3QUEST: Applications & Use in Practice



HEAD acoustics Application Note









3QUEST: Applications & Use in Practice

Contents

1.	INTR	RODUCTION	4
2.	TEST	SETUPS	5
	2.1. 2.2. 2.3.	Used Equipment	5 6 7
3.	RESU	JLTS OF THE TESTS	8
	3.1. 3.2. 3.3. 3.4. 3.5. 3.6.	TMOS SND CHARACTERISTICS / TEST SERIES 1 D-VALUE TESTS / TEST SERIES 1	8 9 11 13 15
4.	SUM	1MARY 1	9
5.	REFE	RENCES	20
6.	ANN	NEX / GRAPHS	21



1. Introduction

Nowadays, the design of telecommunication devices has to cover the fact that their typical case of use is in the presence of background noise. This is especially valid for e.g. mobiles or car hands-free kits. Thus, the necessity of corresponding and appropriate test scenarios to ensure sufficient speech quality is widely accepted.

A classic parameter for tests in the presence of background noise is the so-called *D*-Value according to ITU-T G.111 [1]. A D-Value is calculated as a one-dimensional score, comparing the sensitivity of transmitted speech towards the sensitivity of transmitted background noise. Tests of D-Value are executed basically in three steps: First, speech sensitivity is tested (as in SLR tests). Subsequently, noise sensitivity (without the presence of speech) is measured. Finally, both sensitivities are compared. Due to the fact that sensitivities are tested, the result of the noise sensitivity test is not directly related to the level of the used background noise: In case a linear system is tested, any higher background noise will linearly result in a higher level of transmitted noise, so the sensitivity would be constant. It should be noted that in practice, this is typically never the case due to the fact that modern mobiles do not react linearly on background noise.

The description of how the D-Value is tested shows two main problems: First of all, no test of speech quality *in the presence* of background noise is done. There is no test of *simultaneous* speech and background noise. This also means that any "smarter" noise reduction algorithm will not be represented adequately due to the fact that the algorithm will not be active as long as speech is transmitted (first test step) and will not be affected by any speech transmission because of the isolated noise transmission (second test step). Another disadvantage is that the D-Value test only compares transmitted *levels* of speech and BGN. There is no statement regarding speech quality.

One of the best known parameters for speech quality is the $PESQ^{TM}$ score according to ITU-T P.862 [7]: Although usually developed for tests of electrical connections without presence of any terminals (like e.g. codec testing), the $PESQ^{TM}$ algorithm allows environmental noise as a test factor. However, the suitability of $PESQ^{TM}$ for terminal testing is limited: According to P.862, there is neither a validation for acoustic terminal testing, nor for any effects of echo canceller and noise reduction algorithms. Nevertheless, it should be noted that $PESQ^{TM}$ is sometimes broached from operators and manufacturers for device tests and occasionally used in practice – even without any proved statement of its reliability.

This application note describes the use of D-Values and PESQTM scores for different kinds of mobile phones under different background noises in contrast to the new analysis method 3QUEST (<u>3</u>-fold <u>Quality Evaluation of Speech in Telecommunications</u>) based on ETSI EG 202 396-3 [3]. 3QUEST is explicitly intended to be used for analyzing speech in the presence of background noise. The mobile phones have been selected to cover a wide range of today's market spectrum.



2. Test Setups

All tests described below were made with different mobile phones. Mobile phones were used since they show a big variety of different designs and almost all of them use integrated noise cancellation techniques. Thus, both the influence of the acoustical/mechanical design as well as the influence of noise cancellation techniques could be evaluated.

2.1. Used Equipment

All tests were conducted using a HATS conforming to ITU-T Recommendation P.58 [5] equipped with an artificial ear type 3.3 (ITU-T Rec. P.57 [4]) for the test series and ear type 3.4 for the room equalization. The phones were positioned as described in ITU-T Recommendation P.64 [6].

The D-Value tests were conducted as described in TS 51.010-1 [8]. The 3QUEST tests were conducted with an identical hardware setup and as described by EG 202 396-3. All tests were carried out in 2G (GSM 900) mode. Different types of background noises were used as described in detail.

As background noise playback system, the HEAD acoustics HAE-BGN system (<u>H</u>EAD <u>A</u>utomated <u>E</u>qualization for <u>B</u>ackground <u>N</u>oise), mainly consisting of HAE-BGN software, programmable equalizers PEQ V, speakers and subwoofer was used. Figure 1 shows the system setup.







The background noise simulation was equalized and calibrated as described in EG 202 396-1 [2]. Instead of time aligning the 4 loudspeakers in each path of the playback arrangement, a certain delay was introduced in order to increase the diffusivity of the binaurally recorded background noise from the database of EG 202 396-1 and to avoid standing waves in the rooms. Empirical tests have shown that these delays can be used for typical anechoic rooms in telecommunication scenarios.

The delays inserted in each path are shown in the following table:

	Left	Right	
Front	16 ms	51 ms	
Rear	7 ms	21 ms	

E:	ام ماسم ما ما	- - · · -	f	
rigure z:	inserieu	delays	lor eq	Udiizaiion

After inserting the delays, the equalization and calibration procedure as described in EG 202 396-1 was applied.

All measurements and analyses were carried out with the HEAD acoustics ACQUA system (<u>A</u>d-vanced <u>C</u>ommunication <u>Qu</u>ality <u>A</u>nalysis).

2.2. 3QUEST Test Setup

A test setup for 3QUEST measurements is first of all based on the standard setup as described in the previous chapter. It includes a HATS and Handset Positioner in the noise simulation environment. In addition, 3QUEST tests require a small setup extension: A reference microphone must be placed close to the microphone of the device under test for capturing the recorded speech and background noise as the so-called *unprocessed signal* (according to EG 202 396-3).

In summary, the 3QUEST algorithm requires three signals:

- Unprocessed signal, recorded from the reference microphone;
- **Processed signal**, i.e. the transmitted signal which represents the usual recording in sending direction, including all signal processing;
- Clean speech signal, delivered as source signal by ACQUA (no additional measurement necessary).

Further background information can be found in ETSI EG 202 396-3 as well as in the HEAD acoustics 3QUEST Application Note [10].



2.3. Devices under Test

For the two different test series, different types of phones from different manufacturers were used.

Test series 1

- "Flip Phone 1": Flip Phone with Noise Cancellation (NC)
- "Flip Phone 2": Flip Phone with NC (different manufacturer than "Flip Phone 1")
- "Short Phone": Short Phone with moderate NC, older type
- "Double Mic 1": Short Phone with NC as well as a double microphone solution
- "Double Mic 2": "Business Phone" with NC as well as a double microphone solution

Attention should be drawn to the last two phones: These types are equipped with a new-technology double microphone solution. An additional microphone mounted at the back side of the phone is especially used for noise cancellation. It can be assumed that such type of phones will react most strongly to different types of background noise scenario.

Test series 2

- "Device 1": Larger "Smartphone" (incl. small Keyboard) with a double microphone solution
- "Device 2": "Business Phone" with a double microphone solution
- "Device 3": Small size "Smartphone" (PDA functions)
- "Device 4": Small "Business Phone" (Slider)
- "Device 5": Standard "Business Phone", average size
- "Device 6": "Smartphone" (PDA functions) classic style
- "Device 7": Standard "Business Phone", larger size
- "Device 8": "Smartphone", average size

All phones make use of a noise cancellation capability.



3. Results of the Tests

All test results are presented in detail in the following subchapters. For test series 2, all measured D-Value and 3QUEST results which are the basis for the different result overviews shown are found in Chapter 6 "Annex / Graphs".

3.1. TMOS SND Characteristics / Test series 1

The figure below shows the TMOS SND comparison of all devices under test. The red dashed line shows the minimum TMOS SND score of MOS 3.2, this limit is taken from typical prospects of network operators. A mobile should reach this score at least, so the corresponding bar plot should be higher than the red dashed line of figure 3.



Figure 3: TMOS SND measured for the different phones

Test series 1: Result summary of TMOS SND tests

The TMOS SND test shows that all five phones provide sufficient speech quality in sending direction in case no background noise is present. Thus, all five devices shall be seen as representative for the following D-Value and 3QUEST tests.



3.2. D-Value Tests / Test series 1

In this experiment series, the D-Value tests according to ITU-T G.111 have been conducted with further different background noises. Therefore, besides tests in the presence of Pink Noise, the following six different binaural background noises of the ETSI background noise database [2] have been in use:

- **"Pink Noise"**: Classic pink noise sound field of -24 dBPa(A), equivalent to 70 dB_{SPL}(A). As per HAE-BGN, the noise has been equalized with a single microphone in the absence of a head and torso simulator.
- **"Fullsizecar100"**: ETSI "Fullsize_Car1_100Km/h_binaural" signal, recorded at driver's position, average level approx. 64 dB(A). Very constant level signal, no speech content.
- **"Insidebus"**: ETSI "Inside_Bus_Noise_binaural" recording in passenger cabin, average level approx. 77 dB(A). This signal is the loudest one of the selection, it has a lower speech content.
- "Insidetrain2": ETSI "Inside_Train_Noise2_binaural" signal, average level approx. 63 dB(A). Low speech content, varying signal level vs. time.
- **"Pubnoise"**: ETSI "Pub_Noise_binaural_V2", recording in a pub, average level approx. 74 dB(A). The signal is rather loud and with high speech content.
- **"Cafeterianoise"**: ETSI "Cafeteria_Noise_binaural", recording at the sales counter of a cafeteria. Average level approx. 68 dB(A), ambient speech content mixed with ambient noises.
- **"Kindergarten1"**: ETSI "Kindergarten_Noise1_binaural", recorded in a children's playroom of a kindergarten. Average level approx. 73 dB(A), varying signal level vs. time. Almost pure ambient speech content, very little other ambient noises.

All tests have been executed with *synchronized* playback: Signal analysis has always been 2 s after the first repetition of the background noise file. This procedure is seen as essential to create sufficient reproducibility when testing with such non-linear noise signals. In practice, triggering was realized with trigger pulses that are synchronized via the source signals played back by ACQUA. The trigger pulses are transmitted to HAE-BGN by a HEAD acoustics Pulse Splitter Box. This method allows synchronization accuracy in the range of a few milliseconds. D-Value signal analysis length has always been 10 s.

Figures 4 & 5 show the D-Value results of the different phones captured with different background noises.





Figure 4: D-Value of different phones with Cafeteria Noise & Pub Noise (selected sample)



Figure 5: Comparison of D-Values for all different phones and different noises

Test series 1: Result summary of D-Value tests

- Due to their technology, both "Double Mic" phones deliver the highest D-Value scores for almost all test cases. The achieved results are a magnitude better than for all other phone types.
- The "Short Phone" delivers the lowest D-Value scores for all background noises used. It is inferior due to its size as well as due to its internal old technology.



- The average progression of the D-Values compared to the single noises show a rough tendency downward from left to right listed noise. Pink Noise delivers highest D-Value scores for almost each single phone whereas Kindergarten Noise delivers low D-Value scores for almost each phone. Cafeteria Noise is at the low end side of the measured D-Values.
- However, the "Double Mic 2" phone does not exactly follow the above stated trend.
- As indicated, both "Double Mic" phones achieve highest D-Value scores whereas the "Short Phone" delivers lowest D-Values for almost all noises. In general, the results can be grouped into three quality areas: "Double Mic" phones, "Flip Phones" and "Short Phone".
- The reproducibility of D-Value tests can be impaired by the non-linear and time-variant processing of the devices: This is indicated by the varying result scatters of the D-Value tests.

3.3. 3QUEST Tests (S-/N-/G-MOS) / Test series 1

In a third experiment the different ETSI background noises were used to determine scores according to the new 3QUEST method [3]. Again, the setup as described in chapter 2 was used.



Figure 6: Comparison of 3QUEST G-MOS for the different phones with different noises

In contrast to figure 5, the 3QUEST G-MOS analysis presents "Insidebus" noise and "Pubnoise" as most critical noises for the five tested devices whereas "Cafeterianoise" and "Kindergarten1" noise deliver comparatively higher G-MOS scores.

This circumstance as well as further interpretation possibilities of 3QUEST shall now be observed in detail—first of all by an additional analysis of the S-MOS and N-MOS scores of 3QUEST, second by a direct comparison between D-Value and 3QUEST scores.





Figure 7: Comparison of 3QUEST S-MOS for the different phones with different noises



Figure 8: Comparison of 3QUEST N-MOS for the different phones with different noises

Test series 1: Result summary of 3QUEST tests

- The "Insidebus" and "Pubnoise" results also show the lowest S-MOS and N-MOS scores.
- The S-MOS analysis shown in figure 7 indicates a classification of the phones into two main groups: The two flip phones deliver significantly higher scores than the other three phones. This indicates that for high speech quality in the presence of background noise, the mechanical geometry can be seen as dominant parameter.



- Figure 7 also shows that the (old-fashioned) short phone delivers a speech quality in the presence of background noise comparable with the top modern "double mic" phones.
- Analyzing N-MOS scores, the result shows a different picture: Here, the short phone can clearly be identified as inferior.
- However, the "double mic" phones do not show a significant advantage for any of the noises towards a flip phone in terms of quality of transmitted background noise: Only for "Cafeterianoise" and "Kindergarten1" Noise, the "double mic" phones show significantly higher S-MOS scores.
- The G-MOS can be seen as a combination of the analysis of S-MOS and N-MOS.
- The two flip phones deliver the highest G-MOS scores, followed by the two "double mic" phones. The short phone shows the lowest performance.
- Due to the split of the quality parameters into G-MOS, S-MOS and N-MOS, a detailed analysis of the individual performance of each device under test is possible.
- The fact that the single values of Fig. 6, 7 & 8 are of almost constant distance from each other indicates that a 3QUEST value of one noise type can successfully be used to predict the 3QUEST value of the other noise types. However, this deduction only applies in the given data set. It is not a property of the 3QUEST algorithm but an indication of how the noise cancellers in these particular telephones work. Furthermore, counter examples to this behavior have been observed amongst telephones equipped with more cutting edge noise cancelling algorithms, this can also be verified by test series 2. This shows the need to test a telephone using 3QUEST with various noise types.

3.4. Comparison D-Value vs. 3QUEST / Test series 1

The previous two chapters showed a serious difference in the result evaluation between D-Value and 3QUEST. Now, a direct comparison between 3QUEST scores and D-Values is applied for some of the tests.

As a first sample, figure 9 shows S-/N-/G-MOS as well as the corresponding D-Value for the device "Flip Phone 1". Figure 10 shows the same results for the "Short Phone". Both figures use the same scaling to make comparisons easier.









Figure 10: Comparison 3QUEST vs. D-Value / Short Phone



Test series 1: Result summary of comparison tests D-Value vs. 3QUEST

- Both figures show an incomparable progression between D-Value and S-/N-/G-MOS. It should be emphasized that no correlation is indicated between D-Value and N-MOS (which could eventually be expected).
- The figures indicate that a correlation between 3QUEST and D-Value is not existent: Due to the
 elemental difference in its application method, the scores represent different analyses. While
 the D-Value only is a degree for the noise suppression under the (non-realistic) comparison to
 speech transmission, 3QUEST allows an in-depth analysis of transmitted overall quality, speech
 quality and background noise (reduction) quality.

3.5. Comparison PESQ[™] vs. 3QUEST / Test series 1

Apart from the comparison of D-Value vs. 3QUEST, a comparison between 3QUEST and PESQ[™] is also of interest: Although not validated for this type of application, PESQ[™] is sometimes used in practice to analyze speech in the presence of background noise. Care should be taken of the fact that PESQ[™] is explicitly *not* intended to analyze effects of artifacts from echo cancellers and noise reduction algorithms (see scope of ITU-T P.862, [11]).

Figure 11 shows a comparison of the five different devices under test for three different background noises: Each solid line represents the results of 3QUEST G-MOS whereas each dashed line represents the P.862 score.





Test series 1: Result summary of comparison tests PESQ[™] vs. 3QUEST

- Despite the fact that comparable rough tendencies might be interpretable, the figure shows an incomparable score distribution between PESQ[™] and G-MOS when interpreted overall: Depending on noise and device, the difference between PESQ[™] score and 3QUEST G-MOS varies between +1.27 and -0.56. There is no correlation in value progressions identifiable.
- The figure further reveals that PESQ[™] scores show a significantly lower differentiation between the devices: All PESQ[™] scores are located between 2.3 and 3.4 whereas the 3QUEST scores vary between 1.75 and 4.47. As a consequence, a quality separation of the devices is more ambiguous: PESQ[™] does not clearly separate the good from the poor, the results are almost all in the same range.

3.6. D-Value & 3QUEST Tests / Test series 2

In a second test series, the results completed by test series 1 were verified by another test setup, using different devices under test. The focus for the devices of this test series was using "state of the art" modern phones like business phones or "Smartphones" (PDA). Comparison in this test series has been made for four different background noises according to the ETSI database [2].

The following three figures show a comparison between 3QUEST scores and D-Values which can be seen as representative for the 8 devices under test.



Figure 12: Comparison 3QUEST vs. D-Value / Device 1





Figure 13: Comparison 3QUEST vs. D-Value / Device 2



Figure 14: Comparison 3QUEST vs. D-Value / Device 3



Test series 2: Result summary of comparison tests

- For all three devices, "Road" noise results in lowest S-MOS and G-MOS scores whereas "Cafe" noise results in highest S-/G-MOS scores.
- N-MOS scores are highest for "Car" noise for all three devices: This result can be explained by the fact that "Car" noise is of comparatively high level but of constant character (quasistationary). For constant noises, even for higher levels noise cancellers of the phones provide a comparatively good performance.
- S-MOS scores vary between the devices: Especially device 1 shows a poorer performance in speech transmission than devices 2 & 3. However, all devices show a weak performance for "Road" noise which is loud as well as non-stationary. Such types of noise are most difficult to handle for most noise cancellers. So it is recommended to also include such type of time variant noises when testing noise cancellers.
- G-MOS scores can be interpreted as a combination of S-MOS and N-MOS. However, a comparison of devices 1 and 3 for "Train" noise proves that a resulting G-MOS score can have different magnitudes of S-MOS and N-MOS: For this example, both G-MOS scores are almost identical (2.2 / 2.3) whereas S-MOS (1.3 / 2.4) and N-MOS (4.7 / 3.2) differ significantly. A typical reason for this effect is a different implementation of the "aggressiveness" of built-in noise cancellers: Moderate implementations may have fewer effects on speech quality but will also let pass some noise. Consequently, a detailed analysis as given by 3QUEST is necessary to understand backgrounds and causes of speech quality impairments.
- Figures 12—14 show an incomparable progression between D-Value and S-/N-/G-MOS for the different noises. There is also no correlation between D-Value and N-MOS. Depending on devices and background noises, D-Values vary within a magnitude: Device 1 shows a variation of 28.3 dB whereas the N-MOS keeps an approximately constant score over all background noises, see figure 12!



4. Summary

The experiments conducted lead to the following conclusions:

- The performance of mobile phone background noise cancellers is depending on the type and character of background noises to which the phone is exposed. In this context, the overall level of background noise is of importance, however the *character* of noise like its stationarity, its embedded voice-like contents or its level varieties shows a higher effect.
- A sufficient T-MOS score cannot be taken as indication that the phone quality will also be sufficient in the presence of background noise. The assumption that today's noise cancellation systems will cause any reaction in silent conditions (which is a prerequisite for T-MOS tests) is no longer valid.
- Using the 3QUEST S-/N-/G-MOS values allows a differentiated analysis of speech quality in the presence of background noise: Depending on the implementation of the specific mobile, overall quality (represented by the G-MOS score) may be impaired more by speech transmission quality (S-MOS) or noise transmission quality (N-MOS). This separation of quality areas allows a more comprehensive evaluation of quality aspects and therefore a more targetoriented statement for possible improvements.
- Analysis and expressiveness of 3QUEST is incomparable to the "classical" D-Value. Due to their different origin and application, 3QUEST scores cannot be brought into line with D-Value scores. It should be noted that this point is also valid for the N-MOS score: The N-MOS describes noise quality *in the presence of speech*, so the initial situation is already different to the D-Value. This is reflected by the test results.
- Due to the more realistic test scenario, 3QUEST analysis can be seen as superior to D-Value analysis: 3QUEST corresponds more to a customer's impression and therefore provides more trustful results than D-Value tests.
- In contrast to PESQ[™], the 3QUEST G-MOS score allows a distinct quality separation between different devices. Taking the additional advantage into account that 3QUEST offers a detailed evaluation by its three dimensional analysis of S-/N-/G-MOS, the 3QUEST analysis can be seen as at least predominant to PESQ[™] – the evaluation of terminals and noise cancellers is out of the scope of PESQ[™].



5. References

- [1] ITU-T Recommendation G.111, Transmission Systems and Media, General Recommendations on the Transmission Quality for an Entire International Telephone Connection, Loudness Ratings in an International Connection
- [2] ETSI EG 202 396-1, Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database
- [3] ETSI EG 202 396-3, Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 3: Background noise transmission - Objective test methods
- [4] ITU-T Recommendation P.57, Artificial ears
- [5] ITU-T Recommendation P.58, Head and Torso Simulators for Telephonometry
- [6] ITU-T Recommendation P.64, Telephone Transmission Quality Objective Measuring Apparatus, Determination of Sensitivity/Frequency Characteristics of Local Telephone Systems
- [7] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ[™]): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs
- [8] ETSI TS 51.010, GSM/EDGE Radio Access Network: Digital cellular telecommunications system (Phase 2+); Mobile Station (MS) conformance specification; Part 1: Conformance specification
- [9] HEAD acoustics Application Note, Telecommunications Objective Speech Quality Assessment
- [10] HEAD acoustics Application Note, 3QUEST: 3-fold Quality Evaluation of Speech in Telecommunications



6. Annex / Graphs



Fig. 15: Narrowband device No. 1, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)



Fig. 16: Narrowband device No. 2, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)



Fig. 17: Narrowband device No. 4, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)





Fig. 18: Narrowband device No. 4, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)



Fig. 19: Narrowband device No. 5, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)



Fig. 20: Narrowband device No. 6, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)





Fig. 21: Narrowband device No. 7, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)



Fig. 22: Narrowband device No. 8, D-value vs. S-, N- and G-MOS (left), Quality Pie (right)