

# Speech Quality Index in CDMA 2000

Technical Paper



# 1 Introduction

Mobile telephony is moving fast from being an expensive accessory for the business-man, to the natural choice of communication for people at all professions and ages. In some countries the number of mobile subscriptions are now higher than the number of fixed subscriptions, and the trend is now moving fast towards establishing the mobile as the basic communication device; with camera, music player and soon even TV capabilities.

However, the most important basic service will still be plain old speech, and with more and more of the communication moving from fixed to mobile systems, the higher the quality demands will be. Users will expect a mobile speech quality which is close to the fixed line quality, or, with the future introduction of wide-band speech codecs, even better-than-fixed quality.

At the same time, the ever-expanding mobile usage creates capacity problems in many networks. The wireless spectrum is a limited resource, and when more and more mobiles are being used in a geographic area, there is a substantial risk that the speech quality eventually suffers.

It is therefore important for the operator to be able to measure the end-user speech quality in an accurate and efficient manner, so that preventive actions can be taken if the quality trend is pointing downwards in the network. This paper discusses how such **objective speech quality** measurements can be made, and also how these measurements relate to the **subjective speech quality**.

# 2 Subjective Speech Quality

Subjective speech quality is defined as the result of a subjective test, where a number of test persons listen to and judge the test material. There exist a number of different standardized ways of measuring subjective speech quality, depending on the purpose of the measurement.

In some cases the intelligibility is in focus, for instance in military command systems, while in other cases the amount of degradation or noise level is most important. In yet another case the pleasantness or maybe the ease of conversation might be in focus.

## 2.1 The MOS Scale

A common measure for subjective speech quality is the Mean Opinion Score (MOS) scale, defined in the ITU-T standard P.800 [1]. In a MOS test, the test persons listen to short speech samples, where every speech sample consists of two to five sentences. In practice, two sentences are often used, and the total length of a speech sample is about five to eight seconds, including some silence before, between, and after the sentences.

After listening to each speech sample, the test person shall grade the sample according to the following scale:

<u>Quality of the speech</u>	<u>Score</u>
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The total MOS score is then the mean of all individual results. Due to the absolute nature of the grading, this kind of test is also called an Absolute Category Rating (ACR) test.

Despite the apparent simplicity of a MOS test, there are many factors which in practice influence the results of such a test. The P.800 standard contains a lot of recommendations about speech material, listening levels, talker gender, selection of test persons, randomization of listening order etc., so that the results shall be as general as possible.

Nevertheless it is true that results from different MOS tests can never be directly compared. The results of a MOS test can only be used in a relative sense between the tested conditions inside each test. It is a common misunderstanding that similar MOS values derived from different MOS tests describe the same subjective speech quality.

## 2.2 The MNRU Scale

The ITU-T standard P.810 "Modulated Noise Reference Unit (MNRU)" [2] describes how a speech sample can be distorted in a mathematically deterministic way by adding multiplicative band-limited white noise. Experiments have shown that the subjective impact of adding a certain amount of MNRU noise to different speech samples are relatively similar.

This can be used in a MOS test by including some extra speech samples in the test, where these extra samples has been distorted by adding different amount of MNRU distortions. The MNRU scale is measured in dBQ, which corresponds to the signal-to-noise ratio between the original speech signal and the noise signal. In a MOS test typical MNRU values for creating these extra signals are 0, 6, 12, 18, 24, 30, 36 dBQ.

When a MOS test is finished, the extra MNRU samples now has both an MOS value and an MNRU value associated with them. This makes it possible to produce a relation between the dBQ and MOS values for a test. Typically this relation is produced by a regression of a sigmoid curve.

By using such a relation between MOS and dBQ, the MOS score for every speech sample in the test can be translated to an "equivalent Q" in the dBQ domain. While MOS values from different MOS tests cannot be directly compared, the MOS values transformed into the dBQ domain tend to be more experiment and language independent, and a comparison even between different tests is normally possible, within reasonable limits.

Figure 1 below shows a typical example of such a MOS-to-dBQ mapping, derived from a subjective test for the SMV codec [3]. Note that every mapping is unique for each MOS test.

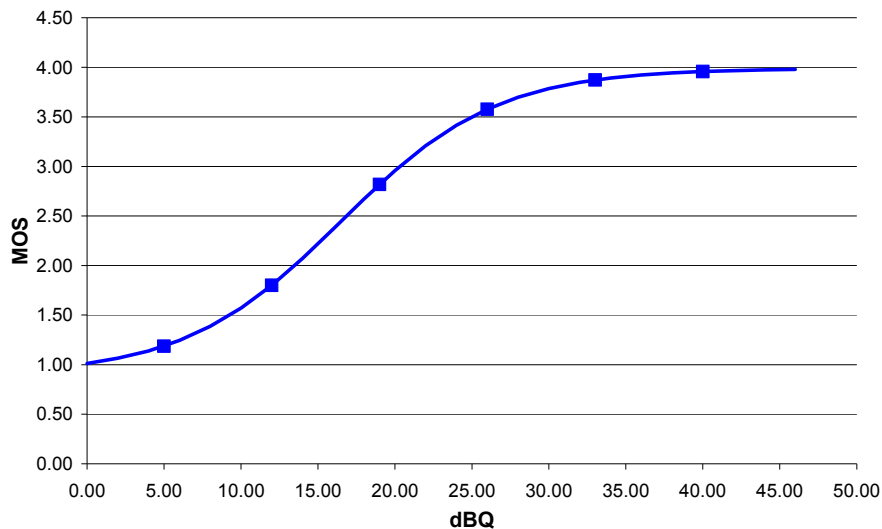


Figure 1 An example mapping between MOS and dBQ

### 3 Objective Speech Quality

The main drawbacks with using subjective tests are that they are expensive, and can only grade a limited number of speech samples. Thus it is impractical to use them for network monitoring purposes, and an operator must rely on other ways of assessing the speech quality in his network.

#### 3.1 PESQ

The ITU-T standard P.862 "Perceptual Evaluation of Speech Quality (PESQ)" [4] implements an algorithm which compares an original speech sample (the "reference signal") with another recorded speech sample (the "degraded signal"). By identifying the differences, and by modelling the characteristics of the human perception, the PESQ algorithm produces a MOS-like score for each speech sample. A simplified block diagram of PESQ is shown in Figure 2 below.

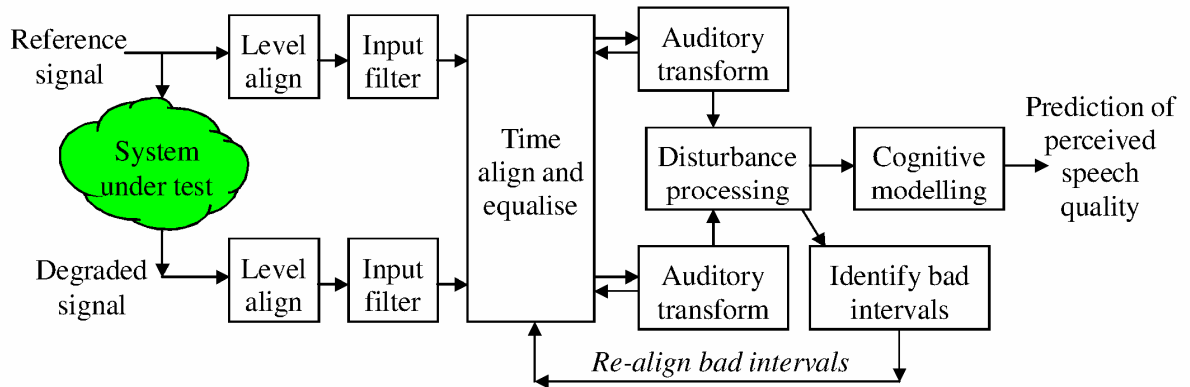


Figure 2 PESQ block diagram

### 3.2 SQI

The Speech Quality Index (SQI) is a patented algorithm which is built upon the fact that almost all speech distortions in a mobile network are due to problems with the radio transmission. By using detailed information from the channel and speech decoder, together with recorded and subjectively graded speech material, it is possible to build an algorithm which translates these errors into their effect on the resulting speech quality. Figure 3 below illustrates the basic SQI concept.

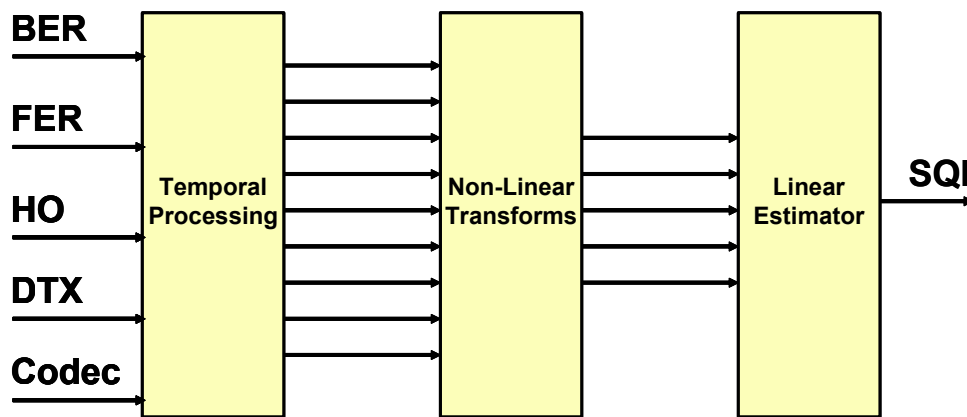


Figure 3 SQI algorithm concept for GSM

SQI was originally deployed as an uplink quality measurement parameter in Ericsson's GSM system for the Enhanced Full-rate, the Full-rate, and the Half-rate codecs, and has subsequently been expanded with the Adaptive Multi-Rate codec family. For the downlink, SQI has been implemented in the TEMS product family, and, with the advent of Enhanced Measurement Reporting, also on the GSM system side.

Due to the inherent non-comparability characteristics of the ordinary MOS scale, the SQI scale has been directly tuned towards the more general dBQ scale. The range of SQI is thus from below 0 dBQ for extremely bad samples and up to about 30 dBQ, which corresponds to the top quality of the Enhanced Full-Rate and AMR 12.2 codecs.

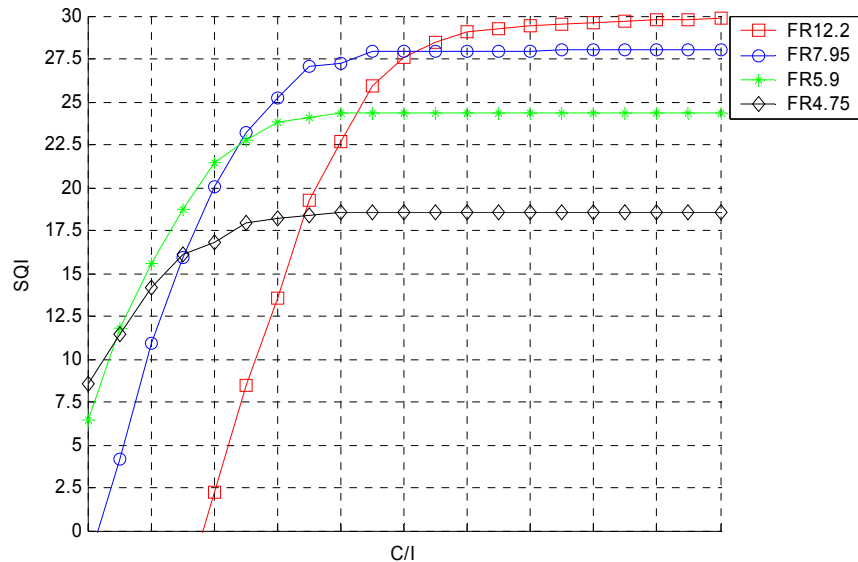


Figure 4 Example of SQI curves for selected GSM AMR full-rate modes

Figure 4 above shows typical SQI curves for GSM AMR FR for four example codec modes. The SQI algorithm use detailed codec and radio information for each speech frame to produce an SQI value covering the last 2.5 seconds of time. If codec mode changes have been done during these 2.5 seconds, these changes are also reflected in the total SQI value, since different SQI codec models will then be used for different parts of the sentence, and then aggregated to a total SQI score for the whole sentence.

### 3.3 SQI in CDMA 2000

The latest addition to the TEMS SQI family are the CDMA 2000 codecs QCELP13K, EVRC, SMV and narrow-band operation of VMR-WB (however, the SMV and VMR-WB codecs are not yet deployed). The different CDMA codecs operate within one of two rate sets and use the following rates:

Rate	Rate set 1 (EVRC, SMV)	Rate set 2 (QCELP13K, VMR-WB)
Full-rate	8.55 kbps	13.3 kbps
Half-rate	4.0 kbps	6.2 kbps
Quarter-rate	2.0 kbps	2.7 kbps
Eighth-rate	0.8 kbps	1.0 kbps

Table 1 CDMA rate sets

Unlike GSM, where the source coding is not related to the speech signal (except for DTX at silent periods), the CDMA codecs selects the rate to be used depending on the speech signal, so that speech parts which are easier to code is coded with a lower rate. The lowest rate, the eighth-rate, is primarily used for coding non-speech, i.e. silence or noise between the active parts of the speech signal.

The QCELP13K and the EVRC are both single-mode codecs, while the SMV and VMR-WB can operate in different modes depending on the operator-selected trade off between quality and system capacity. The modes differ in their rate usage mix, where better modes typically use a higher proportion of full-rate.

The SMV modes are:

- Mode 0: **Premium**, with an average bit-rate equal to EVRC, but better speech quality
- Mode 1: **Standard**, with lower bit-rate than EVRC, but comparable quality
- Mode 2: **Economy**, with still lower bit-rate, but slightly decreased quality
- Mode 3: **Super economy**, with even lower bit-rate, and a noticeable decrease in quality
- Mode 4: **Half-rate 0**, for bandwidth-sensitive applications, a mode which use half-rate as its maximum coding rate
- Mode 5: **Half-rate 1**, as above, but with slightly lower average bit-rate

The VMR-WB modes are:

- Mode 0: **Premium**, providing the best quality
- Mode 1: **Standard**, with lower average bit-rate
- Mode 2: **Economy**, with still lower bit-rate
- Mode 3: **AMR-WB**, with an average bit-rate slightly higher than mode 0, but fully interoperable with the 3GPP AMR-WB codec

Note that due to the source-controlled characteristics of the CDMA codecs, a normal speech signal must be sent in the downlink during the SQI measurements, since otherwise only the eighth-rate silence frames will be transmitted.

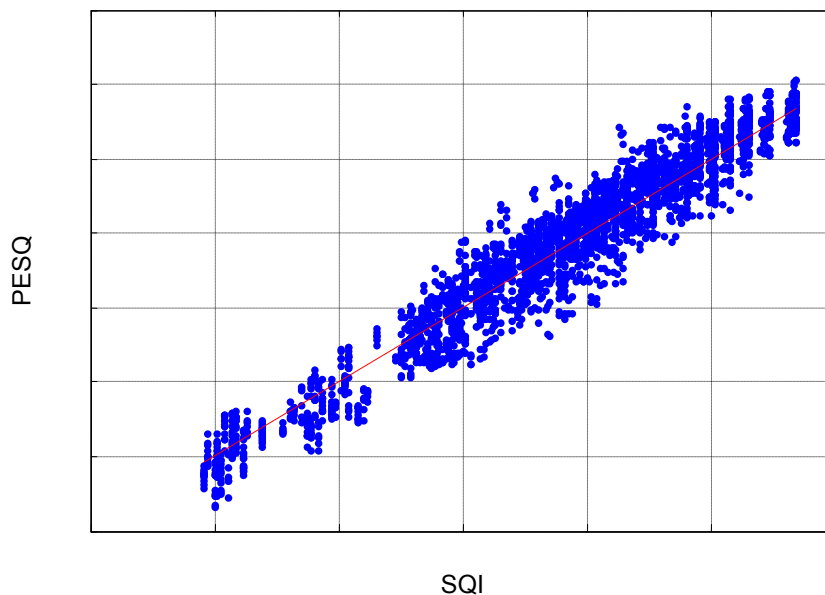
## 4 PESQ vs. SQI

Even though PESQ and SQI both measure the speech quality, the quality is derived in different ways, and thus have different characteristics.

PESQ measure the complete end-to-end speech quality, including any impact from the core network or hardware failures, which gives a more complete test coverage. However, it is not easily possible to see if quality problems are due to radio problems or due to failures in other parts of the call chain.

SQI, on the other hand, only use radio-related parameters to calculated the speech quality. This does not give a complete speech-path coverage, but has the advantage of addressing the speech impact of the radio network, which is often the main concern for an operator.

Figure 5 below shows PESQ and SQI evaluated on a large number of simulated conditions, giving a mutual correlation between the two quality measures of 96%.



*Figure 5 Example relation between SQI and PESQ for the same simulated conditions (for the EVRC codec)*

Since PESQ operates directly on the speech signal, it is important to use a representative set of signals during the measurement. For instance, if only a low-pitched male speech signal is used this gives slightly higher scores than if a high-pitched female voice is used. The reason is that it is in general more difficult to code high-pitched signals.

This difference in scoring is in principle not an error, since the quality of the two signals will be slightly different. However, it means that you need to use a set of carefully selected speech signals during the measurement, and each of these signals will give a slightly different score, even for identical transmission conditions.

To avoid a fluctuating score, the results from the PESQ measurements should be averaged over the complete set of speech signals. This gives a stable score, but due to the averaging over a longer time period the geographical resolution of the averaged score will be rather low.

SQL, on the other hand, has been trained and modelled with a corresponding set of selected speech signals, and the resulting model always predicts the average speech quality, given a certain radio condition. This gives both a stable speech quality value and good geographical resolution. Figure 6 below illustrates the sentence effect with a test case using eight different speech signals, repeated cyclically (1, 2, 3, 4, 5, 6, 7, 8, 1, 2, ...)

This effect can also be seen in the upper right part of Figure 5, where PESQ gives a wider range of quality scores than SQL, due to the use of several sentences.

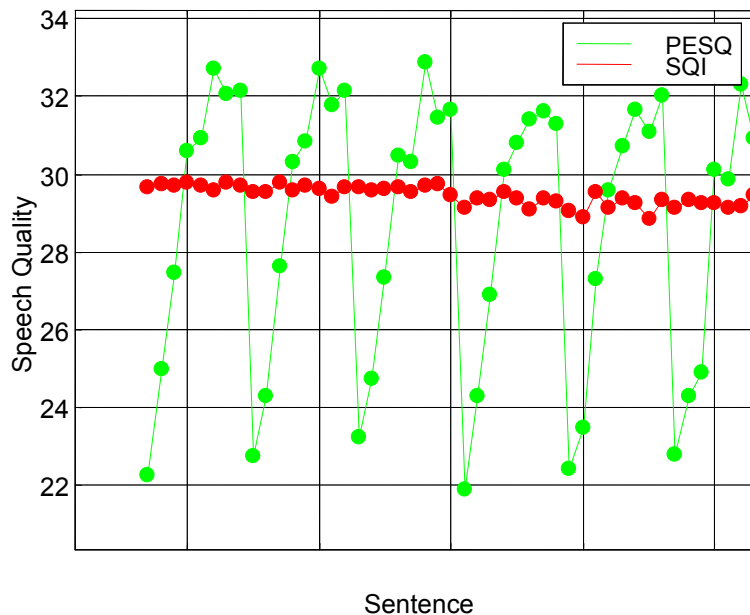


Figure 6 Sentence dependency for PESQ

The SQL algorithm is based on the typical quality characteristics of a fixed-to-mobile call, thus the background noise is assumed to be low, and the effect of noise suppression algorithms are not included in the model. This is also the case for most PESQ measurements, since the original speech signals used are often relatively noise-free. Handling the quality impact of high-level background noise is a difficult topic where more speech quality research is needed.

## 5 References

- [1] ITU-T standard P.800; Methods for objective and subjective assessment of quality
- [2] ITU-T standard P.810; Modulated noise reference unit (MNRU)
- [3] 3GPP2-C11-20010326-003; SMV Post-Collaboration Subjective Test – Final Host and Listening Lab Report; Figure 5-1
- [4] ITU-T standard P.862; Perceptual evaluation of speech quality (PESQ)