International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.340
Amendment 2
(01/2019)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Voice terminal characteristics

Transmission characteristics and speech quality parameters of hands-free terminals

Amendment 2: Annex B: Objective test methods for multi-talker scenarios

Recommendation ITU-T P.340 (2000) – Amendment 2

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	**Series**	**P.30** **P.300**
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50 P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.340

Transmission characteristics and speech quality parameters of hands-free terminals

Amendment 2

Annex B

Objective test methods for multi-talker scenarios

Summary

Recommendation ITU-T P.340 provides audio performance requirements and test methods for conference phones and hands-free terminals. These are limited to one-to-one meeting scenarios.

Amendment 2 to Recommendation ITU-T P.340 replaces Annex B: Objective test methods for multi-talker scenarios, which was introduced in Recommendation ITU-T P.340 (2000) by its Amd.1 (10/2014).

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.34	1980-11-21		11.1002/1000/8001
2.0	ITU-T P.34	1984-10-19		11.1002/1000/5905
3.0	ITU-T P.34	1988-11-25		11.1002/1000/1726
4.0	ITU-T P.34	1993-03-12	XII	11.1002/1000/1727
5.0	ITU-T P.340	1996-08-30	12	11.1002/1000/3632
6.0	ITU-T P.340	2000-05-18	12	11.1002/1000/5078
6.1	ITU-T P.340 (2000) Cor. 1	2004-03-31	12	11.1002/1000/7296
6.2	ITU-T P.340 (2000) Amd. 1	2014-10-29	12	11.1002/1000/12324
6.3	ITU-T P.340 (2000) Amd. 2	2019-01-13	12	11.1002/1000/13841

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.340

Transmission characteristics and speech quality parameters of hands-free terminals

Amendment 2

Annex B

Objective test methods for multi-talker scenarios

1) Scope

This amendment replaces Annex B, Objective test methods for multi-talker scenarios, which was introduced by Amd.1 to this Recommendation.

2) Modifications to ITU-T P.340

2.1) Clause 2, References

Add the following normative references to clause 2:

[22] ETSI TS 103 224 (2014), *Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database*.

[23] ETSI TS 103 281 (2017), *Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*.

2.2) Clause 3

Add the following abbreviations to clause 3 in alphabetical order:

DUT Device Under Test

HATS Head And Torso Simulator

2.3) Annex B

Replace Annex B with the following text after Annex A:

Annex B

Objective test methods for multi-talker scenarios

(This annex forms an integral part of this Recommendation.)

B.1 Introduction

The existing standardized methods for measuring the performance of conference phones and hands-free terminals are dedicated mainly to one-to-one communications, i.e., the situation where only one person in a room is speaking, or when multiple people are speaking, turn-taking is so infrequent that the dynamics of turn-taking can be ignored.

However, the primary use case of conference phones and a frequent use case of hands-free terminals is to capture the voices of several people in a room. In a typical teleconferencing situation people may be speaking one at a time for a prolonged period, such as during a presentation, or they may be taking turns in rapid succession, such as during a discussion. When multiple talkers in a room are having interactive, natural conversations among themselves with a high level of interactivity, i.e., with sudden turn-taking or concurrent talking, there is a chance that remote participants cannot hear all talkers (sometimes without noticing) and may also hear distorted speech, due to the limited room capture-capability of conference phones.

This annex expands the scope to the communication situation where multiple talkers in a room interact frequently. Four performance criteria and their associated measurement techniques are defined. The first characterizes the adaptation time of the device under test (DUT) when it is presented with alternating (i.e., non-overlapping) talk bursts from different angular directions. The second quantifies how well the DUT preserves the level of each of two concurrent (i.e., completely overlapping) talk bursts. The third simulates more complicated turn-taking dynamics between two talkers including switching with partial overlap of talk bursts and abrupt talker switching without overlap. This criterion evaluates how well the level of a talker is maintained immediately after talker switching. The fourth criterion describes the concurrent talkers in the frequency domain to quantify how well the timbre of individual talkers is maintained.

The test methods described in this Annex use real speech signals per [9].

B.2 Test set-up and configuration

The test set-up shall meet the requirements described in clause 5, [ITU-T P.341] and [ITU-T P.581]. The head and torso simulator (HATS) should be calibrated per [ITU-T P.581], and should be positioned according to the test set-up arrangements per clause 4 of [ITU-T P.341], e.g., 80 cm of horizontal and 30 cm of vertical distance from the centre of the DUT to the lip ring centre of the HATS for conference phones as shown in Figure 6 of [ITU-T P.341]. Test signal levels applied acoustically to the DUT should be calibrated following the procedures in clause 5. If the DUT provides only multi-channel voice to its sending direction, the measurement signal at the electrical reference point should be down-mixed to mono.

The HATS should be arranged around the DUT with a certain angle setting throughout the tests. The angle terminology used in this Annex is depicted in Figure B.1. Using the reference zero-degree angle as the control panel of the conference phone, the angle value increases in a clockwise direction.

Because the performance of these tests may depend on the angular separation between the HATSs and the orientation of the HATSs towards the end point, the tests described in this Annex may need to be performed at multiple orientations to completely specify the DUT. It is the responsibility of the experimenter to select the HATS angles and orientations that are appropriate for the DUT. For DUTs that appear symmetric, it is recommended to test the combinations of angles A and B given in Table B.1.

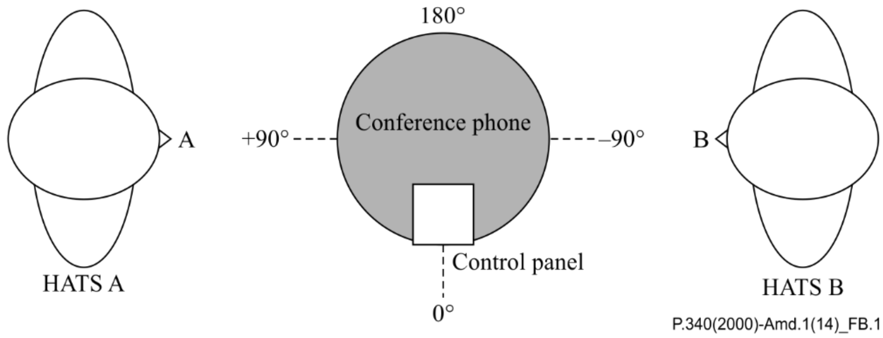

Figure B.1 – Diagram showing a conference phone and HATS angle configuration

HATS A and B are positioned at the angle of 90 degree and -90 degree, respectively, in Figure B.1. Talker angle at zero-degree refers to the direction of the control panel.

Table B.1 – Recommended HATS angle configurations

No	Angle A [degree]	Angle B [degree]		No	Angle A [degree]	Angle B [degree]
1	0°	90°		9	0°	180°
2	45°	135°		10	45°	−135°
3	90°	180°		11	90°	−90°
4	135°	−135°		12	135°	−45°
5	180°	−90°		13	180°	0°
6	−135°	−45°		14	−135°	45°
7	−90°	0°		15	−90°	90°
8	−45°	45°		16	−45°	135°

B.3 Test 1: Adaptation time in talker alternation

Purpose

The purpose of this test is to measure the adaptation performance of the DUT to two alternating (i.e., non-overlapping) speech signals, applied from different angles.

Performance objective

To ensure a stable state for subsequent measures, the DUT should asymptote to a stable state and should not deviate from that target adaptation level by more than 1 dB. This objective should be satisfied for both angles A and B.

A faster adaptation time is preferable to a slower adaptation time. The adaptation period should be equal to or less than 1 period (i.e., the DUT should be adapted in 1.0 s for the test signal characterized in Figure B.2).

Experimenters are recommended to report the actual measurement results of $L_A(n)$ and $L_B(n)$ as well as the quantities described as the requirements.

Method

1) The test arrangement is described in clause B.2.

2) The test signal used for the measurements shall be the short words as described in clause 7.3.4 of [9]. The level of the test signal shall be −4.7 dBPa at the MRP of both HATS. As shown in Figure B.2, the input signal is constructed such that the short words applied to the HATS positioned at angle A (upper channel) and HATS located at angle B

(lower channel) in an alternating manner. The signal shown in Figure B.2 (1 second) is repeated eight times to create the test signal. The simulated two alternating talkers are located at different angles around the DUT.

3) The measurement shall be conducted at the electrical reference point. For each period of the test signal (Figure B.2), the quantity L_A is calculated as active speech level according to [ITU-T P.56] corresponding to the angle A. Given a series of L_A over multiple periods of the measurement signal, i.e., $L_A(n)$, n = 0, 1, …, 7, the target adaptation level, TAL_A, is defined as $L_A(n)$ with the smallest n, if both $|L_A(n) - L_A(n\text{-}1)|$ and $|L_A(n) - L_A(n\text{-}2)|$ are less than 1 dB. The adaptation period is defined as the smallest n such that $| L_A(n) - TAL_A | < 1$. Similarly, L_B, the target adaptation level and adaptation period of angle B, can be defined in the same manner.

4) These measurements should be repeated with different levels of test signal at angle B (−6 dB, −3 dB, 3 dB and 6 dB relative to the original level at angle B) and different combinations of angles A and B given in Table B.1.

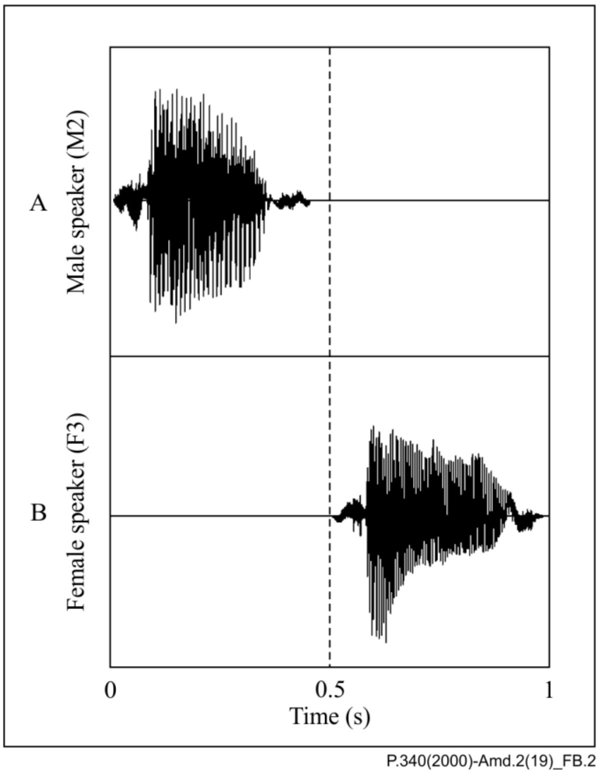

Figure B.2 – One period of input signal for Test 1

B.4 Test 2: Level of completely overlapping (concurrent) talk bursts

Purpose

The purpose of this test is to measure the capture performance of the DUT, when two different talk bursts, applied from different angles, are completely overlapped in time. The level of the transmitted concurrent talk speech signals should reflect the sum of levels of both individual talkers.

Performance objective

The absolute difference of the level L_C (during concurrent talk) and L_D should be less than 3 dB.

The energetic sum of L_A and L_B is calculated as follows:

$$L_D = 10 \log (10\text{^}(L_A/10) + 10\text{^}(L_B/10)) \quad \text{(B.1)}$$

In a transparent system, such as an air path, L_C equals L_D. When L_C does not equal L_D, a smaller absolute difference is preferable to a large one. Differences less than 3 dB are considered perceptually acceptable. The minimum and maximum differences between L_C and L_D should be reported.

Method

1) The test arrangement is described in clause B.2. In order to ensure a reliable activation of the device under test, the DUT shall be adapted according to Test 1 before the actual test.

2) The structure of the test signal is shown in Figure B.3. The upper channel and lower channel are applied at the angles of HATS A and B, respectively. At first, the first male sentence and the first female sentence from British-English single talk sequence of [9] are applied at HATS A (male sentence) and HATS B (female sentence) in alternating manner (i.e., non-overlapping) to simulate a single talk situation. After that, the concurrent signal, which consists of the same male sentence at HATS A (upper channel) and the same female sentence at HATS B (lower channel) is inserted, as shown in Figure B.3.

3) The active speech level measured at the electrical reference point is calculated for each interval: L_A for single talk of A, L_B for single talk of B, and L_C for concurrent talk. The level of concurrent talkers, L_C, is measured in relation to the level of the preceding single talk of HATS A and B.

4) These measurements should be repeated with different levels of test signal at angle B (−6 dB, −3 dB, 3 dB and 6 dB relative to the original level at angle B) and different combinations of angles A and B given in Table B.1.

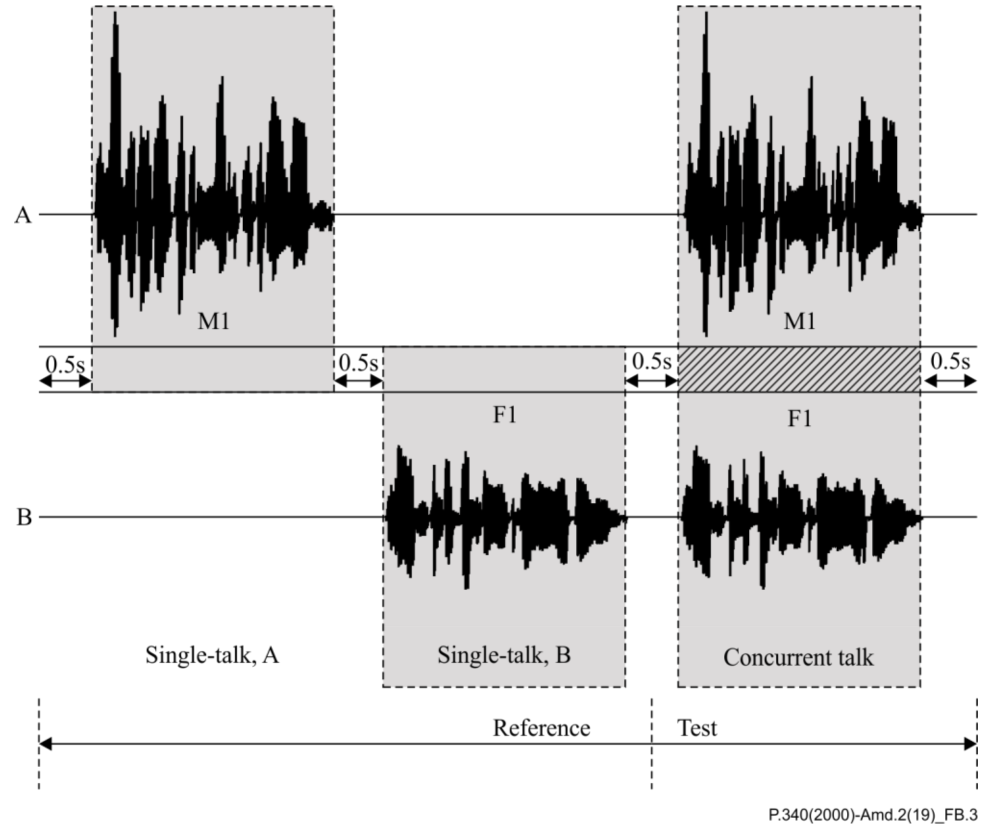

Figure B.3 – Input signal of Test 2

B.5 Test 3: Dynamic turn-taking – switching characteristics

Purpose

The purpose of this test is to measure the performance of the DUT to the dynamic turn-taking between two talkers from different angles. The level measured immediately after talker switching should not demonstrate major distortions. Two different switching scenarios are covered in this test: 1) partially-concurrent talk bursts and 2) abrupt talker switching without overlap.

Performance objectives

The performance is evaluated as long term variation and short term variation:

- Active speech level analysis – the loss or gain of signal level after talker switching:

  The absolute difference between the active speech levels of A_X and A_{REF} (and of B_X versus B_{REF}) should be less than 3 dB, which corresponds to excellent switching performance.

  In an ideal (transparent) system, the difference between the active speech level A_X and A_{REF} (and B_X and B_{REF}) is expected to be 0 dB. Devices providing smaller level differences are preferred.

- Level versus time analysis – the loss or gain of signal levels after talker switching in short time sense:

  The maximum absolute difference between level versus time series of A_X and A_{REF} (and of B_X and B_{REF}) should be less than 6 dB, which corresponds to excellent switching performance.

  In an ideal (transparent) system, the difference of level versus time series of A_X and A_{REF} (and B_X and B_{REF}) is expected to be 0 dB. Devices providing smaller difference between these minimum levels is preferable. Devices providing smaller variations in level versus time differences are preferred.

Method

1) The test arrangement is described in clause B.2.

2) In order to ensure a reliable activation of the device under test, the DUT shall be adapted according to Test 1 before the actual test.

3) A speech sequence consisting of the first male sentence and the first female sentence from British-English single talk sequence of [9] are applied at HATS A and HATS B as shown in Figure B.4. The upper channel and lower channel are applied at the angles of HATS A and B, respectively. In this test signal, the interval A_X represents an active talker switching from HATS B to HATS A with partial overlap. The interval B_X corresponds to an abrupt switching of a talker from HATS A to HATS B without any overlap in time.

4) The level vs. time measured at the electrical reference point is calculated using an exponential weighting filter with a time constant of 35 ms [IV.15]. The time series of level values in the transition interval A_X is compared to the corresponding time series of level values in the interval A_{REF}, which is the corresponding part of A_X when no talker switching happens. In the same manner, the level values in the interval B_X are compared with those in the interval B_{REF}.

5) These measurements should be repeated with different levels of test signal at angle B (−6 dB, −3 dB, 3 dB and 6 dB relative to the original level at angle B) and different combinations of angles A and B given in Table B.1.

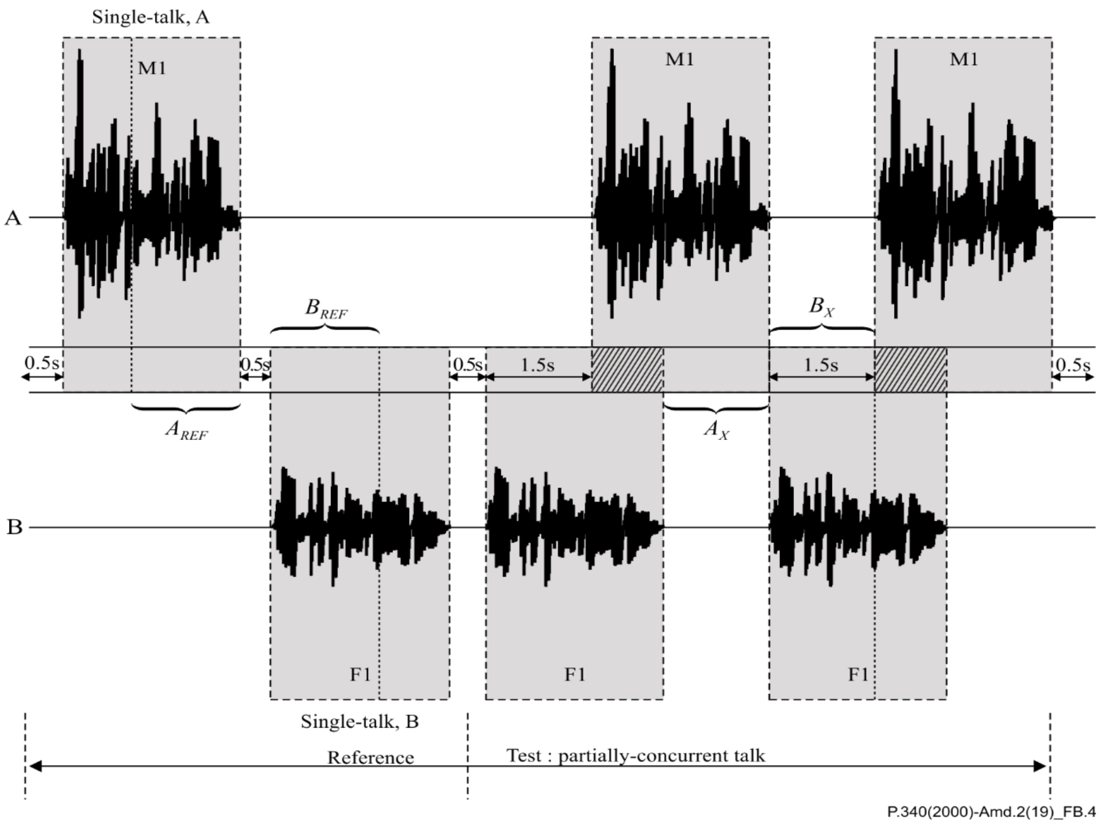

Figure B.4 – Input signal of Test 3

B.6 Test 4: Concurrent talk test with voice-like AM-FM signals

Purpose

The purpose of this test is to measure the capture performance of the DUT in maintaining frequency contents of two concurrent input signals applied from different angles.

Performance objective

The level of all the corresponding frequency bands measured by concurrent talk should not deviate more than 3 dB from the level of the single talk, which would correspond to excellent performance during concurrent talk:

$$|LD_A(k)| = |L_{A,con}(k) - L_{A,single}(k)| \leq 3 \text{ dB for all } k = 1, 2, \ldots, 29 \quad \text{(B.2)}$$

where:

k is the frequency band index for signal A

$$|LD_B(k)| = |L_{B,con}(k) - L_{B,single}(k)| \leq 3 \text{ dB for all } k = 1, 2, \ldots, 28 \quad \text{(B.3)}$$

where:

k is the frequency band index for signal B

In an ideal, transparent system the signal components of HATS A do not affect the signal components of HATS B and vice versa. Therefore, level differences $|LD_A(k)|$ and $|LD_B(k)|$ are expected to be 0 dB for all frequency bands k.

If deviations occur, DUTs with a smaller magnitude of these differences are considered to have better timbre characteristics than DUTs with larger differences.

Method

1) The test arrangement is described in clause B.2. In order to ensure a reliable activation of the device under test, the DUT shall be adapted according to Test 1 before the actual test.

2) To enable the separation of two input signals after processing by the DUT, two voice-like signals modulated in amplitude and frequency for wideband applications (S_A and S_B) are used in this test, see Table 7-6 in clause 7.2.4 of [9]. These input signals are designed such that the frequency contents of the two signals do not overlap. S_A has 29 frequency components starting from 125 Hz (Table 7-6 of [9], column "Sending direction") and S_B has 28 frequency components starting from 180 Hz (Table 7-6 of [9], column "Receiving direction").

3) In Figure B.5, the spectrum in blue shows frequency components of the signal applied at angle A (S_A), and the red frequency components are for the signal applied at angle B (S_B). The blue dotted line represents the filter response used for the analysis of S_A, and the red dotted line is for the analysis of S_B.

4) After the DUT is adapted, the first voice-like signal S_A is applied to HATS A, followed by a pause (0.5 s). After this, the second voice-like signal S_B is applied to HATS B. Then both signals S_A and S_B are applied in a concurrent manner. This step results in three time intervals as shown in Figure B.6: 'Single talk A', 'Single talk B' and 'Concurrent talk'. The signals of talker A and B are applied as shown in Figure B.6. For information, the last period of the adaptation signal is shown for talker A and B first, which is followed by the two single talk intervals and the concurrent talk interval.

5) The measured signal during the 'Single talk A' interval is processed by comb filter using mid-frequencies and bandwidth according to the signal components of the signal (see [ITU-T P.501]) to select the frequency contents of S_A, and the level $L_{A,single}(k)$, is calculated for each k-th frequency band. The measured signal during the 'Single talk B' interval is processed by comb filter using mid-frequencies and bandwidth according to the signal components of the signal (see [9]) to calculate the level, $L_{B,single}(k)$ for each k-th frequency band. These two levels represent the frequency response of the DUT for single talk (no overlap), and are used as references when evaluating the response of the concurrent talkers case.

6) From the measured signal during the 'Concurrent talk' interval, two different levels are estimated. The first one is the concurrent level for S_A, $L_{A,con}(k)$, obtained through analysis for each frequency band using the same comb filter as in the 'Single talk A' interval. Similarly, the concurrent level for S_B, $L_{B,con}(k)$, is obtained using the same comb filter as in the 'Single talk B' interval.

7) These measurements should be repeated with different levels of test signal at angle B (−6 dB, −3 dB, 3 dB and 6 dB relative to the original level at angle B) and different combinations of angles A and B given in Table B.1.

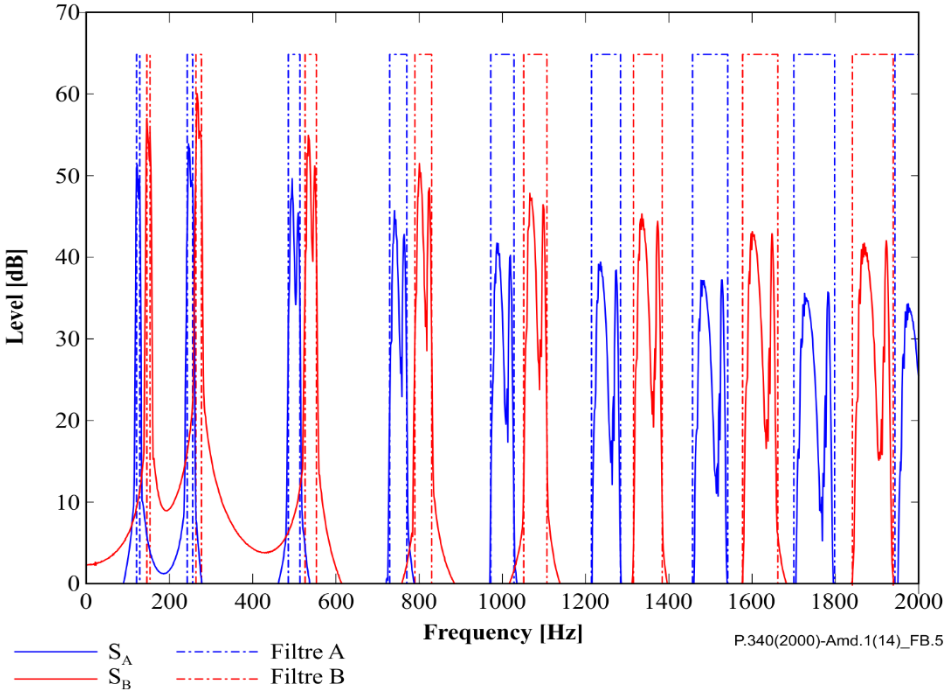

Figure B.5 – Power spectral density of input signals for Test 4 (shown only up to 2 kHz, although the signal contents span to 7 kHz range)

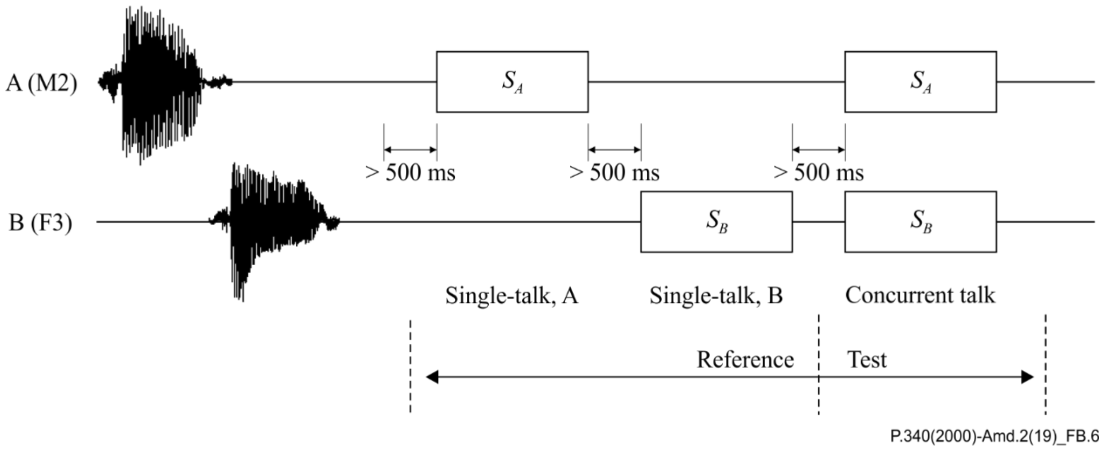

Figure B.6 – Input signal diagram of Test 4

B.7 Test 5: Speech quality in the presence of background noise

Purpose

The purpose of this test is to evaluate the speech quality under ambient noise conditions, which may occur in typical conference scenarios. In addition, this test takes resulting level differences between talker A and B into account.

Performance objective

For each talker A/B, the DUT should provide the following minimum values:

$$S\text{-}MOS\text{-}LQO_f \geq 3.2$$

$$N\text{-}MOS\text{-}LQO_f \geq 3.0$$

$$G\text{-}MOS\text{-}LQO_f \geq 2.6$$

NOTE – The speech quality assessment model predicts in a super-wideband/fullband context. In previous studies, it was shown that results for wideband terminals are highly correlated with auditory test results. Results for narrowband terminals are calculated in a super-wideband/fullband context. In consequence, scores will be lower compared to evaluations in a narrowband context.

Method

1) The test arrangement is described in clause B.2. In addition, a noise field simulation for desktop operated hands-free testing according to clause 5.4 of [22] with eight loudspeakers is set up for each of the eight different positions of HATS A.
2) The full-band speech samples according to Annex E of [23] are used for this test. The length of each sentence is 4.0s and 16 samples are included: two male (m1, m2) and two female talkers (f1, f2), with four sentences each per talker (S1-S4).
3) In order to ensure a reliable activation and convergence, the DUT shall be adapted with the four initial speech samples according to Annex E of [23] before the actual speech sequence. First and third sample are played back via HATS A, the second and fourth sample via HATS B.
4) During the test, the speech samples are alternatingly reproduced by the two HATS, resulting into eight speech segments per talker. The complete test sequence is depicted in Figure B.7, the corresponding permutation scheme for each individual sample is described in Table B.2.
5) An additional recording in silence is conducted prior to all tests with background noise. It is used to calibrate all other noisy recordings according to the procedure of clause 9.5 in [23].
6) The measurement is carried out with triggered and synchronized background noise playback according to [22].
7) The test is carried out for each of the noise scenarios *Conference1*, *Conference2* and *Conference3*. The corresponding recordings are available in the noise database of [22].
8) After calibration of the recordings, the speech quality prediction model A according to [23] is evaluated for each of the eight samples per talker A/B (excluding the convergence sequence). For each talker A/B, the average across the eight samples is calculated and reported.
9) These measurements should be repeated with different levels of test signal at angle B (−6 dB, −3 dB, 3 dB and 6 dB relative to the original level at angle B) and different combinations of angles A and B given in Table B.1.

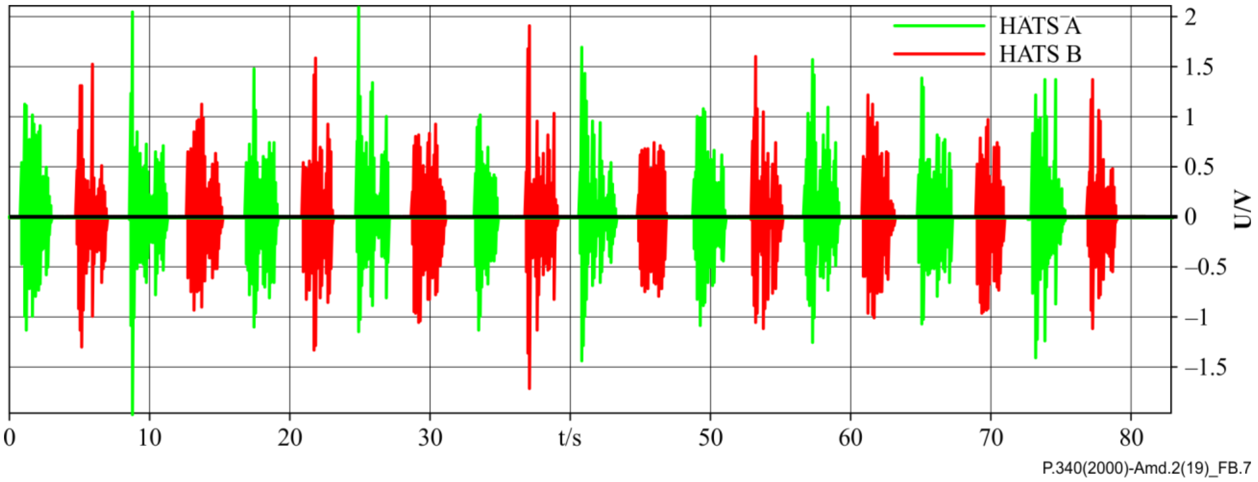

Figure B.7 – Concatenated input signal of Test 5

Table B.2 – Permutation scheme for the test sequence

Segment	Talker	Sentence	HATS
Convergence sequences	m1	S8	A
	f1	S8	B
	m2	S8	A
	f2	S8	B
1	m1	S1	A
2	f1	S1	B
3	m1	S2	A
4	f1	S2	B
5	m1	S3	A
6	f1	S3	B
7	m1	S4	A
8	f1	S4	B
9	m2	S1	A
10	f2	S1	B
11	m2	S2	A
12	f2	S2	B
13	m2	S3	A
14	f2	S3	B
15	m2	S4	A
16	f2	S4	B

2.4) Appendix IV

Add the following reference to Appendix IV:

[IV.15] IEC 61672-1 (2013), *Electroacoustics – Sound level meters – Part 1: Specifications.*

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2019