ESG

Engineering Services Group

PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs

January 2008

80-W1253-1 Rev D

QUALCOMM®

PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs

80-W1253-1 Rev D

QUALCOMM Incorporated
5775 Morehouse Drive
San Diego, CA 92121-1714
U.S.A.

This technical data may be subject to U.S. and international export, re-export or transfer ("export") laws. Diversion contrary to U.S. and international law is strictly prohibited.

Table of Contents

List of Figures

MAY CONTAIN U.S. EXPORT CONTROLLED INFORMATION

List of Tables

This page intentionally left blank.

1. Introduction

1.1 Purpose

This document explains how the objective quality metrics obtained by the Perceptual Evaluation of Speech Quality (PESQ) tool is biased against the Enhanced Variable Rate Codec (EVRC) used in CDMA networks and other codecs in this family (EVRC-B and EVRC-WB).

1.2 Scope

This document evaluates the accuracy of certain Objective Measurement Tools such as PESQ to evaluate Objective Voice Quality of EVRC-family based CDMA networks.

1.3 Revision history

Table 1-1 shows the revision history for this document.

Table 1-1: Revision history

Version	Date	Description
A	August 2007	Initial release
B	August 2007	Revised cover page
C	October 2007	Updated text
D	January 2008	Updated for EVRC-B & EVRC-WB

1.4 Technical assistance

For assistance or clarification on information in this guide, you may send email to cdma.help@qualcomm.com.

1.5 Acronyms

Table 1-2 lists acronyms used in this document.

Table 1-2: Acronyms

Term	Definition
AGC	Automatic Gain Control
AMR	Adaptive Multi Rate Coding

Term	Definition
CDMA	Code Division Multiple Access
CELP	Code Excited Linear Prediction
EVRC	Enhanced Variable Rate Coding
EVRC-WB	Wideband EVRC
GSM	Global System for Mobile Communication
MOS	Mean Opinion Score
MOS-LQO	MOS Listening Quality Objective
NELP	Noise Excited Linear Prediction
PESQ	Perceptual Evaluation of Speech Quality
RCELP	Relaxed Code Excited Linear Prediction
UMTS	Universal Mobile Telecommunication System
VoIP	Voice over Internet Protocol

1.6 References

[1] ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, February 2001.

[2] ITU-T Recommendation P.862.1. Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO, November 2003.

[3] ITU-T Recommendation P.800. Methods for Subjective Determination of Transmission Quality, August 1996.

[4] ITU-T Recommendation P.800.1. Mean Opinion Score (MOS) Terminology, July 2006.

[5] P. Morrissey, “How to measure call quality,” in Network Computing, Digital Convergence, Feb. 17, 2005.

[6] M. Varela, I. Marsh, and B. Grönvall, “A systematic study of PESQ’s behavior (from a networking perspective),” In Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN ’06), Prague, Czech Republic, June 2006.

[7] S. Pennock, “Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm,” in Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN ’02), Prague, Czech Republic, May 2002.

[8] Ericsson Technical Paper-AQM in TEMS automatic –PESQ.

[9] W. Kleijn, P. Kroon, and D. Nahumi, "The RCELP speech-coding algorithm", European Transactions on Telecommunications, vol. 5, pp. 573-582, September/October 1994.

[10] ITU-T Recommendation P.862.3, "Application guide for objective quality measurement based on Recommendation P.862, P.862.1 and P.862.2", November 2005.

[11] 3GPP2 C.S0014-C, “Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems”

[12] 3GPP2/TSG-C1.1, "SMV Post-Collaboration Subjective Test - Final Host and Listening Lab Report," C11-20010326-003

[13] 3GPP2/TSG-C1.1, “Characterization Final Test Report for EVRC-Release B,” C11-20060424-015R2.

[14] 3GPP2/TSG-C1.1, “EVRC-WB Characterization Test Report,” C11-20061030-009r2.

This page intentionally left blank.

2. Problem Description

It is observed that the speech quality measurement tool “PESQ” (an objective way of measuring the speech quality of the audio codecs) is biased against the EVRC family of speech codecs during the estimation of objective Mean Opinion Score.

There are significant limitations in the PESQ algorithm with regards to the time alignment and psychoacoustic modeling. These limitations in PESQ are having much higher/prominent impact on the EVRC family of codecs. Hence, the usage of PESQ for EVRC codecs would impair the speech quality measurement results significantly, because of the way the EVRC codecs are designed.

This page intentionally left blank.

3. Background

The preferred method of calculating the perceived speech quality of cellular telephones is through subjective testing, also known as perceptual testing. In subjective testing, a group of listeners independently rate voice quality. Each listener rates the speech quality of a communication network/device by selecting one of the following five options, each of which has a numeric rating:

- Bad (1)
- Poor (2)
- Fair (3)
- Good (4)
- Excellent (5)

The average of these numeric scores is the Mean Opinion Score (MOS).

However, it is expensive and time-consuming to obtain subjective test scores in this manner.

To address the disadvantages of subjective testing, there is a requirement for the telecommunication industry to design a test methodology capable of predicting speech quality from objective measurements.

The ITU-T has conducted a competition to find a state-of-the-art solution for objective prediction of speech quality. It was intended that this objective method be used by the telecommunication industry to measure perceived quality of network connections. In this competition, the Perceptual Evaluation of Speech Quality (PESQ) algorithm was shown to outperform other objective speech quality models. In February 2001, PESQ was approved as ITU-T recommendation P.862.

The PESQ tool, described in ITU-T Rec. P.862 and its extension 862.1, uses an auditory model that combines a mathematical description of the psychophysical properties of human hearing with a technique that performs a perceptually relevant analysis, taking into account the subjectivity of errors in the received signal. The process compares the original and received signal and determines a rating analogous to the Mean Opinion Score (MOS) described in ITU-T P.800.

The PESQ algorithm produces a value ranging from 4.5 to 1. A PESQ value of 4.5 means that the measured speech has no distortion; it is exactly the same as the original. A value of 1 indicates the severest degradation.

It is important to note that PESQ only measures one aspect of transmission quality. ITU-T Recommendation P.862 states: “It should also be noted that the PESQ algorithm does not provide a comprehensive evaluation of speech quality, it only measures the effects of one-way speech distortion and noise on speech quality. The effects of loudness loss, delay, side tone, echo and other impairments related to two-way interaction are not reflected in the PESQ scores. Therefore it is possible to have high PESQ scores, yet poor quality of connection overall."

The PESQ algorithm consists of two parts:

1. Conversion to the psychoacoustic domain.
2. Cognitive modeling.

The most important steps in each part are depicted in Figure 3-1.

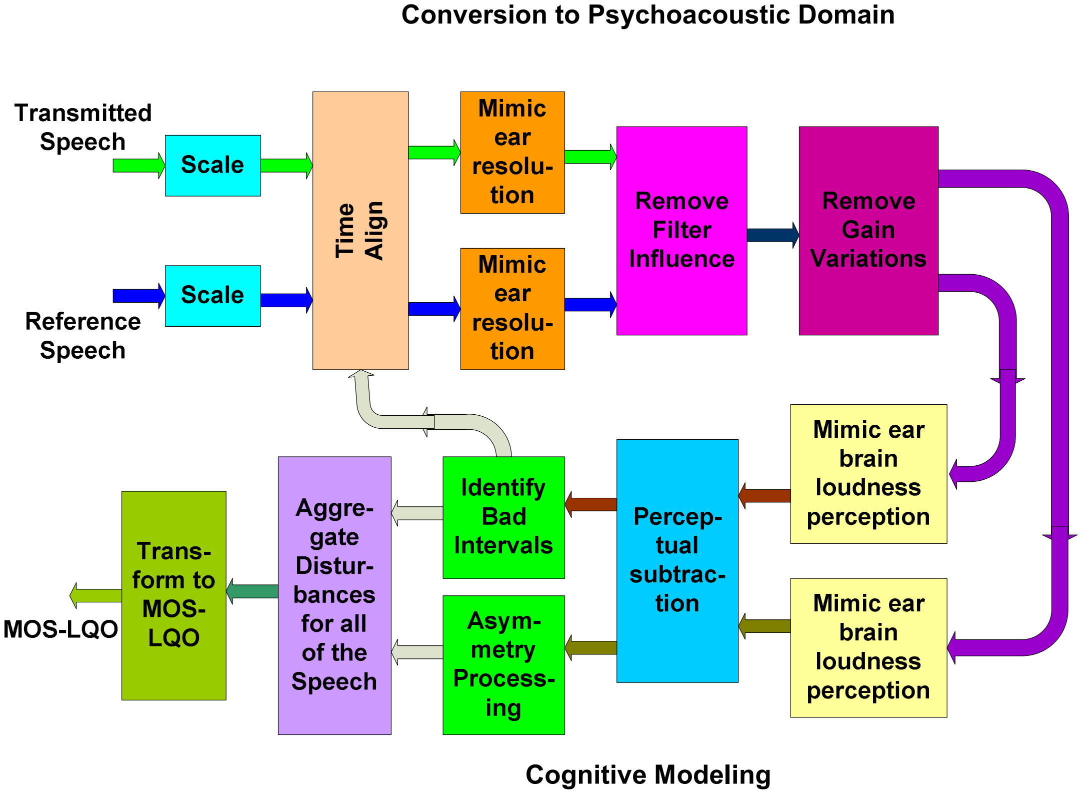

Figure 3-1: Block Diagram of PESQ (Reference [8])

Each block in Figure 3-1 is explained below.

- **Scale**: Both the transmitted and the reference speech are scaled to compensate for the overall gain in the network
- **Time Align**: In a mobile network, the transmission delay can change both between speech references and within a single speech reference. This is due to handovers or Voice over IP (VoIP) delays. The reference and the transmitted speech are time aligned, so all parts of the transmitted speech match the reference.
- **Mimic Ear Resolution**: Transform the speech signal into the frequency domain, and then warp the Hertz scale into the critical band domain. This warping tries to imitate the way the ear treats different frequencies in the signal. Higher frequencies get a lower resolution.

- **Remove Filter Influence**: Remove the effect of filtering. The mobile network and PSTN may have filtering, which would affect the PESQ score more negatively than it should. By measuring the transfer function of the network and using that measure to equalize the reference, filter influence is decreased. This is an improvement over PSQM, which produced excessively bad scores in the presence of filtering, for example the filtering in AMR at lower rates.
- **Remove Gain Variations**: Automatic Gain Control (AGC) units in the network can cause gain variations. The influence of gain variations is removed.
- **Mimic Ear-Brain Loudness Perception**: Warp the intensity of the spectrum to mimic how the human ear transforms intensity into perceived loudness.
- **Perceptual Subtraction**: The loudness representation of the reference and transmitted signals are subtracted, taking into account how the brain perceives differences. The result is a disturbance density signal.
- **Identify Bad Intervals**: If the disturbance signal contains an interval of very bad disturbances, it might be due to an incorrect time alignment for this interval of speech. In this case, the time alignment and the rest of the PESQ processing is redone for the bad interval. If this results in a better disturbance signal, this result is used instead.
- **Asymmetry Processing**: If a speech codec adds noise to the original speech, a clearly audible distortion will result. The asymmetry processing calculates an asymmetric disturbance density signal, which contains the added disturbances.
- **Aggregate Disturbances for all of the Speech**: First, both disturbance signals are summed in the frequency plane. This results in disturbance and asymmetric disturbance signals that represent how distorted the speech is during very short periods of time. These very short periods are summed to 320 ms periods, called split second disturbances. Then a PESQ_MOS score is calculated as a combination of the average split second disturbances and average split second asymmetrical disturbances for the entire speech reference.
- **Transform to MOS-LQO**: To produce a PESQ score, which can be compared to subjective listening tests, the PESQ_MOS is transformed according to ITU P.862.1 into the MOS_LQO score.
- **MOS-LQO**: MOS_LQO resembles the Mean Opinion Score (MOS) scale. MOS_LQO ranges from 4.5 (best) to 1.0 (worst).

Although PESQ is state-of-the-art in terms of the objective prediction of perceived quality, it does not always accurately predict perceived quality. Performance data presented in ITU-T Recommendation P.862 presents a very optimistic view of PESQ accuracy that can be expected by the telecommunications industry. This paper examines the accuracy of PESQ for measuring the speech quality of the EVRC family of CDMA codecs.

This page intentionally left blank.

4. Investigation and Analysis

EVRC family codecs, including EVRC, EVRC-B and EVRC-WB [11], utilize advanced signal processing techniques to enhance performance without impacting perceived speech quality. However, due to limitations of time alignment and the psychoacoustic model in the PESQ algorithm, the evaluation performance of PESQ for testing EVRC family codecs does not accurately reflect the subjective assessment of listeners as measured by real subjective mean opinion scores (MOS).

4.1 Low correlation with subjective MOS score

Table 4-1 shows that PESQ does not accurately predict the quality of EVRC family codecs. The table presents formal subjective MOS test results conducted by 3GPP2 comparing AMR 12.2 kbps with EVRC, and shows the corresponding PESQ scores.

Table 4-1: MOS score comparison

	AMR (12.2 k)	EVRC	Difference
Subjective MOS score from 3GPP2 MOS test	3.932	3.852	0.08
PESQ (P.862.1)	4.114	3.796	0.32

The data in this table is from the formal SMV Post Collaboration MOS Tests officially conducted by 3GPP2 in November 2000; the results are provided in 3GPP2 contribution C11-20010326-003 from the March 2001 meeting [12].

These subjective tests conducted by 3GPP2 used 64 listeners and 8 speakers (4 male, 4 female databases); hence, each of the codecs obtained 512 votes. The reliability of this test is very good. Typically ITU and 3GPP use 256 or 192 votes; 512 exceeds both these figures.

The PESQ scores were obtained based on ITU P.862 and P.862.1, using the identical executables as AMR and EVRC from the above MOS test. The speech database used to compute PESQ scores are also identical to the one used in the MOS tests from the November 2000 3GPP2 formal test.

The 95% confidence interval for this 3GPP2 test is approximately 0.12 MOS. Therefore, Table 4-1 clearly shows that the subjective MOS results for AMR and EVRC are statistically equivalent, while the objective PESQ score indicates a considerable quality advantage (0.318 MOS) for AMR. PESQ tends to “artificially” underestimate the score of EVRC with respect to AMR, which may result in a score reduction of 0.3 PESQ or more for EVRC. This result clearly shows that PESQ fails to accurately predict the objective score for EVRC.

Note: AMR at 12.2 kbps active speech (when there is actual speech) is at a much higher data rate than EVRC at 8.55 kbps for active speech.

4.2 RCELP algorithm in EVRC

EVRC family codecs are based upon the RCELP algorithm [9], appropriately modified for variable rate operation and for robustness in the CDMA environment. RCELP is a generalization of the Code Excited Linear Prediction (CELP) algorithm. Unlike conventional CELP encoders, RCELP does not attempt to match the original speech signal exactly. Instead of attempting to match the original residual signal, RCELP matches a modified version of the original residual that conforms to a simplified piecewise linear pitch contour. The pitch contour is obtained by estimating the pitch delay once in each frame and linearly interpolating the pitch from frame to frame. One benefit of using this simplified pitch representation is that more bits are available in each packet for the stochastic excitation and for channel impairment protection than would be if a traditional fractional pitch approach were used. This results in enhanced error performance without impacting perceived speech quality in clear channel conditions.

4.3 PESQ analysis procedure

PESQ compares an original reference signal and a degraded signal to predict the perceived quality of the degraded signal, using a two-step approach.

1. The original reference signal and the degraded signal are aligned by splitting each signal into a few segments and estimating delay for each segment.

2. The original signal and degraded signal are transformed based on a perceptual model. Then for each frame (256 samples/frame, 50% overlapping), two types of distance measures between the two signals are computed, called "frame disturbance" and "frame asymmetrical disturbance", respectively. These disturbances are aggregated over time to generate the average disturbance value, *d*, and the average asymmetrical disturbance value, *da*. The PESQ score is obtained by:

   PESQ = 4.5 - 0.1*d - 0.0309*da

Hence, larger disturbance values result in lower PESQ scores.

4.4 Inaccuracy of PESQ for RCELP modification

This section presents some experimental data that illustrates how PESQ cannot reflect the perceptual transparency of RCELP, either through time alignment or through the perceptual model it uses. The original speech signal in this experiment is a sentence pair approximately 6 seconds long. Three codecs/modes are used: EVRC, AMR at 12.2 kbps, and AMR at 4.75 kbps.

4.4.1 EVRC versus AMR at 12kbps

According to the formal 3GPP2 MOS tests, the perceived EVRC quality (MOS score: 3.852) is statistically equivalent to AMR at 12.2 kbps (MOS score: 3.932). However, the PESQ score for EVRC is much lower than the PESQ score for AMR at 12.2 kbps. For example, for the sentence pair used in this experiment, the PESQ score is 4.190 for AMR at 12.2 kbps and 3.787 for EVRC, according to ITU P.862.1. But there is no perceptual difference between them.

To better illustrate the PESQ bias against EVRC, Figure 4-1 shows the values of frame disturbance and frame asymmetrical disturbance of each frame for EVRC coded signal and for AMR 12.2 coded signal. The reference signal is the original speech signal.

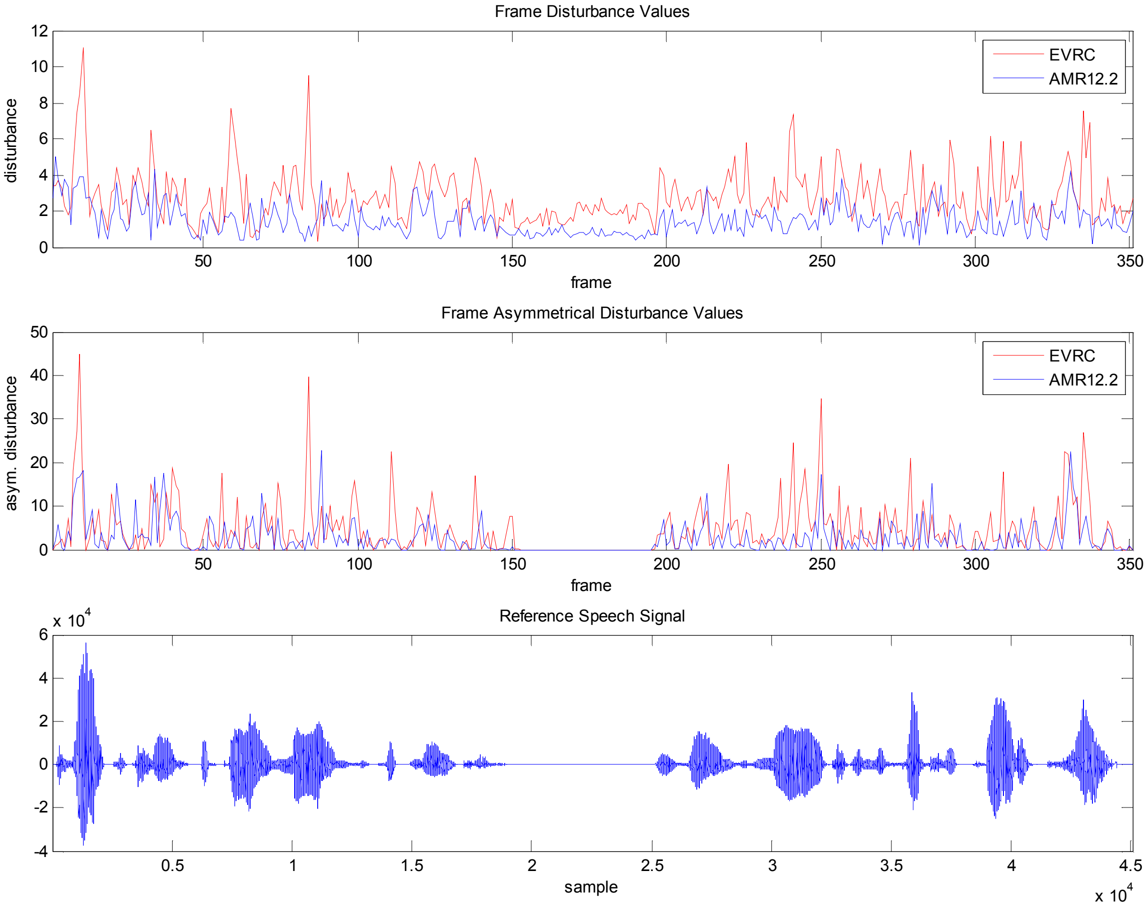

Figure 4-1: Frame disturbance and frame asymmetrical disturbance

For most frames, EVRC gets much higher disturbance values than AMR 12.2, hence the lower PESQ score. The higher disturbance values for EVRC are due to the fact that PESQ cannot align the reference signal and the coded signal correctly because of modifications made by the RCELP algorithm.

Figure 4-2 shows how PESQ aligns the degraded signal with the reference signal for different codecs. The range is from the 79th frame to the 81st frame, which is the beginning of a voiced region.

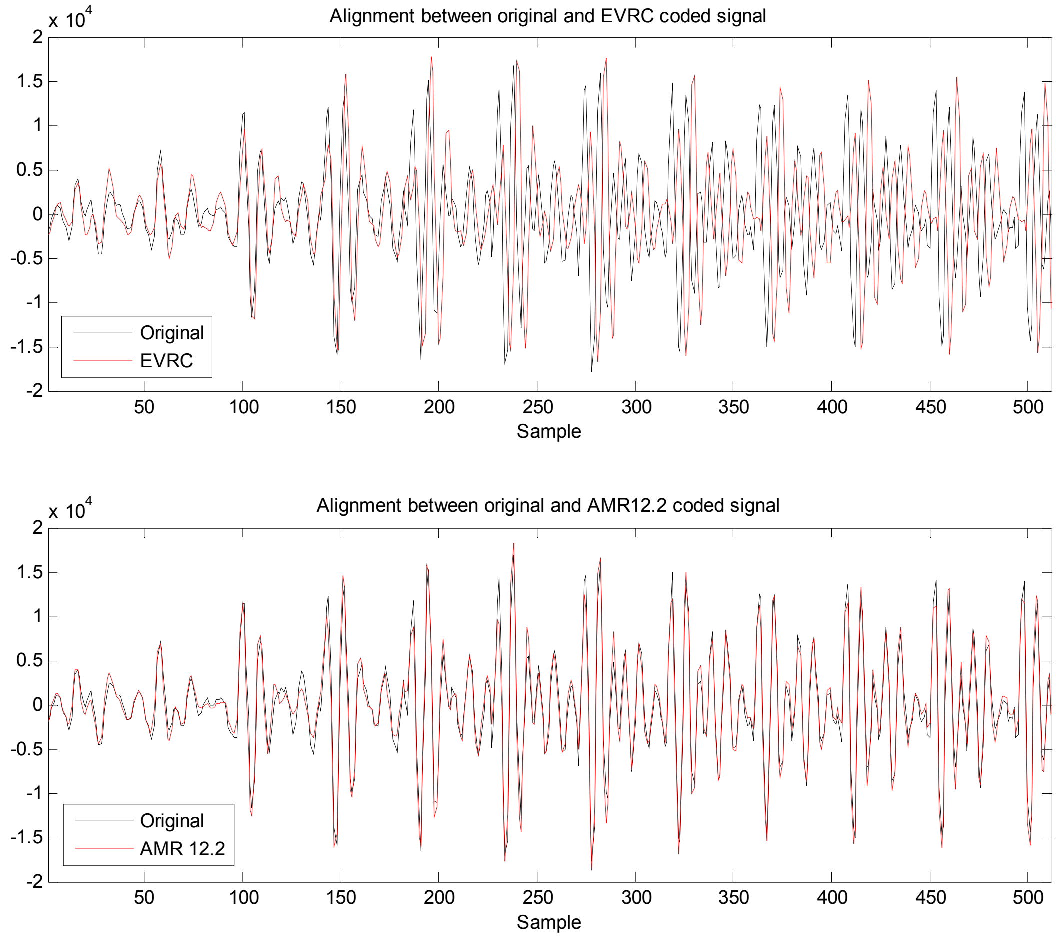

Figure 4-2: PESQ alignments for frames 79, 80, and 81

Table 4-2: Disturbance values for frames 79, 80, and 81

	Frame Disturbance			**Frame Asymmetrical Disturbance**		
Frame	79	80	81	79	80	81
EVRC	4.46	4.52	2.59	4.57	2.41	1.27
AMR 12.2	1.84	1.58	0.80	2.00	0.22	1.66

Figure 4-2 shows that the EVRC coded signal and the original signal are aligned at the beginning. However after a few pitch periods, they are misaligned despite only minor changes in the waveform shape. This is because in EVRC, the signal is modified to generate a linear pitch-period contour. This modification has been shown to be perceptually transparent, but the PESQ algorithm cannot track this change. By comparison, the original waveform and the AMR12.2 waveform are fully aligned.

The time alignment procedure in PESQ does not have sufficiently high resolution for correct alignment after RCELP modification. In the RCELP modification, a speech segment usually is shifted only by a few samples; but in PESQ, the minimal length of a segment for narrow band speech is 2400 samples – i.e., 300ms. (In reality, the resulting shortest segment in PESQ is usually much longer than that, due to other constraints). This resolution is not fine enough to provide good alignment for the EVRC coded signal. Additionally, the perceptual model in PESQ cannot accurately predict the quality for EVRC coded signals when the signal is modified. As shown in Table 4-1, the frame disturbance and frame asymmetric disturbance for EVRC are higher than the values for AMR 12.2 for most of the frames (this can also be seen in Figure 4-1).

PESQ can become even more inaccurate. Due to the poor temporal resolution nature of the delay estimation algorithm in PESQ, the misalignment continues into the steady voiced region. Figure 4 shows the alignment of the waveform from the 83rd frame to 85th frame as determined by the PESQ time alignment procedure for EVRC and AMR 12.2. Table 4-3 compares the disturbance values. The EVRC coded signal is totally misaligned with the original reference signal, and the disturbance values for EVRC are much higher than the corresponding values for AMR 12.2.

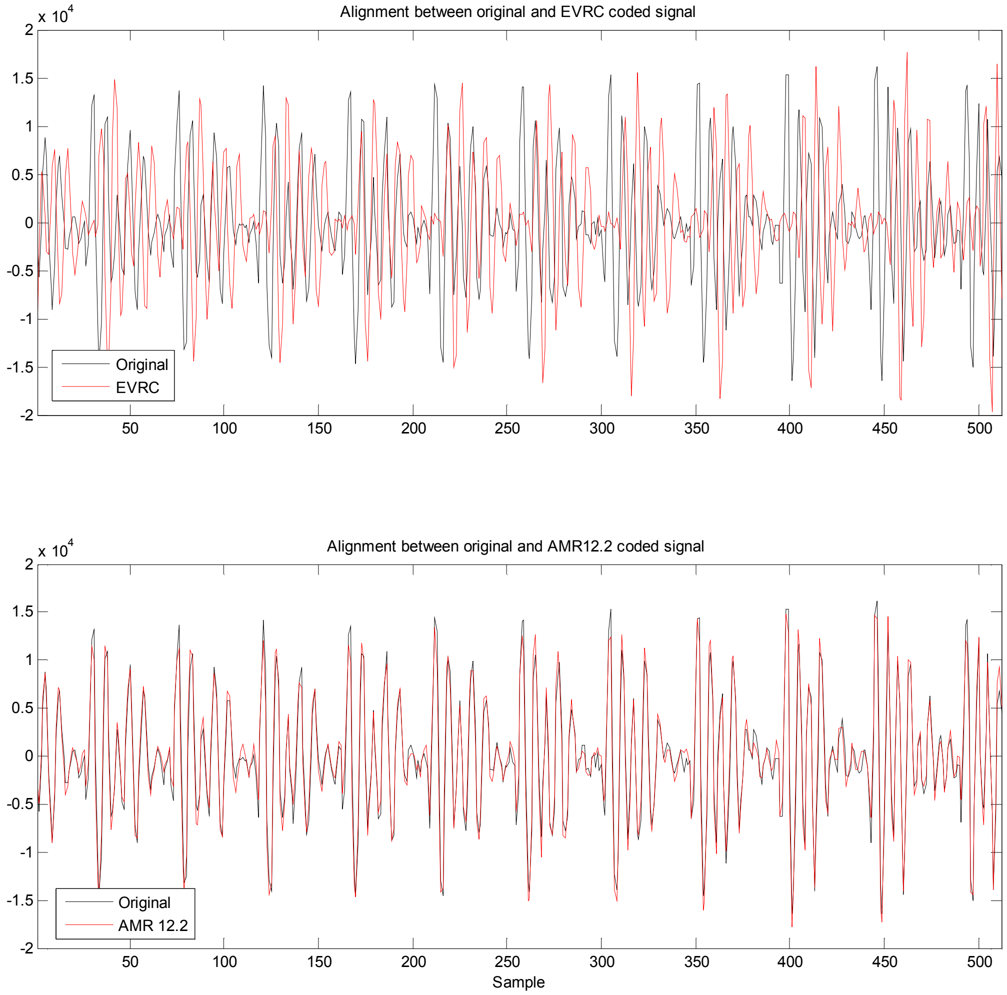

Figure 4-3: PESQ alignment for frames 83, 84, and 85

Table 4-3: Disturbance values for frames 83, 84, and 85

	Frame Disturbance			Frame Asymmetrical Disturbance		
Frame	83	84	85	83	84	85
EVRC	4.58	9.51	3.49	9.48	39.82	9.47
AMR 12.2	0.32	1.13	0.58	0	4.02	1.12

The PESQ application guide ([10], Footnote 11) notes that PESQ results for EVRC depends on the particular alignment of the coding frame boundaries with the input PCM data. However, simply doing frame boundaries alignment as suggested in the PESQ application guide does not solve the problem. Figure 4-4 shows the alignment of the 85th frame by PESQ algorithm (top figure) and by manual adjustment (bottom figure). In the manual adjustment, we align the frames along the right boundaries. However, the left part of the EVRC coded frame is still misaligned with the original speech frame.

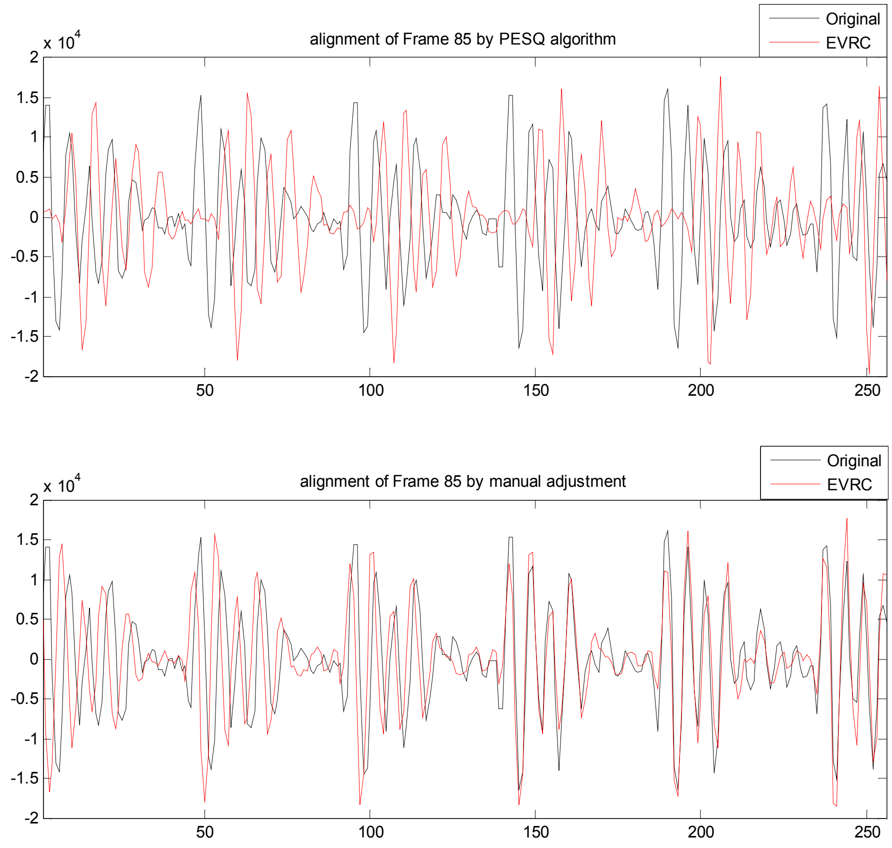

Figure 4-4: Alignment of the 85th frame by PESQ algorithm and by manual adjustment

4.4.2 EVRC versus AMR 4.75

The perceptual quality of EVRC is much better than AMR 4.75. However, the PESQ score of EVRC (3.787) is only slightly higher than the PESQ score of AMR 4.75 (3.562), which is inconsistent with the perceived quality.

The reason again is because PESQ cannot accurately predict quality for EVRC family codecs. Figure 4-5 shows the disturbance values of each frame for EVRC and AMR 4.75. For many frames, PESQ shows even higher disturbance values for EVRC than for AMR 4.75.

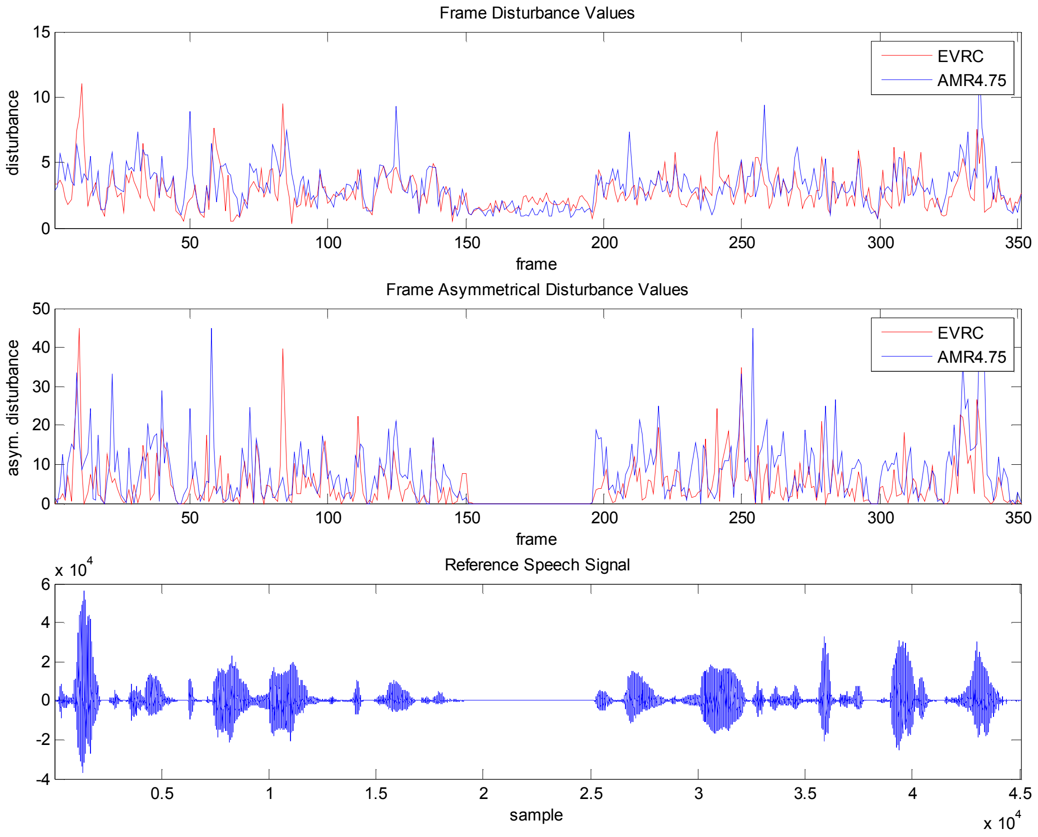

Figure 4-5: Disturbance values for EVRC and AMR 4.75

This page intentionally left blank.

5. More on EVRC-B and EVRC-WB

The EVRC-B and EVRC-WB codecs not only use RCELP techniques, but also introduce other sophisticated signal processing techniques [11], such as Noise Excited Linear Prediction (NELP) and Prototype-Pitch-Period (PPP) waveform interpolation to achieve lower bit-rates while maintaining high quality reconstructed speech. NELP uses a filtered pseudo-random noise signal to model unvoiced speech, rather than a codebook. The PPP coding scheme extracts a representative pitch cycle (the prototype waveform) at fixed intervals and transmits its description, reconstructing the speech signal by interpolating between the proto type waveforms. These techniques have already been proven to be perceptually transparent through formal subjective listening tests. However, the PESQ psychoacoustic model underestimates the quality of these techniques compared to P.800 formal listening test result.

5.1 EVRC-B MOS vs. PESQ

Table 5-1 shows a comparison of PESQ and MOS scores under clean conditions (i.e., no frame erasures) for EVRC-B at different *channel* rates. (Note that AMR12.2 operates at the *source* rate of 12.2kbps.) Table 5-2 shows the scores under 1% frame erasure condition. All the MOS data is from the formal characterization test for EVRC-B conducted by 3GPP2, as documented in [13], except for the first two rows of Table 5-1, which are from [12] and are included for comparison purposes. From both tables, it is obvious that PESQ consistently under-estimates MOS scores for EVRC-B. Furthermore, as the percentage of frames encoded by NELP or PPP increases, the discrepancy between subjective MOS and PESQ also increases. This is because these techniques used in EVRC-B, while perceptually transparent, do not preserve the shape of the original signal, and their perceptual transparency can not be correctly predicted by the psychoacoustic model in PESQ algorithm.

Again, it should be noted that while PESQ under-estimates MOS scores for EVRC and EVRC-B, it overestimates the MOS score for the AMR codec.

These results are shown graphically in Figure 3-1Figure 5-1 and Figure 5-2. Figure 5-1 shows the MOS and PESQ scores for different codecs under clean conditions (i.e., 0% frame erasure). Figure 5-2 illustrates the MOS and PESQ scores for EVRC-B at different rates and clearly shows PESQ's growing under-prediction of MOS as NELP and PPP frames are added.

Table 5-1: PESQ and MOS scores for EVRC-B under 0% frame erasure

Codec	MOS	PESQ (P.862.1)	ΔMOS **	ΔPESQ ***	RCELP	NELP	PPP
AMR 12.2k *	3.932	4.114	0.08	0.32			
EVRC *	3.852	3.796	0	0	✓		
EVRC	3.879	3.796	0	0	✓		
EVRCB at 9.3kbps	3.984	3.823	0.11	0.03	✓		
EVRCB at 6.6kbps	3.887	3.490	0.01	-0.31	✓	✓	✓
EVRCB at 5.8kbps	3.684	3.281	-0.20	-0.52	✓	✓	✓

* All the MOS scores in this table are from the EVRC-B characterization test, except the first two rows, for which the MOS scores are taken from the MOS test in [12].

** ΔMOS = MOS score of the current codec - MOS score of EVRC in the same MOS test

*** ΔPESQ = PESQ score of the current codec - PESQ score of EVRC in the same MOS test

Table 5-2: PESQ and MOS for EVRCB under 1% frame erasures

Codec	MOS	PESQ (P.862.1)	ΔMOS **	ΔPESQ ***	RCELP	NELP	PPP
EVRC	3.727	3.658	0	0	✓		
EVRCB at 9.3kbps	3.883	3.680	0.16	0.02	✓		
EVRCB at 8.4kbps	3.844	3.528	0.12	-0.13	✓	✓	
EVRCB at 7.8kbps	3.883	3.456	0.16	-0.20	✓	✓	
EVRCB at 7.4kbps	3.856	3.396	0.13	-0.26	✓	✓	✓
EVRCB at 7.0kbps	3.793	3.368	0.07	-0.29	✓	✓	✓
EVRCB at 6.6kbps	3.809	3.353	0.08	-0.31	✓	✓	✓
EVRCB at 6.2kbps	3.711	3.303	-0.02	-0.36	✓	✓	✓
EVRCB at 5.8kbps	3.688	3.281	-0.04	-0.38	✓	✓	✓

** ΔMOS = MOS score of the current codec - MOS score of EVRC in the same MOS test

*** ΔPESQ = PESQ score of the current codec - PESQ score of EVRC in the same MOS test

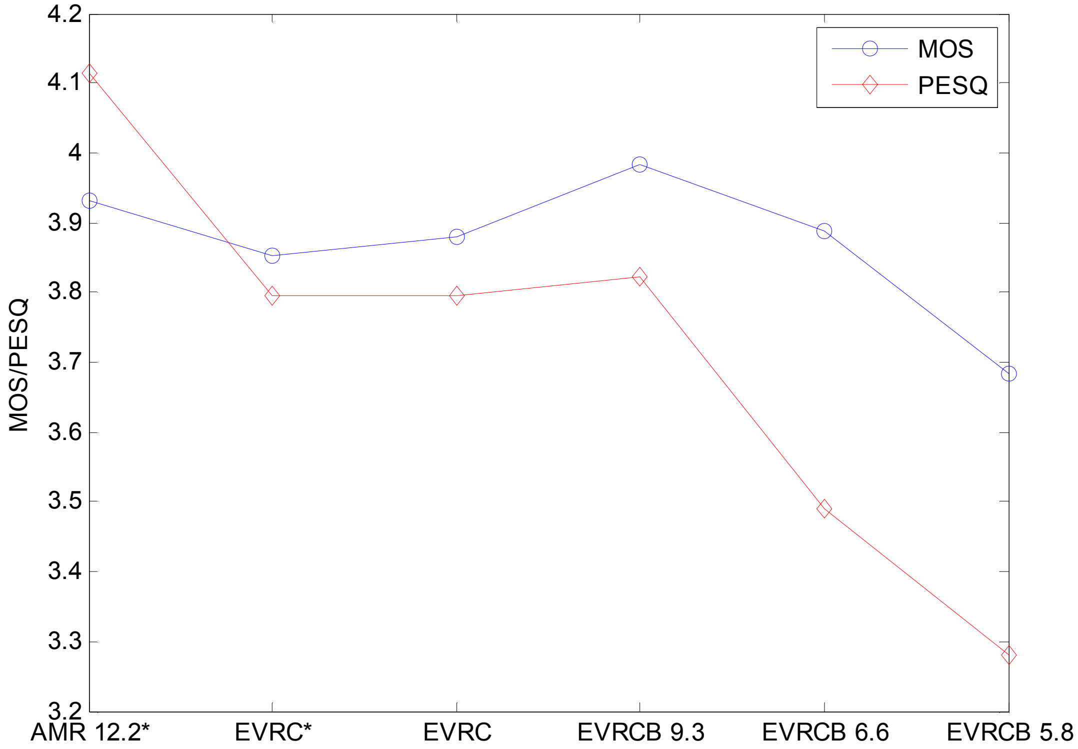

Figure 5-1: Comparison of PESQ and MOS for different codecs under 0% frame erasure

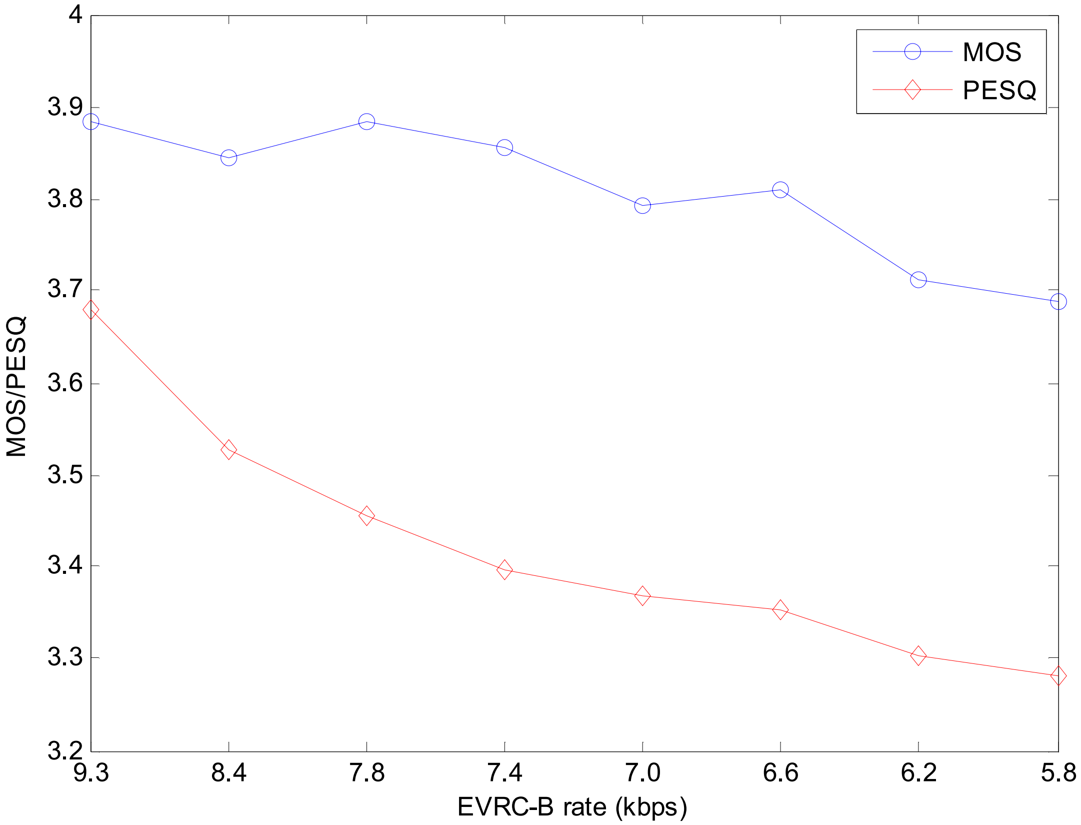

Figure 5-2: Comparison of PESQ and MOS for EVRC-B at different channel rates under 1% frame erasures

5.2 EVRC-WB MOS vs. PESQ

Table 5-3 shows a comparison of PESQ and MOS for EVRC-WB and AMR-WB 12.65kb/s mode. The MOS scores are from the formal characterization test for EVRC-WB conducted by 3GPP2, as documented in [14]. The PESQ scores are computed based on P.862.2. For all conditions, EVRC-WB P.800 MOS scores are statistically equivalent or better than AMR-WB 12.65kb/s mode, but PESQ scores always underestimate the quality of EVRC-WB. For some conditions, the PESQ score of EVRC-WB is more than 0.6 lower than AMR-WB 12.65kb/s mode.

Figure 5-3 shows a scatter plot of MOS and PESQ scores for EVRC-WB and AMR-WB 12.65kb/s mode. A straight line with the slope of 1 is provided as a reference. It is obvious to see the PESQ under-prediction of EVRC-WB in all conditions. Figure 5-4 compares the PESQ difference and MOS difference between EVRC-WB and AMR-WB 12.65kb/s mode under various conditions.

Table 5-3: Comparison of PESQ & MOS for EVRC-WB and AMR-WB (12.65kb/s)

	EVRC-WB		AMR-WB			
Condition	MOS	PESQ	MOS	PESQ	ΔMOS *	ΔPESQ **
clean (nominal level)	4.078	3.130	4.125	3.745	0.05	0.62
clean (low level)	4.012	3.057	4.090	3.423	0.08	0.37
clean (high level)	3.859	3.005	3.867	3.585	0.01	0.58
1% FER	3.867	2.963	3.883	3.488	0.02	0.53
2% FER	3.727	2.820	3.652	3.268	-0.08	0.45
3% FER	3.539	2.743	3.332	3.053	-0.21	0.31
6% FER	3.148	2.429	2.914	2.525	-0.23	0.10
1% D&B+ 1% packet level signaling	3.969	3.109	3.809	3.503	-0.16	0.39
Average Score	3.775	2.907	3.709	3.324	-0.07	0.42

* ΔMOS = AMR-WB MOS score - EVRC-WB MOS score

** ΔPESQ = AMR-WB PESQ score - EVRC-WB PESQ score

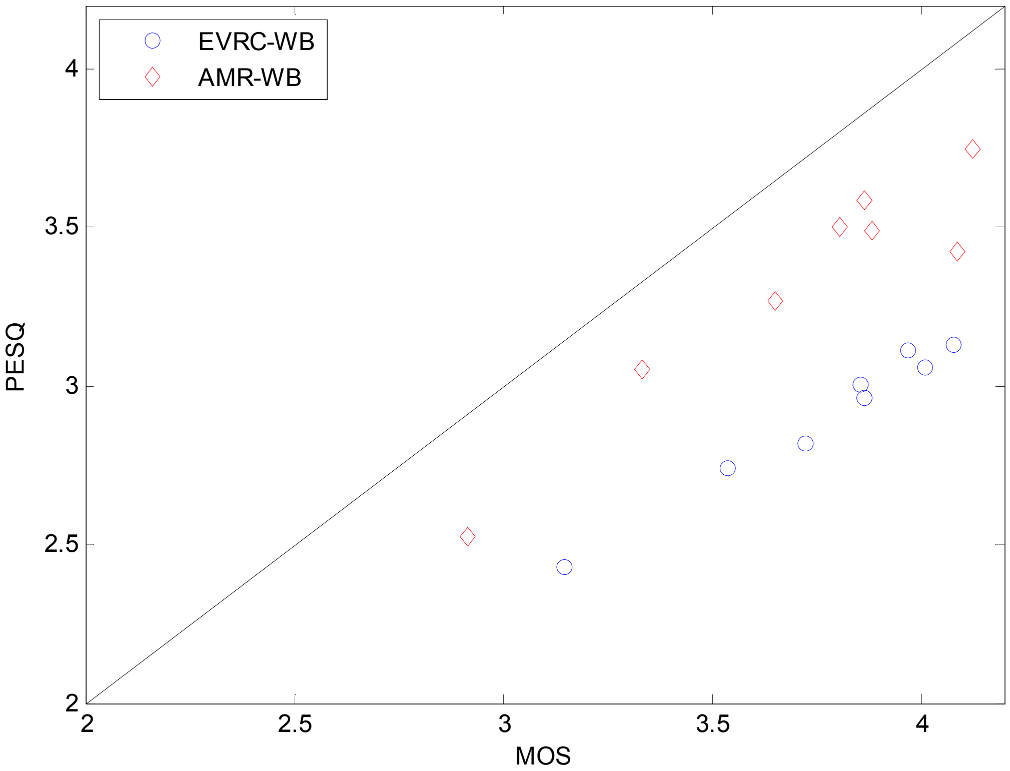

Figure 5-3: PESQ vs. MOS for EVRC-WB and AMR-WB 12.65kb/s mode

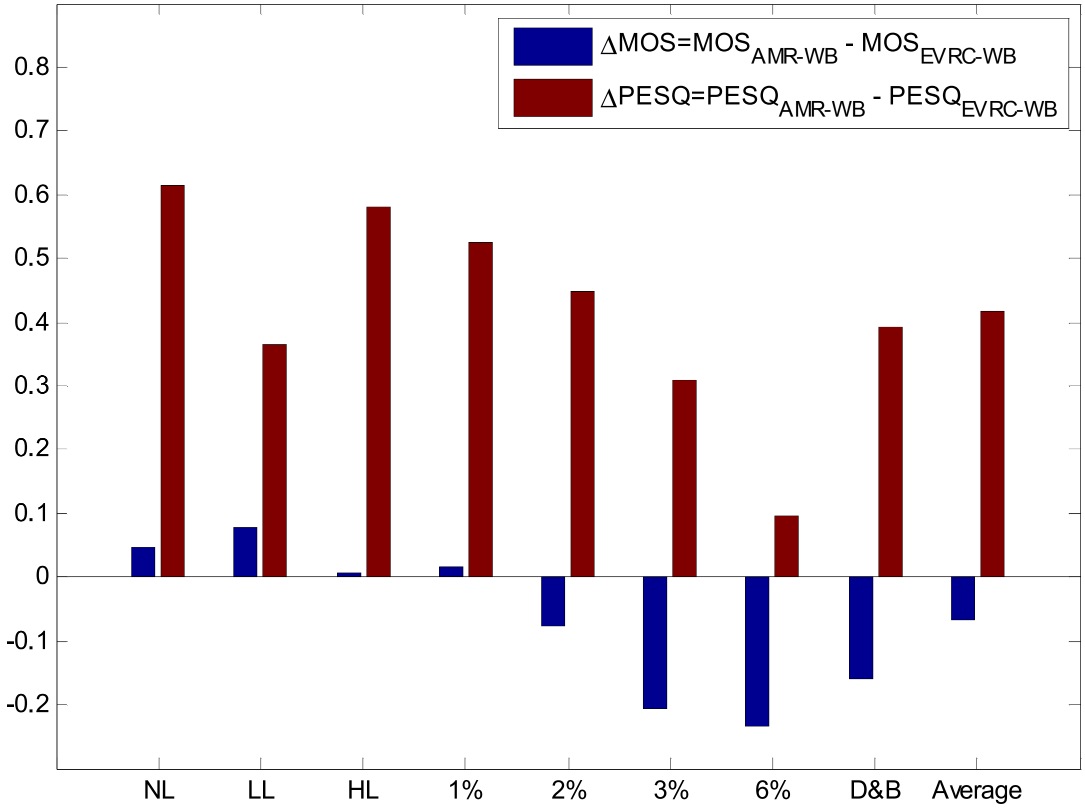

Figure 5-4: ΔMOS and ΔPESQ for EVRC-WB and AMR-WB 12.65kb/s mode

The listed conditions include NL (nominal level: signal level at -22 dB); LL (low level: signal level at -32 dB); HL (high level: signal level at -12dB); 1%, 2%, 3% and 6% frame erasure rates; D&B where the system experiences 1% dim-and-burst and 1% packet-level dimming; and average values of MOS and PESQ.

6. Conclusions

EVRC family codecs, including EVRC, EVRC-B and EVRC-WB, use advanced signal processing techniques, such as RCELP, PPP and NELP, to enhance performance. The perceptual transparency of these techniques is not reflected by the PESQ algorithm due to the limitations in its time alignment procedure and the psychoacoustic model it uses. 3GPP2 test results substantiate this claim. Subjective MOS scores for AMR and EVRC are statistically the same, but the objective PESQ score provides a difference of 0.32.

PESQ objective quality metrics should not be used to compare similar speech codecs that have vastly different algorithms, especially when the algorithms use a wide variety of non-linear signal processing like those in EVRC family codecs, such as noise suppression, residual modification, and waveform interpolation. These speech coding techniques either maintain or improve perceptual speech quality, but also reveal the limitations of objective quality measures.

This page intentionally left blank.