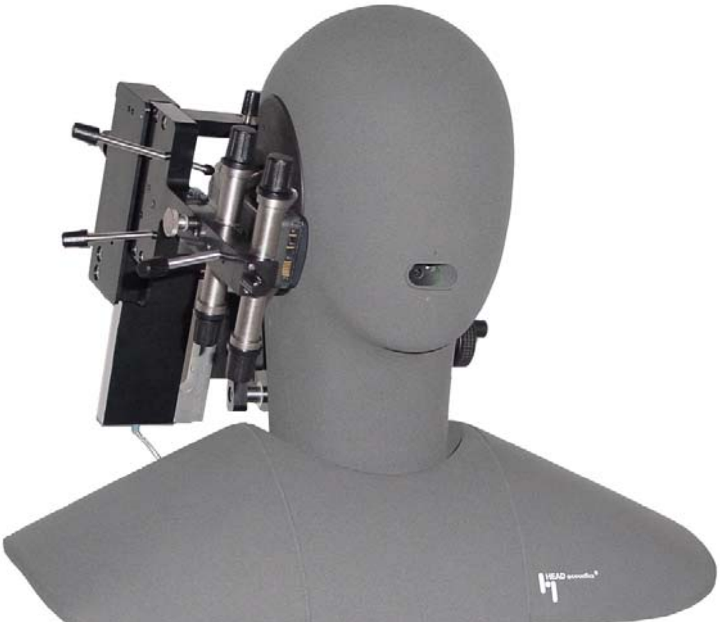
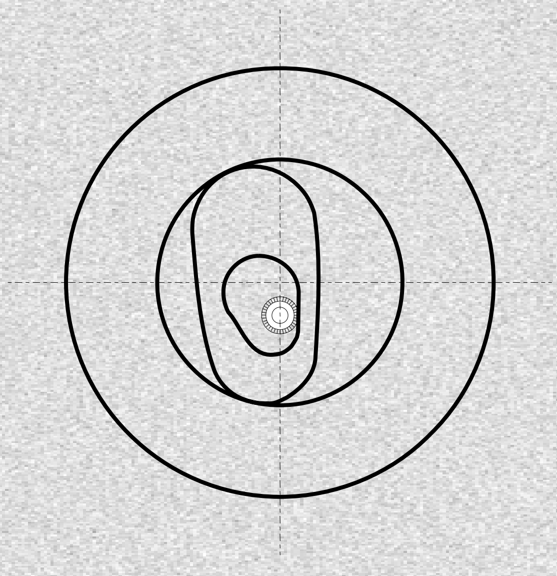
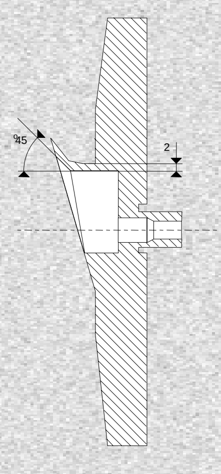
# Option
# TOSQA

## TELECOMMUNICATIONS OBJECTIVE SPEECH QUALITY ASSESSMENT







**HEAD acoustics**
**Application Note**

HEAD acoustics GmbH
Ebertstraße 30a
D-52134 Herzogenrath
Tel.: +49 (0) 2407-577-0
Fax: +49 (0) 2407-577-99
E-mail: sales@head-acoustics.de
WEB: www.head.acoustics.de

# Option TOSQA

## Description of the objective speech quality measurement methods TOSQA and TOSQA2001

### Introduction

Like in other current methods for predicting speech quality TOSQA and TOSQA2001 are used for modeling the results of listening tests, i.e. they provide the result of a listening test that could be expected if the recorded speech signal was subjectively rated by test persons.

### TOSQA and TOSQA2001
### - the principal difference -

In analogy to other known methods such as PSQM according to ITU-T Recommendation P.861 [7] or PESQ according to ITU-T Recommendation P.862 [8] the TOSQA method can be applied to scenarios, where a speech signal is transmitted via a measurement object and recorded at an electrical interface. The assessment of an acoustical recording via a terminal is not possible, because in these methods a terminal is already fully integrated, thus being part of the method itself. This means that terminals cannot be included in the quality assessment.

By contrast, TOSQA2001 allows this assessment of a terminal in the transmission path and the acoustical recording of the transmitted signals as an additional option. While TOSQA, like PSQM [7] or PESQ [8], is limited to electrical recordings, TOSQA2001 allows both: the quality assessment of signals recorded electrically as well as acoustically (via terminals).

The fundamentals of both methods TOSQA and TOSQA2001 are described in more detail below.

### Results

TOSQA as well as TOSQA2001 both provide the following result parameters:

- **TOSQA value**
- **TMOS value**
- **Equipment Impairment Factor $I_E$**
- **Delay**

The result which is probably most important for practical use is the so-called **TOSQA value**, which describes the quality of the transmitted speech signal. From the maximum TOSQA value of 100, subtractions are made depending on the degree of impairment introduced to the transmitted signal. A high speech quality typically results in a value of approx. 90. TOSQA values lower than 50 correspond to an unacceptable speech quality.

From the TOSQA value, a so-called **TMOS value** (***T***OSQA ***M***ean ***O***pinion ***S***core) is derived. This is based on the following: In a listening test the test persons involved give their rating after listening to each speech sample. Typically a 5 point scale according to ITU-T Recommendation P.800 [9] with the following category and point assignment is used:

- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: bad

Arithmetic averaging of the ratings of all test persons for one speech sample results in an MOS value (**M**ean **O**pinion **S**core) for this speech sample. An MOS value of approx. 4.2 typically corresponds to the maximum possible quality via telephone band limited transmission systems. Interestingly, an MOS value of 5.0 is never achieved in a listening test. On the one hand, this is due to the telephone band limited signals, on the other hand it is caused by the “restraint“ of the test persons in giving the best possible rating of 5 points.
TOSQA estimates this value for the analyzed speech signal as TMOS value, representing the quality value to be expected for the analyzed speech sample in a listening test. Therefore this value is also called **estimated MOS value**.

As a further quality parameter TOSQA and TOSQA2001 provide the **estimated Equipment Impairment Factor $I_E$**. This value $I_E$ is a parameter for further processing in the so-called "E-Model" according to ITU-T Recommendation G.107 [2]. The E-Model serves for planning and configuring networks. Impairment of speech quality by non-linear and/or time-variant degradations, such as e.g. speech coding, is taken into account in the E-Model by the so-called Equipment Impairment Factor $I_E$. Under the assumption that all other values (e.g. loudness rating, delay, echo attenuation etc.) of a transmission system are known, the degradation introduced by coding, packet loss etc. can be accounted for by the corresponding $I_E$ value. Based on the estimated impairment of speech quality, TOSQA provides an $I_E$ value which can be used in the E-Model.

Note: During an acoustical recording TOSQA2001 also provides an Equipment Impairment Factor $I_E$. However, this can currently not be used in the E-model, because the E-model only considers the influence of the terminal based on loudness ratings. Additional impairments, in particular the influence of transfer functions on speech quality, are not yet implemented in the E-model.

By comparison of the transmitted signal and the fed-in measurement signal, the algorithms TOSQA and TOSQA2001 furthermore determine the average delay of the measurement object during the signal transmission. The maximum delay which can be measured is 1 s. The value displayed corresponds to the sum of the fixed delay and the average value of any time-variable delays that may be present. Moreover, TOSQA itself requires this delay for adaptation of the transmitted signal to the reference signal (see below for more).

## Principles of the methods

Generally speaking, the working methods of TOSQA and TOSQA2001 are similar: both methods are based on a speech signal which is *fed in as measurement signal* at one point of a transmission path, transmitted via the path (e.g. the network) and *recorded* at another point as *transmitted signal*. The recorded signal and the fed-in signal (also called reference signal) are compared and from this comparison the quality values (TOSQA value, TMOS value, Equipment Impairment Factor $I_E$) are calculated.

If the recorded signal shows a good correlation with the fed-in signal (i.e. it was only slightly “falsified”, distorted or otherwise impaired during the transmission), high TOSQA values and correspondingly high TMOS values are achieved. If the signal was severely imapired by the transmission path, i.e. strongly modified compared to the fed-in signal, the comparison of the transmitted signal to the fed-in signal obviously shows strong differences and the algorithm calculates a loss of quality, represented by lower TOSQA and TMOS values.

Figure 1 shows the basic principle of processing as realized in TOSQA and TOSQA2001.

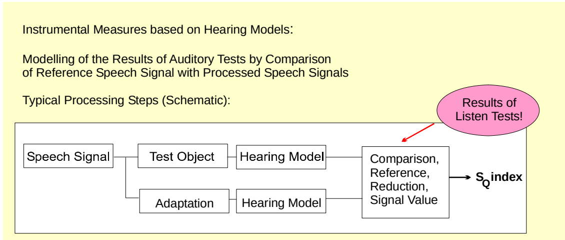



Fig. 1: Principle of models for objective speech quality prediction

The speech signal is transmitted via the test object and analyzed in a hearing model. The same is done with the reference signal. For comparison to the transmitted signal, it has to be adapted with regard to delay and level. The delay compensation is extremely important, because the comparative analysis of both signals in the hearing model is done in very short time intervals which have to be exactly the same in the transmitted signal and in the original signal. This delay compensation is time-variable and thus also takes account of delays in the transmission path which change over time (“jitter“).

Furthermore, TOSQA (like other methods) requires that during the recording at an electrical interface – i.e. without terminal – linear distortions in the transmission channel (frequency response) are not so severe that the quality of the transmitted speech is impaired, so that they can be at least partly compensated. In the signal processing block “adaptation“ the delays are therefore matched and the transfer function of the test object is modeled.

TOSQA2001 provides a more sophisticated evaluation: as TOSQA2001 can take account of terminals, which have a considerable influence on the transmission quality due to their typical transmission characteristics, the compensation is more detailed and aurally-adequate. In addition, TOSQA2001 is no longer limited to narrowband systems but can be applied to wideband systems as well.

The resulting adapted signal in Figure 1 is passed through a hearing model, which is explained in more detail further below (cf. **Processing steps of the algorithms**). From this hearing model the most important feature vectors for both signals are extracted and monitored for similarities. The more similar the feature vectors, the higher the probability that the transmitted speech signal is only slightly impaired. The various error vectors are evaluated and reduced to a numerical value. This single numerical value is called TOSQA value and represents the quality rating.

The subjectively perceived (auditory) speech quality, i.e. the estimated MOS value in a listening test, is assigned by transforming the TOSQA value, so that an estimated MOS value - the TMOS value - can be derived from the internally generated TOSQA speech quality value. This transformation as well as the previous calculation of the TOSQA value itself requires the auditory validation for all impairments caused by the test object. This is basically true for all instrumental measures. Here it also becomes clear that these measures can only be used for the speech quality rating of systems which the method has been auditorily validated for. Like any other method for objective speech quality assessment, TOSQA always requires a new validation for new impairment factors (new speech codecs, channel disturbances etc.).

## Background information for the practical use of TOSQA and TOSQA2001

### Feeding-in of the test signal

Both methods, TOSQA and TOSQA2001, use real speech as test signal. The following nomenclature is applied:

An ***original speech signal*** consists of sentences spoken by test persons under controlled acoustical conditions. These recordings are typically made in an unechoic environment with a measurement microphone.

The ***signal which is fed in*** during a TOSQA measurement is also called ***measurement signal*** or ***reference signal***, because it is used as such by the analysis system. It can either be the original speech signal, e.g. if it is played back by an artificial mouth and fed into the transmission path (e.g. network) via a telephone. It can also be a processed (e.g. pre-filtered) original signal. An application example is the electrical feeding of the measurement signal into the transmission path: in reality, such a signal has already been transmitted via a terminal, otherwise it could not have reached this point in the network. For this scenario, “the terminal which is not present in the transmission path” is considered by an IRS filtering (IRS in sending direction according to ITU-T Recommendation P.48 [4]).

Conclusion: for the electrical or acoustical feeding of measurement signals different test signals are used: the original speech signal, when an artificial mouth is used and a terminal is present as part of the test setup, or a pre-filtered signal (IRS filtering), when feeding in at an electrical entry point. This applies both to TOSQA and TOSQA2001.

The choice of the measurement signal is automatically controlled via the ACQUA measurement descriptor, so one only has to make sure that the correct measurement descriptor is started.

**Recording of the test signal**

The ***recorded signal***, also called ***transmitted signal***, is the signal that is impaired by the transmission.

Like other psycho-acoustically motivated methods (PSQM [7] or PESQ [8]), TOSQA and TOSQA2001 calculate the quality perception in the listening situation. The algorithms with the implemented hearing model are therefore based on a signal which is applied to the “ear of the assessing test person“ (“ear signal“) and compare it to the undisturbed signal. However, the undisturbed signal is also assumed to be applied “to the ear of the assessing person“.

The algorithm thus requires the transmitted signal as an input value and for comparison the fed-in measurement signal as second input value. Both signals are compared to each other as described above. Since it refers to the assessment of a telephone situation, the algorithm consequently requires both the fed-in signal and the recorded signal as ear signals.

This results in two measurement technology possibilities:

If an electrical interface is available for recording, the recorded signal is not an "ear signal". It has not yet been transmitted via a terminal in receiving direction.

In this case the algorithm modifies the signal by adding the typical characteristics of a "terminal in receiving direction at the human ear" via preprocessing blocks. The algorithm thus simulates the terminal by internal processing blocks. These recordings can be rated with TOSQA as well as TOSQA2001: While TOSQA implicitly assumes an electrical recording, TOSQA2001 allows the explicit choice (**Measurement: electrical**). Nevertheless small differences in the results may occur when analyzing with TOSQA and TOSQA2001, because the processing cores of both algorithms differ slightly: the adaptation of linear distortions and level variations over time is optimized with regard to aural adequacy in TOSQA2001 as opposed to TOSQA.

However, if the recorded signal already is an acoustical recording at a terminal by means of an artificial head measurement system, this recording itself already represents an ear signal.

Important: Such a recording can only be analyzed with TOSQA2001. Here, the algorithm is based on a free-field equalized artificial head measurement system, which has to be calibrated so that a sound pressure level of 94 $dB_{SPL}$ corresponds to an output voltage of 0 $dB_V$. TOSQA2001 was validated for recordings depending on pressure force by using an artificial ear Type 3.4 according to ITU-T Recommendation P.57 [5] with an artificial head measurement system as specified in ITU-T Recommendation P.58 [10].

TOSQA2001 can thus be used both for electrical and for these acoustical measurements. This is possible because in the TOSQA2001 algorithm the processing blocks, which filter the recorded signal with the characteristics of a terminal at the human ear, can be deactivated. In the case of an acoustical recording this would result in the error of assessing the signal "through two terminals and two ears": one which was already present physically during the acoustical recording, i.e. as part of the transmission path, and the other one which is simulated by the processing blocks in the algorithm.

Furthermore, the free-field equalization used with the artificial head measurement system is corrected to the ERP (Ear Reference Point).

Conclusion: TOSQA2001 allows the selection of the options "Acoustical Measurement" or "Electrical Measurement" at the press of a button. The version TOSQA, but also other methods like PSQM or PESQ, are not validated for acoustical recordings and must not be used for these ratings.

### “Listening test setup“ when using TOSQA2001

TOSQA and TOSQA2001 simulate the results of a listening test. In a “real“ listening test with test persons listening samples are assessed. Typically the assessment of a listening sample also depends on the playback of all samples as a whole. Although this kind of test has to ensure a balanced collection of all listening samples, a qualitatively rather average listening sample may achieve a relatively good rating, if the samples as a whole are of poor quality. By contrast, if the overall quality of the speech samples is good, the same listening sample will get a poorer rating.

When using the TOSQA method, this listening test setup is fixed.

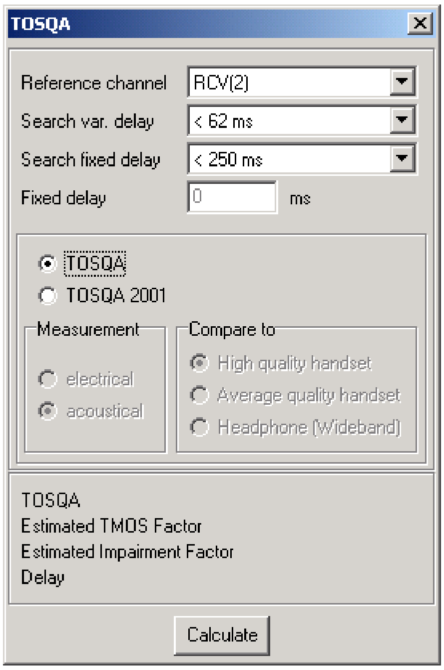



Fig. 2: Typical choice of parameters in ACQUA when using TOSQA. In this example the entry **Reference channel: RCV(2)** in the upper area determines the measurement signal, i.e. the reference, to be on channel 2. The transmitted signal in this case has to be on the other channel **SND(1)**. The parameter **Search var. delay: < 62 ms** determines the maximum expectable time variation of delay during the measurement (in this example 62 ms). The numerical value under **Search fixed delay**: (example: < 250 ms) determines the maximum expectable constant delay of the measured transmission path in order to adjust the reference signal, i.e. time-synchronize it. This value should be somewhat higher than the constant delay.

TOSQA is always based on an electrically recorded signal, the selection frame “**Measurement**“ is therefore deactivated. The same is true for the selection frame “**Compare to**“ (reference situation), as a listening situation with a high-quality conventional handset is always assumed.

The algorithm simulates the result of a listening test, where the reference situation comprises a high-quality conventional handset with dynamic ear capsule, high pressure force (small acoustical leakage) and one ear (monaural). The speech signals are thus perceived and rated by test persons with a high-quality handset. It is assumed that the speech signal is perceived with a level of 79 $dB_{SPL}(A)$ at the ERP (Ear Reference Point) and the speech quality value is calculated accordingly. There are no further selection possibilities for the analysis.

By contrast, TOSQA2001 allows to influence the listening test setup in a limited way by the choice of different parameters. This affects the results in a limited way, with the following background:

1. In the case of an **electrical recording** rated via TOSQA2001 there are two different selection possibilities, each representing one type of listening test:
In one scenario it is assumed that the recorded signal is perceived by the test persons via a high-quality conventional handset with dynamic capsule, one ear (monaural) and high pressure force as in a typical telephone situation. The terminal is modeled according to this situation. Again, it is assumed that the speech signal is perceived with a level of 79 $dB_{SPL}(A)$ at the ERP and the speech quality value is calculated accordingly.

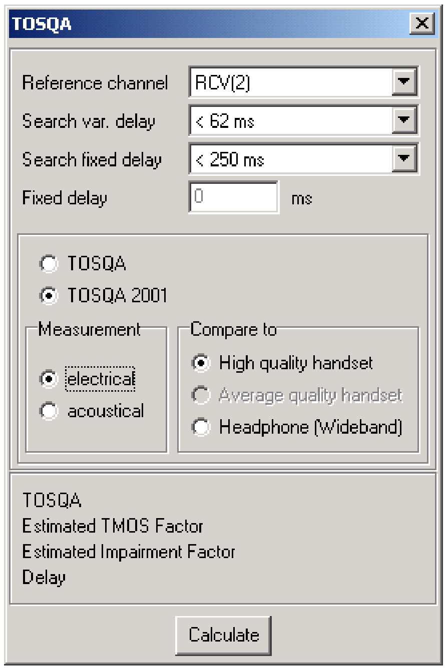



Fig. 3: Choice of parameters in ACQUA when using TOSQA2001 with electrical recording. TOSQA2001 assumes an electrically recorded signal when the option "**Measurement: electrical**" is selected. In the selection frame "**Compare to**" the parameter "**High quality handset**" selects the listening situation of a high-quality conventional handset, while "**Headphone**" selects the listening situation via headphones (diotic) with signals transmitted via wideband.

Optionally, the TOSQA and TMOS values can be calculated if the recorded speech sample is perceived in a listening test via diotic headphones, i.e. identical signals at both ears. This form of analysis in TOSQA2001 is suited for signals recorded via wideband. It is assumed that the signal is perceived with a level of 73 $dB_{SPL}(A)$ at the ERP and the speech quality value is calculated accordingly. This level of 73 $dB_{SPL}(A)$, which is lower than in a monaural situation, is selected because listening occurs via both ears.

2. In the case of an **acoustical recording** of the transmitted signal and analysis by TOSQA2001 three possibilities arise:

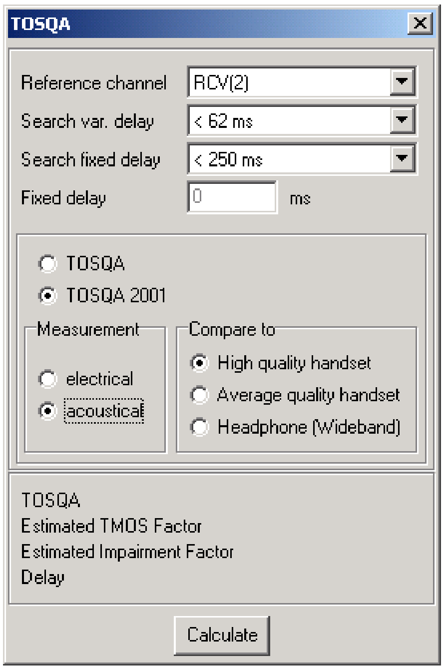



Abb. 4: Choice of parameters in ACQUA when using TOSQA2001 with acoustical recording. TOSQA2001 assumes an acoustically recorded signal when the option "**Measurement: acoustical**" is selected. For the reference - situation ("**Compare to**") three selection possibilities are offered.

All three listening test scenarios assume, that the speech signal was recorded with an artificial head measurement system (HATS, **H**ead **A**nd **T**orso **S**imulator according to ITU-T Recommendation P.58 [10]) with free-field equalization. This has to be ensured during recording. Furthermore, the selection of **High quality handset** or **Average quality handset**, as mentioned above, assumes that the speech sample is monaurally perceived with 79 $dB_{SPL}$(A) in the listening test.

When selecting "**Compare to: High quality handset**" the recorded speech signal is assessed in comparison to a reference situation where the reference is formed by the undisturbed speech signal, which is perceived via a conventional handset with dynamic ear capsule. In this case, there is a high quality reference situation. If the artificial head recording was made with an average quality leakage-sensitive handset, typically with badly coupled low frequencies, the recording gets a poorer rating in this situation.

In a simplified way, this can be explained as follows: The recorded speech material is compared and rated in a listening test, where all other speech samples (which are also assessed in this "fictional" listening test) have been recorded via a high quality handset. If the own speech recording is a recording with an average quality terminal, the rating will be rather poor compared to the other reference samples.

However, if a recording situation with an average quality handset is chosen as comparative situation for the own speech material by selecting "**Compare to: Average quality handset**", a leakage-sensitive handset is assumed for the quality comparison. This situation simulates a listening test where the quality of the recorded speech material is rated by assessing all speech samples via an average handset. In this example, if the own speech material again is a recording via an average quality terminal, it will get a better rating in comparison to the other reference samples.

Recordings made for example via leakage-sensitive handsets should be compared to a corresponding reference situation. In this case the choice of the parameter "**Compare to: Average quality handset**" leads to a better validity of the results.

Of course, TOSQA2001 also allows the **assessment of hands-free terminals** in receiving direction, if these recordings were also made with artificial head measurement systems (free-field equalization, calibration 94 $dB_{SPL} \cong 0$ $dB_{V}$). Up to now, neither TOSQA2001 nor any other method takes account of the binaural signal processing of the human hearing. TOSQA2001 therefore only analyzes one ear signal. TOSQA2001 again compares these recordings with a "standard telephone situation", i.e. telephoning with a handset as reference. The same applies as described above: When selecting "**Compare to: High quality handset**" the artificial head recording of the hands-free terminal is referenced to comparable speech recordings via a high quality handset. Selecting "**Compare to: Average quality handset**" compares the recording to speech recordings via an average quality handset.

As a third option, "**Compare to: Headphone (wideband)**" can be selected as comparative situation for speech transmitted via wideband. Here the recorded signal is rated in comparison to a reference situation where the test persons perceive the speech samples via headphones, assuming diotical perception, i.e. identical signals on both ears.

When using **headsets** for the acoustical recording of telephone band-limited methods, however, it is recommended to select "**High quality handset**" as reference situation.

**Processing steps of the algorithms**

Figure 5 shows a detailed block diagram of the TOSQA algorithm.

The preprocessing block described above is shown here: At first, delay and level are compensated (block-by-block) so that during later analysis ear signal and transmitted signal are available at the same time and at the correct level. Attention: The level for weak signals at the output of the test object can only be compensated in a limited way: If the signals are too weak, a compensation is no longer possible, because the disturbing noise would severely falsify the calculation otherwise.

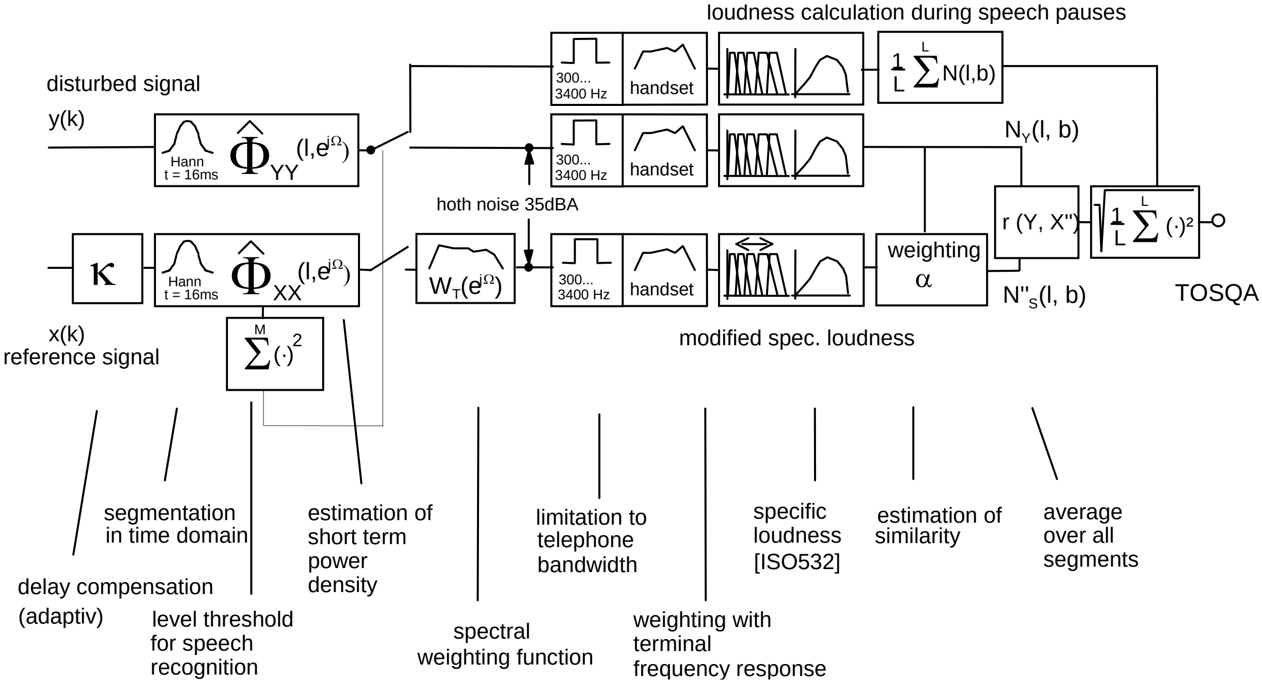



Fig. 5: TOSQA Processing Steps

The next step is block-by-block transformation to the frequency domain. An energy threshold for voice activity detection prevents the inclusion of speech pauses in the speech quality rating. Subsequently, reference signal and transmitted signal are limited to the telephone band and weighted with the power transfer function of a typical handset (similar to ITU-T IRS frequency response). TOSQA and TOSQA2001 differ in these processing blocks as described above.

The subsequent frequency group assignment is dynamic and takes account of frequency shifts that might occur in the transmission system, e.g. by coding. The next step is transformation to the loudness scale. Thus, the sensitivity of human hearing is sufficiently taken into account.

The comparison of feature vectors is achieved by determining the Bark-spectral similarity in each block. The similarity is calculated separately for the lower and upper frequency groups. These Bark-spectral similarities are then averaged for all blocks.

$$BSA = \sqrt{\frac{1}{L}\sum_{l=1}^{L} r_l\left[N_y(l,b), N_s(l,b)\right]}$$

The scale transformation of this average similarity value to a Mean Opinion Score (MOS) is based on the results of extensive listening tests performed by Deutsche Telekom, ETSI and ITU-T.

Figures 6 and 7 show the results of typical speech quality predictions in comparison to the results of listening tests for different test situations using TOSQA. In figure 8 comparative results with TOSQA2001 for acoustical as well as electrical recordings are displayed.

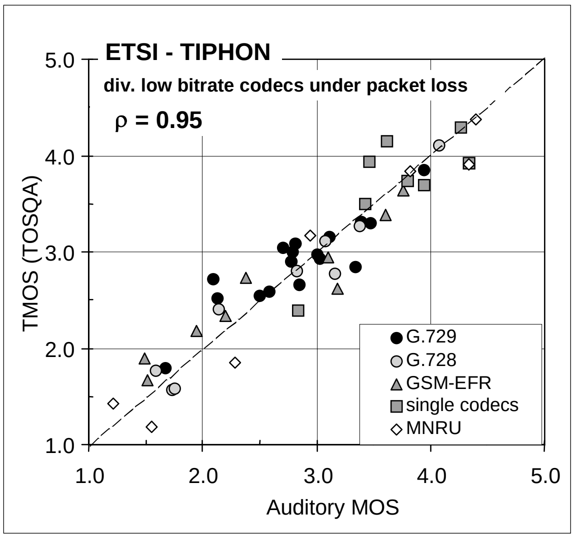



Fig. 6: Different Speech Codecs with Packet Loss in IP Simulations

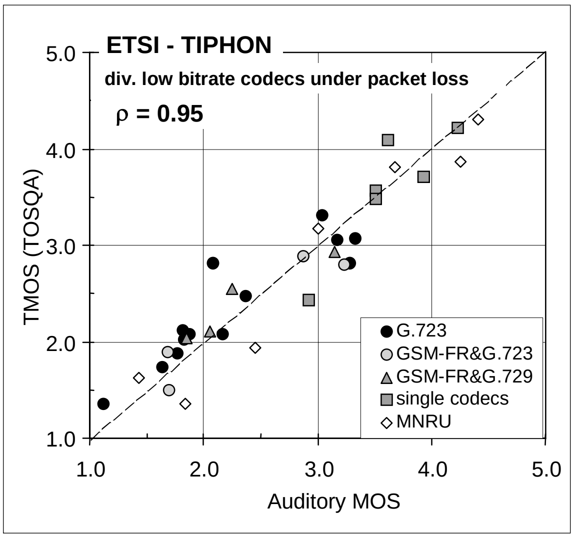



Fig. 7: Different Speech Codecs with Packet Loss in IP Simulations

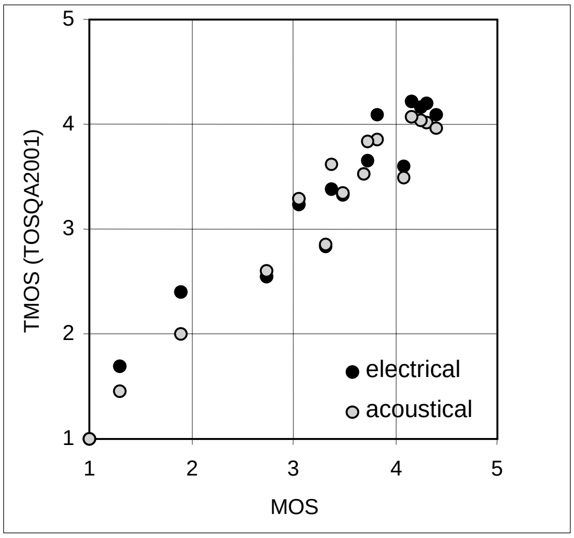



Fig. 8: Comparative results with TOSQA2001 for acoustical and electrical recording [6]

## Application of TOSQA

As with other objective speech quality measures, the application of TOSQA and TOSQA2001 is limited to a purely auditory situation, i.e. only one aspect of speech quality is estimated, which essentially is the “speech transmission quality“ or “speech sound quality“. Currently, no method adequately takes account of problems which affect speech quality in conversational or double talk situations.

The application of both methods is limited to impairments which they have been validated for. These are all currently known ITU-T codecs, typical impairments by jitter and packet loss - as occurring in IP networks with H.323 codecs - as well as, to a certain extent, impairments by clipping.

Evaluation of signals with poor S/N ratio (< 20 dB) is generally problematic.

**TOSQA2001 and MP3 speech files**

TOSQA is also suited for the evaluation of speech recordings in MP3 format. The classical TOSQA serves to evaluate telephone speech, i.e. when choosing the narrowband mode the signal is evaluated as if it was heard via a handset. This is evidently not the case with MP3. However, the extended TOSQA2001 also offers a wideband mode. Here the reference is a wideband signal, played back via a headphone. TOSQA2001 thus can also be applied to MP3 speech files using a wideband (not IRS(send)-evaluated) speech sample as input sample for the coder.

In this regard TOSQA2001 was validated for wideband codecs (e.g. CELP, AAC (Advanced Audio Coding, corresponds to MPEG2) with formal listening tests. In an expert's listening test TOSQA2001 also showed satisfactory results for MP3.

It has to be noted that partial narrowing in the frequency band occurs, especially at the lower bit rates which are of interest for speech coding. At 16kbps and less the transmission frequency band is already limited to 4kHz. Obviously this considerably affects the perceived quality particularly in comparison to wideband speech playback.

**Other validations**

TOSQA is primarily validated for source coding methods (also typical speech coders). This applies to all PCM and ADPCM methods, CELP, RELP and related methods, also in connection with packet loss and PLC. The use of TOSQA is not recommended for ATC (Adaptive Transformation Coding), because the results are too optimistic, especially for low bit rates. Anyway, ATC is hardly used anymore.

Up to now the experiences with MP3 also show a good suitability for audio codecs, if they are used for the transmission of speech.

In addition TOSQA is also validated for GSM network disturbances, clippings (amplitudes and temporal), weak background noise (or background noise in an otherwise noise-free context) and high pass behavior (coupling of terminal to the ear).

TOSQA is currently validated for the use of noise reduction methods, especially for the noise-free case, because this often causes the generation of artifacts.

**Limitations**

TOSQA is not suited for Receiver Distortion measurements. TOSQA cannot be used or can only be used with limitations, if Automatic Gain Control is employed.

## Literature


[1] Berger, J., **Instrumentelle Verfahren zur Sprachqualitätsschätzung - Modelle auditiver Tests**, Ph.D. thesis, Kiel, 1998, ISBN 3-8265-4091-3

[2] ITU-T Recommendation G.107, G Series Recommendation, **The E-Model, a computational model for use in transmission planning**, Geneva, 12/1998

[3] ITU-T Recommendation P.800, P Series Recommendation, **Methods for subjective determination of transmission quality**, Geneva, 8/1996

[4] ITU-T Recommendation P.48, P Series Recommendation, **Specification of an Intermediate Reference System**, Geneva, 8/1996

[5] ITU-T Recommendation P.57, P Series Recommendation, **Artificial Ears**, Geneva, 8/1996

[6] Berger, J., **Results of objective speech quality assessment including receiving terminals using the advanced TOSQA2001**, ITU-T Contribution, Dec. 2000, COM 12-20-E

[7] ITU-T Recommendation P.861, P Series Recommendation, **Objective quality measurement of telephone-band (300-3400 Hz) speech codecs**, Geneva 02/1998

[8] ITU-T Recommendation P.862, P Series Recommendation, **Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs**, Geneva 02/2001

[9] ITU-T Recommendation P.800, P Series Recommendation, **Methods for subjective determination of transmission quality**, Geneva 08/1996

[10] ITU-T Recommendation P.58, P Series Recommendation, **Head and torso simulator for telephonometry**, Geneva 08/1996