ETSI TS 103 557 V1.2.1 (2019-08)

Speech and multimedia Transmission Quality (STQ); Methods for reproducing reverberation for communication device measurements

Reference
RTS/STQ-283

Keywords
quality, reverberation, simulation, speech, terminal, testing

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00   Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

The present document is to be used in conjunction with:

- ETSI TS 103 224 [1] series: "A sound field reproduction method for terminal testing including a background noise database".

The present document describes a sound field recording and reproduction technique which can be applied for all types of terminals but is especially suitable for modern multi-microphone terminals including array techniques. While ETSI TS 103 224 [1] focuses on background noise, the present document considers the reproduction of reverberation.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

Many devices that employ microphones to pick up speech signals are used in a hands-free manner. Since there is usually a larger distance between the talker and the device, the microphone signals contain a significant amount of noise and reverberation.

This includes, e.g. phones in hands-free mode, group-audio terminals or smart speakers with speech recognition capabilities as well as terminals in handset or headset mode. Note that the same issues can also arise for hand-held devices depending on the acoustic conditions, see [i.1].

Testing of these devices requires a realistic reproduction of both the noise as well as the reverberation in a defined and reproducible manner. For background noise reproduction, ETSI has standardized a reproduction method (with an accompanying database of background noise signals) in ETSI TS 103 224 [1].

1 Scope

The present document describes a methodology for recording and reproducing different room characteristics and realistic reverberation under conditions that are well-defined and tailored for a calibrated setup in a lab environment. The individual aspects of the description are:

- Measurement of room impulse responses.
- Processing of test signals.
- Loudspeaker setup, calibration and equalization.

The methodology is fundamentally designed for use without access to internals of the Device Under Test (DUT), e.g. the exact positions and orientations of the device's microphones or the unprocessed microphone signals. The methodology is intended to be used for performance evaluation of all types of devices where the room characteristics may impact the performance.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference/.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

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

[2] Recommendation ITU-T P.58: "Head and Torso Simulator for Telephonometry".

[3] N. Xiang: "Evaluation of reverberation times using a nonlinear regression approach" in Journal of the Acoustical Society of America Vol. 98, 1995.

[4] Recommendation ITU-T P.56: "Objective measurement of active speech level".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant and P. Vary: "Do We Need Dereverberation for Hand-Held Telephony?", International Congress on Acoustics (ICA), Sydney, 2010.

[i.2] Recommendation ITU-T P.341 (03/2011): "Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals".

[i.3] ISO 3382-1: "Measurement of room acoustic parameters -- Part 1: Performance spaces".

[i.4] Recommendation ITU-T P.501: "Test signals for use in telephonometry".

[i.5] ETSI TS 103 738: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

[i.6] ETSI TS 103 740: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

For the purposes of the present document, the following symbols apply:

$C_{t_{CO}}$	Clarity with cut-off time t_{CO}
f_s	Sampling frequency
$g_i(k)$	Reverberant components of the impulse response to microphone i
$h_i(k)$	Impulse response to microphone i
$h_N(k)$	Impulse response to the microphone closest to the HATS
$h_{D,i}(k)$	Direct path component of the impulse response to microphone i
K	Length of the impulse response
M	Number of microphones
RT_{60}	Reverberation time
$s(k)$	Source signal
t_{CO}	Cut-off time for clarity (50 ms for speech, 80 ms for music)
MSA-#	Microphone number # of MSA
MU-#	Microphone number # of MU
MM-#	Microphone number # of MM

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

DRR	Direct-to-Reverberant energy Ratio
DUT	Device Under Test
FFT	Fast Fourier Transform
HATS	Head And Torso Simulator
LRC	Lip Ring Centre
MLS	Maximum Length Sequence
SFR	Send Frequency Response
SLR	Send Loudness Rating
SNR	Signal-to-Noise Ratio
MSA	Microphone Sound Array
MU	Mock-Up phone
MM	Measurement Microphone(s)

4 Rationale

Including reverberation in a realistic manner is of paramount importance for accurate testing of, e.g. phones in hands-free mode, group-audio terminals or smart speakers with speech recognition capabilities. While it is possible to test this simply by using the device in a reverberant room, the present document introduces an alternative approach that is based on the explicit reproduction of the reverberant sound field at several microphone positions.

The present method does not require many rooms or variable acoustics setups for testing multiple acoustic conditions. The reverberant sound field can both be measured directly in a reverberant room or it can be calculated based on a database of impulse responses that is provided in combination with the present document.

NOTE 1: The room acoustics has an impact on three transmission paths that could be considered from a measurement perspective:

- from the mouth of the user to the microphone(s) of the DUT (Sending);
- from the loudspeaker(s) of the DUT to the ears of the user (Receiving);
- from the loudspeaker(s) of the DUT to the microphone(s) of the DUT (Echo-path).

NOTE 2: The methodology described in the present document is intended for the first scenario. Although it might be applicable to the other scenarios as well (possibly with modifications), this has not been verified yet and is subject to further study.

5 Room simulation in Sending

5.1 Impulse Response Measurement

5.1.1 Microphone setup

5.1.1.1 Fixed Microphone setup

A fixed microphone setup should be used for DUTs with smaller form factors (e.g. mobile phones in hands-free operation). The microphone setup shall consist of $M = 8$ microphones and conform to the description in ETSI TS 103 224 [1] (see clause 5, "Recording Arrangement"). Since this setup is device-independent, no new impulse response measurements are necessary when testing a new device.

5.1.1.2 Flexible Microphone setup

For larger DUTs, a flexible microphone setup with a use-case dependent number of microphones M shall be used. The microphone setup shall conform to the description in ETSI TS 103 224 [1] (see clause 7, "Generalization of the method for a more flexible loudspeaker and microphone arrangement"). Since this setup is device-dependent, testing a new device requires new impulse response measurements.

5.1.2 Sound source

Since the testing scenario consists of a human talker in a reverberant environment, a HATS with an equalized artificial mouth according to Recommendation ITU-T P.58 [2] shall be used for sound generation. This applies to measurement of impulse responses as well as to recording of reverberant speech signals.

5.1.3 Measurement procedure

There exist different possibilities for measuring impulse responses, e.g. using Maximum Length Sequences (MLS) or using swept-sines (sweeps). The advantage of sweeps is that non-linearities can easily be observed and that the SNR in lower frequencies is higher than with MLS. Using logarithmic sweeps is therefore recommended for system identification. The sweep should cover a frequency range from 20 Hz to 20 kHz and the length of the sweep should be chosen in such a way that no significant components of the impulse response are truncated. Accordingly, a sweep length of at least 2 seconds should be used for typical rooms while larger rooms might need longer sweeps. While the sweep can be used directly as the measurement signal, it is recommended to construct the measurement signal from the individual sweep by repeating it at least five times. If the repetition is used, the determination of the impulse response should be based on an averaging of all but the first sweep period. The first period is discarded to avoid transient effects and the averaging should be performed in the time domain.

5.2 Loudspeaker setup for reproducing reverberation based on the fixed microphone setup

5.2.1 Introduction and System Overview

In order to correctly reproduce the sound field in a reverberant room, the setup described in ETSI TS 103 224 [1] is not sufficient. Direct sound as well as the reverberant components shall be considered. A combination of the multi-channel loudspeaker setup with the artificial mouth of a HATS is used for the reproduction.

The multi-channel loudspeaker setup shall follow the description in ETSI TS 103 224 [1]. If possible, the HATS should be positioned at the same position (distance, angle, mouth direction) for the sound field reproduction with respect to the microphone arrangement as in the original reverberant room. If this is not possible, the HATS shall be positioned at the same angle with respect to the microphone arrangement as in the original reverberant room and the change in distance shall be considered when adjusting levels and delays between the direct path and the reverberant components.

An overview of the signal processing is given in Figure 1. The two paths are visible here as well: The signal for the HATS is only subject to a level (multiplication by coefficient a) and delay (delay element of length T_{DIFF}) adjustment which is covered in clause 5.2.2.2. The reverberant signal components are reproduced by removing the direct path from the impulse responses (see clause 5.2.2.1) and generating the target signals for the system according to ETSI TS 103 224 [1] by convolving the source signal with the remaining parts of the impulse responses.

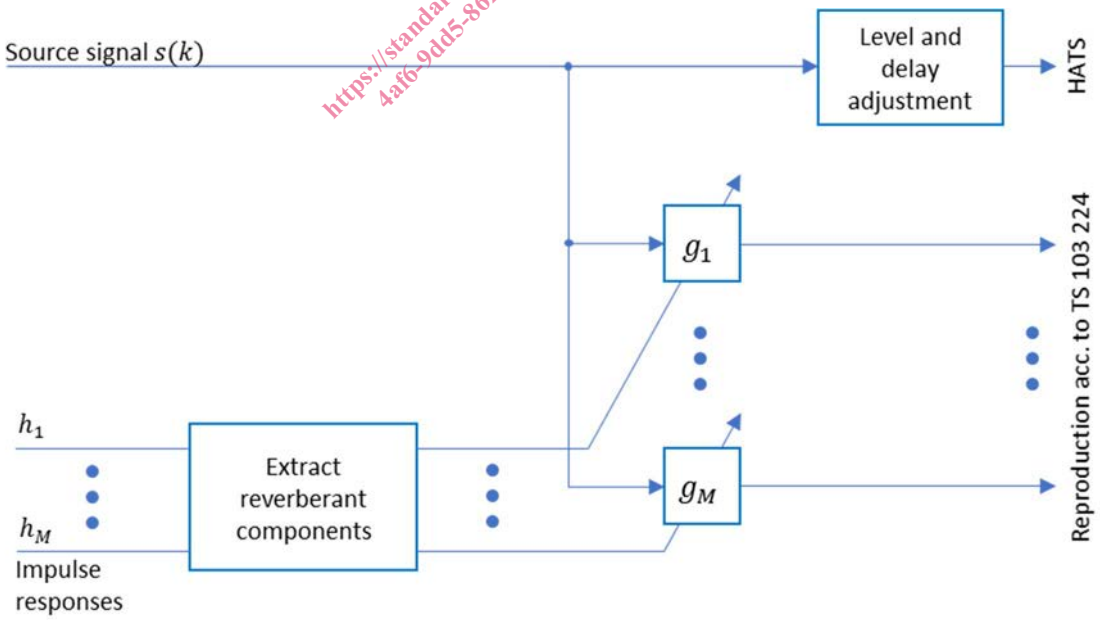

Figure 1: Overview of the signal processing for the reproduction setup

For the fixed microphone setup, the impulse responses can be taken directly from the provided database. For the flexible microphone setup, individual measurements have to be carried out to use the reproduction system.

5.2.2 Preparations

5.2.2.1 Separation of impulse responses

The reproduction system needs two signal parts: the direct sound (single channel that is played over the HATS) and the reverberant components (eight target signals that are fed into the reproduction system according to ETSI TS 103 224 [1]). An example impulse response $h_i(k)$ is shown in Figure 2 with the direct path component $h_{D,i}(k)$ in orange and the remaining reverberant components $g_i(k)$ in blue. These two components constitute the entire impulse response according to:

$$h_i(k) = h_{D,i}(k) + g_i(k)$$

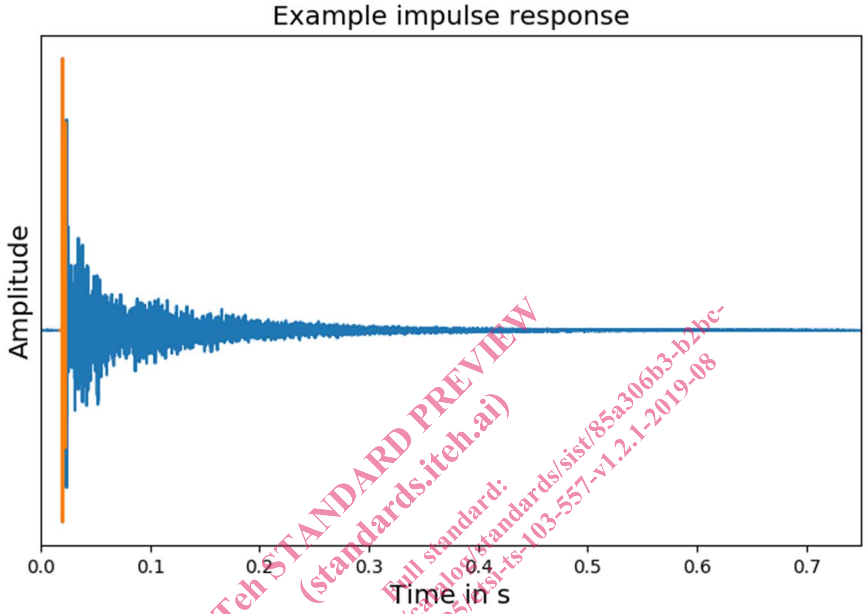

Figure 2: Time domain representation of an example impulse response

An overview of the necessary signal processing is depicted in Figure 5. From the measured impulse responses $h_1(k)$ ... $h_M(k)$, the reverberant components $g_1(k)$ ... $g_M(k)$ are extracted by removing the first few milliseconds of the impulse response, i.e. the direct path. These reverberation filters $g_1(k)$ ... $g_M(k)$ are then used to calculate the input signals for the reproduction system according to ETSI TS 103 224 [1].

The extraction of the direct path from the impulse response is done by searching for the largest absolute amplitude in the impulse response and selecting a window of ±2,5 ms around this position. To avoid signal processing artefacts, squared sinusoidal sections shall be used for fading in and out. The entire 5 ms section shall have 0,25 ms of squared sine fade-in in the beginning and 0,25 ms of squared sine fade-out in the end.

The resulting window function is depicted in Figure 3 and an enlarged view of the fade-in is presented in Figure 4. If the largest amplitude in the impulse response is closer than 2,5 ms to the beginning of the impulse response, the fade-in shall be omitted and the first part of the impulse response (from the start to 2,5 ms after the largest amplitude) shall be used for the direct path. The remainder of the impulse response contains all the reverberant components.

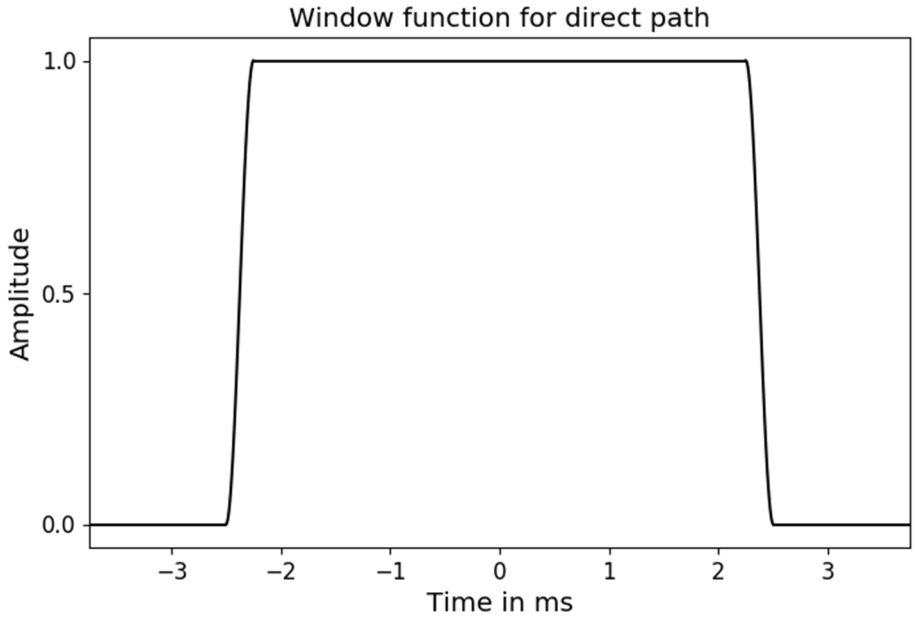

Figure 3: Window function for extracting the direct path from the impulse responses (time scale relative to maximum position of impulse response)

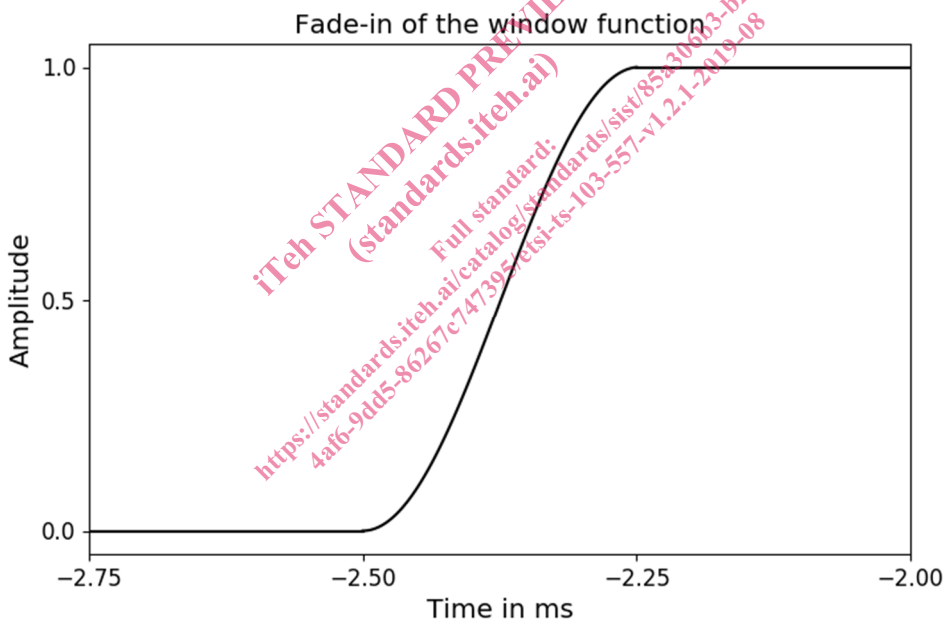

Figure 4: Fade-in of the window function

The remainder of the impulse response contains all the reverberant components.

All channels of the impulse response shall be separated according to the given procedure. As described, the direct paths are used for determining the delay and level differences between the two system components. For calculating the reproduction targets for the reproduction system according to ETSI TS 103 224 [1], however, the reverberant components are used.

5.2.2.2 Delay and level adjustment

For the aforementioned level and delay compensation, the signal from the artificial mouth needs to be adjusted to the signal from the reproduction system. This shall be achieved by comparing the level of and the delay between two signals. A linear or logarithmic sweep signal of at least 2 s covering a frequency range from $\leq$ 100 Hz to $\geq$ 4 000 Hz shall be both played by the artificial mouth and convoluted with the direct path components to get target signals for the reproduction system.