB gives quality equal to or better than G.722 at 56 kb/s at C/I-ratios about 10 dB and above.

Fig. 9 shows an extract of the performance of AMR-WB in GSM full-rate channel under channel errors and with 15 dB

Car background noise. This experiment is taken from the 3GPP characterization phase and it was carried out using the English language [13]. The performance curves of each codec mode are shown. Note that the two highest AMR-WB bit-rates were not included in the test as they are not targeted for GSM full-rate use.

The VAD/DTX/CNG operation has been assessed as transparent to the listener in the 3GPP characterization tests.

IX. Conclusions

AMR-WB extends the audio bandwidth to 7 kHz and gives superior speech quality and voice naturalness compared to existing codecs in fixed line telephone networks and in second- and third-generation mobile communication systems. The introduction of AMR-WB to GSM and WCDMA 3G systems brings a fundamental improvement of speech quality, raising it to a level never experienced in mobile communication systems before. It far exceeds the current high quality benchmarks for narrow-band speech quality and changes the expectations of a high quality speech communication in mobile systems.

The good performance of the AMR-WB codec has been made possible by the incorporation of novel techniques into the ACELP model in order to improve the performance of wideband signals.

The AMR-WB codec has also been selected by the ITU-T in the standardization activity for a wideband codec around 16 kb/s. This is of significant importance since this is the first time that the same codec is adopted for wireless as well as wireline services. This will eliminate the need of transcoding and ease the implementation of wideband voice applications and services across a wide range of communications systems.

References

[1] "AMR Wideband Speech Codec; General Description,", 3GPP TS 26.171.

[2] "Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB),", Geneva, ITU-T Recommend. G.722.2, 2002.

[3] K. Järvinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Adoul, "GSM enhanced full rate codec," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Munich, Germany, Apr. 20–24, 1997, pp. 771–774.

[4] K. Järvinen, "Standardization of the adaptive multi-rate codec," in European Signal Processing Conference (EUSIPCO), Tampere, Finland, Sept. 4–8, 2000.

[5] *AMR Wideband Speech Codec; Source Controlled Rate Operation*, 3GPP TS 26.193.

[6] *Adaptive Multi-Rate Wideband Speech Transcoding*, 3GPP TS 26.190.

[7] *AMR Wideband Speech Codec; ANSI-C Code*, 3GPP TS 26.173.

[8] *AMR Wideband Speech Codec; Test Sequences*, 3GPP TS 26.174.

[9] *AMR Wideband Speech Codec; Voice Activity Detector (VAD)*, 3GPP TS 26.194.

[10] *AMR Wideband Speech Codec; Comfort Noise Aspects*, 3GPP TS 26.192.

[11] *AMR Wideband Speech Codec; Error Concealment of Lost Frames*, 3GPP TS 26.191.

[12] *AMR Wideband Speech Codec; Frame Structure*, 3GPP TS 26.201.

[13] *AMR-WB Speech Codec Performance Characterization*, 3GPP TR 26.976.

[14] *ANSI C-Code for the Floating-Point Adaptive Multi-Rate (AMR) Wideband Speech Codec*, 3GPP TS 26.204.

[15] *Packet-Switched Streaming Services (PSS); Protocols and Codecs*, 3GPP TS 26.234.

[16] *Multimedia Messaging Service (MMS); Media Formats and Codecs*, 3GPP TS 26.140.

[17] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," *IEEE Trans. Vehicular Technol.*, vol. 43, pp. 808–816, Aug. 1994.

[18] Y. Bistritz and S. Pellerm, "Immittance Spectral Pairs (ISP) for speech encoding," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1993, pp. II-9–II-12.

[19] E. Ordentlich and Y. Shoham, "Low-delay code-excited linear predictive coding of wideband speech at 32 kbps," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Toronto, ON, Canada, May 14–17, 1991, pp. 9–12.

[20] R. Salami, C. Laflamme, J. P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of CS-ACELP: A toll quality 8kb/s speech coder," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 116–130, 1998.

[21] R. Hagen, E. Ekudden, B. Johansson, and W. B. Kleijn, "Removal of sparse-excitation artifacts in CELP," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Seattle, WA, 1998, pp. I-145–I-148.

[22] *Adaptive Multi-Rate Inband Control and Link Adaptation*, 3GPP TS 45.009.

[23] S. Bruhn, P. Blocher, K. Hellwig, and J. Sjöberg, "Concepts and solutions for link adaptation and inband signaling for the GSM AMR speech coding standard," in *IEEE Vehicular Technology Conf.*, Amsterdam, The Netherlands, Sept. 19–22, 1999, pp. 2451–2455.

Bruno Bessette received the B.S. degree in electrical engineering from the University of Sherbrooke, QC, Canada, in 1992.

In 1993, he was with SMIS (an R&D company), Sherbrooke, QC, Canada, as a Software Engineer and developed a teletex receiver for the account of HydroQuebec. In 1994, he joined the Electrical Engineering Department of the University of Sherbrooke, where he is currently a Researcher Engineer with the Speech and Audio Coding Research Group. He has taken part in the design and real-time implementation of speech coding algorithms, many of them are currently standardized including the ETSI EFR codec and the more recent 3GPP AMR-WB codec (ITU-T Recommendation G.722.2). He is a cofounder of VoiceAge Corporation, Montreal, QC, Canada. His research interests include speech and audio coding, DSP development systems, and modern digital mobile radio systems.

Redwan Salami (S'87–M'94–SM'99) was born in Tyre, Lebanon, in 1961. He received the B.Sc. degree in electrical engineering from Al-Fateh University, Tripoli, Libya, in 1984 and the M.Sc. and Ph.D. degrees in electronics from the University of Southampton, U.K., in 1987 and 1990, respectively.

In 1990, he joined the Department of Electrical Engineering, University of Sherbrooke, QC, Canada, as a Postdoctoral Fellow, then as a Research Assistant, where he worked on the design and real-time implementation of low bit rate speech coding algorithms. Since 1996 until the present, he has held the position of Adjunct Professor at the University of Sherbrooke. In 1999, he was a cofounder of VoiceAge Corporation, Montreal, QC, Canada, where he currently holds the position of Vice-President, Research and Development. He contributed to several speech coding standards in ITU-T and cellular industry, including ITU-T Recommendation G.729 and G.729 Annex A, 12.2 kb/s enhanced full-rate (EFR) GSM codec, 7.4 kb/s EFR TDMA codec (IS-136) and 3GPP AMR-WB codec (ITU-T Recommendation G.722.2). He has numerous publications and patents in the area of speech coding.

Dr. Salami is a member of the Speech Technical Committee of IEEE Signal Processing Society. He was a member of the technical committees of several IEEE Speech Coding Workshops. His research interests include speech and audio coding, speech communication over packet networks, and digital mobile radio systems.

Roch Lefebvre (M'98) was born in Sherbrooke, QC, Canada, in 1965. He received the B.Sc. degree in physics from McGill University, Montreal, QC, Canada, in 1989 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Sherbrooke in 1992 and 1995, respectively. In 1995, he joined the Department of Electrical Engineering of the University of Sherbrooke as a Research Assistant where he worked on the design of low bit rate wideband speech coding algorithms. Since 1996 until the present, he has held the position of Professor and research director at the University of Sherbrooke. He is a cofounder of VoiceAge Corporation, Montreal. He contributed to recent speech coding standards in ITU-T and the cellular industry, namely 3GPP AMR-WB codec (ITU-T Recommendation G.722.2). His research interests include speech and audio processing including coding, signal enhancement, robust quantization, and real-time streaming applications.

Milan Jelínek (M'00) was born in Brno, Czech Republic, in 1967. He received the degree in electrical engineering from the Technical University of Brno in 1991, the M.Sc. degree from the University Paris XI, France, in 1993, and the Ph.D. degree from the University of Sherbrooke, in 1998.

He joined the Speech and Audio Coding Research Group at the University of Sherbrooke as a Postdoctoral Fellow in 1998 and as a Research Engineer in 1999. Since 2001, he has been an Adjunct Professor at the University of Sherbrooke. He is a part-time employee in research and development at VoiceAge Corporation, Montreal, since 2000. His research interests cover speech and audio compression, speech enhancement and speech communications over packet networks.

Jani Rotola-Pukkila was born in Tampere, Finland, in 1974. He received the M.Sc. degree in electrical engineering from Tampere University of Technology in 1998.

He joined Nokia Research Center, Tampere, in 1996. He is working in the field of speech coding and has contributed to several speech coding standards, including EFR TDMA codec (IS-136), 3GPP AMR codec, and 3GPP AMR-WB codec.

Janne Vainio was born in Pälkäne, Finland, in 1967. He received the M.Sc. degree in computer science from the Tampere University of Technology, Finland, in 1991.

He has been working in the Department of Signal processing in the Tampere University of Technology until 1994, when he joined Nokia Research Center, Tampere. His work has been involved digital image processing, channel, and speech coding. He contributed to several speech coding standards especially in cellular industry, including GSM HR codec, GSM enhanced full-rate (EFR) codec, TDMA EFR codec (IS-136), 3GPP AMR-NB codec, and AMR-WB codec (ITU-T Recommendation G.722.2). Currently his research interests include speech, channel and audio coding and their applications in cellular systems. He has several conference publications and more than ten granted patents on the area of speech and channel coding.

Hannu Mikkola was born in Mänttä, Finland, in 1965. He received the M.Sc. degree in electrical engineering from the Tampere University of Technology, Finland, in 1990.

Until 1995, he worked in speech coding in the Department of Signal Processing in Tampere University of Technology. Since 1995, he has been working for Nokia Research Center. He is working as a Senior Research Engineer and has contributed to standardization of several cellular speech codecs including GSM Enhanced full rate (EFR), 3GPP AMR, and 3GPP AMR-WB. He has several conference publications and granted patents on the area of speech and channel coding.

Kari Järvinen was born in Kajaani, Finland in 1961. He received the M.Sc. degree and the (post-graduate) Licentiate of Technology degree in electrical engineering from Tampere University of Technology, Finland, in 1985 and 1987, respectively.

From 1985 to 1989 he worked in speech coding research at Tampere University of Technology. Since 1990, he has been with Nokia Research Center, Tampere, where he currently holds the position of Research Fellow in the Speech and Audio Systems Laboratory. His work is focused on evolution of digital mobile communication systems and, in particular, on low bit-rate speech coding and transmission quality aspects in digital mobile communication systems. He has carried the responsibility within Nokia for standardization of speech services for digital mobile systems and has held several Chairman positions of speech codec working groups in both ETSI and 3GPP. He participated in GSM speech codec standardization work from 1988 in ETSI. He was Vice-Chairman of ETSI SMG11 (GSM codec standardization working group) from April 1997 until June 1998, when he was elected as Chairman of SMG11. He acted as SMG11 Chairman until the transfer of GSM standardization work from ETSI to 3GPP in July 2000. He was also Chairman of the SMG11 ad-hoc group on AMR (Adaptive Multi-Rate) codec from October 1997 until finalization of the AMR codec standard early 1999. In 3GPP, he acted as the Convenor of TSG SA WG4 (3GPP codec standardization working group, SA4) from December 1998 until March 1999. He was Vice-Chairman of SA4 from March 1999 until June 2000. Since then he has been Chairman of SA4. He has contributed to several speech coding algorithms in standards including the GSM EFR (Enhanced Full-Rate) codec, the TDMA EFR (IS-136) codec, the 3GPP AMR codec and the 3GPP AMR-WB codec (ITU-T Recommendation G.722.2). He has numerous conference publications and more than ten patents in the area of speech and channel coding.

Mr. Järvinen is a founding member of the Median-Free Group International (MFGI).