# ETSI TS 104 063 V1.1.1 (2025-07)

TECHNICAL SPECIFICATION

# Speech and multimedia Transmission Quality (STQ); A nonlinearity measure for distortion analysis of speech communication terminals

Reference
DTS/STQ-314

Keywords
acoustic, distortion, electro-acoustic,
measurement, speech, terminal, transducer

***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 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

# Contents

# Intellectual Property Rights

## Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are 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 IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, 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.

**DECT™**, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

# Foreword

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

# 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

The method specified in the present document is intended to replace distortion tests in test specifications for speech communication terminals that are based on artificial test signals like e.g. sine tones, noise bursts and/or sweep signals. These traditional measures have been used for decades in almost any test specification for speech communication terminals and originate from the electro-acoustic domain. The common basis for all these metrics is the assumption that devices under test are basically linear and time-invariant systems (like e.g. microphones or loudspeakers).

However, due to modern and/or AI-based signal/audio/speech processing and transmission codecs, this assumption is not valid anymore nowadays. In consequence, measurement results obtained with these conventional measurement methods are becoming increasingly unreliable as technology advances.

The present document introduces an alternative analysis method based on spectral coherence to quantify non-linear distortions that are introduced by e.g. electro-acoustic transducers (loudspeakers/microphones) of speech communication terminals in sending and receiving direction. It provides results that are comparable to traditional distortion metrics like e.g. THD or THD+N, but without the limitations of specific and artificial test signals.

The analysis method provides distortion energy versus frequency as a result, which can be aggregated to a single overall measure. The present document also provides guidance on how this measure is correlated with results from traditional distortion measurements.

# 1 Scope

The present document specifies a signal-based analysis method and overall measure for linearity of electrical or acoustical recordings that were transmitted via speech communication terminals. The analysis considers the influence of the full transmission paths (talker to POI and POI to listener), including acoustical paths and components like signal processing and/or speech/audio codecs.

The method specified in the present document is intended to be used for (but not limited to) communication terminals, which are optimized for the transmission of speech/audio signals - but may at the same time significantly distort or even suppress artificial test signals like e.g. sweeps, noise bursts or single-/multi-sine tones, as these are identified as irrelevant or even unwanted signal components. For this reason, source signals that do not correspond to the envisioned use case of the terminal are out of scope.

The method specified in the present document is intended to be used mainly as a technical measure and in a similar way as traditional distortion measures. These assume a mostly linear and time-invariant behaviour of the device under test and are typically used for e.g. detecting defective transducers (e.g. loudspeakers or microphones). Even though a certain correlation can be anticipated, any relation to perceptual quality (independent if assessed auditorily and/or instrumentally) is out of scope.

The specified method requires a sufficiently low idle noise floor in the analysed signals. Usage under ambient noise conditions is out of scope.

# 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 in the ETSI docbox.

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] Recommendation ITU-T P.10 (05/2024): "Vocabulary for performance, quality of service and quality of experience".

[2] Recommendation ITU-T P.501 (04/2025): "Test signals for use in telephony and other speech-based applications".

[3] Recommendation ITU-T P.56 (12/2011): "Objective measurement of active speech level".

[4] Welch P. D.: "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", IEEE Transactions on Audio and Electroacoustics, Volume AU-15 (2), pages 70–73, June 1967.

[5] D'Antona Gabriele and A. Ferrero: "Digital Signal Processing for Measurement Systems", 2006, page 70.

[6] Ramaswamy Sivaramakrishnan: "Frequency sampling digital filters for multirate applications and pipelined implementations" (1994). University of Nevada, Las Vegas, Retrospective Theses & Dissertations. 426.

[7] Steven W. Smith: "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1997.

[8] Alan V. Oppenheim and Ronald W. Schafer: "Discrete-Time Signal Processing" (1989, third edition).

[9] Recommendation ITU-T P.58 (03/2023): "Head and torso simulator for telephonometry".

[10] Recommendation ITU-T P.581 (07/2022): "Use of head and torso simulator for hands-free and handset terminal testing".

## 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 may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] ETSI TS 126 071 (V18.0.0): "Mandatory speech codec speech processing functions; AMR speech codec; General description".

[i.2] ETSI TS 126 171 (V18.0.0): "Speech codec speech processing functions; Adaptive Multi-Rate – Wideband (AMR-WB) speech codec; General description".

[i.3] ETSI TS 126 441 (V18.0.0): "Codec for Enhanced Voice Services (EVS); General Overview".

# 3 Definition of terms, symbols and abbreviations

## 3.1 Terms

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

**distortion:** undesired modification of a signal's original waveform or the relationship between its frequency components, typically resulting in a degradation of the signal

**intermodulation distortion:** special type of distortion occurring when two or more signals of different frequencies pass through a non-linear system, creating frequency components that are the sum and difference of the original ones

**harmonic distortion:** addition of overtones to an audio signal that were not present in the original input signal, and which occur at integer multiples of the original signal's fundamental frequency, typically due to non-linear effects within a system or device

**total harmonic distortion:** measurement of the harmonic distortion in a signal, defined as the ratio of the sum of the powers of a certain number of harmonic components to the power of the fundamental frequency

## 3.2 Symbols

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

| | |
|---|---|
| $C_{XY}(f_j)$ | Spectral coherence |
| $\varepsilon_{min}$ | threshold to avoid division by zero |
| $i$ | segment index of a STFT calculation |
| $i_A$ | segment index of a STFT calculation considered as active |
| $j$ | frequency index |
| $f_j$ | frequency at index j of a STFT calculation |
| $f'_j$ | frequency at index j of a STFT calculation considered as coherent |
| $f_{min}$ | minimum frequency of energy integration range |

| | |
|---|---|
| $f_{max}$ | maximum frequency of energy integration range |
| $\mathfrak{F}(\cdot)$ | function to be applied to obtain non-linear component of any signal |
| $h(k)$ | real impulse response between $x(k)$ and $y(k)$ |
| $\tilde{h}(k)$ | estimate of $h(k)$ |
| $\tilde{H}_0(f'_j)$ | initial estimate of $\tilde{h}(k)$ |
| $\tilde{H}_1(f'_j)$ | final estimate of $\tilde{h}(k)$ |
| $L$ | number of (overlapping) segments of a STFT calculation |
| $L_{lin}$ | average level of linear signal component |
| $L_{nl}$ | average level of non-linear signal component |
| $M_{nl}$ | Non-linearity measure (in dB) |
| $n(k)$ | noise component of $y(k)$ |
| $N$ | number of samples in a segment of a STFT calculation |
| $\tilde{N}(f_j)$ | Estimated spectrum of $n(k)$ averaged versus time |
| $P_{XX}(f_j)$ | power spectral density of $x(k)$ averaged versus time |
| $P_{XX}(i, f_j)$ | power spectral density of $x(k)$ versus time |
| $P_{XY}(f_j)$ | cross-power spectral density between $x(k)$ and $y(k)$ averaged versus time |
| $P_{XY}(i, f_j)$ | cross-power spectral density between $x(k)$ and $y(k)$ versus time |
| $P_{YY}(f_j)$ | power spectral density of $y(k)$ averaged versus time |
| $P_{YY}(i, f_j)$ | power spectral density of $y(k)$ versus time |
| $r_{act}$ | activity ratio |
| $r_{inact}$ | inactivity ratio |
| $T_C$ | coherence threshold (in %) |
| $x(k)$ | source signal used for the measurement and/or calculation of the non-linearity measure |
| $x'(k)$ | source signal used for the measurement but not for the calculation of the non-linearity measure |
| $y(k)$ | acoustically /electrically measured or otherwise generated signal that may be subject to non-linear transmission |
| $\boldsymbol{y}_{L,R}(k)$ | binaurally recorded or otherwise generated version of $y(k)$ containing left and right ear signals |
| $\tilde{y}_{lin}(k)$ | Linear component of $y(k)$ |
| $X(i, f_j)$ | STFT representation of $x(k)$ |
| $Y(i, f_j)$ | STFT representation of $y(k)$ |
| $\bar{Y}(f_j)$ | Averaged spectrum versus time of $y(k)$ |
| $\bar{Y}_{lin}(f_j)$ | Estimate of the linear frequency component of $y(k)$ averaged over time |
| $\bar{\tilde{Y}}_{nl}(f_j)$ | Estimate of the non-linear frequency component of $y(k)$ averaged over time |

## 3.3 Abbreviations

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

| | |
|---|---|
| A/D | Analogue-to-Digital |
| AEC | Acoustic Echo Cancellation |
| AGC | Automatic Gain Control |
| AI | Artificial Intelligence |
| AMR | Adaptive Multi-Rate |
| AMR-WB | Adaptive Multi-Rate for Wideband |
| ASL | Active Speech Level |
| CPSD | Cross-Power Spectral Density |
| D/A | Digital-to-Analogue |
| DRP | Drum Reference Point |
| DSP | Digital Signal Processing |
| DUT | Device Under Test |
| EVS | Enhanced Voice Service |
| FB | Fullband |
| FIR | Finite Impulse Response |
| FSM | Frequency Sampling Method |
| HATS | Head And Torso Simulator |
| LTI | Linear and Time-Invariant |
| MLS | Maximum-Length Sequence |
| MRP | Mouth Reference Point |

| | |
|---|---|
| NB | Narrowband |
| NLM | Non-Linearity Measure |
| NR | Noise Reduction |
| PLC | Packet Loss Concealment |
| POI | Point Of Interconnect |
| PSD | Power Spectral Density |
| STFT | Short-Time Fourier Transformation |
| SWB | Super-Wideband |
| THD | Total Harmonic Distortion |
| WB | Wideband |

# 4 Distortions in speech communication terminals

## 4.1 Overview

Historically, the analysis of distortions in communication terminals focused on the electro-acoustic components, namely the microphone and the loudspeaker. These transducers were the primary sources of non-linearity in otherwise largely analogue and linear systems. The distortions they introduced, such as harmonic and intermodulation distortion, could be reliably characterized using simple, artificial test signals like sine tones. The assumption was that the Device Under Test (DUT) behaved as a Linear and Time-Invariant (LTI) system, where any deviation from this ideal could be quantified.

This LTI assumption is no longer valid for modern speech communication terminals. The signal path in today's devices, from mobile phones to conferencing systems, is dominated by sophisticated and highly non-linear Digital Signal Processing (DSP). Components such as noise suppressors, acoustic echo cancellers, automatic gain controllers, and advanced speech codecs are designed to be adaptive and time-variant. They actively modify the signal based on its content and the surrounding acoustic environment.

While these processes are intended to enhance the perceived quality and intelligibility of speech, they can introduce complex distortions that bear little resemblance to the classic harmonic distortions of analogue systems. These modern distortions are often highly signal-dependent and transient, making them difficult to quantify with traditional metrics that rely on static, artificial signals. Consequently, a comprehensive understanding of distortion in contemporary terminals requires considering the entire processing chain, as detailed in the following clauses.

## 4.2 Signal Processing

Modern terminals apply various real-time signal processing algorithms to enhance speech intelligibility and suppress unwanted signals. While these techniques serve important functions, they often introduce side effects that may be perceived as distortions:

- Automatic Gain Control (AGC): AGC attempts to normalize signal levels but may cause unnatural fluctuations in signal amplitude. Sudden changes in gain, particularly in response to non-speech events, can lead to distortions that are difficult to detect using linear analysis methods.
- Noise Reduction (NR): NR algorithms attenuate background noise, typically based on spectral subtraction or statistical modelling. However, aggressive noise suppression may introduce musical noise artifacts, modulated background sounds, or speech component suppression - particularly during transient speech segments or in fluctuating noise environments.
- Acoustic Echo Cancellation (AEC): AEC systems are designed to suppress acoustic echoes, particularly in hands-free setups. Imperfect modelling or adaptation delays may lead to residual echo artifacts, double-talk suppression effects, or partial speech dropouts.
- Limiter and Clipping Effects: In scenarios with high signal levels or insufficient dynamic range, signal clipping may occur. This results in waveform distortion, which introduces high-frequency energy and harmonic components unrelated to the speech signal.
- Time-Variant Processing: Many algorithms operate in a non-linear and time-variant manner, making the distortion pattern dependent on context, input characteristics, and prior signal history. This dynamic behaviour cannot be fully characterized using static test signals.

These processing-induced distortions are often non-stationary, occur intermittently, and exhibit complex spectral-temporal patterns, which complicates traditional linear analysis.

## 4.3 Codec

The integration of speech and audio codecs is a fundamental part of modern digital communication systems. While codecs are essential for efficient transmission, they inherently introduce quantization noise, bandwidth limitations, and signal reconstruction errors:

- Influence of speech codecs: Speech codecs such as AMR [i.1], AMR-WB [i.2], or EVS [i.3] operate at bitrates where perceptually irrelevant information is discarded. At low bitrates, the signal reconstruction is coarse, and artifacts such as pre-echo, temporal smearing, or spectral warping may appear. In addition, many speech codecs use codebook structures for synthesis. The reconstruction accuracy depends on the selected excitation vector and filter coefficients, leading to signal-dependent distortions that are often non-linear in nature.
- Frame Loss and Packet Loss Concealment (PLC): In network-based communication, frame or packet loss can occur. PLC algorithms aim to reconstruct the missing information using prediction or interpolation, but these approximations may introduce synthetic or repetitive distortions.
- Bandwidth Mismatch and Filtering: When the codec input or output bandwidth differs from the nominal transmission range (e.g. sending wideband speech through a narrowband channel), additional spectral distortion or aliasing artifacts may result.
- Bitrate and Complexity Trade-offs: In low-complexity devices, simplified codec implementations may be used. These may compromise fidelity to reduce computational cost, introducing further distortions not present in the reference implementations.

Overall, codec-induced distortions are tightly coupled to the speech signal characteristics and exhibit non-linear, signal-dependent behaviour. Moreover, they often affect both spectral and temporal properties of the signal, making them difficult to isolate using traditional linear system assumptions.

# 5 Non-linearity measure

## 5.1 Overview & conventions

In general, the Non-Linearity Measure (NLM) is based on a source signal $x(k)$, which is fed acoustically or electrically into a terminal, device or system under test. The signal is transmitted through one or multiple linear or non-linear components/paths, which could be either variant or invariant versus time. The signal recorded at the output is denoted as $y(k)$ and may contain an unknown degree of degradations/distortions compared to the source signal. In addition, $y(k)$ is also subject to additive noise $n(k)$ (e.g. ambient noise of test room, idle noise introduced by D/A or A/D converters, etc.), which is assumed to be uncorrelated to the input signal $x(k)$.

NLM is based on the signal model as defined in equation (1) and illustrated in Figure 1, which assumes that $y(k)$ is composed of a linear component $y_{\mathrm{lin}}(k)$ (corresponding to a convolution of $x(k)$ and an impulse response $h(k)$, i.e. a FIR filter), a non-linear component $y_{\mathrm{nl}}(k)$ (corresponding to a non-linear function $\mathfrak{F}(x)$ applied on $x(k)$) and the aforementioned noise signal $n(k)$.

$$y(k) = y_{\mathrm{lin}}(k) + y_{\mathrm{nl}}(k) + n(k) = x(k) * h(k) + \mathfrak{F}(x(k)) + n(k) \quad (1)$$

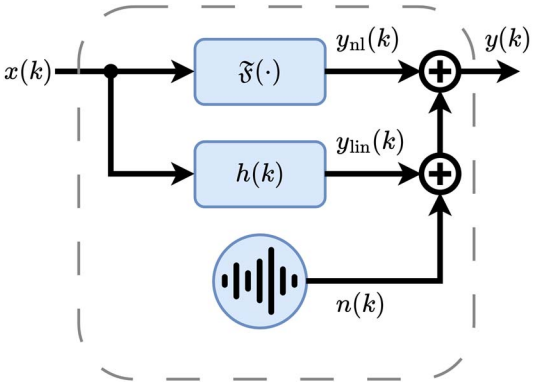


**Figure 1: Signal model of the non-linearity measure**

NOTE: The NLM analysis method is designed to quantify non-linear signal components in a similar way as traditional distortion measurements, which focus on harmonic components introduced by electro-acoustic transducers. However, for the estimate of the non-linear component, the non-linear function $\mathfrak{F}(x)$ is not modelled by any prototype, kernel or basis function(s), as in approaches like e.g. ETSI TS 126 171 [i.2] or ETSI TS 126 441 [i.3].

The analysis method and corresponding parameters specified in the following clauses assume that signals are sampled at 48 kHz. In general, the analysis method can be applied also with different sampling rates, which requires that several values like e.g. frame sizes, have to be adapted.

Several calculation steps in the following clauses utilize a certain analysis range $f_{\mathrm{min}}$ to $f_{\mathrm{max}}$, which should be adapted to the audio bandwidth used by the transmission system during the recording of $y(k)$. Typically, the bandwidth definitions for NB, WB, SWB and FB telephony according to Recommendation ITU-T P.10 [1] are used in such cases. However, since several sub-analyses in the following (like e.g. coherence or impulse response estimation) are sensitive for low signal energy content at very low and/or high frequencies, it is recommended to use slightly decreased frequency ranges for some bandwidths, as exemplarily shown in Table 1.

**Table 1: Frequency ranges**

| Bandwidth | Nominal audio bandwidths [i.1] | | Recommend audio bandwidths | |
|---|---|---|---|---|
| | $f_{\mathrm{min}}$ [Hz] | $f_{\mathrm{max}}$ [Hz] | $f_{\mathrm{min}}$ [Hz] | $f_{\mathrm{max}}$ [Hz] |
| Narrowband (NB) | 300 | 3 400 | 300 | 3 400 |
| Wideband (WB) | 100 | 7 000 | 100 | 7 000 |
| Super-wideband (SWB) | 50 | 14 000 | 100 | 12 000 |
| Fullband (FB) | 20 | 20 000 | 100 | 12 000 |

NLM results shall always be stated in the test report together with the values chosen for $f_{\mathrm{min}}$ and $f_{\mathrm{max}}$.

## 5.2 Preparation and preprocessing

The following requirements and recommendations apply for the source signal $x(k)$ used for testing:

- The signal-to-noise ratio of the source signal $x(k)$ shall be equal to or greater than the signal-to-noise ratio of the British English speech sequence specified in clause 7.3 of Recommendation ITU-T P.501 [2].
- The source signal $x(k)$ should contain speech activity of at least 60 %, measured according to Recommendation ITU-T P.56 [3].
- The source signal $x(k)$ shall contain at least 2,0 s of active speech.
- The source signal $x(k)$ shall provide at least super-wideband audio bandwidth (fullband is recommended). The nominal audio bandwidth of the recording $y(k)$ should be at least narrowband. For definitions of audio bandwidths, see Recommendation ITU-T P.10 [1].

Speech signals from Recommendation P.501 [2] listed in clauses 7.3, 7.4, Annexes C and D are recommended to use, as these already meet the requirements specified above.

NOTE: The usage of the non-linearity measure with signal types other than speech might be possible but is for further study.

Before inserting source signal $x(k)$ and measured signal $y(k)$ into the analysis method, several recommended and required pre-processing steps have to be conducted:

- The average level of the recording $y(k)$ shall be at least 40 dB above the idle noise level.
- The delay between $x(k)$ and $y(k)$ shall be minimized and within a margin of ±1 ms.
- $x(k)$ and $y(k)$ shall contain (or zero-padded to) the same number of samples after compensating delay.

The pre-processing steps described above are not part of the analysis and out of scope of the present document.

## 5.3 Activity detection

Since $x(k)$ (and in consequence also $y(k)$) could possibly contain longer inactive/silence segments, these need to be excluded for some analyses in the following clauses. This is achieved with classification into active and inactive level-vs-time instances.

In a first step, the Short-Time Fourier Transformation (STFT) is calculated for $x(k)$ leading to $X(i, f_j)$. The parameters and window function for this transformation are provided in Table 2.

**Table 2: STFT parameter for estimation of impulse response**

| Parameter | Value |
|---|---|
| Window function | Flat top [5] |
| Number of samples per frame | 16 384 |
| Overlap | 97,5 % |

For the classification of active/inactive frames, the Active Speech Level (ASL) according to Recommendation ITU-T P.56 [3] is calculated for source signal $x(k)$. Depending on the audio bandwidth, the input bandpass filter according to Annex B of [3] for WB or SWB or Annex C of [3] for FB shall be used for the ASL calculation.

The ASL analysis also provides an activity ratio $r_{\mathrm{act}}$ (in %) as an additional result value. A corresponding inactivity ratio $r_{\mathrm{inact}}$ (in %) is defined in equation (2).

$$r_{\mathrm{inact}} = 100\,\% - r_{\mathrm{act}} \tag{2}$$

Then, the power spectral density (PSD) $P_{XX}(i, f_j)$ versus time is calculated for $X(i, f_j)$ according to equation (3).

$$P_{XX}(i, f_j) = \frac{1}{N} X(i, f_j)\, X^*(i, f_j) \tag{3}$$

Where:

- $\mathrm{X}(i, f_j)$ is the STFTs of the *i*-th segment of $x(k)$.
- $X^*(i, f_j)$ is the complex conjugate of $X(i, f_j)$.
- *N* is the number of points in the segment.

Based on the PSD, a level-vs-time representation $L_{XX}(i)$ is calculated according to equation (4).

$$L_{XX}(i) = \sum_{j=0}^{M} P_{XX}(i, f_j) \tag{4}$$

Next, activity threshold $LT_{\mathrm{act}}$ is determined using $r_{\mathrm{inact}}$ as a percentile according to equation (5).

$$LT_{\mathrm{act}} = \underset{i}{\mathrm{percentile}}(L_{XX}(i),\ r_{\mathrm{inact}}) \tag{5}$$

Finally, all frames $i_A$ that are considered as active speech can be defined as per equation (6).

$$i_A = \{\, i \mid L_{XX}(i) > LT_{\mathrm{act}} \,\} \tag{6}$$

An example result of the activity frame classification is illustrated in Figure 2, for which the file *P501_D_EN_fm_SWB_48k.wav* from Annex D of Recommendation ITU-T P.501 [2] was used (ASL = -26 dBov, $r_{\mathrm{act}} \approx 80$ %).

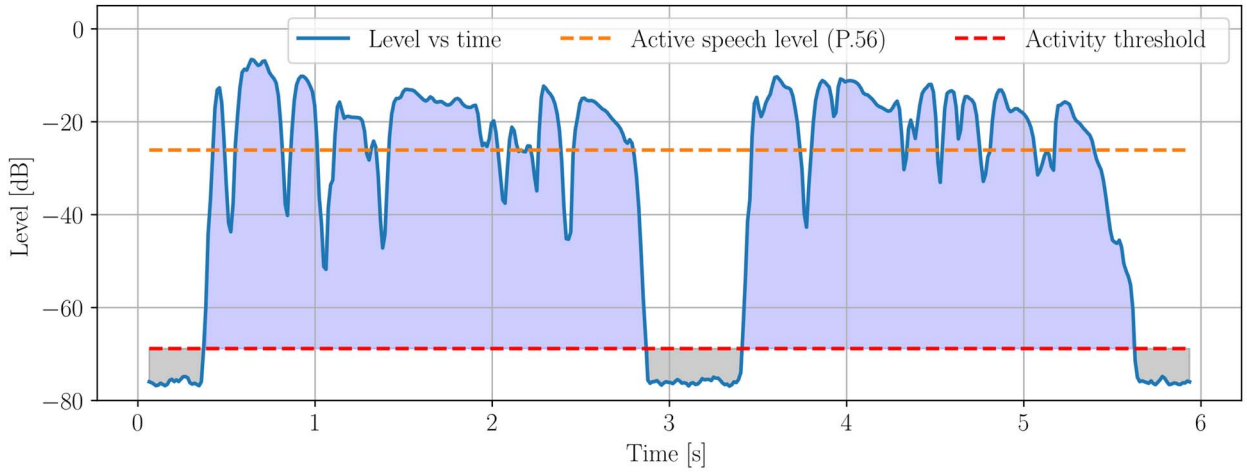


**Figure 2: Example result of activity detection**

## 5.4 Estimate of impulse response

The first step of the analysis is to estimate an impulse response $\tilde{h}(k)$ based on the available signals $x(k)$ and $y(k)$. The calculation method is illustrated in Figure 3.

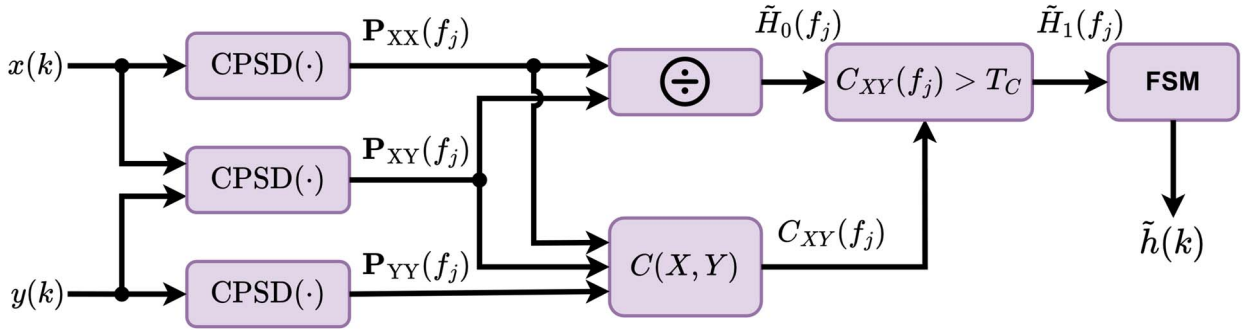


**Figure 3: Estimation of impulse response $\tilde{h}(k)$**

In addition to the STFT $X(i, f_j)$ as already determined in clause 5.3, $Y(i, f_j)$ is calculated accordingly for the measured signal $y(k)$. Welch's averaged, modified periodogram method [4] is then used to calculate cross-power spectral density (CPSD) $P_{XY}(f_j)$ for $X(i_A, f_j)$ and $Y(i_A, f_j)$ according to equation (7).

$$P_{XY}(f_j) = \frac{1}{L_A} \sum_{i_A} \frac{1}{N} X(i_A, f_j) Y^*(i_A, f_j) \tag{7}$$

Where:

- $X(i_A, f_j)$ and $Y(i_A, f_j)$ are the STFTs of the $i_A$-th active segment of signals $x(k)$ and $y(k)$.
- $Y^*(i_A, f_j)$ is the complex conjugate of $Y(i_A, f_j)$.
- $N$ is the number of points in the segment.
- $L_A$ is the number of active frames / overlapping segments.

Accordingly, power spectral densities (PSDs) $P_{XX}(f_j)$ and $P_{YY}(f_j)$ are calculated. Note that in contrast to CPSD, the two PSDs result in real-valued spectra.

A first estimate of the complex transfer function $\tilde{H}_0(f_j)$ is calculated according to equation (8).

$$\tilde{H}_0(f_j) = {P_{XY}(f_j)} \Big/ {\max\left(\varepsilon_{\min}, P_{XX}(f_j)\right)} \tag{8}$$

Where:

- $\varepsilon_{\min}$ is a minimum threshold applied to avoid division by very small values and corresponds to -120 dB.

In order to remove possibly unreliable magnitude values in $\tilde{H}_0(f_j)$, the spectral coherence $C_{XY}(f_j)$ is calculated according to equation (9).

$$C_{XY}(f_j) = {{P_{XY}}^2(f_j)} \Big/ {\max\left(\varepsilon_{\min}^2,\ P_{XX}(f_j)\, P_{YY}(f_j)\right)} \tag{9}$$

Only frequency bins $f'_j$ providing a coherence more than the threshold $T_C$ = 2,5 % are considered in the modified and real-valued transfer function $\tilde{H}_1$, as given by equation (10).

$$\tilde{H}_1(f'_j) = \left|\tilde{H}_0\left(\arg_{f_j} C_{XY}(f_j) \geq T_C\right)\right| \tag{10}$$

Within the specified frequency range, at least 25 % of all frequency bins should be above the threshold $T_C$. If $T_C$ is not met, the calculation steps in the following might still succeed, but could lead to inaccurate overall result.

The impulse response estimate $\tilde{h}(k)$ is determined by the Frequency Sampling Method (FSM) as described in [6], [7] and [8], which designs an FIR filter by specifying the desired magnitude response at discrete frequency points. The filter coefficients are computed by interpolating between these points, ensuring linear phase and performing an inverse Fourier transform to obtain the impulse response.

FSM is applied on the magnitude of $\tilde{H}_1(f'_j)$ and shall only consider frequencies between $f_{\min}$ and $f_{\max}$ (including) in the spectral interpolation step. The impulse response $\tilde{h}(k)$ shall be estimated for 8 192 coefficients.

## 5.5 Estimate of linear signal component

The estimate of the linear component $\tilde{y}_{\text{lin}}(k)$ is calculated according to equation (11) as the convolution of the source signal $x(k)$ with the impulse response estimate $\tilde{h}(k)$.

$$\tilde{y}_{\text{lin}}(k) = \tilde{h}(k) * x(k) \tag{11}$$

## 5.6 Estimate of spectral noise component

For the measured signal $y(k)$, the STFT is calculated, leading to $Y(i, f_j)$. The parameters and window function for this transformation are provided in Table 3.

**Table 3: STFT parameter for spectral analysis**

| Parameter | Value |
|---|---|
| Window function | Hann |
| Number of samples per frame | 16 384 |
| Overlap | 75 % |

The magnitude of the estimated noise component in the frequency domain $\tilde{N}(f_j)$ is calculated as the 3 %-percentile range versus time from the magnitude of the measured spectrum $Y(i, f_j)$ according to equation (12).

$$\tilde{N}(f_j) = \underset{i}{\text{percentile}}\left(\left|Y(i, f_j)\right|, 3\ \%\right) \tag{12}$$

NOTE: The percentile analysis performs a noise estimate based on the lower/lowest magnitude values in the spectrum. It is thus assumed that idle noise can be observed in at least 3 % of the time per frequency, which depends on the signal used for testing. These idle ranges can either be realized individually per frequency band, i.e. by ensuring a sufficient amount of inactive time instances for each frequency - or globally, by ensuring speech pauses of suitable duration.

## 5.7 Spectral analysis

To estimate the average non-linear signal component $\tilde{\bar{Y}}_{\mathrm{nl}}(f_j)$, first the STFT is calculated for estimate of the linear component $\tilde{y}_{\mathrm{lin}}(k)$, leading to $Y_{\mathrm{lin}}(i, f_j)$. Then the average magnitudes of the measured $\bar{Y}(f_j)$ and the linear spectrum $\bar{Y}_{\mathrm{lin}}(f_j)$ are calculated according to equations (13) and (14). The parameters and window function for the STFT are provided in Table 3.

$$\bar{Y}(f_j) = \left|\frac{1}{L}\sum_{i=0}^{L} Y(i, f_j)\right| \tag{13}$$

$$\bar{Y}_{\mathrm{lin}}(f_j) = \min\left(\max\left(\bar{Y}(f_j) - \tilde{N}(f_j),\ 0\right), \left|\frac{1}{L}\sum_{i=0}^{L} Y_{\mathrm{lin}}(i, f_j)\right|\right) \tag{14}$$

Finally, $\tilde{\bar{Y}}_{\mathrm{nl}}(f_j)$ is obtained by subtracting the estimates of the noise (see clause 5.6) and the linear component (see equation (14)) from the measured average spectrum $\bar{Y}(f_j)$, as shown in equation (15).

$$\tilde{\bar{Y}}_{\mathrm{nl}}(f_j) = \max\left(0,\ \bar{Y}(f_j) - \tilde{N}(f_j) - \bar{Y}_{\mathrm{lin}}(f_j)\right) \tag{15}$$

## 5.8 Aggregated non-linearity measure

The non-linearity measure $M_{\mathrm{nl}}$ is determined as the ratio of linear and non-linear signal component levels, calculated in the frequency range from $f_{\mathrm{min}}$ to $f_{\mathrm{max}}$ according to equations (16) to (18).

$$L_{\mathrm{lin}} = \sum_{f_j=f_{\mathrm{min}}}^{f_j=f_{\mathrm{max}}} \bar{Y}_{\mathrm{lin}}(f_j)^2 \tag{16}$$

$$L_{\mathrm{nl}} = \sum_{f_j=f_{\mathrm{min}}}^{f_j=f_{\mathrm{max}}} \tilde{\bar{Y}}_{\mathrm{nl}}^{\ 2}(f_j) \tag{17}$$

$$M_{\mathrm{nl}}\,[\mathrm{dB}] = 10\log_{10}\left({}^{L_{\mathrm{lin}}}/_{L_{\mathrm{nl}}}\right) \tag{18}$$

# 6 Application of the non-linearity measure

## 6.1 Overview

Since NLM is agnostic to the actual recording conditions, the following clauses provide additional guidelines for testing speech communication terminals, in particular in the context of test specifications, which typically refer to sending and/or receiving directions. The naming conventions for both directions are illustrated in Figure 4.

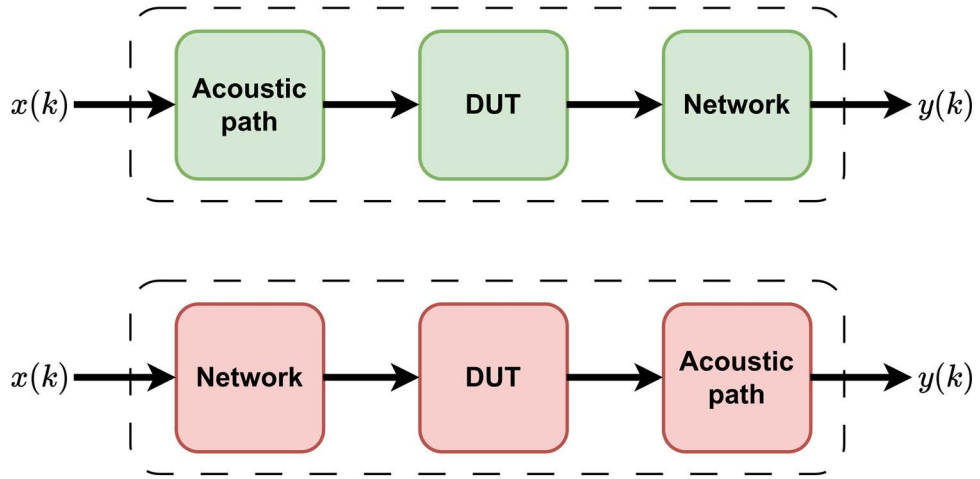


**Figure 4: Signals convention used for sending (green) and receiving (red) direction**

## 6.2 Sending direction

The recording procedure for speech communication devices and acoustical insertion in sending direction is illustrated in Figure 5. If not specified otherwise, a HATS equipped with an artificial mouth simulator according to Recommendation ITU-T P.58 [9] shall be used for the playback of the source signal $x(k)$, which corresponds to the equalized output at MRP. See Recommendation ITU-T P.581 [10] for use of HATS for testing specific device types/form factors (like e.g. handset or hands-free mode). The recording $y(k)$ is captured electrically at the POI.

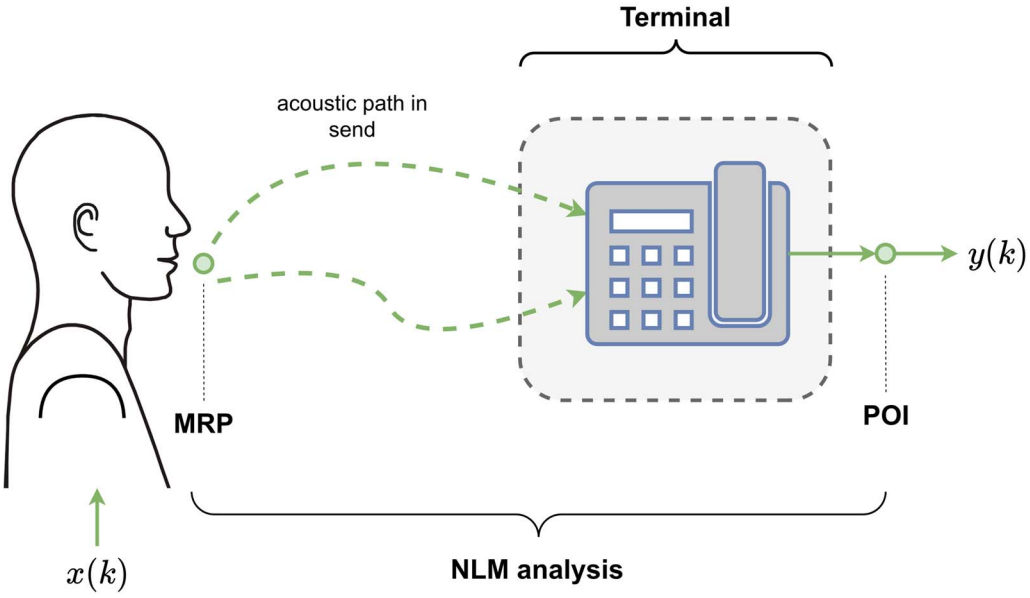


**Figure 5: Measurement in sending direction**

## 6.3 Receiving direction

The recording procedure for speech communication devices and electrical insertion in receiving direction is illustrated in Figure 6. If not specified otherwise, a HATS equipped with artificial ears according to Recommendation ITU-T P.58 [9] shall be used to obtain the signal $y(k)$, which is recorded at the DRP and corrected for diffuse-field. See Recommendation ITU-T P.581 [10] for use of HATS for testing specific device types/form factors (like e.g. handset or hands-free mode). Since the source signal $x(k)$ is expected to be a fullband signal, typically such test signals are pre-filtered and resampled to the audio bandwidth of a specific test setup, which might depend on e.g. terminal, codec or connection/network type. This intermediate signal $x'(k)$ is inserted electrically at the POI. However, for the calculation of NLM, the fullband source signal $x(k)$ shall always be used.

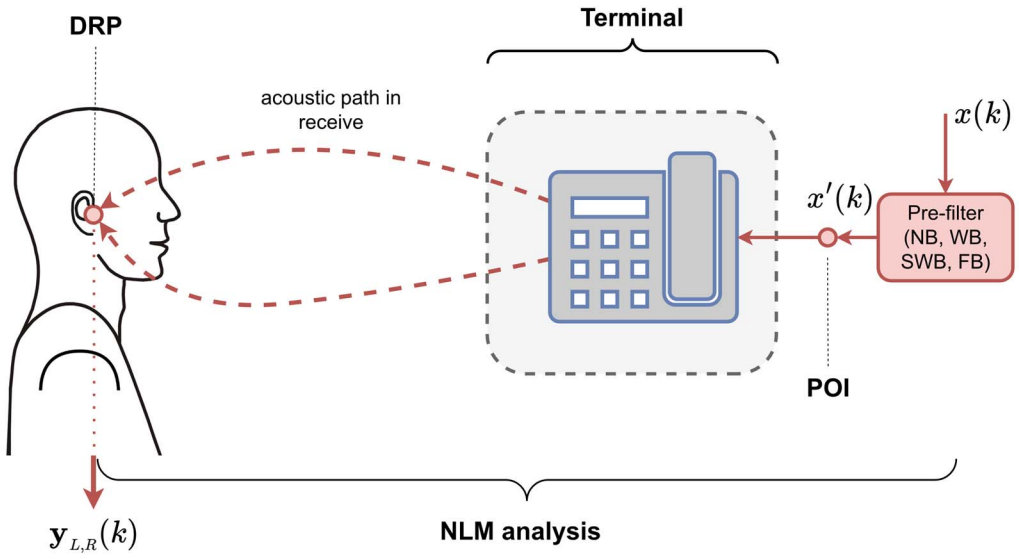


**Figure 6: Measurement in receiving direction**

Two types of recordings with different implications for calculating NLM might be obtained in receiving direction:

1) Monaural signals (captured at a measurement microphone or single ear with e.g. handsets or monaural headsets) shall be analysed and reported in the same way as in sending direction.

2) Binaural signals (captured with two ears, e.g. headset or hands-free mode) shall be analysed and reported individually for left and right ear. In case an indicator for overall performance is needed, the average of both ears may be used.

# Annex A (normative): Reference code

An example implementation of the non-linearity measure as specified in clause 5 is available at https://forge.etsi.org/rep/stq/ts104063-nonlinearity-measure.

The source code repository provides a version tag that corresponds to the version number of the present document. It also contains several examples to demonstrate the use of the non-linearity measure.

# Annex B (informative): Bibliography


- A. Farina: "Simultaneous measurement of impulse response and distortion with a swept-sine technique", 108th Audio Engineering Society Convention, 19-22 February 2000, Paris, France.
- A. Novak, P. Lotton, and L. Simon: "Synchronized Swept-Sine: Theory, Application, and Implementation", Journal of the Audio Engineering Society, Vol. 63(10), pp. 786-798, 2015.
- ETSI ES 202 739 (V1.8.1): "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user".
- ETSI ES 202 740 (V1.8.2): "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user."

# Annex C (informative): Change history

| Date | Version | Information about changes |
|---|---|---|
| 10/2024 | 0.0.1 | First draft |
| 02/2025 | 0.1.1 | Input to STQ#78 |
| 02/2025 | 0.1.2 | Edits/comments during STQ#78 |
| 06/2025 | 0.2.1 | Input to "STQ-Ad-hoc on TS 104 063" |
| 06/2025 | 0.3.1 | Input to STQ#79, uploaded for approval |
| 06/2025 | 0.3.2 | 1st revision during meeting |
| 07/2025 | 0.3.3 | 2nd revision during meeting, for approval |

# History

| Document history | | |
|---|---|---|
| V1.1.1 | July 2025 | Publication |
| | | |
| | | |
| | | |
| | | |