QUALCOMM
CDMA Technologies

Voice Terminal Testing Methodology

White Paper

80-N4402-1 Rev. B

March 4, 2011

Revision history

Revision	Date	Description
A	February 2011	Initial release
B	March 2011	Remove Qualcomm Confidential and Proprietary statements

Contents

Figures

Tables

1 Introduction

1.1 Purpose

This document explains a methodology to test the voice quality of a terminal using any objective speech quality measurement (OSQM) tool, such as Perceptual Evaluation of Speech Quality (PESQ). Due to many factors, PESQ scores vary widely even among good quality terminals. Hence it is possible for both bad terminals and good terminals to have overlapping PESQ scores, making it difficult to classify a test handset as good/bad using its PESQ score. This document proposes a test methodology which constrains the factors that cause wide variations in PESQ scores such that PESQ variability is low for the voice terminals, and hence test terminal voice quality can be classified reliably into good/bad using a single set of thresholds within the set of constraints.

NOTE: The terms *terminal* and *handset* are used interchangeably in this document.

1.2 Scope

This document describes a PESQ-based terminal voice quality test methodology by imposing constraints on factors that cause a wide variation of PESQ scores within voice terminals so that the test terminal can be reliably classified into good/bad.

1.3 Acronyms

List of acronyms used in this document are shown in Table 1-1.

Table 1-1 Acronyms

Term	Definition
AMR	Adaptive Multi Rate Coding
EVRC	Enhanced Variable Rate Coding
MOS	Mean Opinion Score
NELP	Noise Excited Linear Prediction
PESQ	Perceptual Evaluation of Speech Quality
PPP	Prototype Pitch Period
RCELP	Relaxed Code Excited Linear Prediction

1.4 References

[1] PESQ_Limitations_Rev_C_Jan_08, January 2008

[2] 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

[3] 3GPP2/TSG- C1.1-20060424-015R2, “Characterization Final Test Report for EVRC-Release B,” C11-20060424-015R2, April 2006

[4] 3GPP2/TSG-C1.1, “SMV Post-Collaboration Subjective Test – Final Host and Listening Lab Report,” C11-20010326-003, March 2001

2 Problem Description

An objective speech quality measurement tool, such as PESQ, is used to test the voice quality of terminals. Most of the time, the limitations of the objective tools are not considered in the testing process, resulting in incorrect voice quality assessment. This paper describes a methodology of how to use an objective speech quality measurement tool properly for voice quality assessment.

In this paper, a test methodology is proposed based on identifying *well controlled conditions* such that the PESQ scores of voice terminals under the same well controlled conditions do not vary much. Different terminals/handsets can then be compared to each other under the same well controlled conditions. This method of constraining the use of a tool to well controlled conditions, such that the voice quality of terminal can be reliably estimated, is a generic method and can be applied to any objective speech quality measurement tool. PESQ is used only for illustration.

In this paper, the methodology for testing voice quality in terminals is explained using examples and results pertaining to PESQ because it is a widely used objective speech quality measurement tool. The common testing practice is to obtain the PESQ scores from the Device under Test (DuT) and compare it to a reference threshold obtained from one or more other good reference handsets to assess the quality of the DuT. A common pitfall of this method is that people tend to use one threshold to verify the quality of any handset. But PESQ scores vary widely amongst good quality terminals, resulting in overlapping PESQ scores for good and bad terminals; hence using such single threshold can result in large numbers of false positives and false negatives.

Another drawback is that PESQ is not an accurate estimator of MOS, as suggested by much evidence [1]. Voice terminals with equivalent subjective quality can have widely varying PESQ scores. If PESQ is used and interpreted improperly, it may lead to confusing and even wrong voice quality decisions.

The limitations of PESQ, along with other factors such as variability in the voice processing path across different terminals and choice of test speech sequence, can cause wide variation in PESQ scores within good terminals. Hence the voice quality of a handset cannot be assessed directly from PESQ scores without constraining those factors that cause PESQ variations.

3 Limitations of PESQ

Though PESQ is designed as an estimator of subjective MOS [2], due to its limitations [1], PESQ scores are not always consistent with the subjective quality of voice terminals. Two terminals with different speech processing modules (such as different speech codecs) of equivalent subjective quality, can have widely varying PESQ scores. Hence directly comparing PESQ scores between two terminals with different speech processing technologies is not useful in assessing their voice quality.

For example, a terminal with an AMR codec is compared to a terminal with an EVRC codec. All the modules in the voice path of the terminals match except for the codec. It is known that AMR and EVRC give subjectively equivalent MOS scores but PESQ under-predicts the MOS scores of EVRC codecs [1], resulting in a lower PESQ score for the EVRC terminal. Due to this inconsistency of PESQ with terminal voice quality, it is incorrect to conclude that AMR terminal voice quality is better than EVRC terminal voice quality. This inconsistency is due to the limitations of PESQ in time alignment and psycho-acoustic modeling [1].

EVRC family codecs, including EVRC, EVRC-B, and EVRC-WB, use advanced signal processing techniques such as RCELP, PPP, and NELP to maintain or improve the speech quality. But the perceptual transparency of these techniques is not reflected by the PESQ algorithm [1]. Figure 3-1 shows the comparison of MOS and PESQ scores for AMR at 12.2 kbps, EVRC at 8.55 kbps, and EVRC-B codec at different bitrates. The under-prediction of the MOS scores of EVRC family codecs by PESQ is evident in the figure.

Another important observation from this plot is that PESQ does not correctly estimate the subjective MOS scores even with the same codec. As an example, for EVRC-B, the relative PESQ score difference between different capacity operating points does not correctly reflect the difference of their subjective MOS scores.

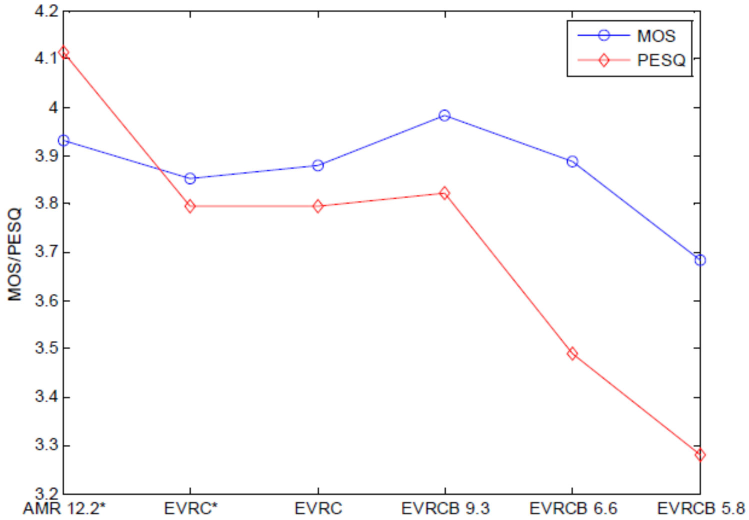

Figure 3-1 Comparison of MOS and PESQ for different codecs. All the MOS scores are taken from the EVRC-B characterization test [3] except for the codecs AMR 12.2 and EVRC which are taken from a different MOS test [4]

Apart from codecs, PESQ also shows inconsistency with MOS for other conditions such as time warping, noise suppression, loudness levels, etc. [2].

The common mistake in using PESQ for voice quality testing is that PESQ scores from different terminals with different speech processing technologies are directly compared with each other for evaluating voice quality. This can lead to incorrect conclusions since terminals with equivalent subjective voice quality can have widely varying PESQ scores.

Chapter 4 explains how to use PESQ properly for reliable terminal voice quality assessment.

4 Well Controlled Conditions

There are many factors which contribute to the large deviation of PESQ scores even among good quality terminals. The factors include choice of input speech, speech codecs, and codec modes and other speech processing modules being used in the voice processing path, etc. Due to this wide range of PESQ scores for good quality terminals, it is possible that a bad terminal and a good terminal have similar PESQ scores making it difficult to classify terminal voice quality into pass/fail with a single PESQ-based threshold. Hence it is necessary to constrain the factors causing large PESQ variations such that it is possible to assess terminal voice quality within the set of constraints that comprise well controlled conditions.

The objective of the proposed methodology is to identify conditions under which PESQ has small variance among all good handsets, so that PESQ-based thresholds can be obtained to reliably classify handsets into pass/fail.

The following sections briefly describe the voice path in a terminal, various factors to be considered in forming well controlled conditions, and a procedure to form them.

4.1 Voice path in a terminal/handset

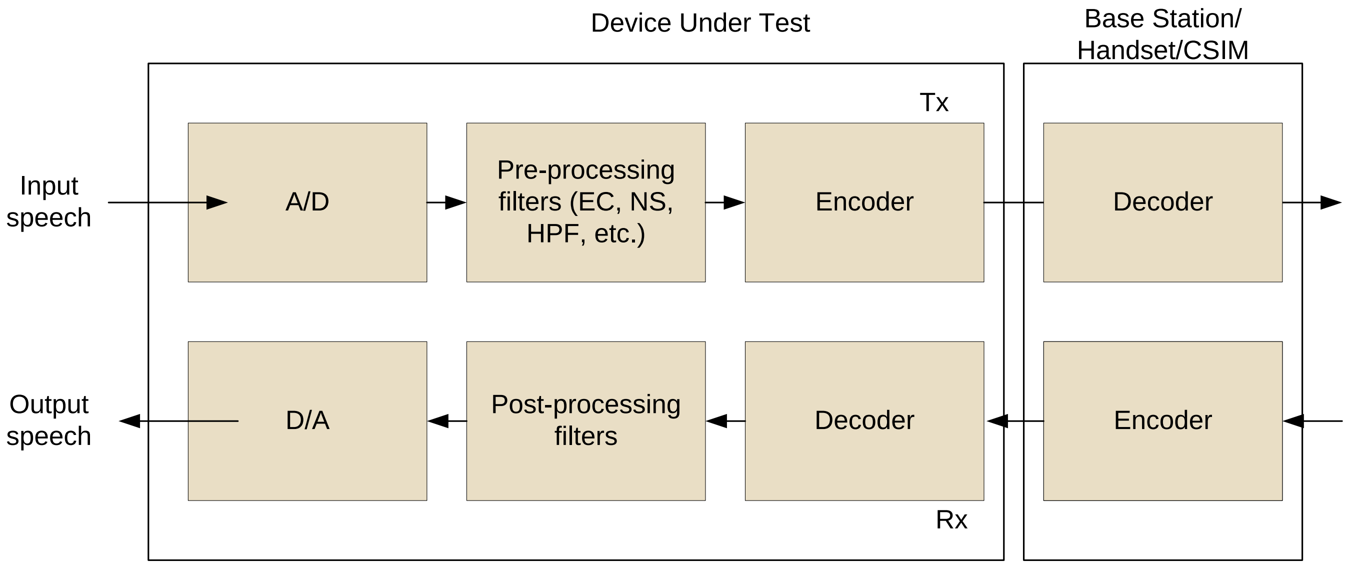

Figure 4-1 Basic block diagram of modules in a handset

Figure 4-1 shows the basic voice modules in a handset. The transmitter (Tx) side is composed of an Analog to Digital convertor, pre processing filters which may include echo canceller, noise suppressor, high pass filter, and an encoder. On the receiver (Rx) side, the encoded bit stream is decoded and processed through a decoder, post processing filters, and a digital to analog convertor.

Usually the Tx and Rx paths of a test handset are tested separately. The test handset is connected to another known good implementation (such as base station simulator, good handset, or offline simulation (CSIM)). Then, voice calls are established to test the Tx or Rx paths independently.

To compute PESQ for handset testing, the reference speech signal is captured at certain point of the handset (for example, captured at the microphone) and the degraded speech signal is captured at another logging point (for example, capture at the speaker on the other side). The PESQ score is calculated using the reference speech signal and degraded speech signal.

Within the scope of this text, we define the voice path as consisting of the reference speech signal, degraded speech signal, and all the elements between them.

4.2 Well controlled conditions

A *well controlled condition* is defined as a particular set of constraints on voice path configuration within which PESQ scores of good handsets show a small variance.

Once a well controlled condition is defined, a test handset can be classified as pass/fail by comparing it to reference handsets (known good handsets) within the same well controlled condition. Otherwise, if the variance is large, two good handsets can have very different PESQ scores, making it difficult to identify whether the low PESQ score of a test handset is due to a bug or its inherent low PESQ score.

A well controlled condition can be constructed by applying constraints on the modules along the voice path (including selection and capture of input and output speech signals) such that the variance of PESQ scores among all the good handsets within this well controlled condition is as small as possible, subject to:

- Practicality of the constraint

  It may be impossible to apply certain constraints in forming a well controlled condition even though it is desirable. For example, ideally, to test a certain module, the logging points for reference speech and degraded speech should be just before and after this module. However, it is generally not possible to have any logging points in a commercial handset other than acoustical or electrical interfaces, even if we know exactly which modules to test. As another example, it may not be possible to disable a certain module on the voice path, even if the disabling of such modules reduce the PESQ variance. Hence practicality of the constraints is a major factor in forming the well controlled condition.

- Test requirements

  Though a well controlled condition can be formed by applying as many constraints as possible, depending on test requirements the constraints may be relaxed. This allows a larger variation in PESQ scores for the handsets in the well controlled condition.

  We can use testing a CDMA handset with EVRC-B codec as an example; Although the recommended practice is to constrain EVRC-B running under a specific COP to reduce variance (since the PESQ scores of the good handsets in a specific COP has a much smaller variance than the PESQ scores from all COPs in EVRC-B, as explained in Section 4.2.3). However, if the purpose of testing is only to capture very big bugs, then it is sufficient to consider all the EVRC-B COPs together to form one well controlled condition. This also allows flexibility for the test handset to run in any COP during testing.

Once a well controlled condition is formed, one can collect a few reference good handsets falling within this well controlled condition, and the test handset quality can be evaluated by comparing its PESQ scores with the threshold values obtained from the PESQ scores of these reference handsets.

Some examples of factors causing widely deviating PESQ scores that should be considered in forming a well controlled condition are provided in the following sections.

4.2.1 Input speech

Generally, the speech signal used for PESQ testing consist of multiple sentence pairs as described in the PESQ application guide [2]. One PESQ score is obtained from each sentence pair. Given these individual scores, the statistics (such as mean value, standard deviation, and minimum score) can be obtained for handset comparison.

The PESQ scores can vary widely from one input speech to another. Hence it is necessary to use the same input speech during handset testing as that used to obtain the reference scores and statistical parameters. Figure 4-2 compares the distribution of the PESQ scores of EVRC-B COP0 for different input speech. We use two different input speech signals in this example:

- The first speech signal is the same sentence pair repeated multiple times.
- The second speech signal consists of different sentence pairs.

Figure 4-2 clearly shows that with different input speech signals, the PESQ scores vary a lot: the mean and standard deviation of PESQ scores using the first speech signal are 3.84 and 0.04; the mean and standard deviation of PESQ scores using the second speech signal are 3.71 and 0.126. Since the PESQ scores vary a lot between different choices of input sequences, it is better to constrain the input speech to be the same when defining a well controlled condition.

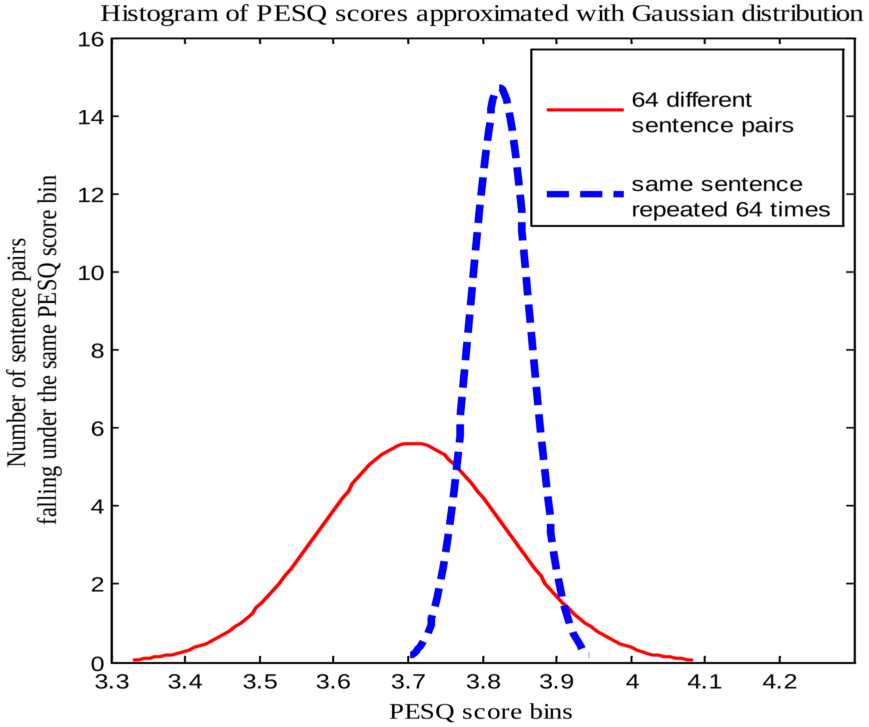

Figure 4-2 Histogram of PESQ scores for different input speech, approximated with Gaussian distribution

The choice of input speech is also important. Different input speech signals cause different extents of variation in PESQ scores. As shown in Figure 4-2, the first speech signal causes a much smaller variance; however the second speech signal covers a larger range of speech syllables because it consists of different sentence pairs. Which one to choose depends on the purpose of the testing. The second speech signal covers a wider range of speech syllables, hence is able to identify some speech-dependent bugs; however the first speech signal causes much smaller variance, making it easier to identify speech-independent bugs. Therefore there is a trade off between the two choices.

4.2.2 Codec module – EVRC vs. AMR

The speech codec module is one of the most important modules along the voice path. PESQ varies a lot among different commonly available codecs, such as EVRC, EVRC-B, and AMR.

For example, EVRC-B COP0 and AMR 12.2 kbps, although being subjectively equivalent, have different PESQ scores [1]. Figure 4-3 shows the distribution of AMR 12.2 kbps and EVRC-B COP0 PESQ scores for an input speech with the same sentence pair repeated 64 times. It can be clearly seen from the figure that, if considering AMR and EVRC-B COP0 separately, the variance is smaller (0.0074 and 0.0158, respectively). However, if combined, the variances are much larger (0.0277). Classification of good/bad handsets is much more accurate when thresholds are obtained separately for EVRC-B and AMR rather than combining them. Obtaining a threshold of the combined distribution can cause a false positive (by passing a bad AMR handset) or a false negative (by failing a good EVRC-B handset).

Therefore, it is better to constrain the codec module in the voice path such that different codecs fall under different well controlled conditions. For example, develop a set of thresholds for AMR related test cases, while developing another set of thresholds for EVRC-B related test cases.

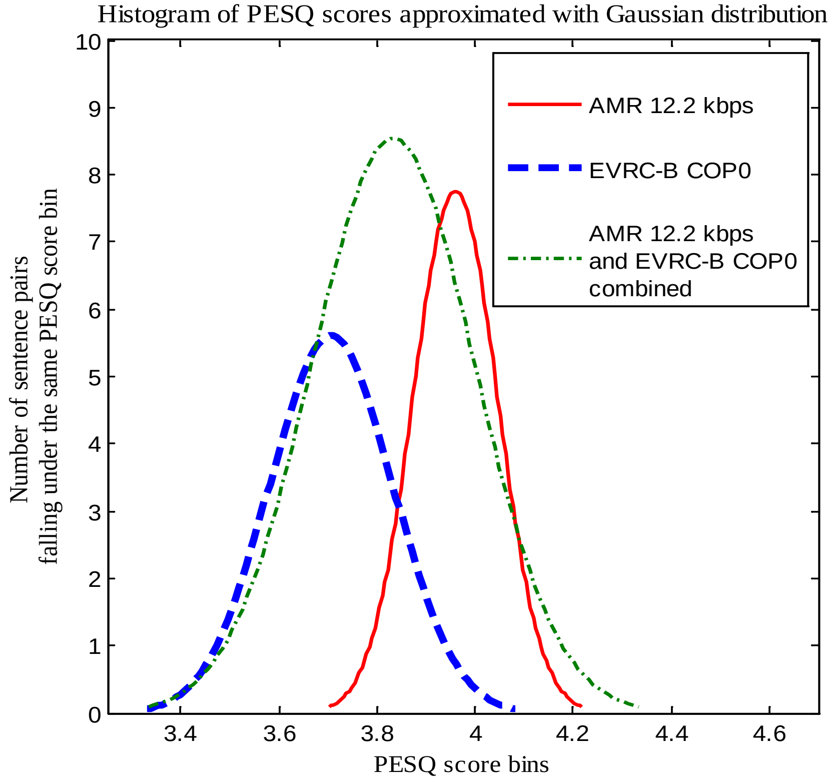

Figure 4-3 Distribution of PESQ scores for AMR and EVRC-B codecs separate and combined

4.2.3 Codec module – EVRC-B COPs

EVRC-B has eight typical Capacity Operating Points (COP). Different COPs are associate with different average bit rates. The COPs (or average bit rates) can be adjusted to balance between capacity and voice quality.

EVRC-B COPs should fall under different well controlled conditions as well. Since different EVRC-B COPs use different proportions of RCELP, PPP, and NELP speech coding techniques, each EVRC-B COP is affected differently by PESQ (though the corresponding deviation in MOS is a lot less). Figure 4-4 shows the PESQ distribution of EVRC-B COP0 and EVRC-B COP4. The variance of EVRC-B COP0 is 0.0016 and the variance of EVRC-B COP4 is 0.0027. If these two COPs are combined, the variance is 0.0172. Obviously, the variance is large when the COPs are combined. Obtaining thresholds from the distribution of combined PESQ can cause false positives and false negatives. For example, a handset operating at a buggy EVRC-B COP0 mode can have a higher PESQ score than another handset which operates at a good EVRC-B COP4 mode.

Higher variance across different COPs in EVRC-B reduces the accuracy of classifying good/bad handsets. Hence the codec mode should be constrained such that different COPs fall under different well controlled conditions.

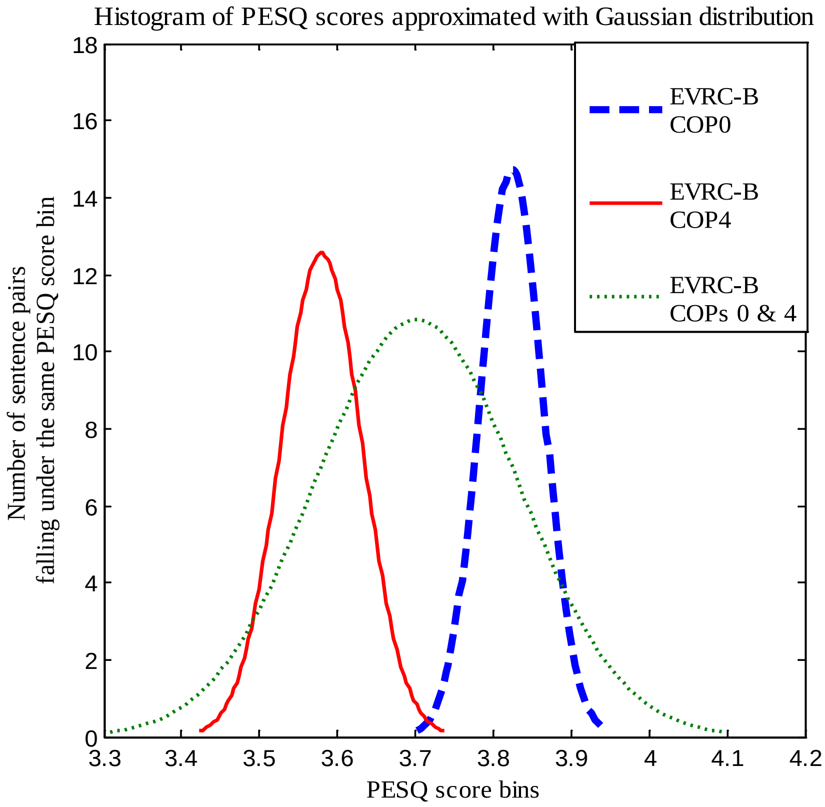

Figure 4-4 Distribution of PESQ scores for EVRC-B COP0 and EVRC-B COP4 codecs separated and combined

4.2.4 Acoustic/Electric interfaces

Insertion/capture of the input/output speech is one of the factors that can cause a large deviation in PESQ scores, and hence a major factor to constrain when forming a well controlled condition.

Acoustic insertion/capture generally results in lower PESQ scores than electrical insertion/capture. Hence when forming a well controlled condition, how to insert/capture input/output speech should be explicitly specified so that all the handsets are compared using the same method of insertion/capture.

Acoustic insertion usually causes much larger variances of PESQ scores than electrical insertion. Hence an electrical interface is preferred, unless the acoustical path is one element for testing.

4.2.5 Logging locations

Ideally, we would like to tap the reference and degraded signals immediately before and after the modules to be tested in order to limit the variance of PESQ scores. Note that this may not be practical in some testing environments. In those cases, the logging is generally restricted to either electrical or acoustical interface.

4.2.6 Modules in the voice processing path

There are many blocks in the whole voice path. Some of the modules, such as AGC and time-warping, can cause a larger deviation in PESQ scores. Hence if these blocks are not being tested, it is better to disable or constrain these blocks in the voice processing path such that the PESQ scores have a small variance, and form a well controlled condition. The simplified block diagram of voice processing path is shown in
Figure 4-1.

4.3 Procedure to form a well controlled condition

As explained in Section 4.2, a well controlled condition is formed by applying constraints on the voice path based on the knowledge of the test handset, practicality of the constrain, and test requirement.

The procedure to form a well controlled condition can be summarized as follows:

1. Decide on the insertion interface. The options are,
   - Electrical
   - Acoustical
2. Decide the logging point of the reference and degraded speech. The options are,
   - Electrical
   - Acoustical
   - Logging point within the software/firmware if possible
3. Choose the input speech according to the test requirements. Some of the choices are,
   - Same sentence pair repeated multiple times – to capture speech-independent bugs
   - Different sentence pairs concatenated – to capture speech-dependent bugs

   Note the first option offers a smaller PESQ variance.
4. Examine and constrain each module in the voice path based on practicality and test requirements whenever the constraint reduces the variance of the PESQ scores. (For example, apply constraints by choosing codec modes, disabling/enabling certain modules, and by choosing the configuration parameters, etc.)

4.3.1 Example for forming well controlled conditions

In this example, the test handset is a CDMA handset with EVRC-B enabled. Well controlled conditions are formed by applying the procedure explained in Section 4.3. Note that the result shown in Figure 4-5 is obtained from handset simulation data. Hence steps 1, 2, and 3 are only assumptions and the numbers in this simulated example are just illustrative purpose.

- Electrical insertion is used since it is not intended to test the acoustical path in this example (electrical insertion causes less PESQ variance than acoustical insertion).
- Logging at electrical interfaces is used to dump reference and degraded speech (since in this example scenario, it is assumed that there is no access to internal modules).
- The same sentence repeated 64 times is chosen in order to test speech-independent bugs only.

- In the assumed scenario, the tester can only access and control the codec module (for example, by changing the settings in base station simulator). The tester can configure the COPs of EVRC-B, hence must decide whether to constrain the COP to form a well controlled condition.

Figure 4-5 shows the distribution of PESQ scores of the COPs separately and combined.

The PESQ scores with all COPs combined has much larger variance than that of the PESQ scores for each single COP. Therefore, to improve the accuracy of identifying a bad handset, the tester decides to use single COP for forming well controlled conditions.

Ultimately, eight different well controlled conditions are formed, each one containing a different COP in EVRC-B.

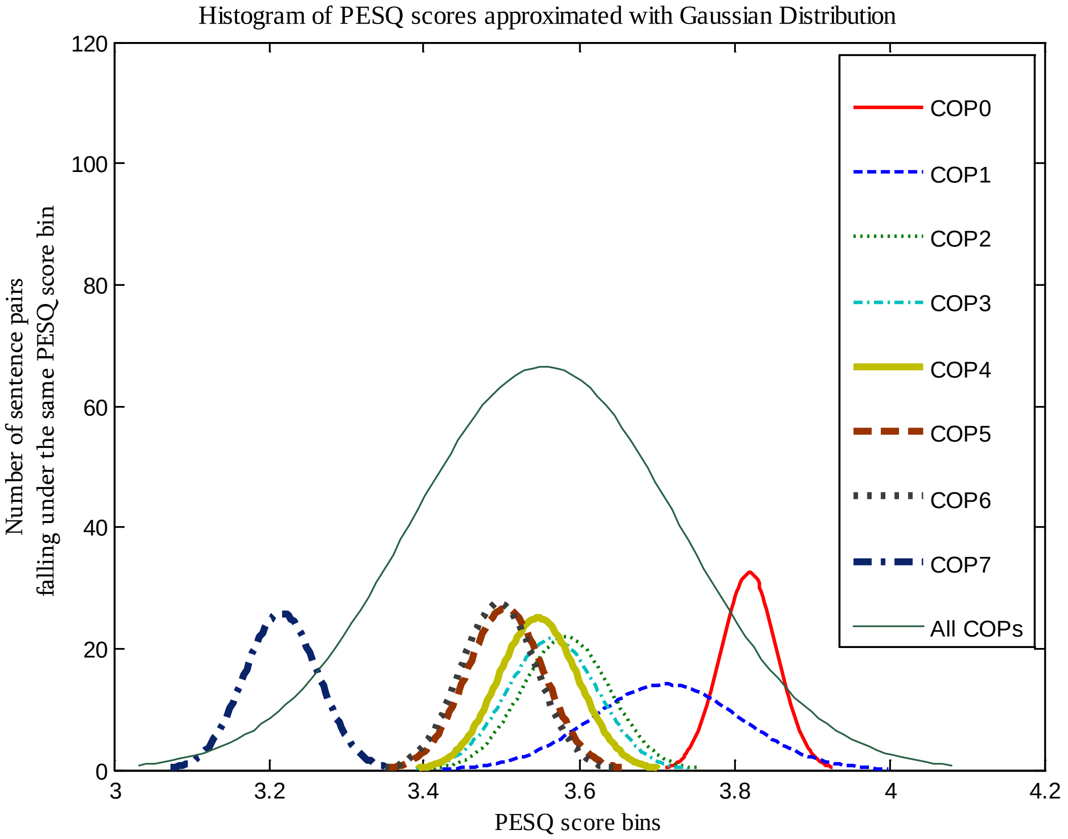

Figure 4-5 Distribution of PESQ scores for each of the EVRC-B COPs separate and combined

5 Training and Testing

For each well controlled condition, PESQ-based statistical parameters are obtained from the reference and test handsets which are then used for testing. The training and testing methodology is described in this section.

5.1 Proposed methodology

The objective of forming a well controlled condition is to choose suitable *reference* handsets for testing the *test* handset in a well controlled condition. Figure 5-1 shows an overview of using well controlled conditions for testing.

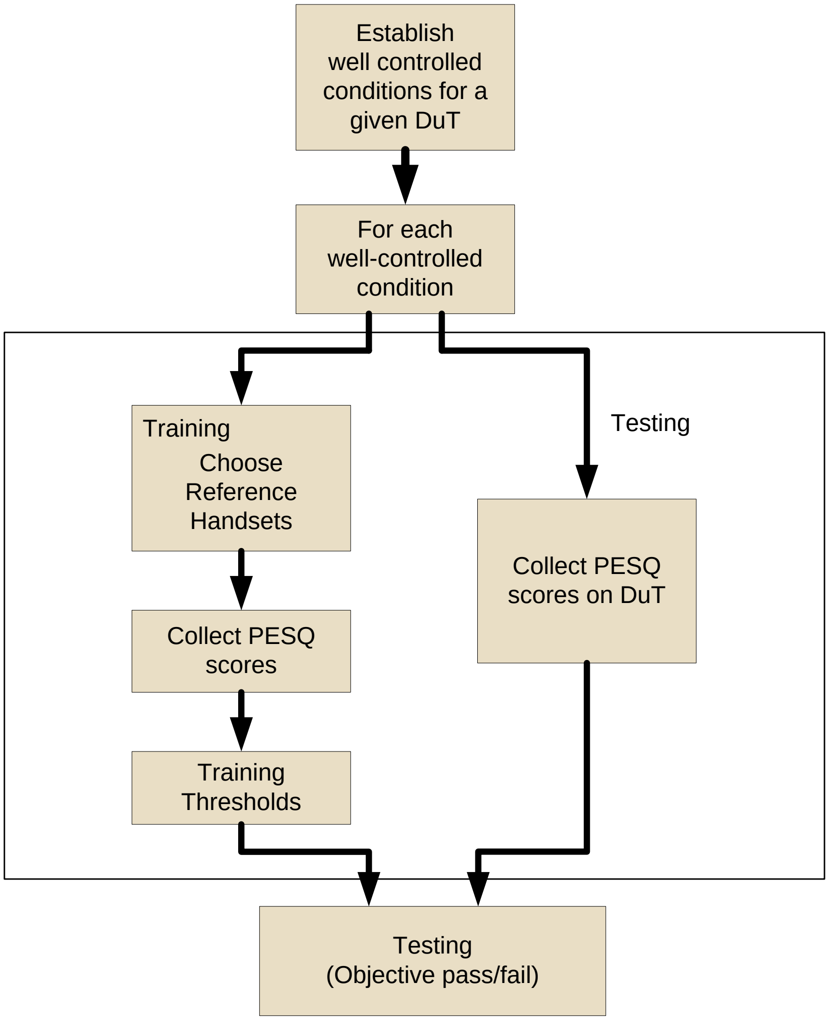
Establish well controlled conditions for a given DuT
For each well-controlled condition
Training
Choose Reference Handsets
Collect PESQ scores
Training Thresholds
Testing
Collect PESQ scores on DuT
Testing (Objective pass/fail)

Figure 5-1 Block diagram of the complete training and testing process

Given a handset for testing, well controlled conditions are established based on the knowledge of the test handset, the practicality of the constraints, and the test requirements. (Refer to Chapter 4 for more details.) Training and testing is performed for each well controlled condition as described below.

Reference handsets are chosen according to the well controlled condition. PESQ scores are collected from the reference handsets operating under the well controlled condition. The scores are then used for training and obtaining thresholds. Note that the training can be done off-line.

When testing a handset, PESQ scores are collected from the DuT under the well-controlled condition.

In the testing block, the test handset PESQ scores are compared with the thresholds for objective classification of the handset quality into good/bad.

Section 5.2 and Section 5.3 explain the training and testing methodology in detail.

5.2 Training methodology

The steps for training are shown below.

For a given well controlled condition (formed as described in Section 4.3):

1. Choose a few reference handsets which can operate under the given well controlled condition. The selected reference handsets should be good handsets.
2. Collect PESQ scores from the reference handsets based on the given well controlled condition (including input speech, insertion interface, logging location and constraints on voice path configuration).
3. Extract mean, standard deviation and minimum per-sentence-pair value of PESQ scores for each handset under the well controlled condition.

   The equation for mean is $Mean(m) = (1/N)\sum_{i=1}^{N} PESQ\_SP(i,m)$ -- (5.1)

   PESQ_SP(i,m) is the PESQ value of the i^{th} sentence pair in the m^{th} voice terminal among M terminals. For each terminal *m*, the mean value is computed.

   Similarly the standard deviation is computed for each voice terminal m as,

   $$std(m) = \sqrt{(1/N)\sum_{i=1}^{N} (PESQ\_SP(i,m) - mean(m))^2} \quad \text{-- (5.2)}$$

   The minimum per-sentence-pair PESQ score for each voice terminal m is computed as,

   $$\min(m) = \min(PESQ\_SP(i,m)) \quad \text{-- (5.3)}$$

4. Among all the reference handsets, store the minimum-most of the mean value *min(mean(m)),* and the minimum-most of minimum per-sentence-pair PESQ value *min(min(m))*. Also store the maximum standard deviation value *max(std(m))*. These values are the thresholds to represent the minimum performance criteria for handsets operating in the given well controlled condition.

5.3 Test methodology

The steps to test handset quality are shown below.

For a given well controlled condition,

1. Collect PESQ scores from the test handset based on the given well controlled condition (including input speech, insertion interface, logging location and constraints on voice path configuration). These scores are denoted as $Test_{PESQ}(i)$, where i is the index of sentence pairs.
2. The mean T_{mean}, standard deviation T_{std} and minimum per-sentence-pair value T_{min} of the PESQ scores are computed for the test voice terminal.
3. If (T_{mean}) < *min(mean(m)),* or if (T_{min}) < *min(min(m)),* or if (T_{std}) > *max(std(m)*; then the test handset is classified as an objective fail. Otherwise it is classified as an objective pass.
4. Subjective listening for verification of the objective pass/fail decision is preferred in order to eliminate any false positives or false negatives. This is especially useful when the number of the reference handsets is limited.

   To verify the objective test results, it is sufficient to listen to only a few sentence pairs. The following metrics are obtained to decide which sentence pairs to subjectively listen. Below are the steps to find out the sentence pairs for subjective listening:

   a. The average value of the PESQ score is calculated for each sentence pair, across the reference handsets. For i^{th} sentence pair, the average PESQ score is computed as

   $$avg_{PESQ}(i) = (1/M)\sum_{m=1}^{M} PESQ\_SP(i,m) \qquad \text{-- (5.4)}$$

   b. The average reference PESQ values avgPESQ are subtracted from the test handset PESQ values for each sentence pair, TestPESQ. For i^{th} sentence pair, the difference is defined as

   $$\Delta PESQ(i) = Test_{PESQ}(i) - avg_{PESQ}(i) \qquad \text{-- (5.5)}$$

   c. It is recommended to do subjective listening verification on the sentence pairs corresponding to the lowest $\Delta PESQ$ scores and the sentence pairs corresponding to the lowest $Test_{PESQ}$ scores. (An AB listening test between the degraded speech signals from reference handsets and test handset is recommended.)

The flowchart of the training and testing methodology for a given well controlled condition is shown in Figure 5-2. The training and testing procedures are also shown in the sample Python script attached in Appendix A.

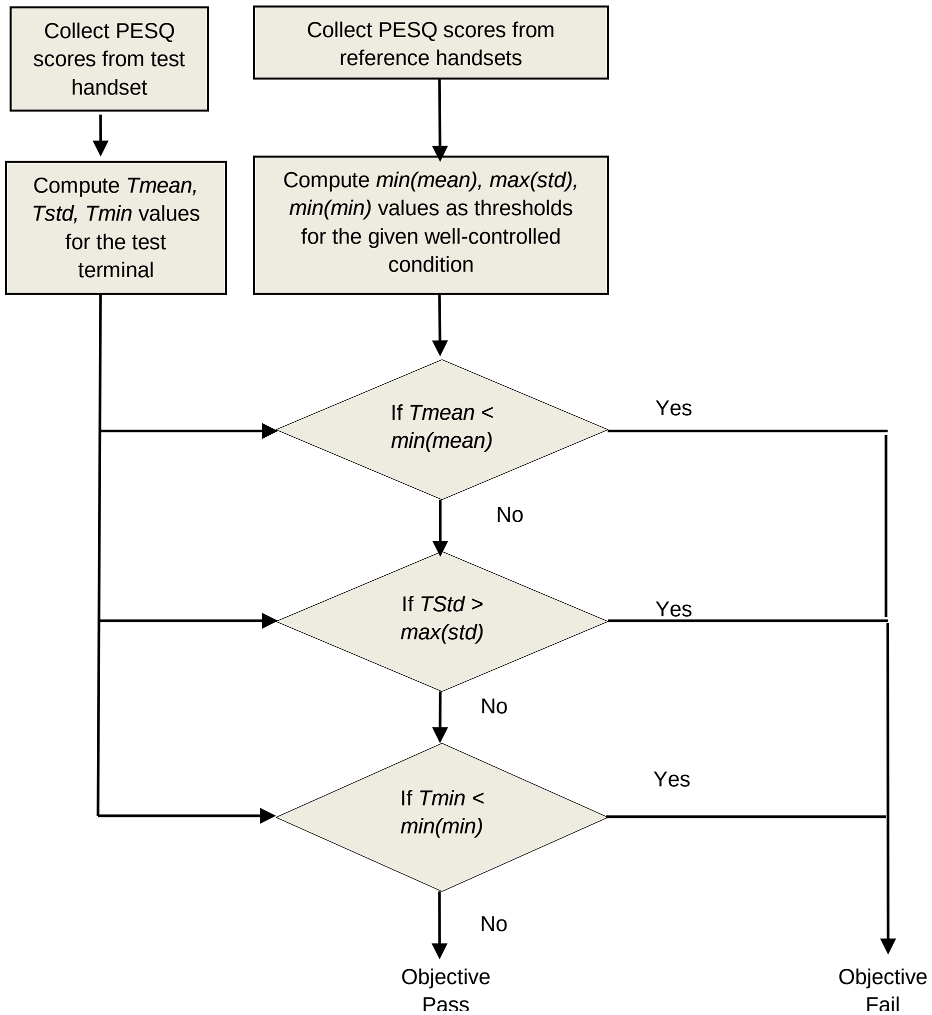
Collect PESQ scores from test handset
Collect PESQ scores from reference handsets
Compute Tmean, Tstd, Tmin values for the test terminal
Compute min(mean), max(std), min(min) values as thresholds for the given well-controlled condition
If Tmean < min(mean)
Yes
No
If TStd > max(std)
Yes
No
If Tmin < min(min)
Yes
No
Objective Pass
Objective Fail

Figure 5-2 Flow chart of Training and Testing methodology to get an objective pass/fail decision

5.4 Example for training and testing methodology

A simulated example

Assume that the test handset is a CDMA handset with EVRC-B codec. A bug is simulated in the test handset with 3% FER.

1. First, well controlled conditions are established for the test handset. Using the procedure explained in Section 4.3, it has been decided to put the constraints on the COPs of EVRC-B. Hence there are eight well controlled conditions (COP0 to COP7). Other constraints (such as input speech, logging, and insertion) are also defined in establishing these well controlled conditions. More details can be found in Section 4.3.

2. For any given well controlled condition, the training steps are as follows: (COP-0 is used as an example here.)

   a. Eight reference handsets which are capable of running EVRC-B with COP-0 are chosen for training the thresholds of the well controlled condition.

b. PESQ scores are collected according to the given well controlled condition.

c. Mean, minimum per-sentence-pair PESQ value, and the standard deviation are computed for each reference handset. The statistical parameters for the reference handsets are shown as red squares in the 3D plot of Mean vs. Minimum vs. Standard deviation in Figure 5-3.

d. The threshold values to represent the well controlled condition are:

- *min(mean) – 3.64*
- *min(min) – 3.54*
- *max(std) – 0.045*

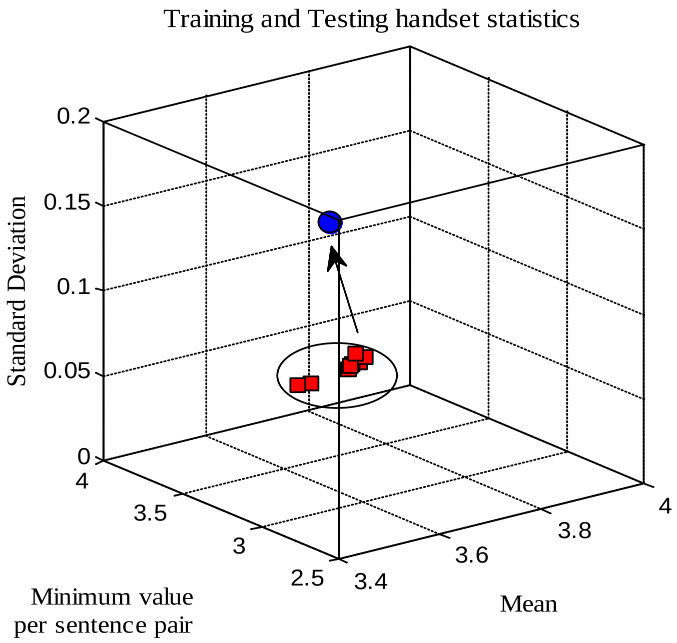

Figure 5-3 Mean vs. Minimum Value vs. Standard Deviation for the EVRC-B COP0 reference handsets (red box) and the EVRC-B COP0 test handset (blue circle). The test handset statistics are degraded and well separated from the training handset statistics.

3. The steps for testing are,

a. Operate the test handset under EVRC-B COP0 and collect PESQ scores.

b. The EVRC-B COP0 with 3% FER (simulation data) test handset statistics are obtained,

- T_{mean} – *3.52*
- T_{min} – *2.95*
- T_{std} – *0.172*

NOTE: The test handset statistical parameters are shown as the blue circle in Figure 5-3.

c. The test handset statistical parameters are compared with the threshold values. It is seen that T_{mean} < *min(mean)*, T_{min} < *min(min)*, and T_{std} > *max(std)*. The test handset fails all the three thresholds, hence it is classified as a fail handset (failing one threshold is enough to be classified as a fail handset).

5.4.1 Testing in a controlled environment using Metrico Wireless system and CMU200

The block diagram of the Metrico Wireless system is shown in Figure 5-4.

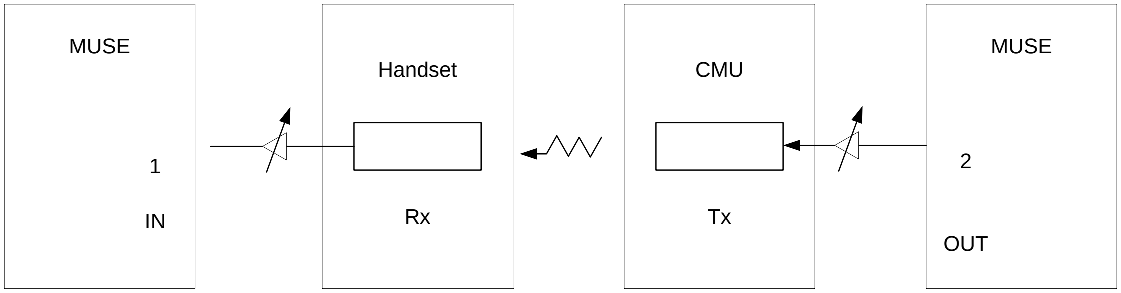

Figure 5-4 Block diagram of the downlink (Rx) test setup in Metrico Wireless system

NOTE: In the block diagram, MUSE is the name of the Metrico box.

There are two separate setups for the Tx and Rx paths of a handset:

Tx: When testing the Tx path of the test handset, the setup is such that the input sequence stored in MUSE is played into the microphone of the handset. The handset encodes the sequence and transmits it to the CMU. The CMU receives the packets, decodes them and sends them to the MUSE. Using the original input sequence and the decoded sequence in MUSE, PESQ measures the degradation due to the Tx path in the handset.

Rx: In the Rx path, the setup is such that MUSE sends the input sequence to CMU. CMU encodes the sequence and transmits the bit-stream to the handset. The handset receives the packets and decodes them. The resulting decoded sequence is electrically captured from the handset by MUSE through the headset interface. PESQ uses the original input sequence and the decoded sequence to measure the degradation in the Rx path.

In our example, we focus on measuring the voice quality degradation in the Rx path.

a. Forming a well controlled condition

Constraints are imposed on the configuration in CMU and the handset to form a well controlled condition.

Constraints imposed:

- The Artificial Speech Test Stimulus (ASTS) pre-stored in the Metrico box is used as the input sequence in all the experiments, and it is repeated 64 times in a single established Rx path.
- Lossless channel conditions are maintained in the communications between the handset and CMU for a controlled network environment.

- Electrical capture is used in the handset in the Rx path.
- Codec in the handset is fixed for each experiment for both reference and test handsets; When EVRC-B is tested, constraint on coding mode is achieved by setting the COP in CMU (the COP is specified as average bit rate in CMU).
- The speech level of the packets received at the handset is calibrated to be at a nominal level (-26 dBov). This is achieved by using a handset which supports packet logging.
- The capture gain in MUSE is also calibrated to avoid saturation.

Good reference handsets are chosen and the above constraints are imposed on these reference handsets to form a well controlled condition. Three reference handsets are used in the experiments.

It can be seen in Figure 5-5 that, while testing EVRC-B codec, variance of the PESQ scores increases if a constraint is not imposed on the COP of the codec as shown in Figure 5-5.

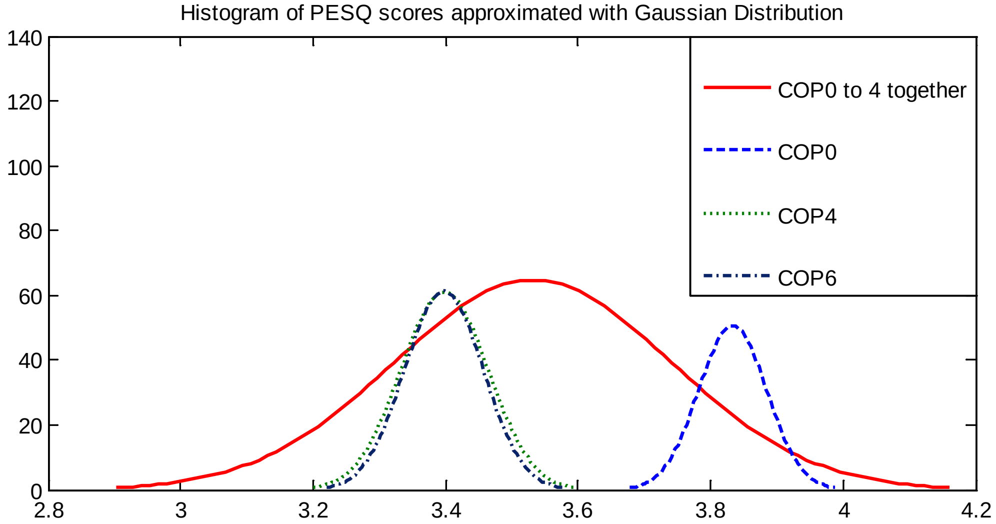

Figure 5-5 Distribution of PESQ scores from reference handsets for each of the EVRC-B COPs 0, 4 and 6, separate and combined.

b. Training and testing procedures

Training thresholds are obtained from the reference handsets separately for each codec and coding mode. Three reference handsets are used. The constraints listed in 5.4.1.a are used to form well controlled conditions. The statistics obtained from training and testing handsets and the pass/fail result for each test handset are shown in Table 5-1. The pass/fail result is obtained using the comparative analysis described in Section 5.3.

Table 5-1 PESQ Statistics obtained from training and testing handsets, and the pass/fail result for each test handset when tested in Metrico Wireless system

Codec	Reference HS Statistics	Representative Thresholds	Test HS Statistics	Pass/Fail Result
EVRC	Ref HS1- Mean: 3.72 SD: 0.047 Min: 3.63	Min(mean): 3.72 Min(min): 3.59 Max(SD): 0.059	Test HS1- Mean: 3.27 SD: 0.134 Min: 2.99	Test HS1- Fail
	Ref HS2- Mean: 3.75 SD: 0.047 Min: 3.62		Test HS2- Mean: 3.31 SD: 0.27 Min: 2.63	Test HS2- Fail
	Ref HS3- Mean: 3.73 SD: 0.059 Min: 3.59		Test HS3- Mean: 3.43 SD: 0.16 Min: 2.85	Test HS3- Fail
			Test HS4- Mean: 3.81 SD: 0.04 Min: 3.67	Test HS4- Pass
EVRC-B COP0	Ref HS1- Mean: 3.81 SD: 0.05 Min: 3.70	Min(mean): 3.81 Min(min): 3.70 Max(SD): 0.05	Test HS1- Mean: 3.41 SD: 0.167 Min: 2.97	Test HS1- Fail
	Ref HS2- Mean: 3.86 SD: 0.042 Min: 3.74		Test HS2- Mean: 3.51 SD: 0.063 Min: 3.29	Test HS2- Fail
	Ref HS3- Mean: 3.92 SD: 0.043 Min: 3.81			
EVRC-B COP4	Ref HS1- Mean: 3.38 SD: 0.063 Min: 3.19	Min(mean): 3.38 Min(min): 3.14 Max(SD): 0.063	Test HS1- Mean: 3.06 SD: 0.11 Min: 2.84	Test HS1- Fail
	Ref HS2- Mean: 3.42 SD: 0.07 Min: 3.28		Test HS2- Mean: 3.20 SD: 0.057 Min: 3.06	Test HS2- Fail
	Ref HS3- Mean: 3.39 SD: 0.075 Min: 3.14			

Codec	Reference HS Statistics	Representative Thresholds	Test HS Statistics	Pass/Fail Result
EVRC-B COP6	Ref HS1- Mean: 3.39 SD: 0.061 Min: 3.28 Ref HS2- Mean: 3.40 SD: 0.058 Min: 3.21 Ref HS3- Mean: 3.40 SD: 0.073 Min: 3.21	Min(mean): 3.39 Min(min): 3.21 Max(SD): 0.073	Test HS1- Mean: 2.99 SD: 0.14 Min: 2.63 Test HS2- Mean: 3.20 SD: 0.055 Min: 3.08	Test HS1- Fail Test HS2- Fail

The objective pass/fail results agree with subjective listening. The log from Test HS1 contains echoes and noises. The log from Test HS2 has unexpected frame erasure-like artifacts.

5.4.2 Testing in a controlled environment using ACQUA Audio Analyzer and CMU200

Another test setup based on an ACQUA Audio Analyzer and CMU200 is used for voice quality evaluation. This example is used to illustrate the difference in PESQ scores and corresponding statistics between different well controlled conditions (i.e., with different testing setups which use different input sequences). Though the reference and test handsets used are the same as those used in the previous example, the PESQ scores and the corresponding statistics are different. The test setup used in this example is shown in Figure 5-6.

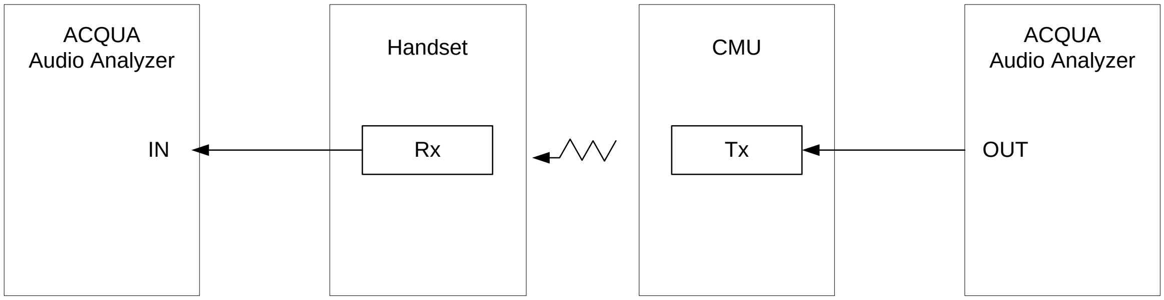

Figure 5-6 Block diagram of the downlink (Rx) test setup formed using ACQUA Audio Analyzer and CMU200

In this example, only the downlink (Rx) path is tested in the controlled environment. The input sequence is sent from the ACQUA Audio Analyzer to the CMU. The CMU encodes the sequence and transmits it to the handset. The handset decodes the received bit-stream. The decoded sequence is electrically captured from the handset by the ACQUA Audio Analyzer.

The overall degradation of voice quality in the Rx path is measured using the input sequence and the decoded output sequence received by ACQUA.

a. Forming a well controlled condition

Constraints are imposed on the configuration in the CMU and the handset to form a well controlled condition.

Constraints imposed:

1. An American English ITU-T P.501 input sequence stored in the ACQUA software is used in all the experiments, and it is repeated 64 times in a single established Rx path.
2. Lossless channel condition is maintained in the communications between the handset and CMU for a controlled network environment.
3. Electrical capture is used in the handset in the Rx path.
4. Codec in the handset is fixed for each experiment for both reference and test handsets; when EVRC-B is tested, constraint on coding mode is achieved by setting the COP in CMU (the COP is specified as average bit rate in CMU).
5. The capture gain in the ACQUA system is also calibrated to avoid saturation.

Good reference handsets are chosen and the above constraints are imposed on the handsets to form a well controlled condition. Three reference handsets are used in all the experiments.

Figure 5-7 shows that while testing EVRC-B codec, variance of the PESQ scores increases if a constraint is not imposed on the COP of the codec.

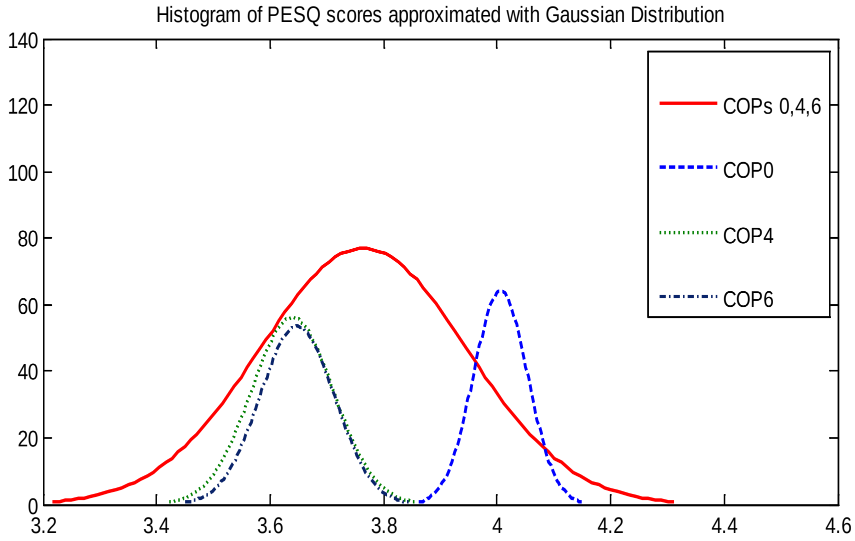

Figure 5-7 Distribution of PESQ scores for each of the EVRC-B COPs 0, 4, and 6, separate and combined. PESQ scores are obtained from the reference handsets

b. Training and Testing procedures

Training thresholds are obtained from the reference handsets separately for each codec. Three reference handsets are used in all the experiments. The constraints listed in Section 5.4.2.a are used to form a well controlled condition. The statistics obtained from training and testing handsets, and the pass/fail result for each test handset are shown in Table 5-2. The pass/fail result is obtained using the comparative analysis described in Section 5.3.

Table 5-2 PESQ Statistics obtained from training and testing handsets, and the pass/fail result for each test handset when tested in the system composed of ACQUA Audio Analyzer and CMU

Codec	Reference HS Statistics	Representative Thresholds	Test HS Statistics	Pass/Fail Result
EVRC	Ref HS1- Mean: 3.8 SD: 0.07 Min: 3.6 Ref HS2- Mean: 3.95 SD: 0.049 Min: 3.78 Ref HS3- Mean: 3.97 SD: 0.049 Min: 3.82	Min(mean): 3.8 Min(min): 3.6 Max(SD): 0.07	Test HS1- Mean: 3.68 SD: 0.117 Min: 3.37 Test HS2- Mean: 3.24 SD: 0.052 Min: 3.11 Test HS3- Mean: 3.80 SD: 0.14 Min: 3.43 Test HS4- Mean: 3.8 SD: 0.042 Min: 3.73	Test HS1- Fail Test HS2- Fail Test HS3- Fail Test HS4- Pass
EVRC-B COP0	Ref HS1- Mean: 3.98 SD: 0.046 Min: 3.87 Ref HS2- Mean: 4.02 SD: 0.038 Min: 3.95 Ref HS3- Mean: 3.99 SD: 0.044 Min: 3.88	Min(mean): 3.98 Min(min): 3.87 Max(SD): 0.046	Test HS1- Mean: 3.09 SD: 0.101 Min: 2.63 Test HS2- Mean: 3.38 SD: 0.047 Min: 3.11	Test HS1- Fail Test HS2- Fail
EVRC-B COP4	Ref HS1- Mean: 3.62 SD: 0.076 Min: 3.46 Ref HS2- Mean: 3.65 SD: 0.067 Min: 3.45 Ref HS3- Mean: 3.59 SD: 0.048 Min: 3.48	Min(mean): 3.59 Min(min): 3.45 Max(SD): 0.076	Test HS1- Mean: 3.42 SD: 0.11 Min: 3.1 Test HS2- Mean: 3.24 SD: 0.06 Min: 2.89	Test HS1- Fail Test HS2- Fail

Codec	Reference HS Statistics	Representative Thresholds	Test HS Statistics	Pass/Fail Result
EVRC-B COP6	Ref HS1- Mean: 3.63 SD: 0.066 Min: 3.48 Ref HS2- Mean: 3.67 SD: 0.058 Min: 3.55 Ref HS3- Mean: 3.62 SD: 0.053 Min: 3.5	Min(mean): 3.62 Min(min): 3.48 Max(SD): 0.066	Test HS1- Mean: 2.91 SD: 0.11 Min: 2.58 Test HS2- Mean: 3.22 SD: 0.049 Min: 3.05	Test HS1- Fail Test HS2- Fail

The objective pass/fail results agree with subjective listening. The log from Test HS1 has echoes and noises. The log from Test HS2 has unexpected frame erasure like artifacts.

5.4.3 Observations made in the Metrico and ACQUA experiments:

The following observations were made from the experiments:

1. The PESQ scores and PESQ-based statistics from the Metrico results are different from the ACQUA results, although the same handsets are used in both experiments. One reason is that different input speech materials are used in these tests. This emphasizes the importance of constructing well controlled conditions (including selection of input sequences) when doing a comparison. The scores/thresholds obtained from different test setups should not be compared without close examination.
2. Since a source controlled variable bitrate codec such as EVRC-B takes time to converge to its average bit rate (the COP selected), it is a good idea to use multiple sentence pairs similar to that used in the experiments (64 sentence pairs).

6 Conclusions

This document proposes a methodology for voice terminal quality testing. The methodology overcomes the limitations of existing objective speech quality measurement tools (such as PESQ) in voice quality assessment. The idea of a well controlled condition is proposed to limit the variation of PESQ scores. Voice quality can be reliably tested by comparing the test handset to reference handsets within the same well controlled conditions. The training and testing procedures for testing handset quality have been described in detail in this document. The training and testing sample Python script is shown in Appendix A.

A Appendix

The sample Python script for training and testing is in the attached zip file, along with simulation results for the example given in Section 5.4. It requires additional xlrd, xlwt libraries for reading from and writing to an Excel spreadsheet. The script reads the training, testing handset data from the spreadsheet, and writes the results into another spreadsheet. The input data has to be arranged in the spreadsheet's 'Scores.xls' such that the first row contains the handset details, and the following rows contain the PESQ scores for each sentence pair for each corresponding handset in row one. The last column is for test handset data and the other columns are for the training handset data.

Double click on each script to open and save, if desired.