Speech Quality Measurement with SQI

Technical Paper

ascom

Contents

1. Introduction

TEMS products offer the quality measure SQI (Speech Quality Index) for estimating the downlink speech quality in a GSM, WCDMA, or CDMA cellular network as perceived by a human listener. SQI has been developed by Ericsson.[1]

Computing SQI for GSM and WCDMA requires data collected with Sony Ericsson phones. SQI for CDMA can be based on data from any CDMA phone that is connectable in TEMS Investigation.

2. Background

2.1. SQI for UMTS

SQI for GSM and WCDMA is a long-standing feature of TEMS products. However, in TEMS Investigation 9.0, the SQI algorithm was completely reworked, although its fundamental function remains similar to that of the old algorithm. The focus of this document is to describe the new algorithm (called “SQI-MOS” in the application; see chapter 4). Reference is made to the previously used algorithm (the “old SQI”), and attention is drawn to certain important differences between the algorithms, but no comprehensive point-by-point comparison is made.

As wideband speech codecs will soon be available in mobile phones and networks, the SQI-MOS algorithm includes a model for rating wideband speech.

2.2. SQI for CDMA

SQI for CDMA is introduced in this version of TEMS Investigation. It uses an SQI-MOS algorithm similar to those for GSM and WCDMA.

SQI for CDMA currently does not support wideband.

3. Input to the SQI-MOS Algorithm

3.1. UMTS

SQI-MOS for UMTS takes the following parameters as input:

- *Frame error rate (FER),* i.e. the percentage of frames that are lost on their way to the receiving party, usually because of bad radio conditions.

[1] The TEMS business was owned by Ericsson until 2009, when it was acquired by Ascom.

Frame errors also occur in connection with *handover*, and these are treated like any other frame errors by the SQI-MOS algorithm. It should be noted that in WCDMA, handover frame errors can usually be avoided thanks to the soft handover mechanism. In GSM, on the other hand, every handover causes a number of frames to be lost.

Handovers are not modeled independently in any way by SQI-MOS.[2] More generally, the current algorithm also does not consider the *distribution* of frame errors over time.

- *Bit error rate (BER).* This is available in GSM only; no such quantity is reported by UEs in WCDMA mode.
- The *speech codec* used. The general speech quality level and the highest attainable quality vary widely between codecs. Moreover, each speech codec has its own strengths and weaknesses with regard to input properties and channel conditions. The same basic SQI-MOS model is used for all supported speech codecs, but the model is tuned separately for each codec to capture its unique characteristics.

SQI-MOS for UMTS is implemented for the following codecs:

- GSM EFR, GSM FR, and GSM HR
- all GSM AMR-NB and AMR-WB modes up to 12.65 kbit/s:
  - for narrowband, 4.75 FR/HR, 5.15 FR/HR, 5.9 FR/HR, 6.7 FR/HR, 7.4 FR/HR, 7.95 FR/HR, 10.2 FR, and 12.2 FR;
  - for wideband, 6.60, 8.85, and 12.65
- all WCDMA AMR-NB and AMR-WB modes up to 12.65 kbit/s:
  - for narrowband, 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, and 12.2;
  - for wideband, 6.60, 8.85, and 12.65.

3.2. CDMA

SQI-MOS for CDMA closely resembles WCDMA SQI; compare section 3.1. Input parameters are:

- Frame error rate
- Speech codec used, including bit rate information

The general discussion of these parameters in section 3.1 applies equally to CDMA (with the term “handoff” substituted for “handover”).

[2] In contrast, the old SQI algorithm included a special “handover penalty” mechanism lowering the SQI score whenever a handover occurred.

SQI-MOS for CDMA is implemented for the following codecs:

- QCELP13K
- EVRC
- SMV
- VMR-WB (narrowband input only)

4. SQI-MOS Output

The output from the SQI-MOS calculation is a score on the ACR[3] MOS scale which is widely used in listening tests and familiar to cellular operators. The score is thus a value ranging from 1 to 5.

The SQI-MOS algorithm produces a new quality estimate at intervals of

- *(UMTS)* approximately 0.5 s
- *(CDMA)* 2–4 s

Such a high update rate is possible thanks to the low computational complexity of the algorithm.

4.1. Narrowband vs. Wideband SQI-MOS (UMTS)

It is necessary to point out that narrowband and wideband SQI-MOS scores are not directly comparable. The same MOS scale and range are used for both (as is the custom in the field of speech quality assessment); however, a given MOS score indicates, in absolute terms, a higher quality for wideband than for narrowband. This is because wideband speech coding models a wider range of the speech frequency spectrum and is thus inherently superior to narrowband coding. The highest attainable quality is therefore markedly better for wideband. It follows from this that when interpreting a figure such as SQI-MOS = 4.0, it is necessary to consider what speech bandwidth has been encoded. A further complicating circumstance is that there is no simple mapping between wideband and narrowband SQI-MOS, for reasons sketched in section 5.1.

4.2. SQI-MOS vs. Old SQI (UMTS)

The old SQI was expressed in dBQ.[4] It should be stressed that SQI-MOS cannot be derived from these dBQ scores; the two algorithms are distinct (even if similar in general terms), and no exact mapping exists in this case either.

[3] ACR stands for Absolute Category Rating: this is the “regular” MOS test where speech samples are rated without being compared to a reference.

5. Alignment of SQI-MOS and PESQ

The SQI-MOS algorithm has been designed to correlate its output as closely as possible with the PESQ measure (Perceptual Evaluation of Speech Quality).[5] In fact, the SQI-MOS models have mostly been trimmed using PESQ scores, rather than actual listening tests, as benchmarks.[6] The exception is the wideband modes, where adjustments to the models have been made using the results of external listening tests. Regarding the latter, see section 5.1.

Note carefully that PESQ and SQI-MOS do not have the same scope. PESQ measures the quality end-to-end, that is, also taking the fixed side into account, whereas SQI reflects the radio link quality only. This means that PESQ and SQI values may differ while both being accurate in their respective domains.

Also bear in mind that PESQ and SQI-MOS use fundamentally different approaches to quality measurement:

- PESQ is a *reference-based* method which compares the received degraded speech signal with the same signal in original and undistorted form.
- SQI-MOS, on the other hand, is a *no-reference* method that works with the received signal alone and extracts radio parameters from it (as described in chapter 3).

Both methods try to assess to what degree the distortions in the received signal will be audible to the human ear; but they do it in completely different ways.

PESQ scores need to be averaged over a range of speakers in order to eliminate speaker bias, i.e. variation stemming from the characteristics of individual speakers. Such averaging is not required in the case of SQI-MOS, since the speaker-contingent variation is already built into the model (it has been trained with a large number of speakers).

5.1. Notes on PESQ for Wideband (UMTS)

(This subsection is relevant for UMTS only, since CDMA SQI currently does not extend to wideband.)

The PESQ algorithm for wideband (8 kHz) speech coding – as opposed to that for narrowband (4 kHz) – is afflicted with certain recognized shortcomings. The use of PESQ as a benchmark therefore complicated the development of SQI-MOS for wideband. Below is a brief discussion of this topic.

---

[4] The old SQI is still accessible in TEMS products (TEMS Investigation, TEMS Presentation), side by side with SQI-MOS.

[5] See www.pesq.org.

[6] This is completely different from the old SQI algorithm, which was trained using listening tests alone. At the time that work was done, no objective speech quality measure of the caliber of PESQ was yet commercially available.

One relevant fact is that, in certain circumstances, wideband PESQ has been found to produce lower scores than narrowband PESQ, even for clean speech.[7] This difference in output range would not in itself be problematic if wideband PESQ behaved similarly to narrowband PESQ as a function of FER; a mapping could then be applied to align the wideband scores to narrowband.

Unfortunately, things are not that simple. Wideband PESQ is much more sensitive to speaker bias than is narrowband PESQ (compare the introduction of chapter 5): at a fixed FER, wideband PESQ scores for different speakers show a spread of more than one point on the MOS scale. For narrowband, this variability is limited to a few tenths of a MOS point.

The upshot of this is that no straightforward mapping between wideband and narrowband PESQ can be constructed, and consequently outputs from the two are not directly comparable.[8] Attempts have been made within ITU to develop such a mapping, but so far with no satisfactory results. (It is probable that the task of assessing wideband speech quality requires further refinement of the mathematical models used.)

For the reasons explained above it was necessary to resort to other reference material besides PESQ scores in order to avoid biasing the wideband SQI-MOS model. The material used was the results from listening tests conducted during standardization of the AMR speech codec; see ref. [1]. Only clean speech ratings from these tests were used.

This tuning resulted in an adjustment of the SQI-MOS model that is linear as a function of FER. The largest correction was applied to the clean-speech SQI-MOS score (i.e. at zero FER), while the rock-bottom SQI-MOS (the worst possible score, attained at very high FERs[9]) was left unchanged.

[7] This is a phenomenon independent of the circumstances described in chapter 4.

[8] This is explicitly stated in [2]. Further comment on ITU Recommendation P.862.2 and on the difficulty of applying a uniform speech quality measurement model to both narrowband and wideband is found in [3].

[9] FER = 60% was selected as endpoint. Samples with FER > 60% were excluded from the SQI-MOS modeling, since PESQ (as is well known) sometimes judges severely disturbed speech in a misleading manner: certain very bad (almost muted) samples receive high PESQ scores.

6. Comparison with Other Radio Parameters

GSM

In the past, speech quality in GSM networks was often measured by means of the RxQual parameter (which is also available in TEMS products). Since RxQual is merely a mapping of time-averaged bit error rates into a scale from 0 to 7 (see 3GPP TS 45.008, section 8.2.4), it cannot of course provide more than a rough indication of speech quality.

7. References

[1] 3GPP TR 26.975, “Quality in Clean Speech and Error Conditions”, version 7.0.0.

[2] “Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs”, ITU document number “P.862.2 (11/2007)”.

[3] “Report of the meeting of Working Party 2/12 (Geneva, 2 - 10 October 2007)”, ITU document number “COM12 - R19 - E”.