Auditory and instrumental evaluation of conference phones

Jan Reimes, Christoph Nelke
HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: telecom@head-acoustics.de

1 Introduction

The usage of conference phones is getting more important in today's business life. As an alternative to face-to-face meetings, conference calls can save traveling time and cost. However, for good user experience, conference devices have to provide good quality regarding speech communication.

On the other hand, there are only a few terminal measurement specifications available which can be used for quality evaluations of such devices. Annex B of ITU-T Recommendation P.340 [1] defines speech quality performance requirements of conference phones, but so far in a provisional state. Several other specifications for hands-free telephony are available as well, but do not regard multi-talker / conference scenarios, i.e. only one-to-one communication is considered.

Obviously, conference phones are generally used in multiple-talker scenarios at the near-end. Moreover, the influence of typical ambient noises (e.g., fan noises or keyboard typing) is also not yet part of any quality measurement.

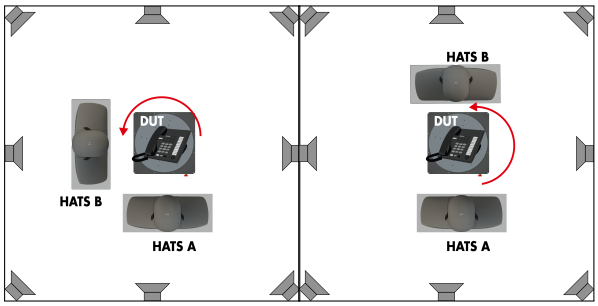
In order to investigate the impact of those factors on speech quality of conference devices, a large measurement series in sending direction was conducted, including multi-talker scenarios and typical noise conditions. Different positions and angles between the talkers as well as multiple terminals were taken into account. Subsequent to the collection of these recordings, a substantial auditory evaluation according to ITU-T P.835 [2] was carried out. The outcome of this evaluation allows pointing out individual shortcomings of the terminals. Finally, the results of the listening tests are compared to common instrumental speech quality metrics.

2 Recording procedure

Measurement Setup

The test setup used is described in [3] and is depicted in Figure 1. The device under test (DuT) is placed on a table in a semi-anechoic room with two head and torso simulators (HATS) (A and B) according to [4] in order to simulate two concurrent talkers.

Each HATS is equalized and calibrated according to [5] and positioned such that the distance between the lip ring center of HATS and the edge of DuT becomes 80 cm horizontally and 30 cm vertically. For the measurements, both HATS A and B are arranged around the DuT with a certain angle configuration. E.g., Figure 1a shows the configuration of HATS A at 0° and HATS B at 90°, where 0° defines the front of the device given by the display or control panel. On the other hand, Figure 1b shows HATS B at 180°.



(a) 90° **(b)** 180°

Figure 1: Positioning of HATS A and B

The background noise is generated by the eight loudspeakers in Figure 1 according to [6]. Here, the setup for desktop operated hands-free terminals is applied, which reproduces a realistic sound field around the DuT. The noise signals *Conference1* and *Conference2* are taken from the desktop hands-free database of [6].

All conditions represent typical background noise scenarios for the use of conference phone testing containing fan noise of a projector, keyboard typing noise, and noise generated by movements of people sitting at the conference table. *Conference1* primarily is a stationary signal dominated by the fan noise. *Conference2* additionally contains an increasing amount of transient components in the recorded signals. Besides the two noisy scenarios, recordings with no background noise are taken into account, too.

To reduce workload during the measurements caused by the moving of the HATS, the DuT is placed on a turntable and eight measurements in steps of 45° are carried out for each angle difference between the two HATS. The generated background noise field is then turned accordingly by using eight independent loudspeaker equalizations for the eight positions of the DuT. In the current evaluation, two angles between the HATS are investigated: HATS A is always located at 0°. The position of HATS B is either 180° (see Figure 1a, opposite to HATS A) or 90° (see Figure 1b).

Altogether, eight measurements in steps of 45° are obtained per talker setup and background noise. For the listening test, only a subset of three measurements with different positions is selected. These excerpts are considered to cover best, worst, and average quality of the transmitted speech signals.

Speech corpus

The full-band speech material used for playback via artificial mouth consists of German sentences taken from the database presented in [7]. The length of each sentence is 4.0 s and 16 samples were included (four male and four female talkers, two sentences per talker) for the test. During the recordings with the conference phones, the speech samples were alternately reproduced by the two HATS as depicted in Figure 2.

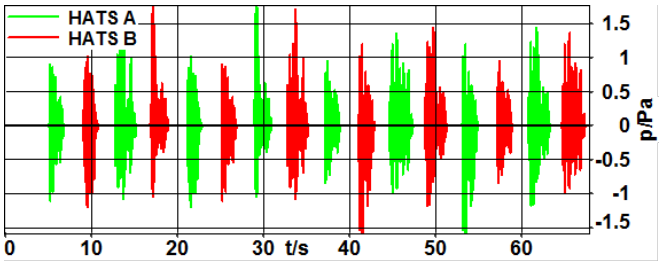


Figure 2: Alternating speech signals from HATS A and B

The alternating speech signals from two different positions simulate a realistic conference scenario, which might be more challenging for DuTs if they apply speech enhancement method by means of a spatial processing of the recorded signals.

It could be argued that this structure of speech signals does not represent a typical conference scenario, since there is no overlap and longer pauses between the sentences. However, for this initial study, this sequence was chosen for two reasons:

1. For the listening test, the speech material was divided into 4.0 s segments of each HATS. Each segment is evaluated separately and allows a dedicated per-talker / per-position evaluation.
2. Overlapping speech samples usually are not validated as reference signals for instrumental quality prediction models, which are also part of the current study (see section 5).

Devices under Test

For the present investigation, three test devices were used. All devices operated in wideband (WB) mode via Bluetooth transmission. They represent different quality classes of conference phones and are named DuT 1, DuT 2, and DuT 3 in the following:

- DuT 1: A simple low-price speaker phone with limited amount of signal processing
- DuT 2: A middle-class speaker phone with a rather aggressive signal processing
- DuT 3: An upper-class conference phone with considerable signal processing capabilities and especially designed for multi-talker scenarios.

3 Auditory Evaluation

The auditory evaluation of the recorded speech material is realized by an ITU-T P.835 [2] listening test with 48 normal-hearing participants with an age distribution between 18 and 64 years.

During the listening test, single speech segments of 4.0 s of the recordings are presented and the participants were asked to rate the speech, noise, and overall quality on a 5-point scale resulting in S-MOS (speech distortion), N-MOS (noise intrusiveness), and G-MOS (global/overall quality) values.

In total, 54 test conditions are obtained by a combination of the following test cases:

- 3 devices
- 3 noises (Silence, *Conference1*, *Conference2*)
- 3 turntable positions
- 2 talker positions (HATS A and B).

In addition, 12 reference conditions according to the processing scheme of [7] were created and used in the auditory evaluation. These conditions provide a uniform distribution of speech, noise and overall quality on the complete 5-point scale. Artificially generated signals with different signal-to-noise ratios (SNRs) and different amount of speech degradation are used for this purpose. The noisy reference files were generated using the *Conference2* noise instead of car noise.

Each participant listened two times to 132 samples (two sessions of two balanced blocks of 66 samples, containing one sample per conditions) within one test, resulting into 12 votes per sample and 192 votes per condition (16 sentences per condition).

4 Auditory Results

Overall View

The main goal of an auditory evaluation of conference terminals is a rating between different conditions, i.e., different devices, noise conditions, or the position of the talkers in the conference scenario. First, a more broad investigation is carried out for a general comparison between the performances of the DuTs. For this purpose, auditory results are provided in box plot representation.

Figure 3 shows the results of the listening test for the three DuTs without background noise. DuT 1 shows the best average S-MOS of about 4.2 MOS. This device obviously does not apply any signal processing impacting speech quality and therefore the speech signal is transmitted without noticeable degradations. Some of the lower N-MOS ratings for the silent conditions (especially for DuT 1 and DuT 2) can be explained by the strong idle noise of these devices and a slight residual noise generated by the turntable.

Figure 4 and 5 depict the ratings for the noisy conditions *Conference1* and *Conference2*, respectively. As expected, lower ratings for the two noisy conditions are obtained than in the noise-free scenarios of Figure 3. The higher amount of transient components in the *Conference2* condition seems not to influence the noise quality ratings here. The now N-MOS scores observed for DuT 1

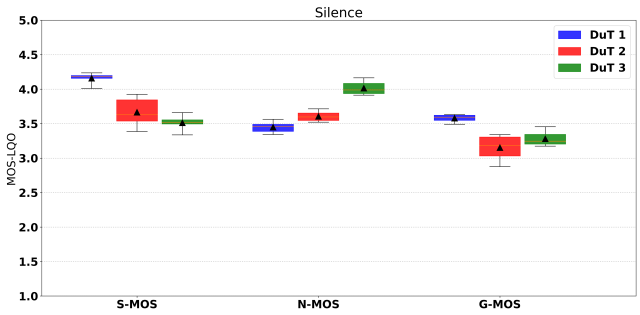


Figure 3: Auditory results under silent conditions

further motivates the assumption that no or very limited noise cancellation is applied here. The unaltered noise in the transmitted signals leads to the poorest ratings among the three devices. Some noise cancellation can be assumed for DuT 2 and DuT 3 where a higher noise suppression can be observed, resulting in better N-MOS values especially for DuT 3. For the S-MOS attribute, DuT 2 shows the most varying results of all devices (indicated by larger spread in box plot), which can be explained by rather aggressive signal processing and directional dependent behaviour, i.e. insufficient tracking of talkers or directional dependent directivity.

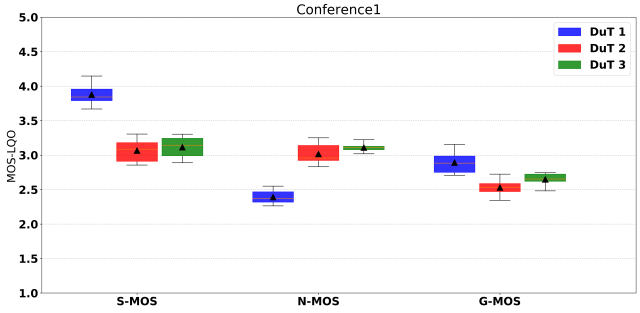


Figure 4: Auditory results for noise *Conference1*

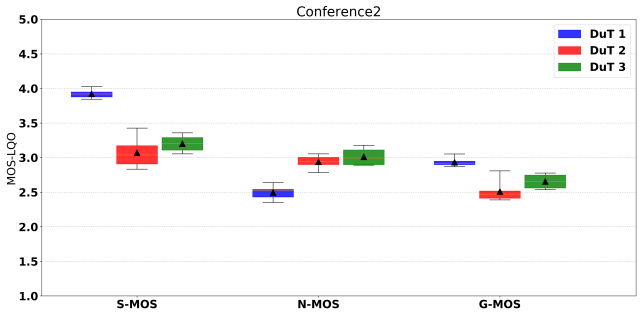


Figure 5: Auditory results for noise *Conference2*

Detailed Analysis

The overall results of the previous section do not show strong outliers and/or extreme outliers. As an intermediate conclusion, DuT 2 exhibits some issues with speech transmission from different directions. Thus, this device is investigated in more detail in the following paragraphs.

In Table 1, the results for S- and N-MOS are shown for talker A, which is rotated around DuT 2. While the position 0° provides very low speech degradation (S-MOS = 4.0), quality decreases by turning to the opposite side (180°, S-MOS = 3.4). N-MOS remains constant in all cases (≈ 3.6), which indicates a constant noise level.

Table 1: DuT 2 in silence

Position talker A	S-MOS	N-MOS
0°	4.0	3.6
90°	3.6	3.7
180°	3.4	3.5

Another example for the inappropriate speech transmission from several directions is shown in Table 2. Here, the scenario with talkers facing each other in the presence of *Conference1* noise is discussed in more detail. The talkers A and B are located at 90° and −90°. A decrease of 0.4 in S-MOS can be observed, while N-MOS again remains constant at 3.0. This asymmetric performance may either be caused by an asymmetric signal processing and/or nonuniform microphone directivity, i.e. left and right side of DuT 2 seem to be transmitted differently.

Table 2: DuT 2 in noise *Conference1*

Position talker A	S-MOS	N-MOS
0°	3.2	3.0
90°	2.8	3.0

Finally, in Table 3 DuT 2 and DuT 1 are compared in a certain position scenarios (A=180°, B=90°) with noise type *Conference2*. Even though both directions seem to be transmitted with a similar quality (S-MOS ≈ 3.0), DuT 1 performs about 1.0 MOS better. Both, the signal processing not affecting the speech quality as well as the omnidirectional characteristics of DuT 1 contribute to higher performance. However, for the same reason, N-MOS is about 0.5 MOS higher for DuT 2 than for DuT 1.

Table 3: DuT 1 vs. DuT 2 with *Conference2* noise

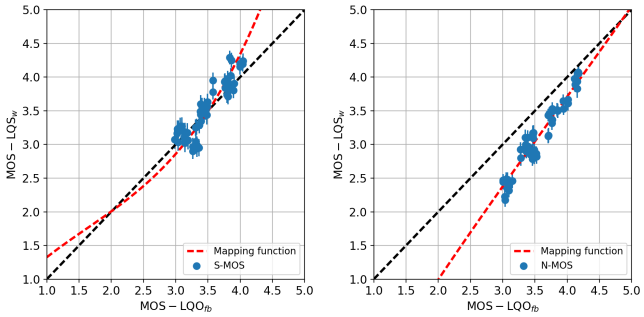
	S-MOS		N-MOS	
Talker	A	B	A	B
DuT 1	3.8	3.9	2.4	2.4
DuT 2	3.0	2.9	2.9	2.9

5 Instrumental Results

In a further investigation, the auditory ratings of the listening are compared to the instrumental speech quality measure according to ETSI TS 103 281 [7]. This recently developed method is intended for a more generalized usage of handset and hands-free devices in super-wideband (SWB) and fullband (FB) mode and is based on an extremely large training database. Here the *Model A* of [7] is applied for the prediction of instrumental S-, N- and G-MOS. In contrast to many other instrumental speech quality prediction models, the proposed method is capable of taking the absolute level information of a listening

test sample into account. This also includes the varying speech level differences between consecutive samples of one measured sequence (due to alternating talkers).

Figures 6a and 6b provide the per-conditions results of instrumental and auditory results in the form of a scatter plot. Here only S- and N-MOS are depicted, since G-MOS is mostly a combination of these two attributes. The 95% confidence intervals (CI95s) of the auditory data are indicated as vertical markers.



(a) S-MOS **(b)** N-MOS

Figure 6: Auditory vs. instrumental results

As performance metrics, Pearson correlation coefficient $r_{Pearson}$, root-mean square error ($rmse^*$) and maximum absolute error ($maxabs^*$) are calculated per attribute. The metrics $rmse^*$ and $maxabs^*$ take the uncertainty of auditory data (CI95) into account, as described in [8]. They are evaluated after 3rd order mapping of the instrumental results (function indicated in red color in Figure 6). Even though the proposed method works in SWB/FB context, it provides highly accurate performance metrics as shown in Table 4.

Table 4: Performance metrics of predicted results

	S-MOS	N-MOS	G-MOS
$r_{Pearson}$	0.901	0.966	0.841
$rmse^*$	0.109	0.057	0.124
$maxabs^*$	0.228	0.156	0.277

The also well-known speech quality measure according to ITU-T Recommendation P.863 [9] is not applicable here, since it is not designed for the evaluation of short speech segments, which are required for testing conference phones in realistic scenarios including alternating talkers. In addition, the usage of this method with acoustic hands-free setups was recently declared for further study.

6 Conclusion

The listening test results presented in this contribution demonstrate that an evaluation of conference phones in the presence of background noise is crucial for a quality comparison between different devices. Not only the performance in different noise scenarios but also the capability of speech transmission in a dual talker case with different talker positions is a key factor for the usability of a conference phone.

The instrumental speech quality measure according to ETSI TS 103 281 is highly correlated with the auditory results, even though the method was not explicitly developed for this purpose. For extensive tests with a broader variation of speech segments, noise signals, and talker positions, this quality metric could be used instead of large listening tests, in order to predict the performance of conference phones in realistic multi-talker scenarios.

For future work, additional aspects may be considered. More sophisticated devices (including e.g. advanced signal processing, beamforming, etc.) should be evaluated to analyze direction-dependent speech transmission. Even though some detailed results indicated better/poorer performance, all devices used in the experiment provided at least average quality, no strong outliers could be observed.

In addition, the impact of room reverberation should be investigated. The impact of additional signal processing like e.g. dereverberation algorithms on speech quality is certainly of interest at higher distances between talker and microphone(s). Finally, also concurrent talkers (near-end double talk) may be considered.

References

[1] *Objective test methods for multi-talker scenarios*, ITU-T Recommendation P.340 Amendment 1, Annex B, Oct. 2014.

[2] *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, ITU-T Recommendation P.835, Nov. 2003.

[3] *Transmission characteristics and speech quality parameters of hands-free terminals*, ITU-T Recommendation P.340, May 2000.

[4] *Head and torso simulator for telephonometry*, ITU-T Recommendation P.58, May 2013.

[5] *Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals*, ITU-T Recommendation P.341, Mar. 2011.

[6] *A sound field reproduction method for terminal testing including a background noise database*, ETSI TS 103 224 V1.3.1, Jul. 2017.

[7] *Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*, ETSI TS 103 281 V1.2.1, Jan. 2018.

[8] *Statistical analysis, evaluation and reporting guidelines of quality measurements*, ITU-T Recommendation P.1401, Jul. 2012.

[9] *Methods for objective and subjective assessment of speech quality*, ITU-T Recommendation P.863, Sep. 2014.