International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.700

(06/2019)

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

Measurements related to speech loudness

Calculation of loudness for speech communication

Recommendation ITU-T P.700

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** **P.700**
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.700

Calculation of loudness for speech communication

Summary

Recommendation ITU-T P.700 describes a unified method required for calculating loudness, allowing comparison of narrowband (NB) (300-3.4k Hz), wideband (WB) (100-8k Hz), super-wideband (SWB) (50-14k Hz) and fullband (FB) (10-20k Hz) telephony, for all types of terminals including handset, hands-free and conference terminals.

The model described in this Recommendation is consistent when switching from one bandwidth to another and independent of the listening situation (e.g., handset, headset, hands-free) with regards to producing a constant perceived loudness.

Compared to loudness rating (LR) models, like the one presented in ITU-T P.79, the present method predicts the *absolute* loudness, considers auditory masking and is applicable to a wide range of acoustic levels.

This Recommendation incorporates a number of annexes that hold test vectors for validation of loudness model implementations as well as of the descriptions and results of the loudness experiments that form the basis for this Recommendation.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.700	2019-06-29	12	11.1002/1000/13931

Keywords

Loudness, speech, telephony.

* 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

Introduction

The loudness of speech in communication systems is one of the main parameters relevant for good user experience in telephone calls. A loudness model predicts the loudness of speech perceived by the user, independent of the listing situation and independent of the type of terminal used. In contrast to loudness rating as defined in ITU-T P.79, the loudness described in Recommendation ITU-T P.700 focuses on the absolute loudness of speech and not on attenuation or amplification introduced by the various components involved in the transmission of speech from mouth to ear. Therefore, loudness does not replace loudness rating, but is used in a complementary manner to describe the perceptually perceived loudness. The loudness model has been validated on a range of conditions that include various types of signal processing used in terminals for which loudness ratings calculations have not been validated.

Recommendation ITU-T P.700

Calculation of loudness for speech communication

1 Scope

This Recommendation describes a loudness model which applies for all telephony audio bandwidths from narrowband (NB), wideband (WB), super-wideband (SWB) to fullband (FB), and for all types of terminals, including handset, headset, hands-free and conference terminals.

It is foreseen that the current ITU-T loudness rating (LR) model, as described in [ITU-T P.79] will be used in parallel with the loudness model given in this Recommendation for an extended period of time, since it is anticipated that network planning will continue to have a preference towards loudness rating calculations according to ITU-T P.79.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T G.160] Recommendation ITU-T G.160 (2012), *Voice enhancement devices*.

[ITU-T G.191] Recommendation ITU-T G.191 (2019), *Software tools for speech and audio coding standardization*.

[ITU-T P.48] Recommendation ITU-T P.48 (1988), *Specification for an intermediate reference system*.

[ITU-T P.56] Recommendation ITU-T P.56 (2011), *Objective measurement of active speech level*.

[ITU-T P.57] Recommendation ITU-T P.57 (2011), *Artificial ears*.

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

[ITU-T P.64] Recommendation ITU-T P.64 (2019), *Determination of sensitivity/frequency characteristics of local telephone systems*.

[ITU-T P.76] Recommendation ITU-T P.76 (1988), *Determination of loudness ratings; fundamental principles*.

[ITU-T P.78] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76*.

[ITU-T P.79] Recommendation ITU-T P.79 (2007), *Calculation of loudness ratings for telephone sets*.

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

[ITU-T P.381] Recommendation ITU-T P.381 (2017), *Technical requirements and test methods for the universal wired headset or headphone interface of digital mobile terminals*.

[ITU-T P.382] Recommendation ITU-T P.382 (2016), *Technical requirements and test methods for multi-microphone wired headset or headphone interfaces of digital wireless terminals*.

[ITU-T P.501] Recommendation ITU-T P.501 (2017), *Test signals for use in telephonometry*.

[ITU-T P.581] Recommendation ITU-T P.581 (2014), *Use of head and torso simulator for hands-free and handset terminal testing*.

[ITU-T P.830] Recommendation ITU-T P.380 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

[ITU-T P.1401] Recommendation ITU-T P.1401 (2012), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 loudness [b-ITU-T P.10]: Loudness belongs to a category of intensity sensations. Loudness is that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud. Loudness takes into account the spectral and temporal sensitivity of the human ear. Generally, masking effects in time and frequency are taken into account. The loudness level measure, according to [b-Zwicker], was created to characterize the loudness sensation. The loudness calculation procedure for stationary signals is defined [b-ISO 532]. For the calculation of the loudness of time variant signal different models are known.

3.1.2 phon [b-ITU-T P.10]: Loudness can also be expressed in phon, knowing that phon scale is equal to scale of dBSPL for a pure tone of 1000 Hz.

3.1.3 sone [b-ITU-T P.10]: Loudness is a subjective scale expressed in sone. By convention, the value of 1 sone is attributed to the loudness of a pure tone of frequency 1000 Hz at 40 dBSPL. Thus, a sound with loudness equal to 2 sone will be perceived with a "strength", or sensation, twice more important than a sound with a loudness of 1 sone.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

AGC Automatic Gain Control

ANOVA Analysis of Variance

DECT Digital Enhanced Cordless Telecommunications

DF Diffuse Field

DRP ear-Drum Reference Point

ERB Equivalent Rectangular Bandwidth

ERL Equal Reference Level

ERP Ear Reference Point

FB Fullband

FF	Free Field
FIR	Finite Impulse Response
HATS	Head and Torso Simulator
JLR	Junction Loudness Rating
LR	Loudness Rating
LTL	Long-Term Loudness
OLR	Overall Loudness Rating
NB	Narrowband
RL	Receiving Loudness
RLR	Receive Loudness Rating
RMSE	Root Mean Square Error
SLR	Sending Loudness Rating
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
STL	Short-Term Loudness
SWB	Super-Wideband
WB	Wideband

5 Conventions

The term P.Loudness is used as an abbreviation for the loudness prediction model that are defined in this Recommendation. This term is used extensively throughout the appendices of this Recommendation. The term is introduced to maintain clarity in the text, especially when comparisons with other loudness prediction models are performed.

6 Existing loudness model structures

All of the main existing loudness models are completely or partly based on the original Zwicker loudness model [b-Zwicker 2]. Zwicker proposed a sophisticated model that predicts loudness perception (in sone), not only as a function of intensity, but also depending on the spectral shape of a stationary sound, using findings from both physiological acoustics and psychoacoustics. This model takes into account the following facts:

- hearing threshold;
- the change in loudness with level;
- spectral masking of frequency components;
- the effect of spectral loudness summation.

All models use the same overall structure detailed in Figure 1. The general algorithm is summarized by the following steps: first, pre-filtering to account for outer and middle ear transmission, then construction of excitation patterns according to an auditory frequency scale (Bark/equivalent rectangular bandwidth (ERB)), next transformation of the excitation into specific loudness and finally summation of the specific loudness across the auditory frequency scale.

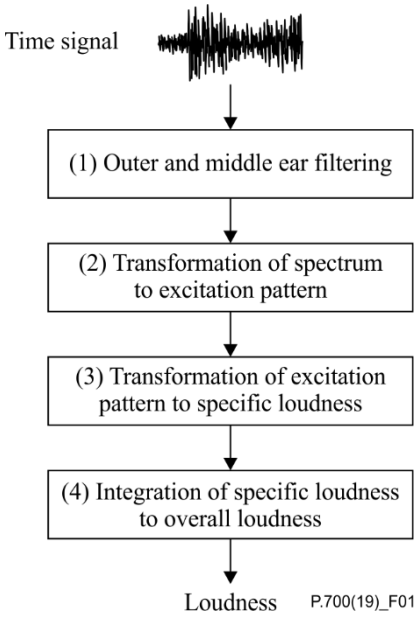


Figure 1 – General structure of a loudness model based on the model proposed by Zwicker

Depending on the type of signal being processed by the models, there exist two main families of models: the ones for stationary sounds and the ones for non-stationary sounds.

For the non-stationary sounds, the basic principle is the same as for stationary sounds. However, the models have been extended to better cope with time-varying sounds. This is done by modelling the post-masking effect and the temporal integration of loudness [b-Zwicker], [b-Widmann]; hence the loudness is calculated as a function of time and not in a global way.

In the following clauses the methods that have been considered during the development of this Recommendation are briefly explained.

6.1 Zwicker model for stationary sound

This is the original model of Zwicker [b-Zwicker 2], [b-Zwicker 3], also adopted as an international standard by ISO [b-ISO 532].

6.2 Moore and Glasberg model for stationary sound

Moore and Glasberg [b-Moore 1], [b-Moore 2], [b-Glasberg 1] have updated the Zwicker model to incorporate more recent findings in psychoacoustics. The three main differences with Zwicker model are: calculation of auditory filter, outer and middle ear filtering and excitation pattern calculation. The model was adopted in the American standard [b-ANSI S3.4-2007]. This model is denoted as **ANSI-S3.4** in the remainder of this Recommendation.

6.3 Zwicker and Fastl model for non-stationary sound

This model was elaborated by Zwicker and Fastl in 1999 [b-Zwicker 1]. In this model, a new stage was added to model how loudness varies with time and the post-masking effect. To derive a single value of loudness for the overall signal, Zwicker and Fastl recommend to use some statistical indicator such as N4, N5 or N7, which is the loudness value reached 4, 5 or 7% of the time, respectively. Zwicker recommend using N7 for speech signals [b-Zwicker].

6.4 Glasberg and Moore for non-stationary sound

This model was elaborated by Glasberg and Moore in 2002 [b-Glasberg 2]. The model was modified for stationary sound to get loudness as a function of time, called instantaneous loudness, and it would correspond to the overall activity inside the auditory nerve measured for a very short period of time.

Then, the short-term loudness (STL) was calculated from the instantaneous loudness by taking into account the temporal masking and the temporal integration. STL corresponds to the loudness perceived during a short segment of sound (a syllable for example). Derived also from the STL, the long-term loudness (LTL) which is used to describe the loudness sensations that are built rather slowly. According to the literature [b-Glasberg 2], the overall loudness could be calculated by considering the maximum of STL or the average of LTL. [b-Rennies 1] and [b-Rennies 2] showed that the averaged LTL better describes the speech loudness

7 Reasons for a new universal loudness model

There are several reasons why it is necessary to assess the loudness, and not only loudness rating for speech/audio terminals:

- the technology of speech and audiovisual/multimedia terminals has strongly evolved during recent years. In particular, most of this type of equipment implement speech enhancement functions which are nonlinear and time variant such as background noise cancellation, improved double talk behavior, automatic gain control (AGC), providing stereophonic and/or three dimensional (3D) acoustical outputs. The real speech signals or specific sound programs also impact the output level of the terminal, in particular on the level perceived by the user;
  - loudness ratings are determined according to [ITU-T P.79]. The algorithm was initially determined for narrowband speech and adaptations have been done for wideband speech. It is not the intention to develop such adaptations for super-wideband and fullband speech;
  - loudness rating is a fundamental parameter for the transmission plans and is determined with a reference input signal at a reference level;
  - the loudness rating is currently assessed on a long-term signal spectrum (currently 30 s of speech).
- some test signals (see [ITU-T P.501]) also give the possibility to compute loudness rating on a per utterance basis (in order to check the variation of LR over time), however there is no standardized approach about this use;
- even if the terminal provides binaural listening, the loudness rating is based on a monaural model;
- the use of loudness rating is a fundamental parameter for transmission planning, while absolute loudness measurement is of interest for other purposes.

8 Loudness model for receiving loudness

8.1 Recording of speech signals

For the calculation of receiving loudness, appropriately captured speech signals are indispensable. In order to ensure high-quality source material and reproducible results, only speech samples of clause 7.3 (including compressed speech), Annex B.3, Annex C or Annex D of [ITU T P.501] shall be used.

If not stated otherwise, the uncompressed, British English single-talk sequence described in clause 7.3 of [ITU-T P.501] shall be used.

For the recording of transmitted speech signals in handset and hands-free mode, a head and torso simulator (HATS) according to [ITU-T P.58] equipped with artificial ears of type 3.3 or 3.4 according to [ITU-T P.57] shall be used. Measurements of handsets shall follow the guidelines in [ITU-T P.64]. Measurements conducted in hands-free mode shall comply with the guidelines in [ITU-T P.581] or

[ITU-T P.341]. Measurement of headset shall comply with the guidelines in [ITU-T P.381] or [ITU-T P.382].

If not stated otherwise, the HATS signals shall be diffuse field (DF) equalized using a filter following the inverse of the ear-drum reference point (DRP) to diffuse-field response in Annex A of [ITU-T P.58].

NOTE – Since the loudness model according to [b-ISO 532-1] also supports free field (FF) equalized recordings, this equalization is also possible, provided the free-field option is selected for the loudness calculation.

8.2 Loudness calculation

The full loudness model is described in [b-ISO 532-1]. Before inserting the signal into the model, it has to be calibrated to the physical unit Pascal.

The loudness analysis is carried out over the entire recording, including possible speech pauses, resulting in a short-term loudness vs. time in sone. In order to obtain a single loudness value per recording, the curve is aggregated vs. time by calculating the average across active speech parts.

If the recording contains speech pauses, leading or trailing silence periods, these shall be excluded from the sone vs. time data. An automated speech activity classification algorithm for pulse-code modulation (PCM) audio is described in Appendix II of [ITU-T G.160] and shall be used for sone vs. time data in this case. This classifier assumes that speech pauses of duration less than 400 ms are still considered as active speech.

The aggregated loudness N (unit sone) shall be reported as the main result per evaluated speech signal. Optionally, the result may also be reported as loudness level LN (unit phon) by transforming loudness N via equation (1) to (3) in clause 5.3 of [b-ISO 532-1].

9 Loudness model for sending loudness

The sending loudness is calculated by assuming a connection with a second terminal having the following receive characteristics:

- 6.28 dBPa/V for all applicable frequencies, for calculation of narrowband sending loudness;
- 4.83 dBPa/V for all applicable frequencies, for calculation of wideband, super-wideband and fullband sending loudness.

NOTE – Loudness calculations can, in principle, only be made for signals in the acoustical domain. The sending loudness of a terminal can however be defined by assuming a connection with a nominal receiving terminal, see Appendix IV.

Test procedure:

- the test setup is the same as for sending loudness rating (SLR) measurements for the respective terminal type (e.g., [ITU-T P.311] for wideband handsets and headsets and [ITU-T P.341] for wideband digital loudspeaking and hands-free telephony terminals);
- as a test signal, the British English single-talk sequence described in clause 7.3 of [ITU-T P.501] shall be used at an [ITU-T P.56] active speech level of –4.7 dBPa at the mouth reference point (MRP);
- the test signal is applied to the terminal using the equalized artificial mouth and recorded at an electrical reference point corresponding to the point of interconnection (POI), represented in Volts. The signal is scaled to representation in Pascal, using the reference receiving path described above, scaling factor either:
  - $10^{\frac{6.28}{20}} = 2.06$ for narrowband; or
  - $10^{\frac{4.83}{20}} = 1.74$ for wideband, super-wideband and fullband;

- the sending loudness is now calculated according to the description for receive loudness.

10 Loudness model for sidetone loudness

For further study.

11 Loudness model for listener sidetone loudness

For further study.

Annex A

Validation and test vectors for loudness model implementation

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

A.1 Introduction to test vectors for validation

This annex provides reference results for speech samples scaled at certain active speech levels according to [ITU-T P.56]. In addition, several band-pass filters are applied in order to simulate typical bandwidth limitations in NB, WB, SWB and FB mode. Examples with and without active speech classification are provided.

A.2 Validation of loudness model implementation

A.2.1 Introduction

The examples provided in this clause illustrate the usage of the loudness model for different bandwidth limitations and (active) speech levels. Note that in contrast to clause 8.2, the entire signal is analysed, hence no speech pauses are excluded from the calculation.

A.2.2 Speech material

For all reference loudness values, the speech signal *FB_male_female_single-talk_seq.wav* of clause 7.3.2 of [ITU-T P.501] is used. The speech signal includes short speech pauses which are longer than 400 ms. However, the analysis is carried out over the entire signal (including trailing/leading pauses).

A.2.3 Processing

For each bandwidth mode, the source speech signal is filtered with one or more finite impulse response (FIR) filters according to [ITU-T G.191]:

- NB: LP35 and MSIN (16 kHz sampling rate);
- WB: P341 (16 kHz sampling rate);
- SWB: 14KBP (32 kHz sampling rate);
- FB: 20KBP (48 kHz sampling rate).

After filtering, each signal is calibrated to active speech levels according to [ITU-T P.56] ranging from 40 dBSPL to 90 dBSPL in six steps of 10 dB.

Then the loudness model which have been implemented according to clause 8 is tested with each of the 24 resulting level/bandwidth conditions with diffuse-field correction. Diotic presentation is assumed for the mono signal.

A.2.4 Reference results

Table A.1 shows the loudness results in sone and phon for the signal processing described in clause A.2.3. Loudness values [sone] that are within +/–4% of the values stated in Table A.1 and loudness levels that are within +/–0.5 phon of the values stated in Table A.1 are considered to be compliant with the specifications of this recommendation.

Table A.1 – Reference results for loudness and loudness level

Active speech level [dBSPL]	Filter	Loudness [sone]	Loudness level [phon]
40	NB	0.93	38.96
	WB	1.03	40.46
	SWB	1.05	40.76
	FB	1.05	40.75
50	NB	2.17	51.17
	WB	2.44	52.85
	SWB	2.50	53.23
	FB	2.50	53.21
60	NB	4.57	61.93
	WB	5.15	63.65
	SWB	5.30	64.07
	FB	5.30	64.06
70	NB	9.06	71.80
	WB	10.24	73.56
	SWB	10.55	73.99
	FB	10.54	73.97
80	NB	17.42	81.23
	WB	19.74	83.03
	SWB	20.31	83.44
	FB	20.29	83.43
90	NB	32.98	90.43
	WB	37.38	92.24
	SWB	38.35	92.61
	FB	38.33	92.60

A.3 Validation of speech pause detection and loudness model implementation

A.3.1 Introduction

In general, if the recording contains speech pauses, leading or trailing silence periods, these shall be excluded from the temporal aggregation. An automated speech activity classification algorithm for this purpose is described in Appendix II of [ITU-T G.160] and shall be used in this case (see clause 8.2). This classifier assumes that speech pauses of duration less than 400 ms are still considered as active speech.

The source code for ITU-T G.160 Appendix II is available as an electronic attachment to [ITU-T G.160]. The main program for ITU-T G.160 Appendix II, named "g160app2.c", can be generalized by utilizing a fixed frame size of 10ms (instead of hardcoded frame length of 80 samples, which refers to a fixed sampling rate of 8 kHz).

For validation of the loudness model described in this Recommendation, the default setting shall be used for all parameters.

A.3.2 Speech material

For all reference loudness values, the first 20.0 s of the speech signal FB_male_female_double-talk_seq.wav of clause 7.3.5 of [ITU T P.501] is used. This signal contains short words only and provides a speech activity of approximately 25 percent.

A.3.3 Processing

In contrast to clause A.2.3, the source speech signal is considered only in FB mode, i.e., the filter 20KBP (48 kHz sampling rate) according to [ITU-T G.191] is applied.

After filtering, the signal is calibrated to active speech levels according to [ITU-T P.56] ranging from 40 dBSPL to 90 dBSPL in six steps of 10 dB.

Then the loudness model according to clause 8 is carried out for each of the 6 resulting level conditions with diffuse-field correction. Diotic presentation is assumed for the mono signal.

A.3.4 Reference results

Table A.2 shows the loudness results in sone and phon for the signal processing described in clause A.3.3. Loudness values [sone] that are within +/–4% of the values stated in Table A.1 and loudness levels that are within +/–0.5 phon of the values stated in Table A.1 are considered to be compliant with the specifications of this recommendation.

Table A.2 – Reference results for loudness and loudness level (excluding speech pauses)

Active speech Level [dBSPL]	Filter	Loudness [sone]	Loudness level [phon]
40	FB	1.45	45.36
50	FB	3.35	57.46
60	FB	6.97	68.02
70	FB	13.69	77.75
80	FB	26.18	87.10
90	FB	48.87	96.11

Appendix I

Result of loudness experiment A

(This appendix does not form an integral part of this Recommendation.)

I.1 Introduction to experiment A

This appendix presents the result of a subjective loudness experiment performed for clean speech, music as well as noisy speech stimuli. A mobile phone supporting narrowband, wideband as well as super-wideband has been used for recording the stimuli. This experiment should be seen in the context of similar experiments which has been conducted during the process of developing this Recommendation. Some of the characteristics of these experiments are summarized in the following part of the introduction.

Orange used one-critical band noises centered at 1 kHz as the reference stimuli and compared the subjective rating of speech samples with that of the reference stimuli [b-AES E-Library] [b-Edjekouane 1]. The ANSI S3.4 [b-ANSI S3.4-2007] implementation with the modification of the loudness growth function, i.e., P.Loudness, is suggested for predicting the loudness of speech samples. The suggested loudness growth function is dependent on the telephone mode, i.e., hands-free and handset. HEAD acoustics utilized 3-Bark band noises centered at 1 kHz including a roll-off characteristic as the reference in order to minimize the perceptual interaction between loudness and annoyance created by the tonal perception of the one-critical-band noise [b-AES E-Library].

These investigations used one session for the reference stimuli and the other for speech samples. Separating the two sessions based on the type of stimuli may cause the range effect as well as the stimulus spacing and frequency. The experiment with the 3-Bark band noise cannot come up with the subjective loudness level directly, and rather the method requires a loudness equalization procedure between the reference and the speech samples. If there is a systematic offset introduced in the objective loudness calculation, the loudness equalization process introduces almost the same offset in both type of stimuli. The resulting comparison between the subjective responses and the objective loudness values may not reveal this systematic offset.

For this reason, the current investigation included references as well as speech samples in the same listening test session. Diotic 1 kHz tones are used as the reference, and it allows to estimate the subjective loudness level directly from the listening test results. Since this approach may discover the systematic offset when calculating the loudness of different stimuli, the calculated error between the subjective and the objective loudness levels may be greater than those reported in [b-AES E-Library].

I.2 Subjective experiment

The detail procedure for the subjective evaluation is already reported in [b-AES E-Library]. Due to the limited time assigned for the listening test, the number of speech samples are reduced to four for the clean speech samples and two for the noisy speech samples. During the preparation of the noisy speech samples, the level of noise is calculated at the ear drum position rather than at the center of the head in the free field. This resulted in increasing the signal-to-noise ratios (SNRs) in the noisy speech samples by approximately 9 dB. Thus, the actual SNRs used for the listening test are 9 and 19 dB instead of 0 and 10 dB.

I.2.1 Subjective loudness function

Reimes et al., [b-AES E-Library] compared the predicted loudness values using the individual loudness function and the averaged one. They concluded that there is no significant difference between them. Thus, the average loudness function is used in the current investigation. Figure I.1 shows the average subjective loudness values as a function of the 1-kHz tone level. The error bars represent the 95% confidence intervals across 18 subjects. A two-factor analysis of variance

(ANOVA) with the telephone mode (handset and hands-free) and the reference levels all constituting within-subjects factors confirmed the significant main effect of telephone mode ($p = 0.009$) and the significant interaction with the reference level ($p < 0.001$). This may be triggered by the perceived loudness difference between hands-free mode and handset mode in average. When a stimulus is presented in a session with more louder stimuli, there is a tendency of judging the perceived loudness of the stimulus to be softer. Consequently, the loudness functions are fitted separately for hands-free and handset mode. The solid lines in Figure I.1 show the fitted loudness functions.

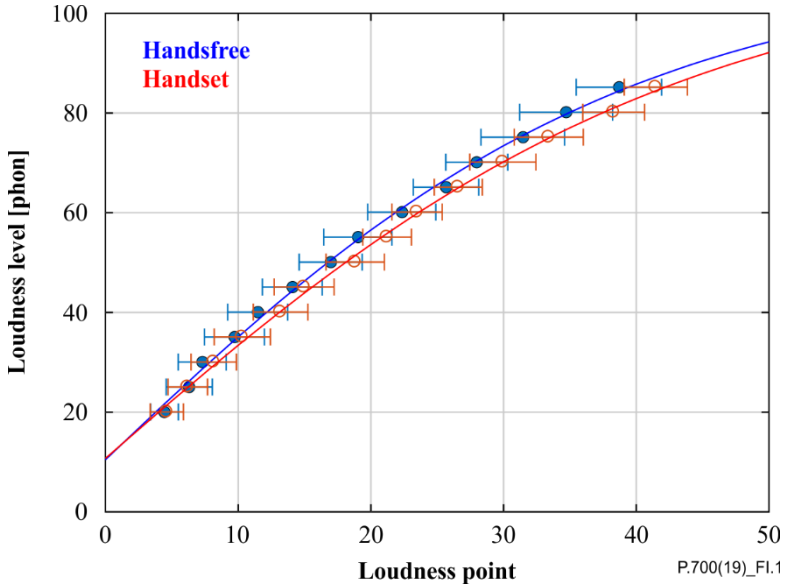


Figure I.1 – Subjective loudness growth function and the corresponding fitted sigmoid function

The sigmoid function used for the fitting is defined in equation I.1, and it is the same as the one suggested in [b-AES E-Library].

$$L(p) = p_{min} + \frac{(p_{max} - p_{min})}{1 + e^{a \cdot (p_0 - p)}} \tag{I.1}$$

Table I.1 summarizes the parameters estimated by the non-linear regression. The minimum value of Pmin is set to –100 for the regression.

Table I.1 – Fitted parameters for the sigmoid function in hands-free mode and handset mode

	Hands-free	**Handset**
Arithmetic mean	Pmin: –100.0 Pmax: 109.3 a: 0.049 p0: –2.3	Pmin: –100.0 Pmax: 111.1 a: 0.044 p0: –2.2

The subjective loudness point values of the speech and music samples from the listening test are transformed to the loudness levels using equation I.1 and the parameters in Table I.1.

I.2.2 Loudness of clean samples

Figure I.2 shows the effect of different presentation levels in the perceived loudness of clean sound samples. Due to limited space in the report, only results for the EVS244 codec is displayed here. The

first four sound samples are speech samples. Sample 5 is the classical music, and sample 6 is the pop music. A four-factor ANOVA with the telephone modes, the codecs, the sound pressure levels (SPLs), and the sound samples all constituting within-subjects factors confirmed the highly significant main effect of SPLs ($p < 0.001$).

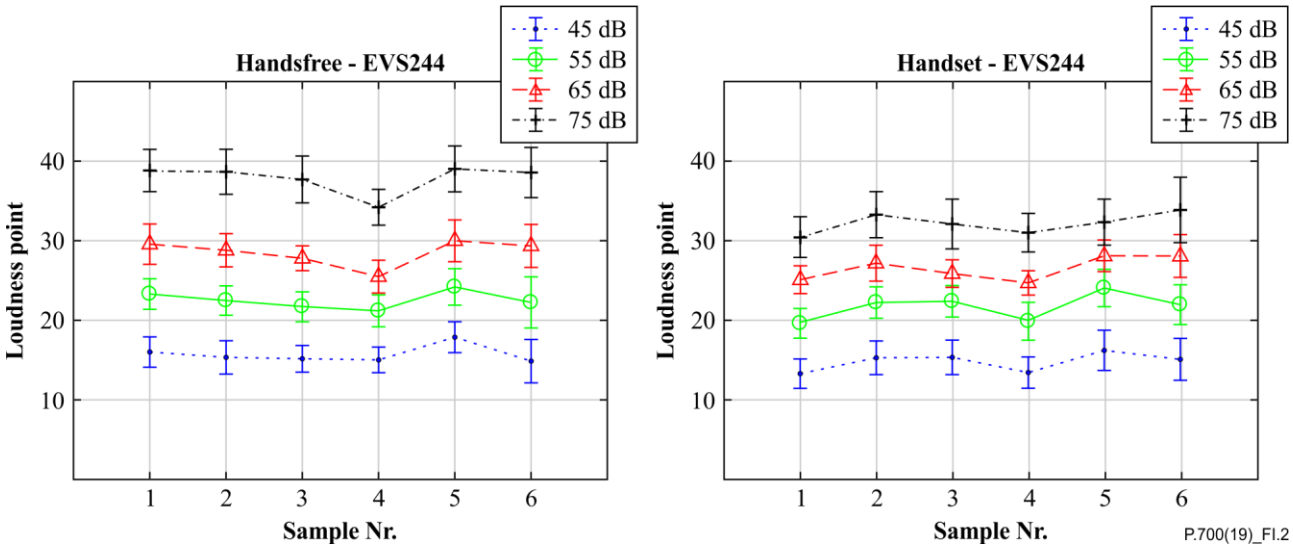


Figure I.2 – Subjective loudness ratings in different presentation levels

The effect of telephone mode is displayed in Figure I.3, and there is a highly significant main effect ($p < 0.001$). This effect of telephone mode is mainly caused by the fact that in handset mode, the subjects are exposed to the stimuli monaurally through their preferred ear. The typical range of binaural gain, i.e., the loudness advantage by listening with two ears compared to one ear, is between 3 and 10 dB [b-Moore 3]. Apart from the effect of binaural listening, the spectral shape of the stimuli is dependent on the telephone mode since the transfer function from the phone to the ear drum is affected by the telephone mode. For example, in handset mode, the outer-ear transfer function of the HATS is not present.

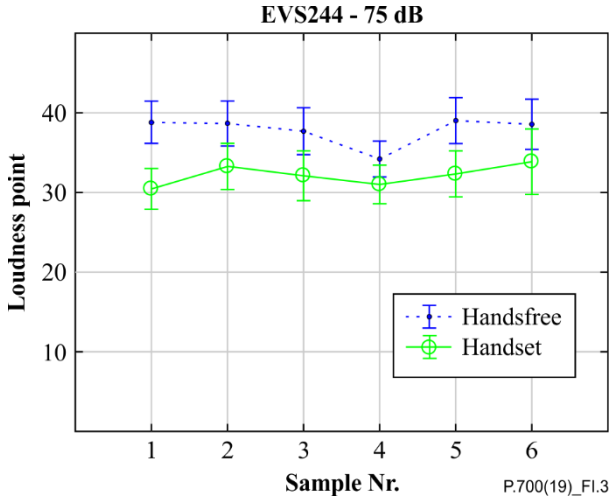


Figure I.3 – Subjective loudness ratings in the two telephone modes, i.e., hands-free and handset, when the level of stimuli is 75 dB

The effect of codecs is less visible compared to the other factors. Figure I.4 compares the subjective results across different codecs. It is noticeable that the FB samples are judged to be softer compared to the other samples in hands-free mode ($p < 0.001$). The FB samples are the original samples without

going through the phone, and therefore they don't include the outer-ear transfer function, which typically increases the perceived loudness of speech samples.

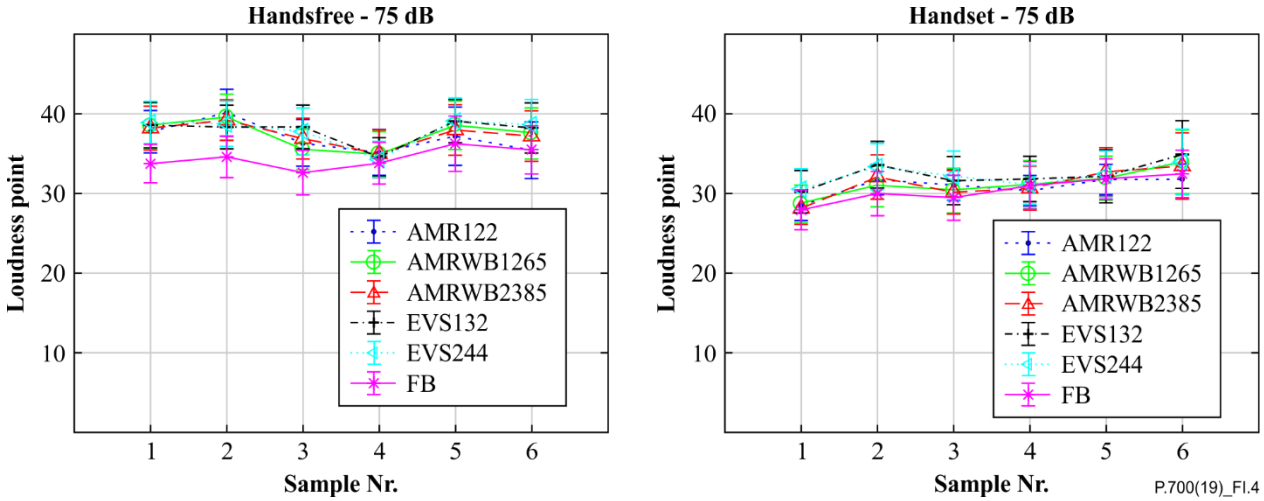


Figure I.4 – Subjective loudness ratings in different codecs when the level of stimuli is 75 dB

I.2.3 Loudness of noisy speech samples

When judging the loudness of a speech sample in noise, i.e., partial loudness, subjects are instructed to focus on the loudness of the target speech while ignoring the background noise. The perceived loudness is expected to decrease when the level of background noise increases. Figure I.5 displays the partial loudness ratings of speech sample 1 as a function of codecs. In hands-free mode, the perceived partial loudness seems to be almost the same for the two SNR conditions. However, in handset mode, i.e., monaural listening, the effect of SNR seems to be more visible. This effect of telephone mode in partial loudness is confirmed by a five-factor ANOVA with the SNRs, the telephone modes, the codecs, the sound pressure levels (SPL), and the sound samples all constituting within-subjects factors. The interaction between the SNRs and the telephone modes is significant ($p = 0.001$).

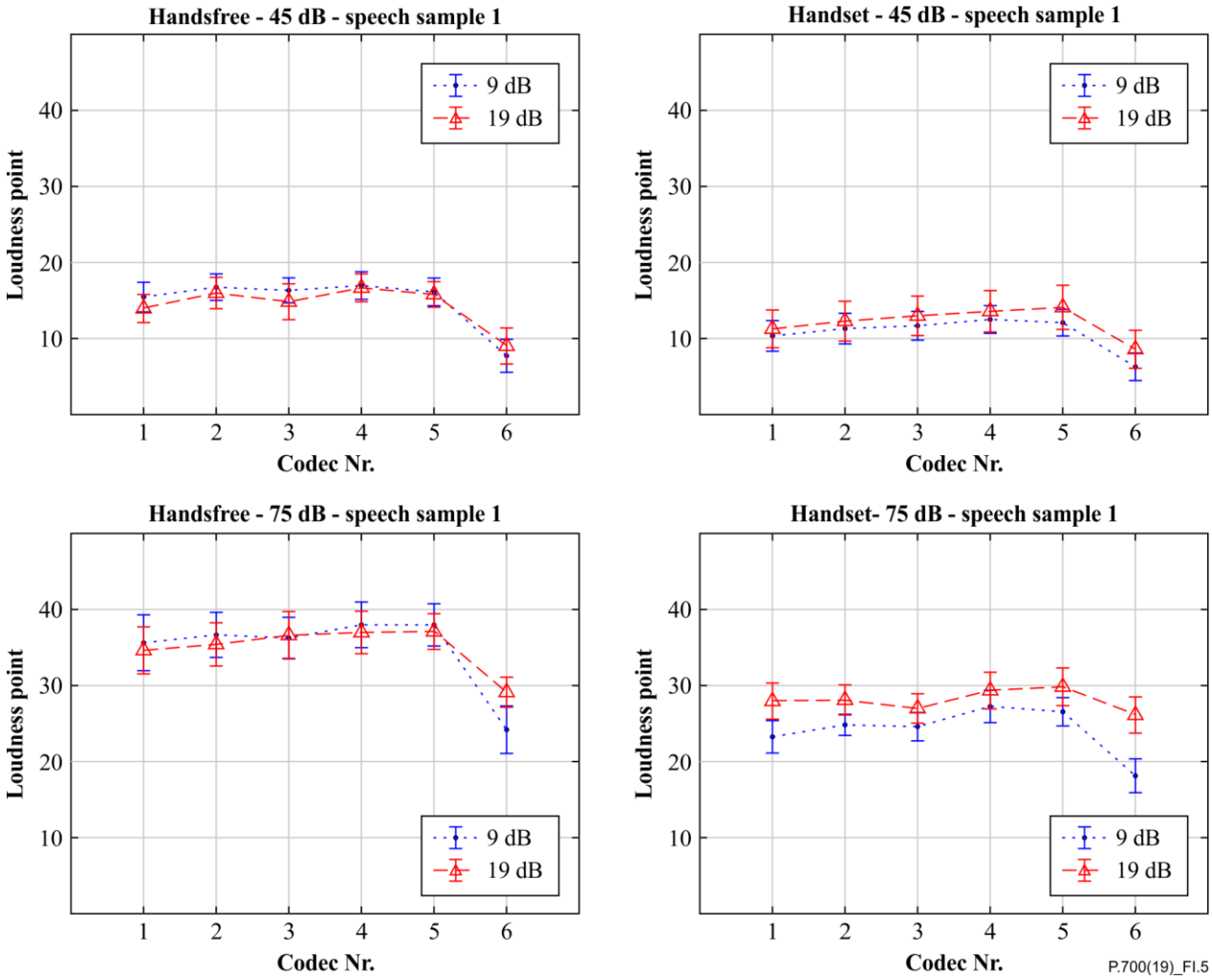


Figure I.5 – Subjective loudness ratings in different codecs for speech sample 1

I.3 Comparison results

I.3.1 Loudness algorithms

The following describes the objective models used for predicting the perceived loudness.

Diotic listening (hands-free mode):

- P.Loudness [b-AES E-Library]: MATLAB implementation provided by Orange;
- [b-ISO 532-1]: based on Zwicker's loudness model and supports both stationary and time-varying model;
- [b-ISO 532-2]: based on Moore-Glasberg model and intended only for stationary signals;
- short-term time varying loudness model from Moore-Glasberg: see [b-Glasberg 2].

Monaural listening (handset mode):

- P.Loudness: perfect loudness summation rule;
- [b-ISO 532-1]: perfect loudness summation rule;
- [b-ISO 532-2]: supports the latest development on binaural loudness including the concept of binaural inhibition [b-Moore 3];
- short-term time varying loudness model from Moore-Glasberg: the latest development on binaural loudness including the concept of binaural inhibition, [b-Glasberg 2] and [b-Moore 3], is implemented.

Target loudness in noise, i.e., partial loudness, the partial loudness model proposed by Moore et al. [b-Moore 2] is implemented in the following methods:

- [b-ISO 532-2];
- short-term time varying loudness model from Moore-Glasberg.

For the time-varying loudness models, the arithmetic mean and N5 are used to determine a single value from the time history of calculated loudness values. The first 200 ms of samples are skipped for this aggregation in order to avoid the influence of temporal loudness integration, i.e., the slow monotonic increase of loudness in the first part of loudness time history.

I.3.2 Clean stimuli result

The general tendency of music samples on perceived loudness does not seem to be very different from that of speech samples according to the previous clause. Hence, the comparison between the subjective and objective loudness values includes both speech samples and music samples in the same analysis.

In the current investigation, the subjective loudness level is measured by including diotic 1-kHz tones in the experiment. The goal of the analysis is to quantify the errors in terms of loudness level in an absolute manner compared to [b-AES E-Library], in which the relative loudness errors are calculated by employing the concept of equal noise level. For example, the method in [b-AES E-Library] cannot estimate the absolute loudness difference between N5 and arithmetic mean of time-varying loudness values since the iterative loudness equalization process compares the calculated loudness of the reference noise and the speech samples with the same algorithm. The loudness difference between N5 and arithmetic mean affects both the reference noises and the speech samples with almost the same degree. In the current investigation, root mean square error (RMSE) is used to quantify the absolute loudness error.

Figure B.6 compares the subjective loudness level with the predicted loudness level using P.Loudness model provided by Orange. In hands-free mode, the free-field pressure is used as the input to the loudness, and it is calculated by applying the HATS free-field equalization function to the stimuli. In handset mode, the stimuli, which are recorded by the HATS, are directly used as the input to the loudness calculation because the algorithm requires the ear-drum pressure. In general, the model overestimates the loudness values for hands-free mode and underestimate them for handset mode. The RMSE value of hands-free mode is slightly higher compared to handset mode.

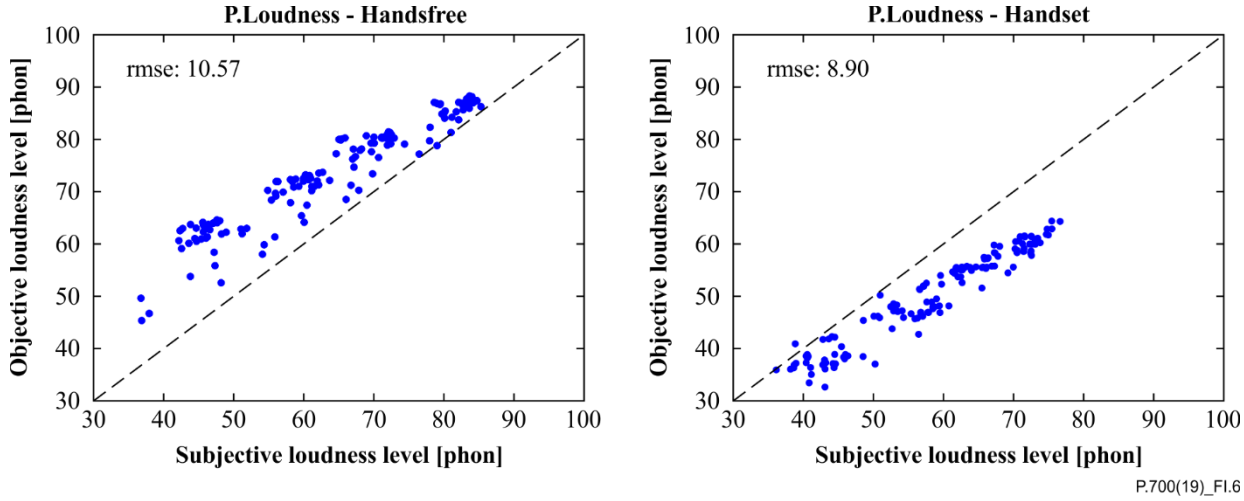


Figure I.6 – Subjective loudness level vs. predicted loudness level using P.Loudness

Figure I.7 and Figure I.8 demonstrate how the ISO 532-1 algorithm works. Since the model requires the input signal to be free-field equalized, the HATS free-field equalization function has been applied

to the signal prior to feeding it to the model. For handset mode, the calculated loudness is divided by 2 for the perfect loudness summation rule. The arithmetic mean aggregation seems to predict the subjective loudness level better compared to N5, and it performs better than the stationary loudness model. In general, the model is more accurate for handset mode compared to hands-free mode.

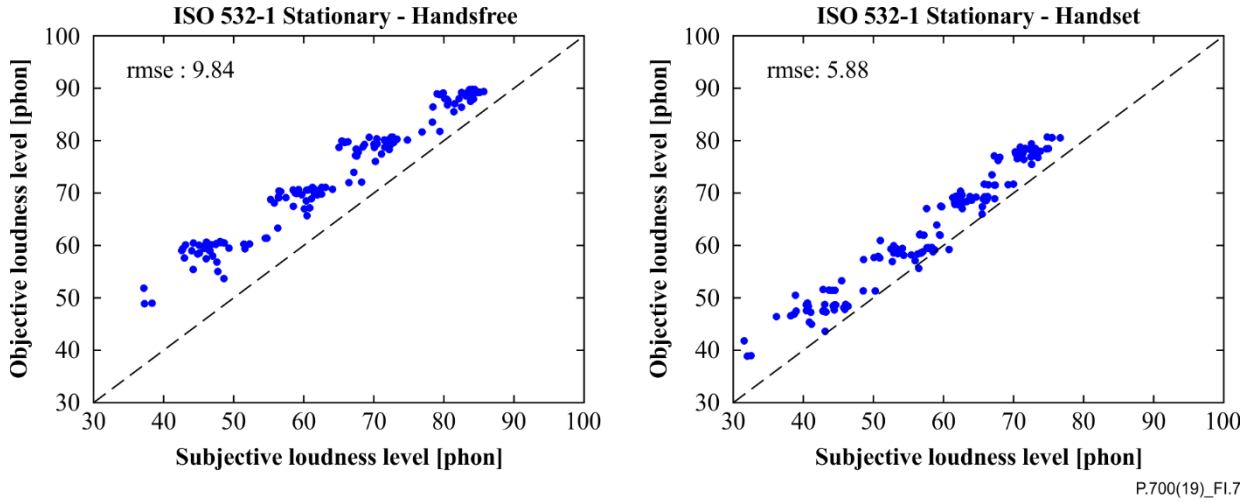


Figure I.7 – Subjective loudness level vs. predicted loudness level using ISO 532-1 stationary model

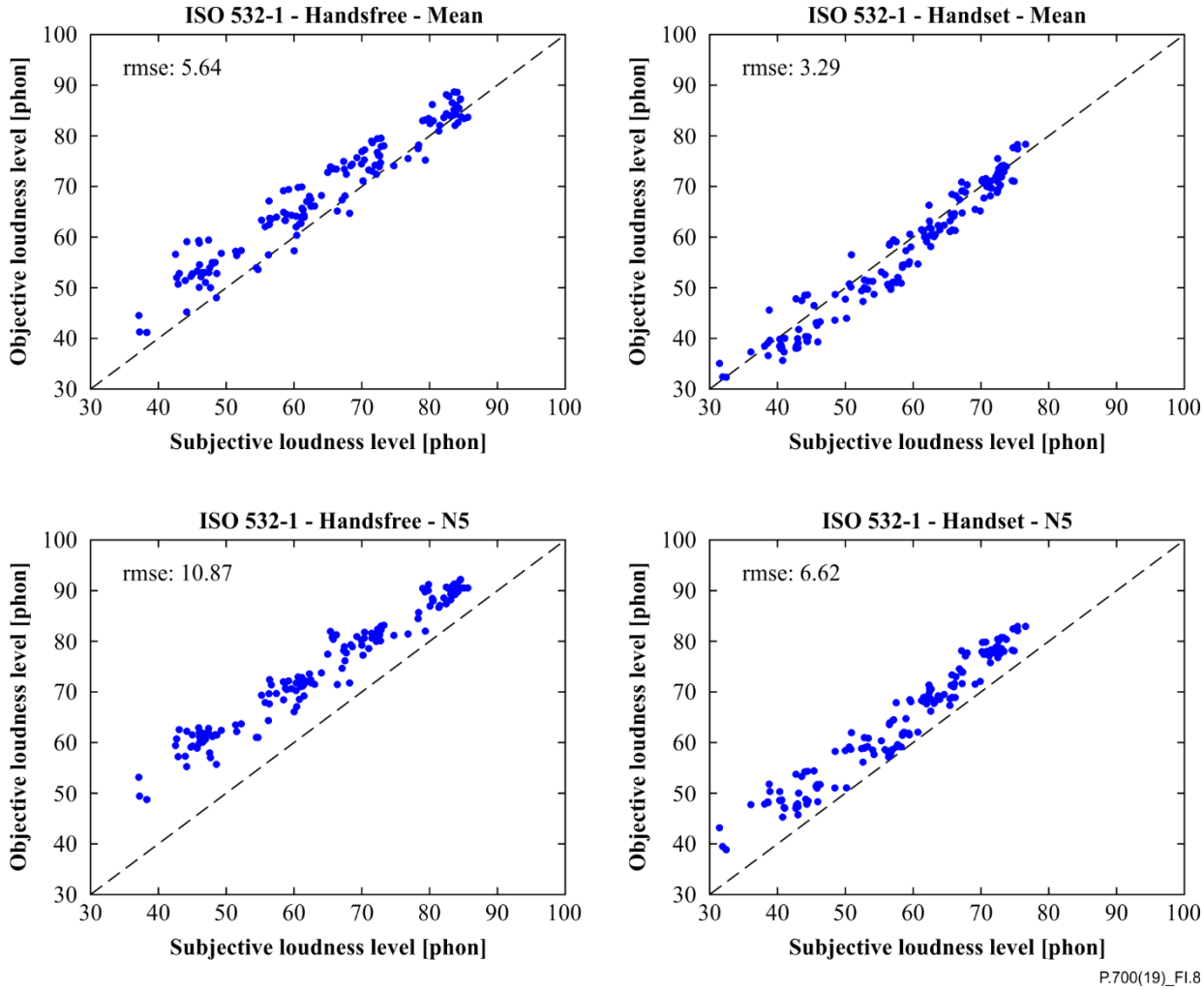


Figure I.8 – Subjective loudness level vs. predicted loudness level using ISO 532-1 time-varying model

Figure I.9 and Figure I.10 display the performance of Moore-Glasberg model on predicting the perceived loudness values. Moore-Glasberg model can take the ear-drum pressure as the input signal, and therefore the stimuli are used directly for the calculation. For handset mode, the model automatically calculates the binaural loudness by providing one of the ear signals with zero values. [b-ISO 532-2] is based on Moore-Glasberg stationary loudness model and does not perform well compared to other algorithms. The short-term Moore-Glasberg time-varying model performs the best among the Moore-Glasberg models when it is combined with the arithmetic mean aggregation. The model seems to overestimate the loudness values.

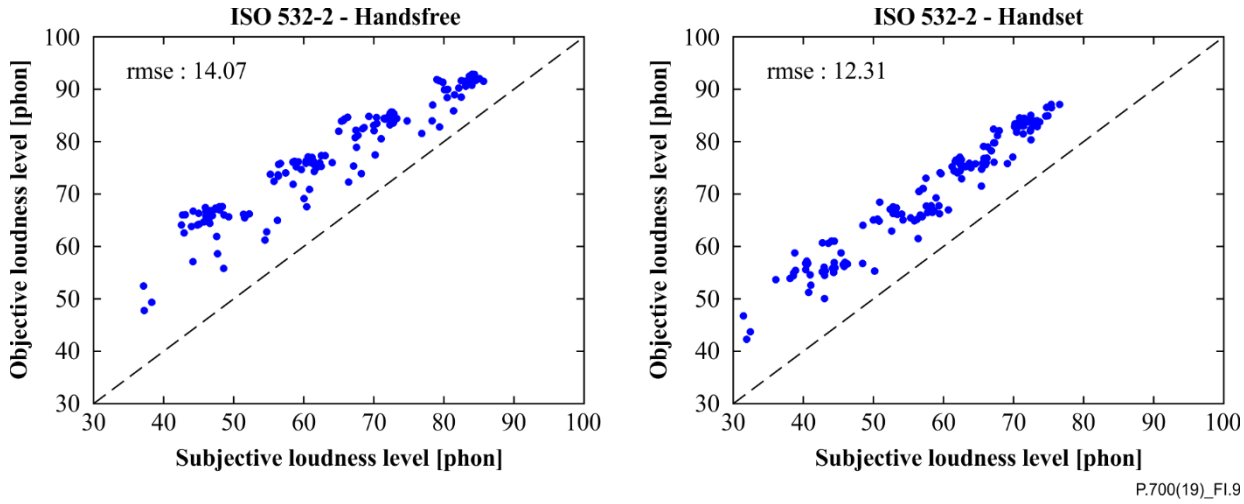


Figure I.9 – Subjective loudness level vs. predicted loudness level using ISO 532-2

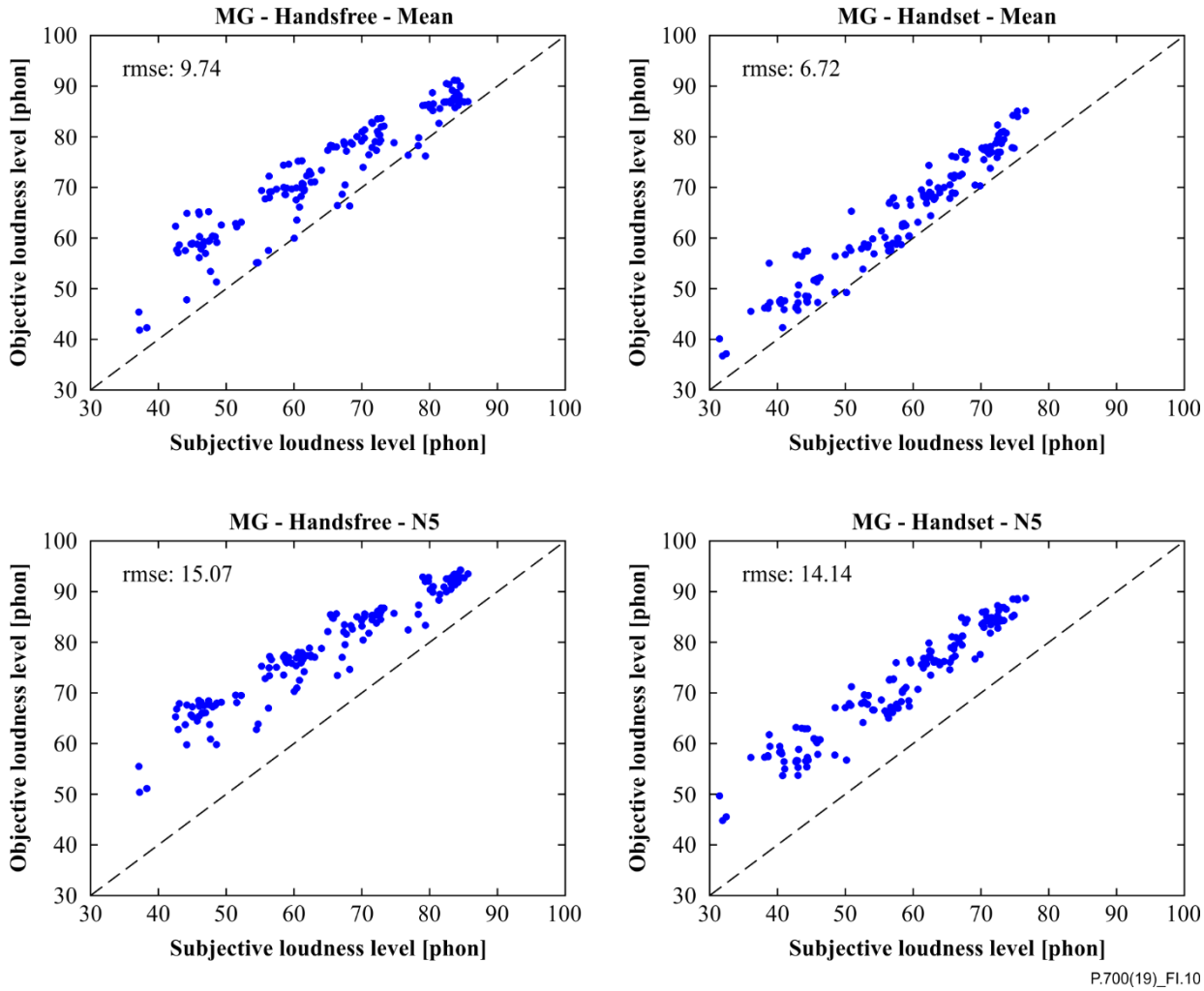


Figure I.10 – Subjective loudness level vs. predicted loudness level using Moore-Glasberg short-term loudness model

I.3.3 Noisy speech result

The partial loudness algorithm requires two input signals, i.e., a clean speech signal and a noise alone signal. For the current investigation, the noise alone signals are not recorded by the mobile phone rather the raw background noise signal is directly mixed with the recorded speech signals. This is mainly due to the unknown effect of automatic gain control in the mobile phone. For this reason, the partial loudness scenario in the investigation fits better with a situation where there is background noise present in the receiving part.

The general tendency of the partial loudness model is similar to the prediction results of Moore-Glasberg model for the clean stimuli. As shown in Figure I.11 and Figure I.12, the arithmetic mean aggregation of the short-term time-varying loudness seems to perform the best in the prediction. The overestimation is clearer for the stationary model and for N5 of the time-varying model. The results confirm that the partial loudness algorithm works as good as the loudness models for the clean stimuli and thus can be used practically.

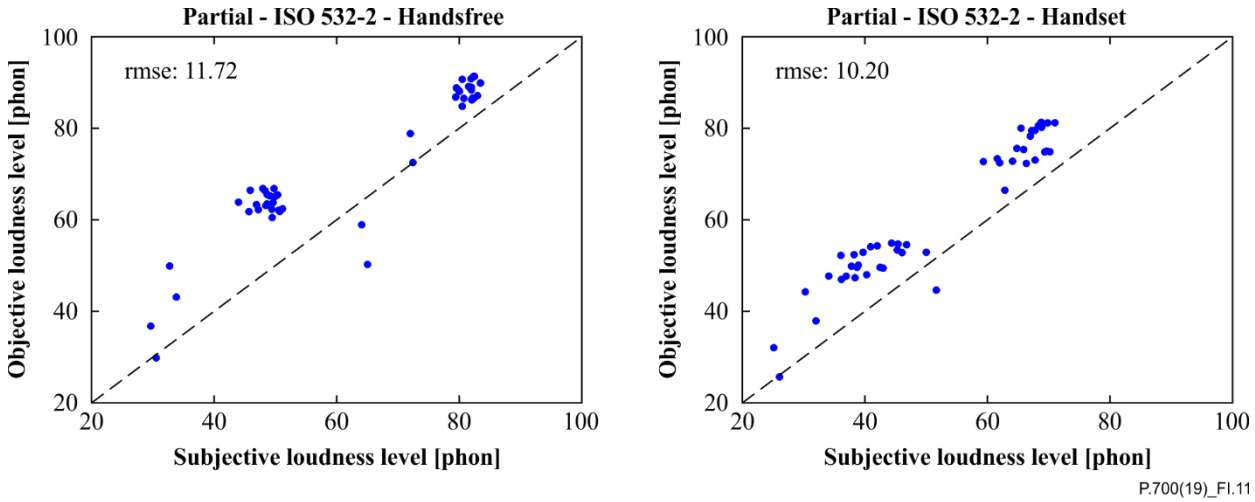


Figure I.11 – Subjective loudness level vs. predicted loudness level using the partial loudness algorithm implemented in ISO 532-2

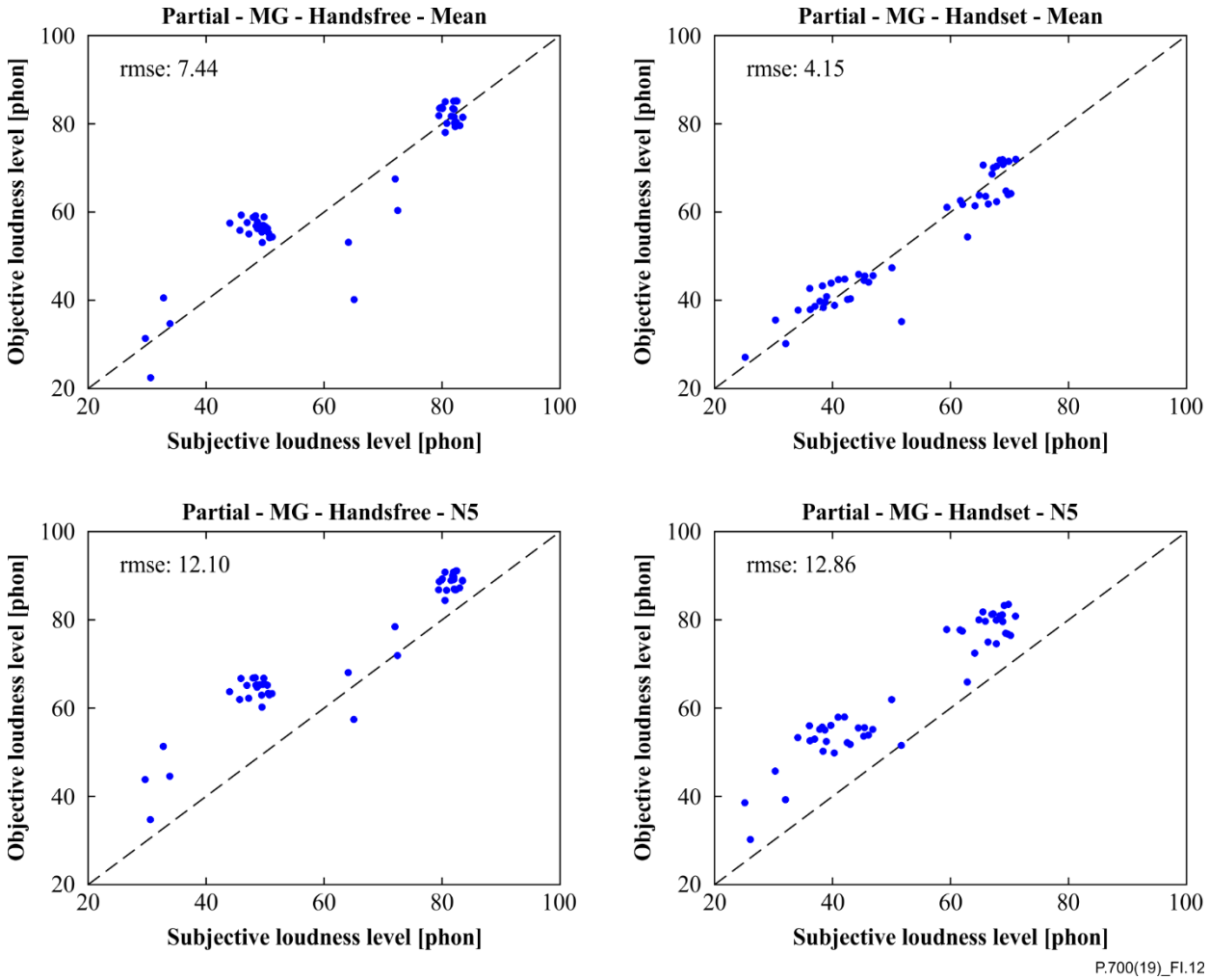


Figure I.12 – Subjective loudness level vs. predicted loudness level using the partial loudness algorithm implemented in Moore-Glasberg short-term time-varying loudness

I.4 Conclusion

A subjective loudness experiment is performed both for clean speech and music stimuli as well as noisy speech stimuli. A mobile phone supporting the option of NB, WB and SWB telephony audio bandwidths is used for recording the stimuli. The level of the stimuli and the background noise is

controlled to simulate different exposure levels of the stimuli. For converting the subjective loudness judgements into loudness level, 1 kHz tones having different levels are presented diotically.

For the clean stimuli, ISO 532-1 time-varying model having the arithmetic mean aggregation seems to perform the best in comparison with other models. The N5 aggregation overestimates the loudness values slightly. This may be caused by the fact that N5 is the 95th percentile and may give rise to noticeably higher values compared to the arithmetic mean for non-stationary time signals.

For the noisy stimuli, only high SNRs are tested. For these SNRs, the arithmetic mean of Moore-Glasberg short-term loudness seems to achieve the best performance, and the resulting RMSEs are comparable with the ones for the clean stimuli.

The [b-ISO 532-1] model can be recommended for testing the loudness of clean sound samples in telecommunication devices based on experimental conditions employed in the current investigation. However, for the partial loudness paradigm, the arithmetic mean of Moore-Glasberg short-term loudness model can be recommended despite of the fact that the method is not standardized.

In most of the conditions, hands-free mode resulted in higher RMSEs compared to handset mode. Therefore, it may be interesting to investigate the influence of binaural loudness summation. The current investigation focused on evaluating absolute loudness errors, but it would be necessary to discuss in the working group whether relative loudness errors are good enough for practical applications.

Appendix II

Result of loudness experiment B

(This appendix does not form an integral part of this Recommendation.)

II.1 Introduction to loudness experiment B

One of the initial proposals for P.Loudness included loudness calculation methods for handset and hands-free scenarios. However, the original experiments presented in [b-AES E-Library] for handset and hands-free only contained coded speech signals. Neither acoustic playback via loudspeaker systems (which may include strong non-linear effects) nor any other typical signal processing (e.g., like speech compression) were involved. Additionally, the analysis of terminals evaluated using common HATS setups as used in terminal testing was not yet considered.

In previous contributions [b-AES E-Library], two measurement series were introduced to cover these issues. Handset and hands-free application modes were evaluated as typical downlink scenarios. Multiple technologies like digital enhanced cordless telecommunications (DECT), voice over IP (VoIP), 3G and 4G as well as certain state-of-the-art codecs were covered by the measurements.

All recordings were then evaluated in listening tests obtaining absolute loudness values. An evaluation framework as described in [b-AES E-Library] was used here. This contribution presents results including 20 test subjects per mode (previously only 15). Additionally, several suggestions made in [b-AES E-Library] are taken into account in order to obtain comparable results over different experiments, e.g., by using multiple reference sounds.

In this update, the instrumental loudness assessment is extended with regard to the additional reference sounds. In addition, a simplified method for the assessment on the proposed equal-reference-level scale is introduced.

II.2 Measurement setup

According to [b-AES E-Library], the measurement setup should contain a wide range of real devices, including the audio bandwidths NB, WB, SWB and FB.

Since there are only a few real terminals currently available capable of SWB and FB, a mobile mock-up in handset and (handheld) hands-free mode was used for the recording procedure additionally. Figure II.1, exemplarily depicts the usage of the mock-up device for both scenarios. The mock-up consists of two different micro-loudspeakers on front (handset) and back (hands-free) side. In these cases, the test signals are coded and decoded by certain codecs in advance to simulate a typical device. In addition, a modified version of the mock-up was used. The frequency response of the device was equalized to obtain a flat frequency response in handset and hands-free mode.

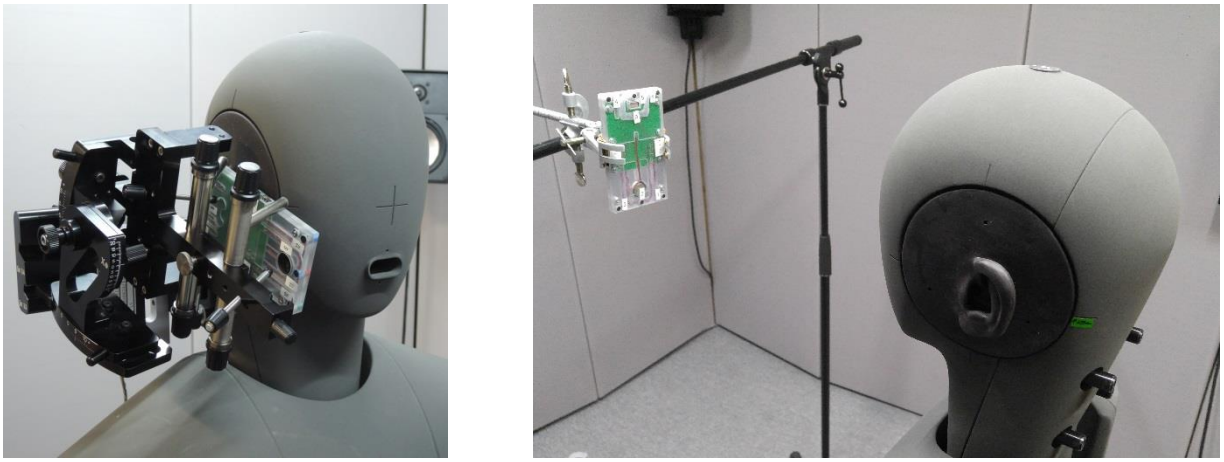


Figure II.1 – Mock-up mounted in handset (left) and handheld hands-free (right) mode

As described in [b-AES E-Library], a wide but also typical level/loudness range should be considered in the evaluation. Varying levels should mainly be obtained by using volume control of the terminal. The electrical level for the insertion into the device was fixed to –16.0 dBm0.

For the current evaluation, each device was first measured with maximum volume setting. To avoid too extreme playback levels in the listening test, the volume was decreased to 90 dBSPL for handset/monaural respectively 85 dBSPL for hands-free/binaural recordings. These conditions are labelled as MAX in the following.

Per device, two additional volume settings were recorded. Similar to the original work in [b-AES E-Library], level offsets of –5 dB and –15 dB relatively to the MAX condition were collected. Since not all devices were able to reach the maximum level of 85 /90 dBSPL, a wide and continuous range of loudness conditions is obtained by this method.

In contrast to the original work, a modified active speech level calculation according to [ITU-T P.56] is used for determining physical level in dBPa or dBSPL. This method is capable of excluding speech pauses from the level calculation and thus does not require exact cropping of samples. Since the calculation of level according to [ITU-T P.56] is only defined for signals including an amplitude of +/–1.0 (digital full scale), a simple modification is applied:

- if the signal to be analysed includes samples of absolute amplitude lower than 1.0 Pa, the standard ITU-T P.56 algorithm is utilized to obtain active speech level directly in dBPa or dBSPL;
- if the signal to be analysed includes samples of absolute amplitude higher than 1.0 Pa, a scaling factor *a* is determined to reduce the maximum amplitude to 1.0. Then the standard ITU-T P.56 algorithm is evaluated on the scaled version of the signal. The resulting active speech level is corrected by 20*log10(*a*) to obtain a value in dBPa or dBSPL.

In this work, neither music/audio material nor noisy speech is used. Even though both stimuli type will become interesting especially in (mobile) SWB/FB scenarios, the current investigations are intentionally limited to this certain scenario (noise-free transmitted speech).

C.2.1 Speech sequence

As described in [b-AES E-Library], German speech material according to [ITU-T P.501] was used for the recordings. The language was selected according to the mother tongue of the test subjects.

To include another typical processing, compression of the input signal was also evaluated. [ITU-T P.501] provides exemplarily compressed speech for British English but not for German material. Since the origin of the processing as well as an adequate description of the compression method is missing, an open source implementation was used for this purpose [b-github].

Figure II.2 illustrates the speech sequence used for the measurement for all devices (handset and hands-free). To collect a reasonable number of samples for an auditory evaluation, four samples per condition were used. For the handset listening test, only the even (s2) and for the hands-free scenario the odd (s1) sentences were selected.

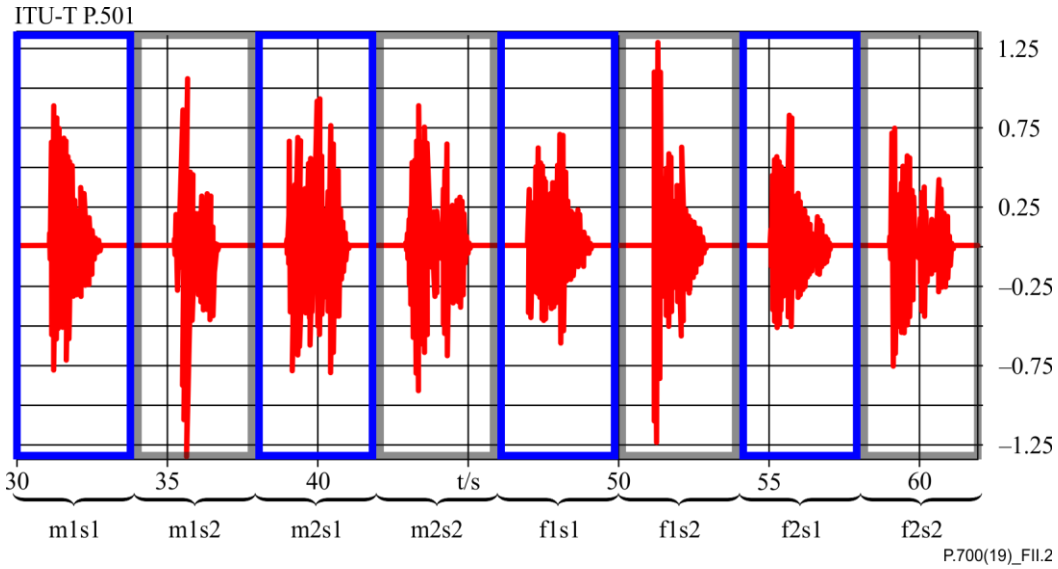


Figure II.2 – Sentences used for handset (grey) and hands-free (blue)

In contrast to the work in [b-AES E-Library], in this series rather short speech samples were used. The source file used for the measurements was composed according to the guidelines of [b-3GPP SA4]. Here the sentences are centred in a time window of 4.0 s. Each sample includes a speech portion between 2.0 and 3.0 s (activity between 50% and 75%).

II.2.2 Handset mode

Several terminals according to Table II.1 were evaluated for generating the handset conditions. As references to the previous experiments in [b-AES E-Library], the pure speech files were also included in the corpus.

The handset recordings were conducted using DF equalization for both artificial ears. Even though no signal is active on the left ear, it was recorded and played back in the listening test to obtain the same / realistic idle noise on both ears. Similar to [b-AES E-Library], also diffuse-field equalized playback was used for the presentation in the listening test.

Table II.1 – Devices and operational modes used for handset evaluation

DUT No.	Type/Net	Codec	Normal/Compressed	Vol. setting
1	DECT	G.726	✓ / ✓	MAX, –15
2	DECT	G.722	✓ / ✓	–15
3	3G	AMR-NB	✓ / ✓	–5, –15
4	3G	AMR-WB	✓ / ✓	–5, –15
5	4G	EVS-SWB	✓ / ✓	MAX, –5, –15
6	VoIP	G.711u	✓ / ✓	MAX, –5, –15
7	VoIP	G.722	✓ / ✓	MAX, –5, –15
8	Mock-up	no Codec-NB	✓ / ✓	MAX

Table II.1 – Devices and operational modes used for handset evaluation

DUT No.	Type/Net	Codec	Normal/Compressed	Vol. setting
9	Mock-up	no Codec-WB	✓ / ✓	MAX
10	Mock-up	Opus-SWB	✗ / ✓	MAX
11	Mock-up	EVS-SWB	✓ / ✗	MAX
12	Mock-up	Opus-FB	✗ / ✓	–5
13	Mock-up	EVS-FB	✓ / ✗	–5
14	Mock-up equalized	no Codec-NB	✓ / ✓	MAX
15	Mock-up equalized	no Codec-WB	✓ / ✓	MAX
16	Mock-up equalized	EVS-SWB	✓ / ✓	MAX, –5, –15
17	Mock-up equalized	Opus-FB	✓ / ✓	MAX, –5, –15
18	Reference	no Codec	✓ / ✓	45, 50, 55, 60, 75

II.2.3 Hands-free mode

Similar to handset scenarios, several terminals were evaluated according to Table II.2. Since for this application there are even less SWB/FB-capable devices available, a loudspeaker in conjunction with pre-coded signals was used here additionally. As for the handset scenario, also the clean speech sentences were included as a reference.

For the binaural hands-free recordings, free-field equalization was selected for both artificial ears.

Table II.2 – Devices and operational modes used for hands-free evaluation

DUT No.	Type/Net	Codec	Normal/Compressed	Vol. setting
1	DECT	G.726	✓ / ✗	MAX
2	DECT	G.722	✗ / ✓	MAX, –5
3	Mobile 3G	AMR–NB	✓ / ✓	MAX, –5, –15
4	Mobile 3G	AMR–WB	✓ / ✓	MAX, –5, –15
5	Mobile 4G	EVS–SWB	✓ / ✓	–5, –15
6	Desktop VoIP	G.711u	✗ / ✓	MAX, –15
7	Desktop VoIP	G.722	✓ / ✓	MAX, –15
8	Loudspeaker	no Codec–NB	✓ / ✓	MAX
9	Loudspeaker	no Codec–WB	✓ / ✓	MAX
10	Loudspeaker	Opus–SWB	✓ / ✗	–15
11	Loudspeaker	EVS–SWB	✓ / ✓	MAX, –5
12	Loudspeaker	Opus–FB	✓ / ✗	MAX
13	Loudspeaker	EVS–FB	✓ / ✓	–5, –15
14	Car HFT 3G	AMR–NB	✓ / ✓	MAX, –15
15	Car HFT 3G	AMR–WB	✓ / ✓	MAX, –15
16	Mock–up	no Codec–NB	✓ / ✗	MAX

Table II.2 – Devices and operational modes used for hands-free evaluation

DUT No.	Type/Net	Codec	Normal/Compressed	Vol. setting
17	Mock–up	no Codec–WB	✕ / ✓	MAX
18	Mock–up	Opus–SWB	✕ / ✓	MAX
19	Mock–up	Opus–FB	✓ / ✓	MAX, –5
20	Mock–up equalized	no Codec–NB	✓ / ✕	MAX
21	Mock–up equalized	no Codec–WB	✓ / ✕	MAX
22	Mock–up equalized	Opus–FB	✓ / ✕	MAX
23	Mock–up equalized	EVS–FB	✕ / ✓	MAX
24	Reference	no Codec	✓ / ✓	40, 55, 70 dB

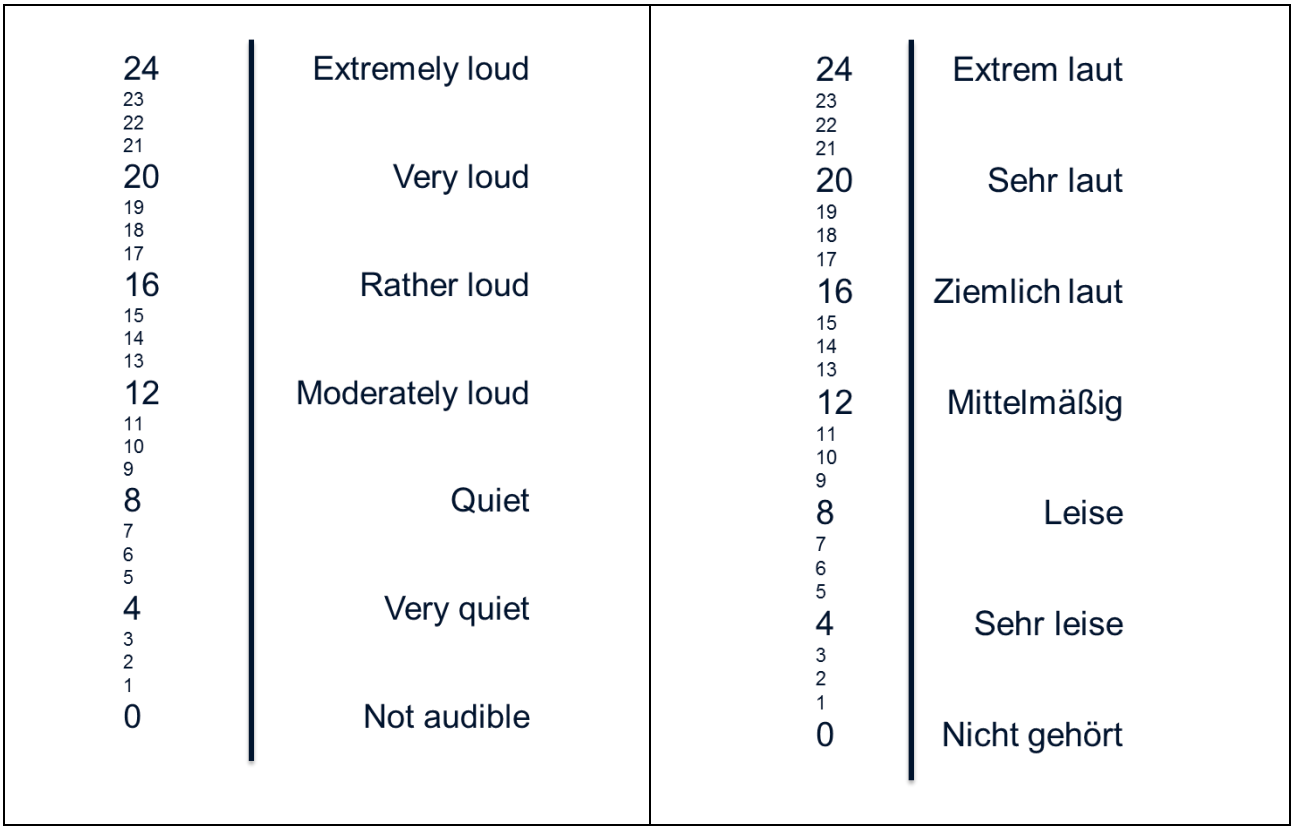


Figure II.3 – The English loudness scale (left) and the German loudness scale (right) for auditory assessment

II.3 Auditory loudness assessment

The categorical 24-point loudness scale as shown in FigureII.3 and already used in [b-AES E-Library] was used for the loudness function assessment as well as for the main evaluation of the speech samples.

II.3.1 Reference sounds and individual loudness functions

The principle of loudness functions is to provide a mapping between auditory results (in averaged points) and a level and/or loudness unit. This transformation is conducted with a reference sound and not with speech stimuli themselves.

In the discussion on recent loudness experiments, the selection of this reference sound was identified as a crucial issue in the auditory assessment procedure. In the original work of [b-AES E-Library], a noise burst of 1-Bark bandwidth at 1 kHz was introduced. The basic idea here was to avoid the usage of a sine tone, but to use a signal where the level directly corresponds to the unit phon. However, as already discussed in [b-Hots 2] and [b-Hots 1], the relation between sub-band level and phon seems unsustainable. In addition, the noise signal includes steep spectral slopes, which cause edge tones and a strong tonal character in the time domain.

In the work presented in [b-AES E-Library] as well as in previous experiments [b-AES E-Library], a more suitable noise signal of 3-Bark bandwidth at 1 kHz was used for the determination of loudness functions. The intention here was to use a sound, which has a similar loudness impression as speech and is easier to judge for the listener. The disadvantage of this reference sound is that there is no relation given between level and loudness. Auditory and instrumental results can only be provided and compared on an equal-reference-level scale (introduced as "equal-noise-level" in [b-AES E-Library]).

The original work in psychoacoustics and loudness assessment always refers to a sine tone as the reference. A sine tone of frequency 1 kHz at 40 dBSPL level is defined as 1 sone or 40 phon. Thus, it seems reasonable to include this reference sound as well. The disadvantage of a sine tone is the tonal character and the increased annoyance at higher levels, which may diverge from the perceived impression of louder speech signals.

In consequence and as an extension to the work already conducted in [b-AES E-Library], all three reference sounds are considered in this contribution for the analysis. Figure II.4 shows the spectra of the three sounds.

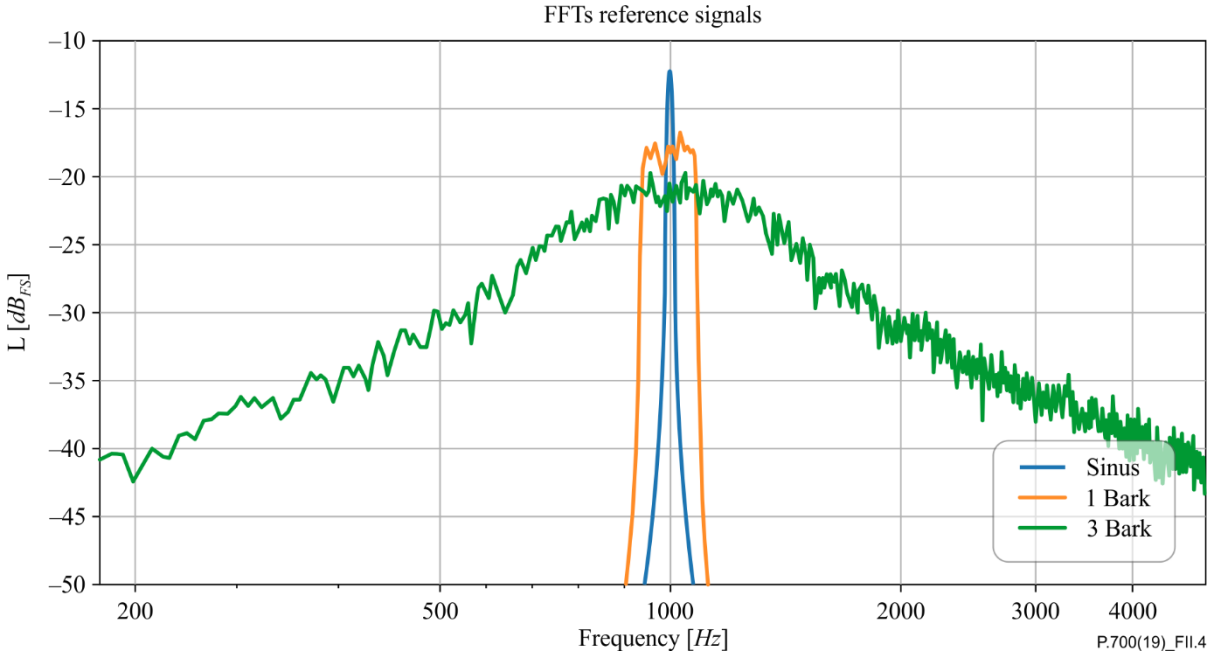


Figure II.4 – Reference stimuli of current (green) and original (red) work

For the determination of loudness functions per test subject, reference sounds with a duration of 4.0 s were used. In overall, ten discrete noise levels between 45 and 90 dBSPL (handset) respectively 40 and 85 dBSPL were used in steps of 5 dB for each reference stimulus. The ten levels were

evaluated six times, which led to 60 trials for the loudness function assessment. An additional initial training phase with four samples according to [b-AES E-Library] were included.

The reference sounds were randomized in a special presentation order. The absolute level difference between two consecutive stimuli was ensured to be lower than 20 dB, which prevents too large jumps in level.

For the transformation from points to equal noise level, a sigmoid function according to equation II.1 is fitted against the averaged auditory data:

$$L(p) = p_{min} + \frac{(p_{max} - p_{min})}{1 + e^{a \cdot (p_0 - p)}} \quad \text{(II.1)}$$

As already mentioned in [b-AES E-Library], the sigmoid function always provides a positive gradient and thus is more reliable than a polynomial fit. The obtained fitting curves for all reference sounds and both modes are shown in Figure II.5 (with inverted axes). All curves provide a reasonable extrapolation to the minimum and maximum point scale. The parameters determined finally for all reference sounds are given in Table II.3.

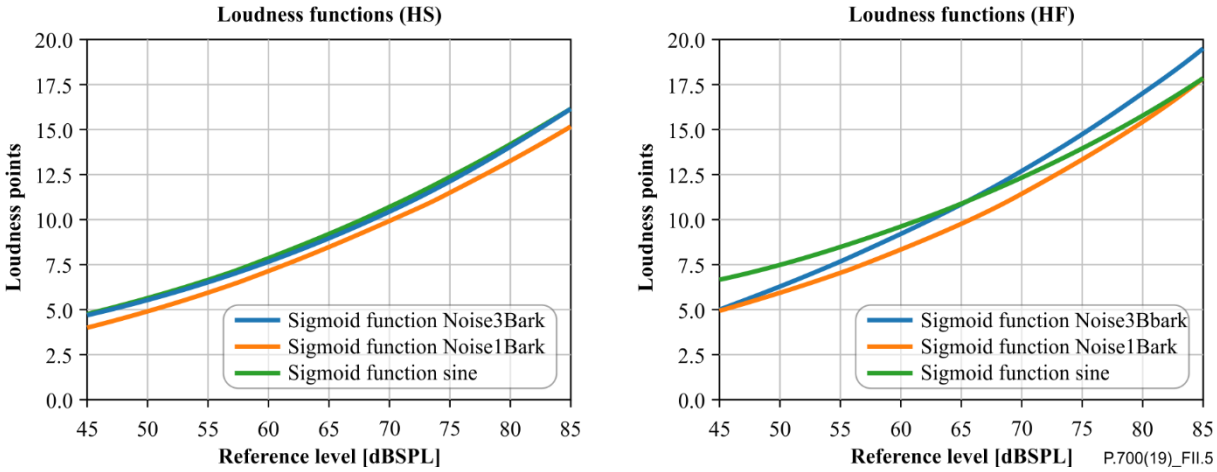


Figure II.5 – Derived sigmoidal functions for handset (left) and hands-free (right) scenarios

II.4 Instrumental loudness methods

Several common loudness methods and temporal aggregations are evaluated in the following analysis. Recent studies like [b-Schlittenlacher] indicate that aggregation of loudness vs. time can improve the prediction performance of loudness models. Thus, especially methods that produce loudness-vs-time curves are investigated in detail.

Since the speech samples are centred and arranged in windows of 4.0 s (see Figure II.2), initial and trailing silence/idle noise portions are included in the samples. For the calculation of level, the method according to ITU-T P.56 is used to exclude these silence periods. For all loudness calculations, only active parts were taken into account. The voice activity detection used is based on the frame classification algorithm of [ITU-T G.160] Appendix II.

Table II.3 – Sigmoid function parameters for fitting handset and hands-free experiments

Parameter	1-Bark noise		3-Bark noise		Sine tone	
	HS	HF	HS	HF	HS	HF
p_{min}	–180.0	–170.0	–19.092	–78.325	–180.0	–3.062e–17
p_{max}	112.083	99.495	101.020	104.489	112.380	93.02
p_0	–9.012	–7.277	3.404	–2.685	–8.424	6.875
A	0.094	0.113	0.150	0.096	0.093	0.211

II.4.1 Loudness according to Zwicker

The time-varying loudness for non-stationary sounds according to Zwicker was standardized e.g., as DIN 45631/A1 [b-DIN 45631] or recently as [b-ISO 532-1]. Only one Zwicker method should be studied in this evaluation, so the latest official version of ISO 532-1 is used. It is expected that the difference to DIN 45631/A1 is negligible.

The current ISO 532-1 model does not include a special binaural loudness assessment. Left and right channel are calculated separately, then an ideal loudness summation is conducted.

II.4.2 Loudness according to Moore/Glasberg

Several methods according to the approach of Moore/Glasberg can be considered for loudness calculation.

II.4.2.1 ANSI S3.4-2007

The loudness for stationary sounds according to Moore/Glasberg is standardized in [b-ANSI S3.4-2007] and in [b-ISO 532-2]. In contrast to the time-varying loudness of Zwicker, this method does not provide a loudness-vs-time curve, so the output of this analysis is directly a loudness value in sone and/or phon. The P.Loudness proposal is based on a modification of this method.

The loudness according to ANSI S3.4-2007 could also be evaluated per block, since an average fast Fourier transform (FFT) is needed anyway for the calculation. In the following, a block size of 4096 samples and 50% overlap is used to calculate a loudness-vs-time curve. As in [b-ISO 532-1], no binaural processing is included for the loudness assessment.

II.4.2.2 Loudness for non-stationary sounds – version 2002

The well-known time-varying loudness of Moore/Glasberg according to [b-Glasberg 2] is not standardized, but the reference implementation provided at [b-APG] can be used to evaluate loudness for short- and long-term frames over time.

As in the ISO 532-1 standard, no binaural processing is included for the loudness assessment.

II.4.4.3 Binaural loudness assessment – version 2016

A recent enhancement of the aforementioned calculation method is provided in [b-Glasberg 3]. While the basic loudness assessment method in this new approach is almost identical to the one presented in [b-Glasberg 2], this model considers binaural inhibition between left and right ear. For this study, the reference implementation provided at [b-APG] is used.

II.4.3 Temporal aggregation of loudness vs. time

In this evaluation, three temporal aggregation functions for a loudness-vs-time curve are used to obtain a single value from the loudness distribution vs. time:

1) arithmetic mean (Avg.)
2) 95% percentile (N5)
3) average loudness level LL(p) according to [b-Fiebig] (N(i) represents loudness in sone at i-th frame):

$$LL(p) = \frac{1}{\log_{10}(2)} \cdot \sqrt{\frac{1}{K} \cdot \sum_{k=1}^{K} N(i)^{\frac{1}{\log_{10}(2)}}}$$

II.4.4 Binaural loudness summation

All analyses described above are in general carried out in binaural mode:

- in the handset mode, the left channel/ear only includes an idle channel. For the instrumental calculation, only the right channel was evaluated for the loudness models that only support perfect loudness summation. Whenever the loudness method under test implicitly applies loudness summation (i.e., assumes diotic listening of single channel file), the sone output value was multiplied by 0.5 in order to address this issue. In case of Moore/Glasberg loudness with binaural inhibition, the left channel containing idle noise was explicitly inserted into the signal under test;
- in hands-free mode, left and right channel are evaluated. In some implementations, each channel was regarded as a diotically presented signal. In these cases, loudness in sone for both channels was multiplied by 0.5 and then summed in order to obtain the overall loudness. Again, Moore/Glasberg loudness with binaural inhibition was the only model directly supporting two-channel recordings.

II.5 Auditory results

Overall, 20 normal-hearing test persons conducted the listening tests for handset and hands-free modes. Each participant listened to all samples, which led to 20 votes per sample, respectively 80 votes per condition (four samples per condition). Experts, experienced listeners, as well as naïve subjects participated in the test. Due to time and privacy constraints, no audiometric pre- or post-screening or data rejection was performed.

II.5.1 Equal reference level

The basic idea behind the assessment of the loudness functions is to obtain auditory votes for a signal with known loudness (either sone or phon). The original work in [b-AES E-Library] assumed that the level (in dBSPL) of a noise signal of 1-Bark bandwidth around 1.0 kHz can be directly interpreted in the unit phon. As already discussed in [b-AES E-Library], this approach does not reflect up-to-date research on loudness perception.

Due to this fact and to the different reference signals used for the determination of loudness functions, the assessed points per test stimulus cannot be directly transformed into sone or phon anymore. Results of the proposed test design are provided on an equal reference level (ERL) scale.

Based on the results found in [b-AES E-Library], listening test results in points are first averaged per sample. Then these average loudness values are transformed with the sigmoid functions according to equation 1 and Table II.3 to the equal reference level domain. All results in the following clauses will thus be provided on the dB ERL scale.

NOTE – In case of the sine tone as the reference signal, the loudness results could also be interpreted directly in unit phon (when strictly following the definitions of psychoacoustics). However, to keep text and graphs in this document consistent over all reference sounds, dB ERL will be used for all further investigations.

C.5.2 Equal reference level of arbitrary loudness models

In order to compare results of this test design with instrumental loudness models, which usually provide output values in phon or sone, a transformation to the ERL domain must be conducted. In previous work [b-AES E-Library], it was proposed to scale the reference sound iteratively and target at the same loudness model output as the signal-under-test.

Even though this method provides correct results, it may cause unnecessary computational effort. As an enhancement, a pre-computed mapping is proposed to convert between model output in sone/phon and auditory results on ERL scale.

The transformation function is determined by calculating sone/phon values of a certain reference sound, a loudness model and a temporal aggregate for a given level range (e.g., in steps of 1 dB). For

the transformation, the mapping function according to equation II.2 is proposed. Here *N* denotes the calculated loudness in sone of a given model, single value aggregate and reference sound.

$$ERL\ [dB] = y_{min} + (y_{max} - y_{min}) \cdot \ln\big((N - x_m) \cdot s\big) \qquad \text{(II.2)}$$

The left graph of Figure II.6 illustrates the principle for handset mode, 3-Bark noise as reference signal and time-varying loudness model of Moore/Glasberg with short-term smoothing. Here the coefficients can be determined by least-mean-square-error optimization as:

$$y_{min}, y_{max}, x_m, s = [69.647, \quad 88.542, \quad -0.826, \quad 0.103]$$

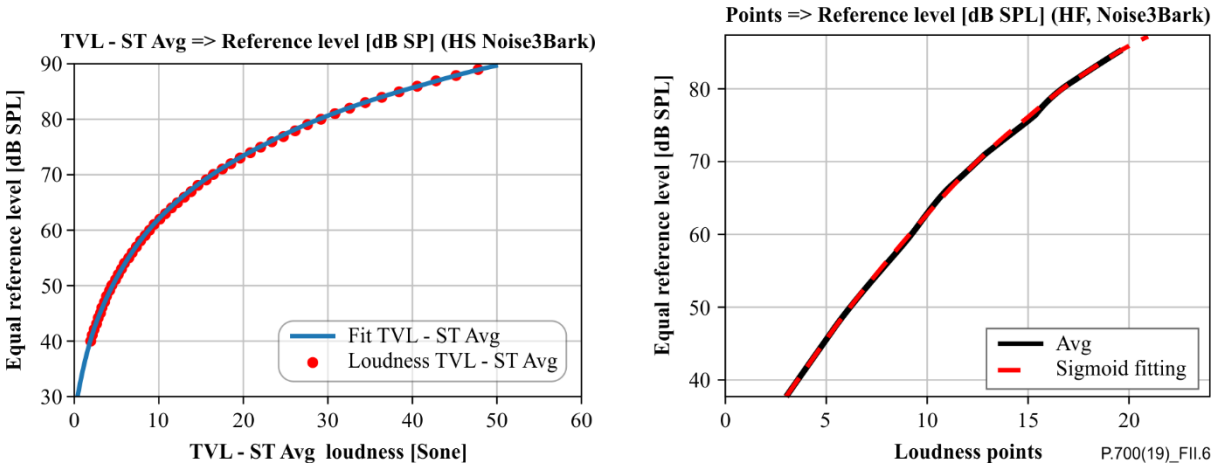


Figure II.6 – Transformation of instrumental (left) and auditory (right) results to ERL scale

As introduced in clause II.3, the auditory result points can be transformed to the ERL domain by an inverse loudness function. For the example described above, the corresponding curve is shown in the right side of Figure II.6.

II.5.3 Comparison of auditory and instrumental results

For each mode, results can be evaluated by multiple loudness models, single value aggregation methods and reference sounds. In order to provide an adequate overview, only the best-performing combinations for each loudness model are provided.

For the determination of the prediction accuracy, epsilon-insensitive root mean square error (RMSE*) according to [ITU-T P.1401] is used. This epsilon-insensitive error metric takes the uncertainty of the auditory data into account, i.e., using the 95% confidence interval as a threshold for the error calculation. Note that due to the non-linear transformation of the auditory points to dB ERL, the confidence interval is asymmetric.

As introduced in clause II.2, the active speech level can also be evaluated as a measure of loudness. For handset and hands-free scenario, the sine tone equal reference level provides best results (2.7 dB for HS, 2.6 dB for HF). Scatter plots of auditory vs. instrumental data are provided in Figure II.7.

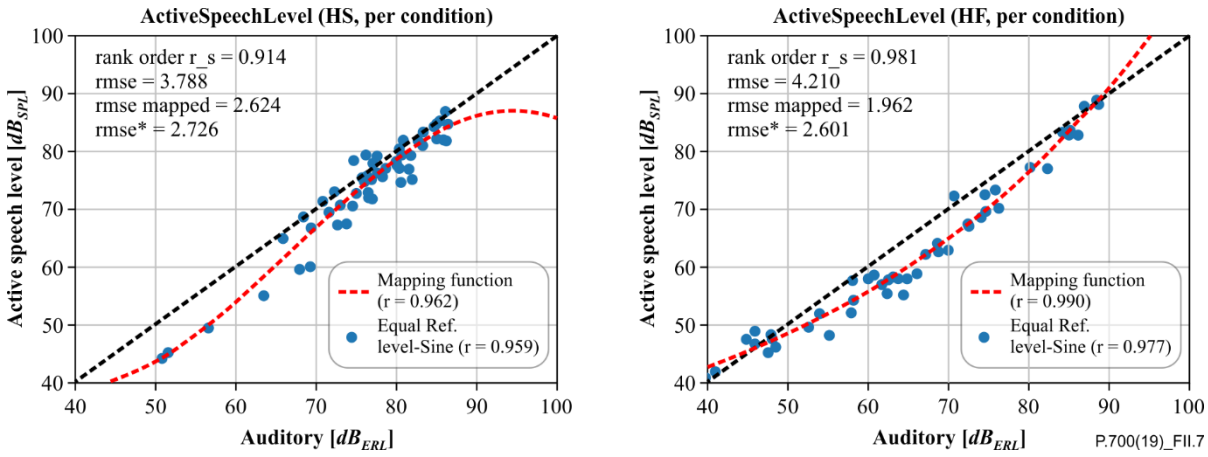


Figure II.7 – Comparison of auditory data vs. ASL ITU-T P.56 of HS (left) and HF (right)

II.5.3.1 Handset mode

Figure II.8 provides the best results for the handset mode:

- ISO 532-1 (upper left):
  - RMSE* = 1.42 dB
  - Single value calculation: average
  - Reference sound: sine tone
- P.Loudness (lower left):
  - RMSE* = 1.42 dB
  - Single value calculation: n/a
  - Reference sound: 3-Bark noise
- Moore/Glasberg binaural loudness (LT), 2016 (upper right):
  - RMSE* = 1.97 dB
  - Single value calculation: average
  - Reference sound: 3-Bark noise
- Moore/Glasberg monaural loudness (ST), 2002 (lower right):
  - RMSE* = 1.73 dB
  - Single value calculation: average
  - Reference sound: 1-Bark noise

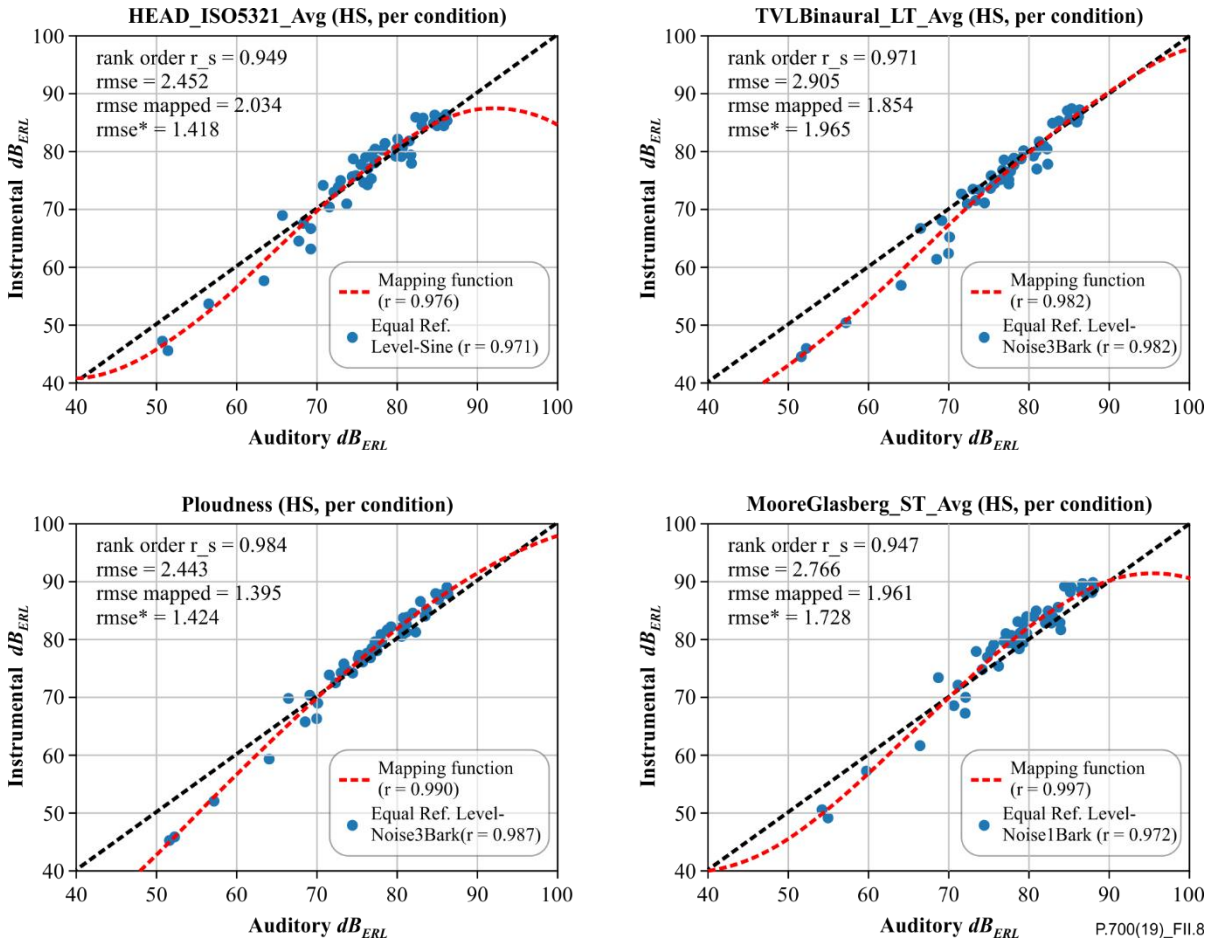


Figure II.8 – Comparison of auditory data vs. loudness models (HS)

II.5.3.2 Hands-free mode

Figure II.9 provides the best results for the hands-free mode:

- ISO 532-1 (upper left):
  - RMSE* = 1.73 dB
  - Single value calculation: average
  - Reference sound: sine tone
- P.Loudness (lower left):
  - RMSE* = 1.70 dB
  - Single value calculation: n/a
  - Reference sound: 3-Bark noise
- Moore/Glasberg binaural loudness (LT), 2016 (upper right):
  - RMSE* = 1.86 dB
  - Single value calculation: LL(p)
  - Reference sound: 3-Bark noise
- Moore/Glasberg monaural loudness (LT), 2002 (lower right):
  - RMSE* = 1.91 dB
  - Single value calculation: N5
  - Reference sound: 3-Bark noise

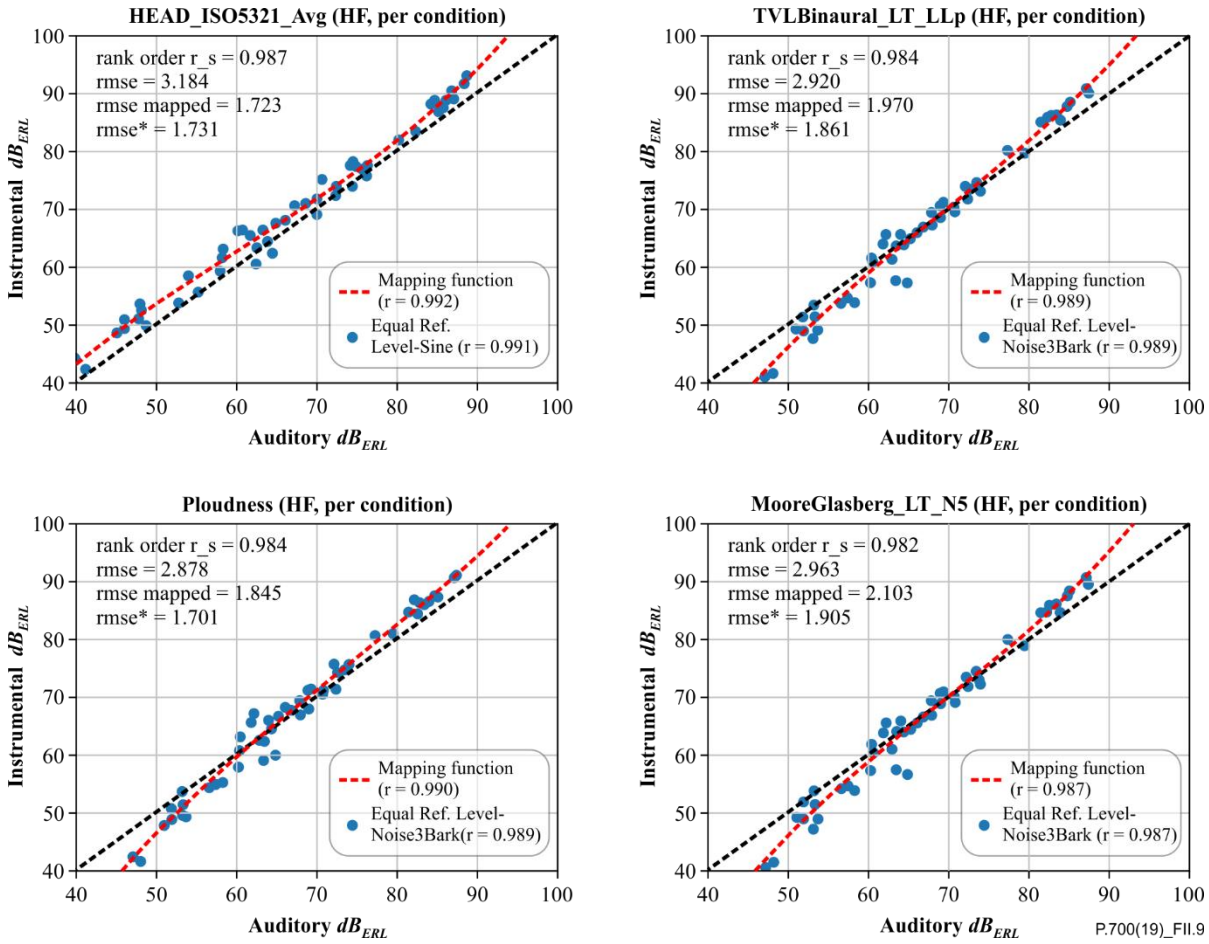


Figure II.9 – Comparison of auditory data vs. loudness models (HF)

II.6 Direct loudness model output

In the discussion of previous studies [b-AES E-Library], it was noted that representation of results on an equal-reference-scale is not the preferred one for several parties. The usage of the psychoacoustic unit phon seems to be more convenient here. Considering the original idea of the individual loudness functions in [b-AES E-Library], i.e., directly converting points to phon, all result comparisons can also be represented directly with the model output in phon.

The psychoacoustic definition (40 dBSPL at 1 kHz corresponds to 40 phon) implies that only the equal-reference-level calculation of the sine tone can be used to convert from points to phon. However, bypassing the transformation of the model output (and thus, the context of the auditory test) may obviously lead to similar correlations, but also to some shifts and offsets.

Since all results presented in the previous clauses were calculated in sone, a generic function according to Figure II.10 is used to convert from sone to phon. Certain loudness models internally may use slightly different implementations, but in general, the illustrated relation is a valid approximation for most model outputs.

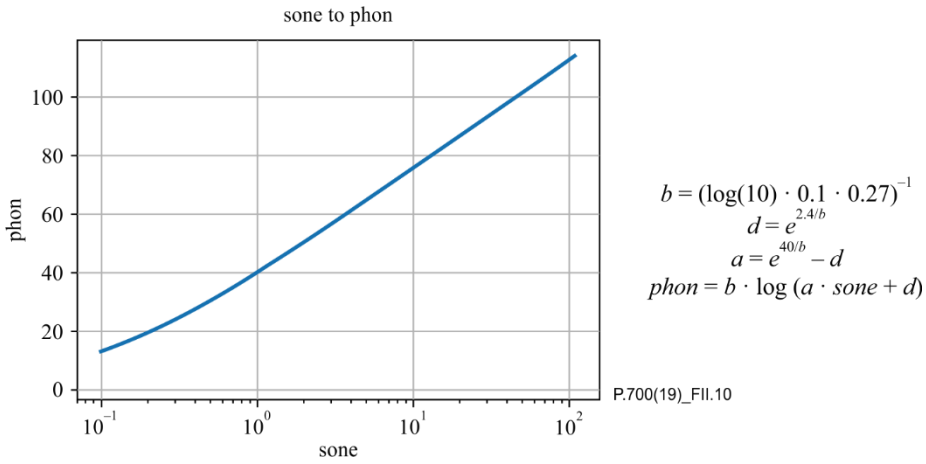


Figure II.10 – Conversion from sone to phon

II.6.1 Handset mode

Figure II.11 provides the best results for the handset mode in phon. For the ISO 532-1 and binaural Moore/Glasberg model, prediction performance slightly decreases. For monaural Moore/Glasberg, the comparison on phon scale performs even better than on the equal-reference-scale. On the other hand, the P.Loudness model obtains much too pessimistic sone/phon values.

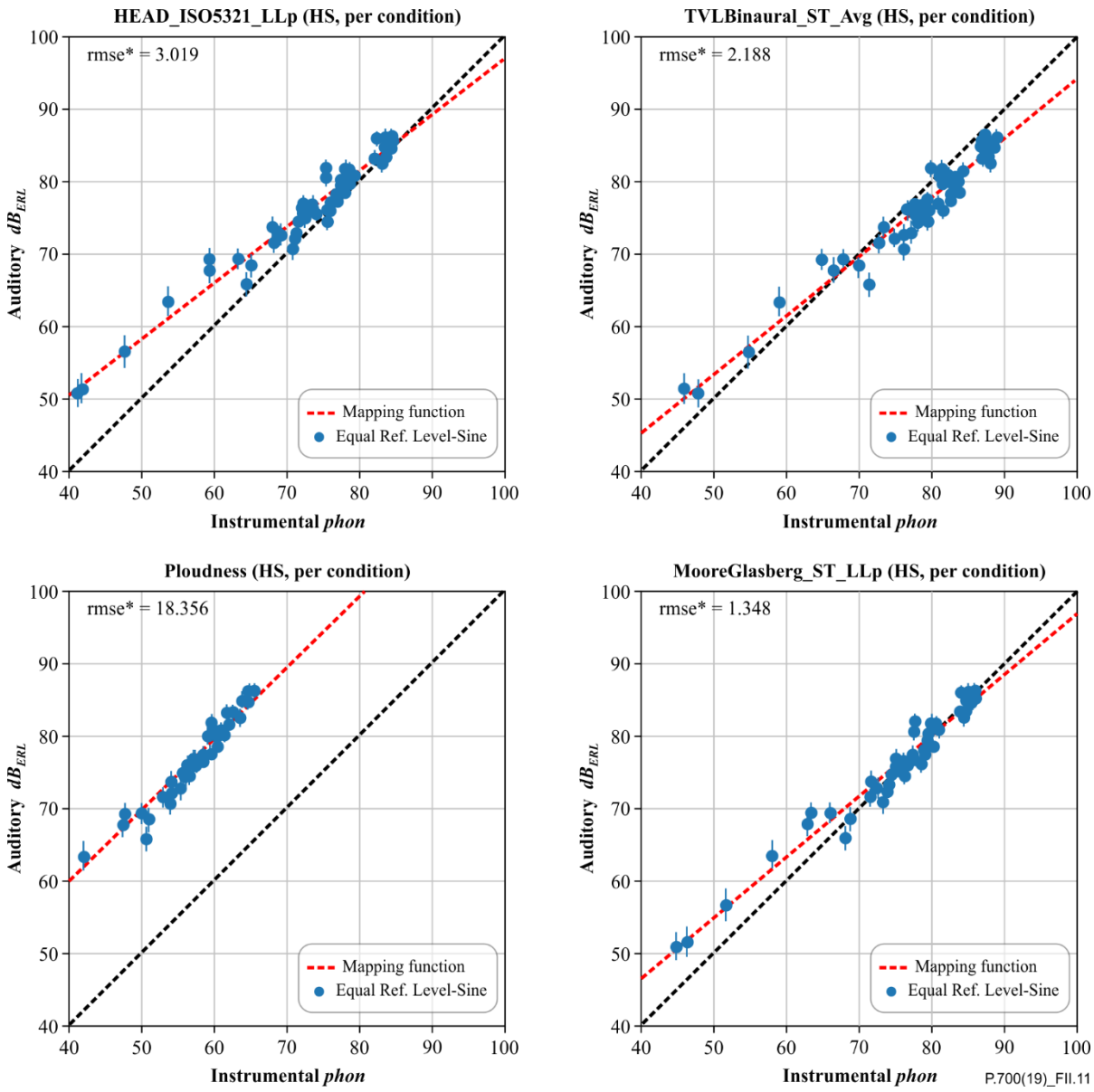


Figure II.11 – Comparison of auditory data vs. loudness models in phon (HS)

II.6.2 Hands-free mode

Figure II.12 provides the best results for the hands-free mode in phon. Again, prediction accuracy of ISO 532-1 slightly decreases, but still is adequate. Since most of the listening test conditions did not include many level differences between left and right ear, it is not surprising that monaural and binaural Moore/Glasberg loudness performs very similar. Both models over-predict loudness in a similar way. The binaural model performs slightly better, possibly due to the binaural inhibition algorithm. The P.Loudness model for hands-free over-predicts loudness, too.

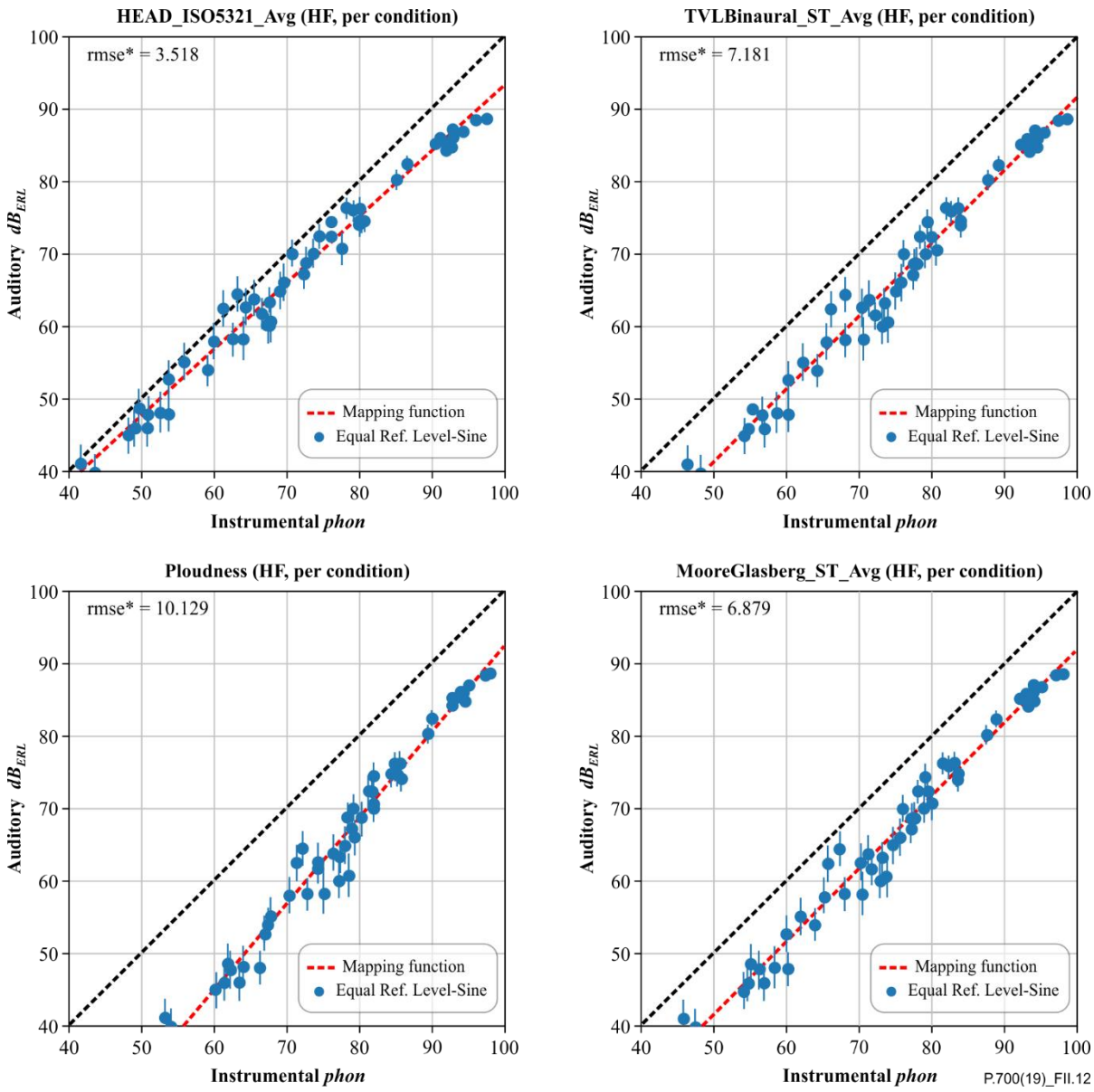


Figure II.12 – Comparison of auditory data vs. loudness models in phon (HF)

II.7 Conclusions

This appendix presents updated auditory and instrumental results of recently conducted listening tests. The test corpus is based on realistic binaural recordings of several terminals including an acoustic bandwidth from NB to FB.

The number of test subjects was increased from 15 to 20, which led to more accurate auditory results. In addition to the recently introduced 3-Bark reference noise, sine tone and 1-Bark noise were added to the assessment of individual loudness functions. In combination with the proposed transformation of the equal-reference-level, multiple single value calculations are possible for each loudness model.

As a prediction performance indicator, RMSE* was introduced in this contribution in order to address the auditory uncertainty of the data. In the evaluation of loudness models, the well-known active speech level according to [ITU-T P.56] provides already a "baseline" performance of approximately 2.7 dB for both modes.

The best-performing configurations of each loudness model category (Zwicker approach, Moore/Glasberg monaural and binaural, P.Loudness) provide accurate predictions with RMSE*

between 1.4 and 1.9 dB. However, compared to the active speech level calculation, the "gain" of loudness methods is only in the range of approximately 1 dB.

Even though the P.Loudness proposal provides adequate prediction results (at least on the equal-reference-scale), the source prefers the usage of the standardized ISO 532-1 loudness model. Based on the observations presented in this study, several advantages are obvious:

- best RMSE* for ISO 532-1 loudness model with average single value and sine tone as reference sound;
- performs best on equal-reference-scale as well as on absolute phon-scale;
- performs best for both handset and hands-free mode. Thus, this method would solve the issue of not having two but just one single loudness model that works for both applications;
- in addition, this specific method is already widely tested by other standardization groups. It provides reference C-code, which is available for standard-conforming implementations.

As an alternative model the binaural loudness model according to Moore/Glasberg described in [b-ISO 532-2] from 2017 could be considered. Even though the performance metrics were slightly worse, this model is currently the only one that considers binaural inhibition. Possibly, it is best prepared for more advanced hands-free scenarios, which contain increased spatial aspects. Such conditions were not part of the current evaluation.

Appendix III

Result of loudness experiment C

(This appendix does not form an integral part of this Recommendation.)

III.1 Introduction to loudness experiment C

III.1.1 Background

The P.Loudness work item aims at specifying a loudness prediction model which is suitable for speech communication systems. It is not intended as a direct replacement of ITU-T P.79 loudness ratings (SLR, receive loudness rating (RLR), junction loudness rating (JLR), overall loudness ratings (OLR)), which are a measure of sensitivity, but intended to predict the absolute loudness perceived by e.g., a user of a telephone.

Generic loudness algorithms for a variety of signal types have existed for a long time and the focus of ITU-T SG12 is to validate their applicability for specifically speech. In particular, to study ISO 532-1.

III.1.2 Motivation

One interesting usage of a loudness predictor is for assessing the output capability of mobile phones in hand-held hands-free mode (speakerphone), where low frequencies are difficult to reproduce due to size constraints. The situation of comparing two mobile phones in speakerphone mode was addressed at a similar sound pressure level but having different frequency response. Or to compare two tunings of the same mobile phone.

Loudness balancing has been used extensively in the past, however, such experiments were repeated but with consideration of:

– FB context;

– no send-side filtering;

– high-enough cut-off frequency for high-pass filters.

Recent studies do have these conditions covered, but in a mix with other conditions and not separately studied.

If an objective model can correctly predict results using also the loudness balancing test paradigm, it provides additional validation for that model.

III.1.3 Earlier experiments within the P.Loudness work item

Loudness of speech and other signals with standard narrowband, wideband and super-wideband filters/codecs was reported by Orange in [b-AES E-Library] and other contributions. This gives e.g., the useful information about the effect of the low-pass at 14, 7 and 3.5 kHz and also the effect of high-pass at ~300 Hz. It also provides data at different listening levels.

HEAD acoustics reported on loudness of speech in [b-AES E-Library] and other contributions, for a variety of speech bandwidths and also with real terminals and speech with dynamic compression. This gives useful validation data for a wide set of conditions.

Brüel & Kjaer, Orange and Delta reported on loudness of speech and noisy speech for a variety of conditions in [b-AES E-Library] and other contributions, also giving useful validation data.

The subjects in the auditory experiments above rated the perceived loudness on a numeric loudness scale with an additional category labelling. Reference sounds of 1 kHz (tone or narrowband noise) were included for aligning with the phon scale directly or indirectly.

At an earlier stage, Ericsson reported on a loudness balancing experiment to assess the difference between narrowband, wideband, super-wideband and fullband in [b-AES E-Library]. The user interface was MUSHRA-like with faders for the volume of each sample. The present study is similar but with plus/minus buttons instead of faders, and other high-pass conditions.

II.1.4 Historical survey

For the purpose of putting the study into context, a small part of telephony-loudness history is provided in the following, including analysis into the Geneva CCITT/ITU-T library.

In the early days of telephony, achieving adequate loudness was a major concern and not always a simple task. The assessment of terminals was made by comparing the terminal to a reference connection with controlled characteristics. A talker would speak with reasonably consistent speech level and a listener would adjust an attenuator until equal loudness was perceived when comparing to the reference. The final value of the attenuator would be the test result, or a combination attenuator reading from various experiments. For more information on such methodologies, see [ITU-T P.78].

Instrumental measurements were later introduced using artificial mouths and ears.

One interesting study is found in IUT-T SG12 contribution 56 from 1977, by France [b-CCITT 1977 SG12-C-0056]. The loudness balancing concept was implemented instrumentally using the 1973 Zwicker loudness model ISO 532B. Good match to subjective data was reported, see Figure III.1.

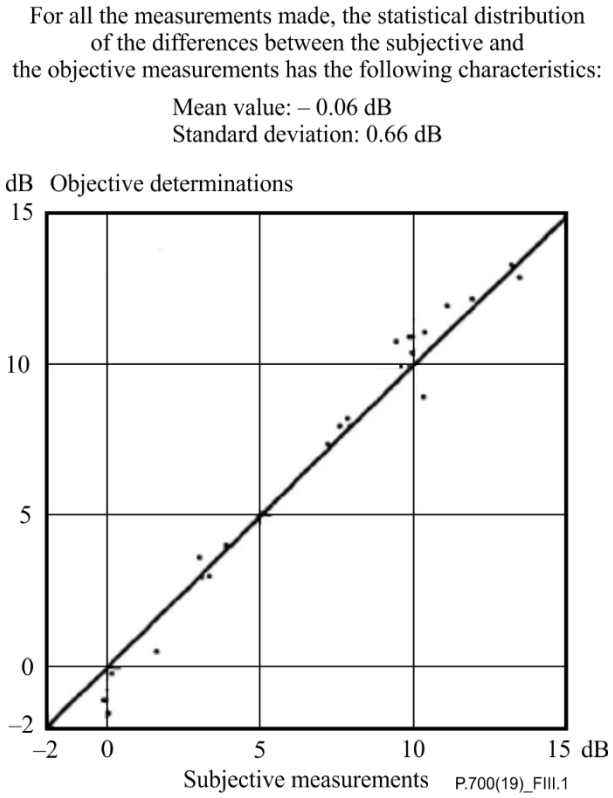


Figure III.1 – Result of the loudness balancing concept by ITU-T SG12

This is continued in e.g., SG12 contribution 15 from 1982, from France [b-CCITT 1982 SG12-C-0015]. Again, good match to subjective data was reported over a variety of handsets. However, at this time, speakerphone with micro-speakers was, for natural reasons, not included.

It was found that although the ISO 532B method did produce a good result, it was considered "overkill" at the time. [ITU-T P.79] still states that "comprehensive models" are "unnecessarily complicated":

"This model does not claim to represent accurately all the features that relate to perception of the loudness of speech; for example, the effects of inter-frequency masking are ignored, and it does not predict the increasing importance of the lower frequencies as the intensity of the sound is increased from the threshold. It is possible to construct models that represent more of the features fairly well, but no completely comprehensive model is known. Such models are unnecessarily complicated for calculating loudness ratings. The most important restriction with respect to this model is that it should be used to make comparisons at the constant listening level indicated in [ITU-T P.76]."

With various reasonable assumptions, it was possible to reduce the Zwicker-based calculations to what is now known as SLR, RLR, JLR, OLR. The sound pressure in each 1/3rd-octave band presented to the ear of a user is predicted by measuring the device under test and assuming certain network and far-end terminal characteristics, the weighting factors W_i are used to consider a typical speech spectrum, the B-party frequency response. The frequency-dependent loudness perception (at the levels of interest) and the *m* (loudness growth) factor is used to convert to a quantity that allows the contributions from different frequency bands to be properly summed (assuming certain typical presentation levels). With a minimum of calculations, a simplified version of the critical band summation in ISO R532B and other such loudness models was created in ITU-T P.79. Finally, the result was converted back to decibels telling us how much loss is needed to match a certain reference connection, as in the original subjective loudness balancing method.

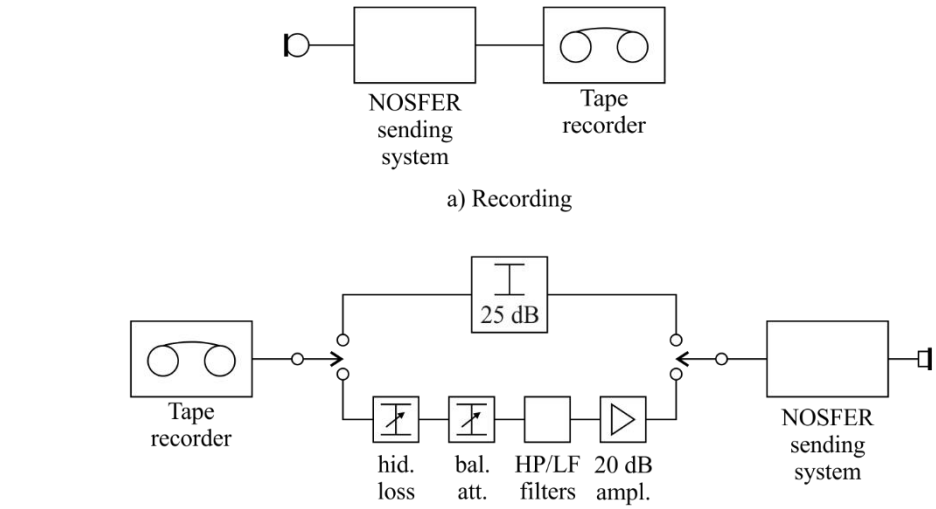
$$LR = -\frac{10}{m} \cdot \log_{10} \sum_{i=N_1}^{N_2} 10^{0.1 \cdot m(S_i - W_i)}$$

Figure III.2 – ITU-T P.79 equation 5-1

To determine what weighting factors to be used, loudness balancing experiments were conducted. One example is found in SG12 contribution 10, from 1981, where China reported on experiments with high-pass and low-pass filters inserted in series with send and receive NOSFER characteristics.

In conclusion, a small part of telephony-loudness history has been covered here. It is also worth mentioning that much early speech research is not described here, e.g., by Stevens, Zwislocki and by Fletcher and Galt, see e.g., [b-Fletcher].

In recent times, advanced loudness models based on Zwicker and Moore/Glasberg including not only spectral masking, but also temporal masking effects have been published. Recently, ISO published such model in ISO 532-1, based on Zwicker.



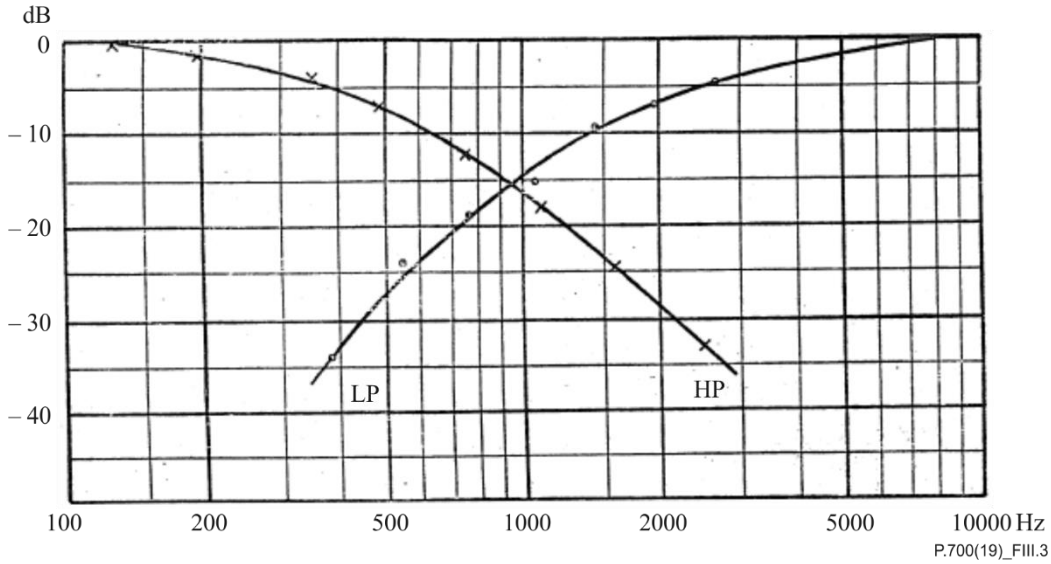


Figure III.3 – From contribution 10, 1981, P.R. China

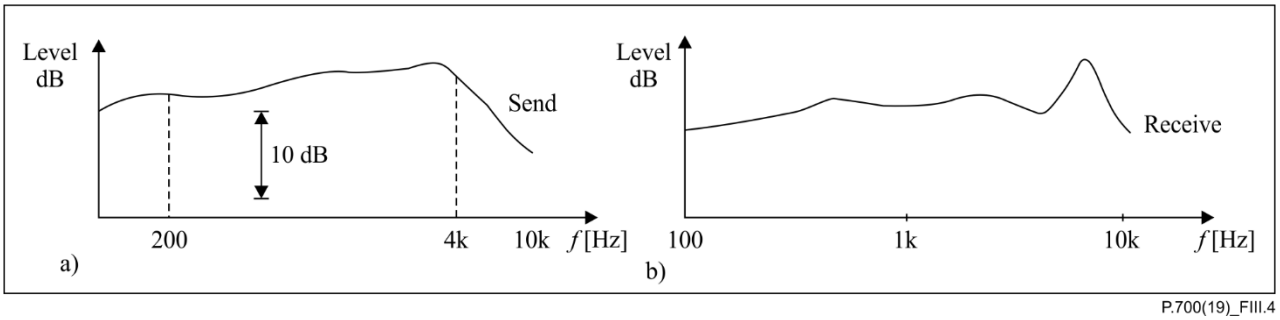


Figure III.4 – NOSFER send and receive characteristics, from loudness rating and other measures of loudness, B&K

III.2 Subjective experiment

III.2.1 Test design

For each sub-test, four sound files were available for playback by the subject:

– original fullband (reference);

– three high-pass filtered versions (3rd order Butterworth at 250, 500 and 1000 Hz).

The reference was placed on the left of the user interface screen and the 250, 500 and 1000 Hz conditions followed from left to right, in randomized order for each subtest.

A graphical user interface provided Play buttons for all files and +/– gain buttons for the high-pass filtered files. The resulting gain was not visible to the test subject.

The reference file was fixed at 65 dB active speech level (as per ITU-T P.56).

The filtered files were first presented at a random level and then adjusted by the test subject in 1 dB steps until equal loudness of all four samples was perceived. The subject could repeat playback as wanted. The final gain for each file was logged. The peak level was also logged for each rated file.

The presentation was diotic (headphone presentation with same signal to both ears).

III.2.2 Initial level randomization

The goal was to randomize the initial level so that the user sometimes needed to raise and sometimes needed to reduce the level of the samples, avoiding guiding the listener in a certain direction.

To this end, sone values were calculated for each file using ISO 532-1 [b-Head acoustics], from 40 to 90 dBSPL presentation level. For conversion, back and forth between sone and SPL, a sigmoid function and a logarithmic function were fitted for each stimulus, to allow interpolation and conversion in both directions as needed, following [b-AES E-Library].

$$N(Level)\ [sone] = p_{min} + \frac{(p_{max} - p_{min})}{1 + e^{a \cdot (p_0 - \mathrm{Level})}} \quad \text{(III.1)}$$

$$Level(N)\ [dBSPL] = y_{min} + (y_{max} - y_{min}) \cdot \ln\big((N - x_m) \cdot s\big) \quad \text{(III.2)}$$

The non-active parts of the files were removed using the speech activity detection from ITU-T P.56, before calculating the overall mean sone value. The ISO 532-1 parameters were:

– type: 'time_varying'

– field: 'D'

From the baseline of expected "correct" level alignment (based on ISO 532-1), a further random gain was applied, up to ±6 dB uniform distribution. The reference file was always kept at 65 dB active speech level.

The result was that samples were initially presented with a random loudness offset and the task of the test subject was to press "+" and "–" to bring the samples back to equal loudness, without seeing how much gain or loss they inserted. Hence, the protocol was similar to [b-AES E-Library] but with invisible gain/loss values.

III.2.3 Headphone equalization and level adjustment

Open-back headphones were placed on an ITU-T P.58 compliant HATS and the frequency response was measured using programme simulation noise (see [ITU-T P.381]), in 1/12th octaves. [ITU-T P.58] Annex A DRP to diffuse-field correction was used. The headphones were re-seated five times and the curves were dB-averaged. The curves were then smoothed, averaging five bands (two below, two above) to avoid creating sharp peaks by the equalization. A 1000-tap FIR filter was fitted to the desired response. As a validation, the equalized headphone was then also measured using speech files and the response curve checked to be flat with 1/3rd band analysis.

Finally, the playback level was adjusted to give 65 dB active speech level after diffuse-field equalization. (It was first planned to run the experiment at 73 dB active speech level but after a pilot test and to better match hands-free use cases, it was decided to use 65 dB ASL, as seen in some other contributions within P.Loudness.)

III.2.4 Test environment

Tests were conducted in an office-type lab room with ambient noise level below 30 dBSPL(A) background noise (typical 25 dB(A)). A computer with RME soundcard, and a graphical user interface for user interaction, was used.

III.2.5 Test sequence

The ID, age and gender was collected. The instructions were given with text and graphics. A familiarization test preceded the actual test. The complete test duration was about 20 minutes, with some variation depending on the subject.

III.2.6 Speech samples

[ITU-T P.501] clause 7.3 British English samples. Three male plus three female single sentences plus one sentence for familiarization. The materials were selected from clause 7.3, rather than from Annex B, to include talkers with a low fundamental frequency. It is also the same material used in some other related studies, and in the currently specified loudness rating test in many standards.

III.2.7 Assessors and screening

Twenty-two naïve assessors participated in the test. They had no known hearing impairment, based on self-reporting. The subjects were all fluent in the English language and intelligibility was not considered to be an issue in the testing (clear and clean speech at high SNR).

Post screening criteria were based on the assumption that a 1000 Hz high-pass filter should warrant a higher compensation gain than 250 Hz.

The confidence intervals were slightly reduced after screening. Remaining assessors are described in Tables III.1 and III.2 below. Heatmaps of results are presented with and without screening, in Figure III.7 and Figure III.8.

Table III.1 – Assessors

No of assessors	Total	Female	Male
	18	7 (39%)	11 (61%)

Table III.2 – Age

Age of assessors	Mean	Std dev	Min	Max
	26.3	7.6	19	44

III.2.8 Instructions and familiarization

Only written instructions were given. Subjects could ask questions at any time, but no questions were asked.

The test included a test round for familiarization, using a different talker and sentence, from the same ITU-T P.501 corpus.

III.2.9 Comments from subjects

Subjects could comment after the test. Some expressed difficulty in deciding whether they perceived equal loudness or not. In a pilot test using trained subjects, a comment was made that for a high-pass filtered stimulus with the same overall loudness as the flat sample, the mids and highs were detected to be on a higher level and such trained subject could then be tempted to rate this higher band level rather than the overall loudness. This could apply to loudness comparisons in general, whether or not it is a side-by-side comparison of actual products or whether it is a lab test such as this.

III.3 Results and comparison to ISO 532-1 objective results

III.3.1 Test results

Like in the classical loudness balancing studies, the result is the gain needed for a filtered file to match the loudness of the original un-filtered file. The results are presented on a sample-basis rather than condition basis, since the high-pass filtering affected different source material differently (e.g., female and male talkers).

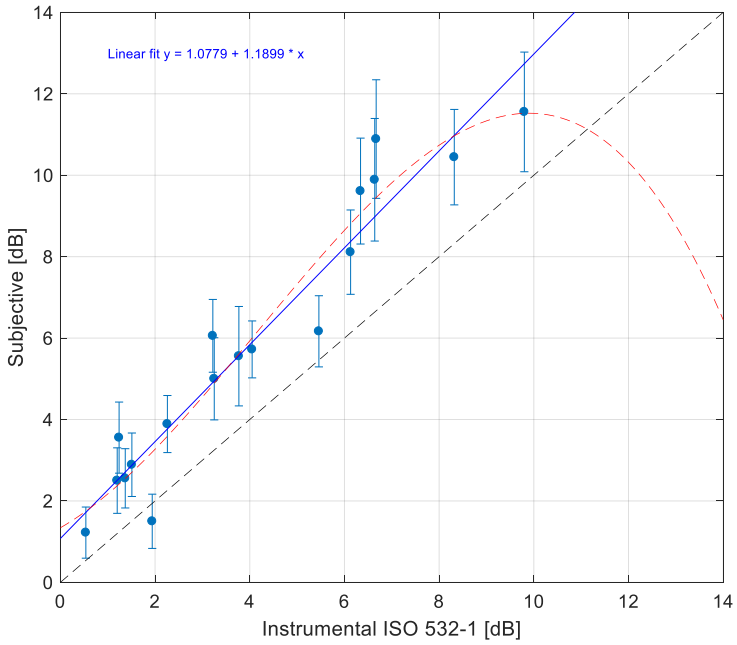


Figure III.5 – Gain applied when loudness-balancing high-pass filtered speech against the unfiltered reference. Subjective results with 95% confidence intervals are plotted vs ISO 532-1 instrumental results. Blue line: linear fit, red curve: 3rd order polynomial fit, dashed black line: perfect fit

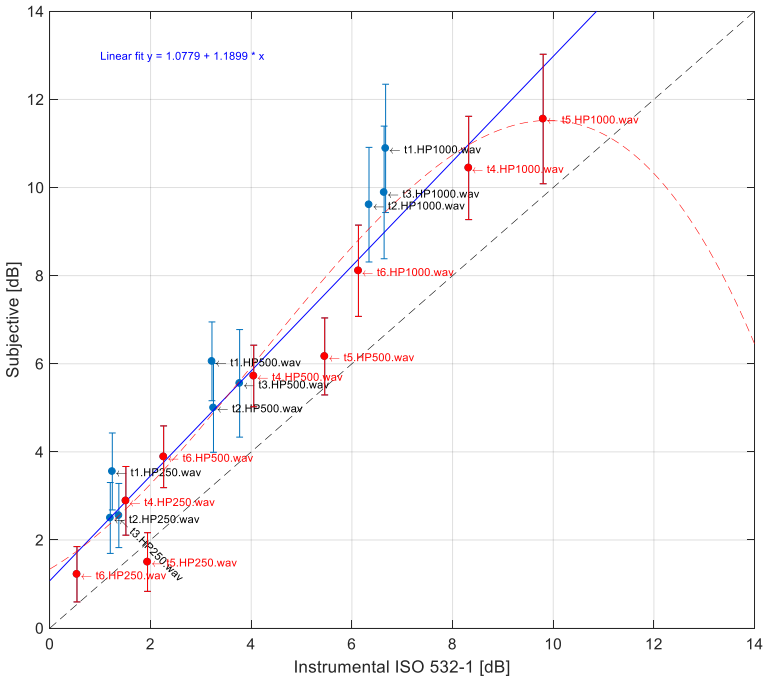


Figure III.6 – Gain applied when loudness-balancing high-pass filtered speech against the unfiltered reference. Subjective results with 95% confidence intervals are plotted vs ISO 532-1 instrumental results. Same as the previous figure but with stimuli labels added. Red markers represent female talkers, blue markers represent male talkers. Blue line: linear fit, red curve: 3rd order polynomial fit, dashed black line: perfect fit

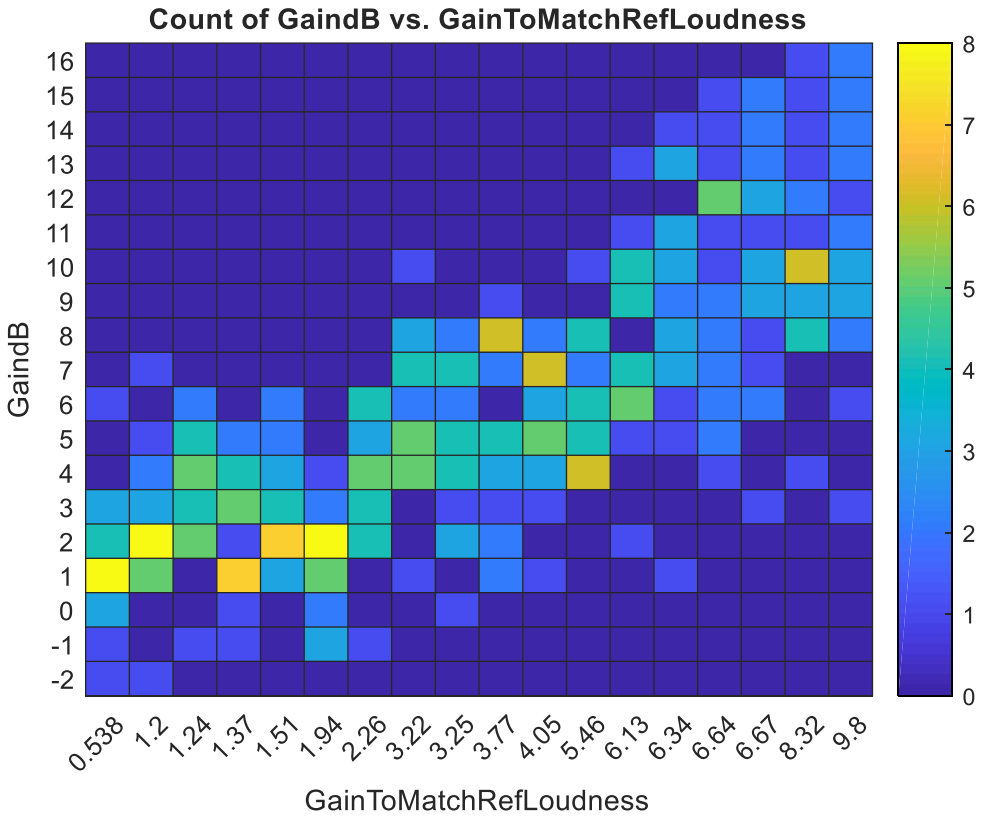


Figure III.7 – Heatmap of gains applied when loudness-balancing high-pass filtered speech (subjective vs. ISO 532-1), all subjects. Lighter colours represent higher occurrence

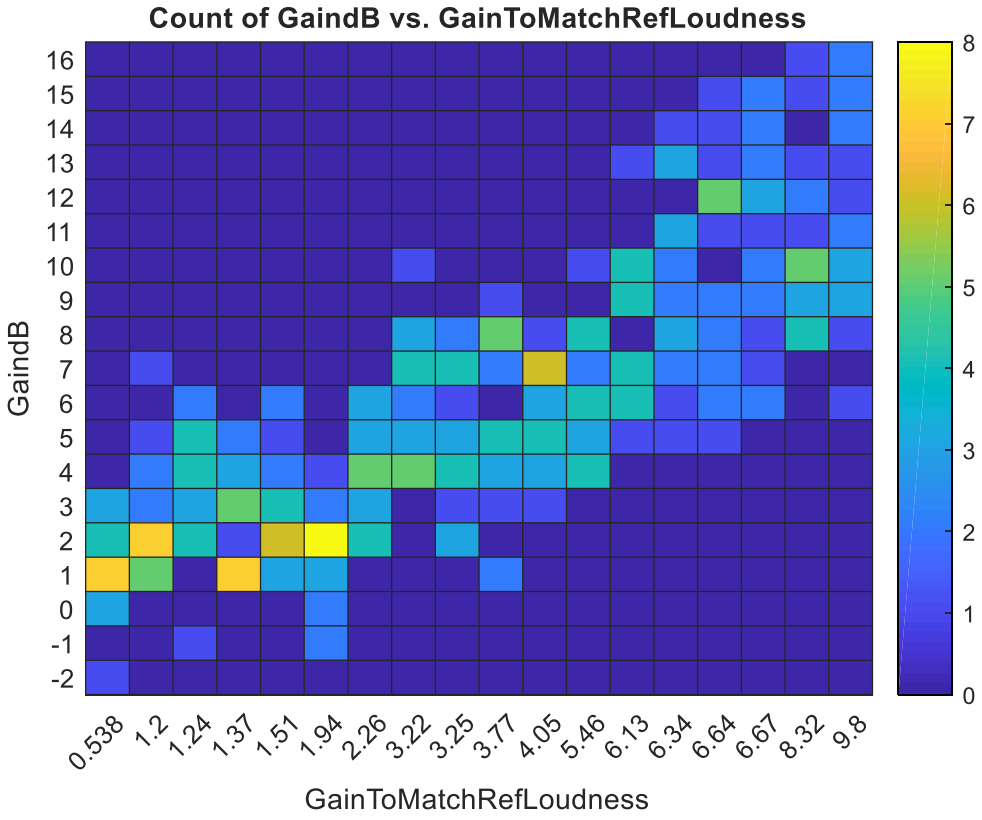


Figure III.8 – Heatmap of gains applied when loudness-balancing high-pass filtered speech (subjective vs. ISO 532-1), non-rejected subjects. Lighter colours represent higher occurrence

Statistics based on [ITU-T P.1401], unmapped results:

– Pearson's correlation coefficient: 0.96

– RMSE: 2.2 dB

– RMSE*: 1.2 dB

For mean opinion score (MOS) experiments it is described in [ITU-T P.1401] how to further map the data using up to 3rd-order polynomial fit, before comparing to the objective data. This is motivated by the *context effect* in each experiment.

In this experiment, the task for the subjects is different, namely, to balance two sounds to each other. It is assumed that there is no similar context effect in such experiment and therefore not attempted to re-map the data.

Talker gender analysis:

– the under-prediction of the model seems to depend on talker gender. (Hypothesis that talker gender does not matter is rejected in a Student's t-test, p-value 3.9e-4);

– the high-pass filtering causes a larger drop in loudness for male talkers, than for female talkers. Since this drop-in loudness is underestimated by ISO 532-1, the prediction error for high-pass filter conditions is larger for male talkers.

The data from this experiment is in reasonably good agreement with data from previous experiments, as summarized in Table III.3.

Table III.3 – Comparison to other studies with high-pass filtered speech. The conditions were not identical to the present study why comparisons mush be made with caution

Other study	Condition	Other study value	This subjective experiment, 3rd-order Butterworth -3dB point, 65 dBSPL	ISO 532-1, 3rd-order Butterworth -3dB p oint, 65 dBSPL
[b-CCITT 1981 SG12-C-0010]: steep high-pass, using NOSFER in send and receive, unknown presentation level, loudness balancing	250 Hz steep high-pass	2.5 dB	2.4 dB	1.3 dB
	500 Hz steep high-pass	7.5 dB	5.4 dB	3.7 dB
	1000 Hz steep high-pass	16.5 dB	10.0 dB	7.3 dB
[ITU-T G.191]: mod IRS filter, 67dBSPL, point scoring with ref 1kHz tones	300 Hz to 3.4 kHz (using flat receive-side modified IRS)	3 phon	(most similar condition 2.4 dB)	(most similar condition 1.3 dB)
[b-AES E-Library]: ITU-T G.191 filter, preferred presentation level, loudness balancing	MSIN filter	3.36 dB (female) 2.9 dB (male)	(most similar condition 2.4 dB)	(most similar condition 1.3 dB)

III.3.2 Check of impact of initial presentation gain offset

As described in clause III.2.2, the initial balance in loudness was randomized, before the subjects started their adjustment. When analysing the relation between subjective scores and the initial presentation gain offset, a very low correlation was found (Pearson's correlation coefficient 0.0471). This suggests that bias in the results due to the initial balance, could indeed be avoided.

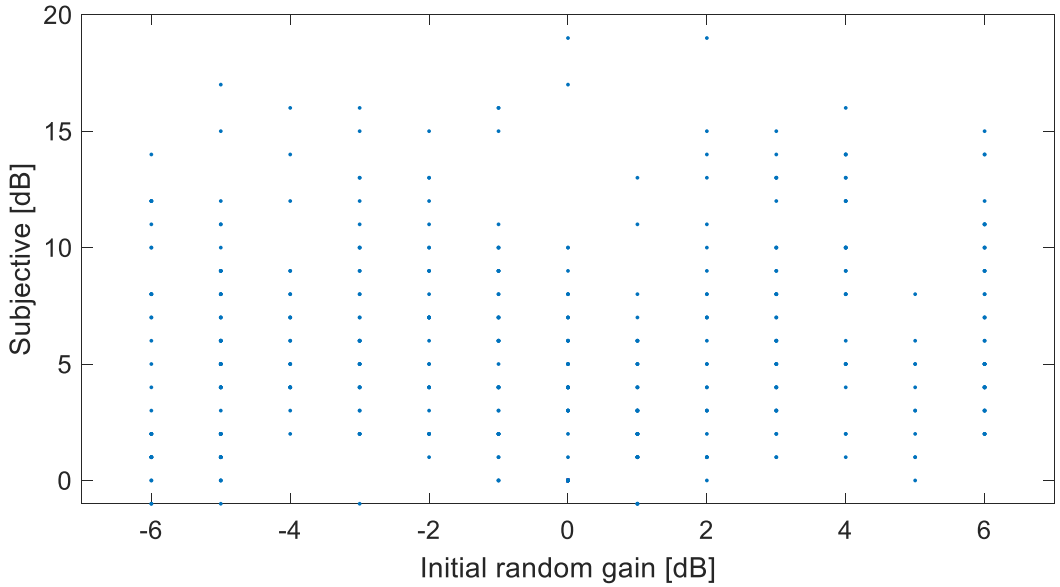


Figure III.9 – Subjective scores as a function of initial random gain – randomness is desirable

III.4 Conclusion

The gain necessary to restore the loudness of speech after high-pass filtering was investigated using a loudness balancing test paradigm. The amount of gain needed to restore the loudness after high-pass filtering was in general higher for the auditory results compared to the ISO 532-1 algorithm, at the presentation level of 65 dBSPL (for the reference sample). The difference between subjective and instrumental scores is statistically significant (p-value 1.27e-6).

A linear fit of the subjective data (y) against the instrumental data (x) gives:

$$y = 1.1 + 1.2x$$

However, as expected with a larger dataset, such data fitting curve should become bound to origo (0,0), since the balancing of two identical sound files should be rated as 0 dB by both subjective and instrumental methods.

Our overall interpretation is that when using loudness balancing of speech at 65 dBSPL, the results suggest that the loudness drop due to a high-pass filter is detected by ISO 532-1, with a fair prediction accuracy.

It should however be noted that the loudness balancing test paradigm is likely not used in the training of ISO 532-1. It should also be noted that the observed differences between auditory and instrumental results are not severe and that the dataset is small. The results do not, on their own, warrant to disqualify the ISO algorithm for the P.Loudness purposes.

Appendix IV

Nominal transmission paths

(This appendix does not form an integral part of this Recommendation.)

IV.1 Nominal receive paths

Nominal receive paths are defined for the purpose of calculating the sending loudness:

– diffuse-field corrected sensitivity/frequency characteristics:
  - narrowband: 6.28 dBPa/V for all applicable frequencies;
  - wideband, super-wideband and fullband: 4.83 dBPa/V for all applicable frequencies.

These receive characteristics correspond to RLR = 2 dB when using [ITU-T P.58] correction between DF and DRP, [ITU-T P.57] correction from DRP to ear reference point (ERP) and RLR calculation according to [ITU-T P.79]. The quantities for the calculation are found in Tables IV.1 and IV.2.

NOTE – The IRS [ITU-T P.48] and modified IRS [ITU-T P.830] paths have not been assumed in this case. Rather a narrowband sending terminal is assumed to be connected to a receiving terminal which is not restricted in bandwidth but where the overall sensitivity is adjusted.

Table IV.1 – Calculation of RLR for the nominal narrowband receive path

Band No.	Mid-frequency (Hz)	Receive char. DF (dBPa/V)	DF-to-DRP (dB) from Table 3/P.58	DRP-to-ERP (dB) from Table 2a/P.57	Receive char. ERP (dBPa/V)	Receive W_{ri} from Table 1/P.79	x_i $10^{\frac{0.175}{10}(S_i - W_i)}$
1	200	6.28	0.0	0.0	6.28	85.0	4.14e-02
2	250	6.28	0.5	–0.3	6.48	74.7	6.32e-02
3	315	6.28	0.5	–0.2	6.58	79.0	5.34e-02
4	400	6.28	1.0	–0.5	6.78	63.7	9.97e-02
5	500	6.28	1.5	–0.6	6.68	73.5	6.69e-02
6	630	6.28	2.0	–0.7	7.08	69.1	8.12e-02
7	800	6.28	4.0	–1.1	9.18	68.0	9.23e-02
8	1000	6.28	5.0	–1.7	9.58	68.7	9.12e-02
9	1250	6.28	6.5	–2.6	10.18	75.1	7.22e-02
10	1600	6.28	8.0	–4.2	10.08	70.4	8.69e-02
11	2000	6.28	10.5	–6.5	10.28	81.4	5.63e-02
12	2500	6.28	14.0	–9.4	10.88	76.5	7.02e-02
13	3150	6.28	12.0	–10.3	7.98	93.3	3.17e-02
14	4000	6.28	11.5	–6.6	11.18	113.8	1.58e-02
$RLR = -\frac{10}{0.175} \cdot \log_{10} \sum_{i=1}^{14} x_i = 2\ dB$							

Table IV.2 – Calculation of RLR for the wideband nominal receive path (also used for super-wideband and fullband)

Band No.	Mid-frequency (Hz)	Receive char. *DF* (dBPa/V)	DF-to-DRP (dB) from Table 3/P.58	DRP-to-ERP (dB) from Table 2a/P.57	Receive char. *ERP* (dBPa/V)	Receive W_R from Table A.2/ P.79	x_i $10^{\frac{0.175}{10}(S_i - W_i)}$
1	100	4.83	0.0	0.0	4.83	152.8	2.57e-03
2	125	4.83	0.0	0.0	4.83	116.2	1.12e-02
3	160	4.83	0.0	0.0	4.83	91.3	3.07e-02
4	200	4.83	0.0	0.0	4.83	85.3	3.91e-02
5	250	4.83	0.5	–0.3	5.03	75.0	5.96e-02
6	315	4.83	0.5	–0.2	5.13	79.3	5.04e-02
7	400	4.83	1.0	–0.5	5.33	64.0	9.40e-02
8	500	4.83	1.5	–0.6	5.23	73.8	6.31e-02
9	630	4.83	2.0	–0.7	5.63	69.4	7.66e-02
10	800	4.83	4.0	–1.1	7.73	68.3	8.71e-02
11	1000	4.83	5.0	–1.7	8.13	69.0	8.61e-02
12	1250	4.83	6.5	–2.6	8.73	75.4	6.81e-02
13	1600	4.83	8.0	–4.2	8.63	70.7	8.20e-02
14	2000	4.83	10.5	–6.5	8.83	81.7	5.31e-02
15	2500	4.83	14.0	–9.4	9.43	76.8	6.62e-02
16	3150	4.83	12.0	–10.3	6.53	93.6	2.99e-02
17	4000	4.83	11.5	–6.6	9.73	114.1	1.49e-02
18	5000	4.83	11.0	–3.2	12.63	144.6	4.90e-03
19	6300	4.83	8.0	–3.3	9.53	165.8	1.84e-03
20	8000	4.83	6.5	–16.0	-4.67	166.7	1.00e-03
$RLR = -\frac{10}{0.175} \cdot \log_{10} \sum_{i=1}^{20} x_i = 2\ dB$							

Bibliography

[b-ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience.*

[b-AES E-Library] Audio Engineering Society, *AES E-Library, Subjective and Objective Measurements of Speech Loudness in Hands-Free Telephony—Toward an Extended Loudness Model for Telephonometry*. <http://www.aes.org/e-lib/browse.cfm?elib=17996>

[b-ISO 532] ISO 532:1975 (2012), Acoustics – *Method for calculating loudness level.*

[b-ISO 532-1] ISO 532-1:2017, *Acoustics – Methods for calculating loudness – Part 1: Zwicker method.*

[b-ISO 532-2] ISO 532-2:2017, *Acoustics – Methods for calculating loudness – Part 2:Moore-Glasberg method*.

[b-CCITT 1977 SG12-C-0056] COM12 – C 56 – E, France (1977): *Objective method for determining the send "relative equivalents" of telephone sets*.

[b-CCITT 1981 SG12-C-0010] COM12 – C 10 – E, Ministry of posts and telecommunications of the people's republic of China (1981): *High-pass and low-pass filters tests and parameters m, G and S' estimated therefrom.*

[b-CCITT 1982 SG12-C-0015] COM12 – C 15 – E, France (1982): *Objective determination of local system sensitivity for send, receive and sidetone, and comparison with the results of subjective measurements of relative equivalents.*

[b-DIN 45631] DIN 45631/A1:2010-03, *Calculation of loudness level and loudness from the sound spectrum – Zwicker method – Amendment 1: Calculation of the loudness of time-variant sound*.

[b-3GPP SA4] DIN 45631/A1:2010-03, *Calculation of loudness level and loudness from the sound spectrum – Zwicker method – Amendment 1: Calculation of the loudness of time-variant sound 3GPP SA4, DESUDAPS-1: Common subjective testing framework for training and validation of SWB and FB P.835 test predictors (v1.1), available in S4-151492, San Jose del Cabo, Mexico., 26-30 October 2015.*

[b-APG] Auditory Perception Group (2019), *Auditory demonstrations and useful software*. <https://www.psychol.cam.ac.uk/hearing/auditory-demonstrations-and-useful-software-1>

[b-Buus] Buus, Florentine (2001), *Modifications to the power function for loudness*. In: Summerfield E, edited by Kompass R, Lachmann T, Fechner Day. Proceedings of the 17th Annual Meeting of the International Society for Psychophysics. Berlin: Pabst, pp. 236-241

[b-Edjekouane 1] Edjekouane, Plapous, Quinquis, Meunier (2015), *Loudness of speech transmitted via handsfree telephone systems –*

Perceptual measurements and loudness models in free field listening. Acta Acustica united with Acustica, pp. 1130-1144, 101(6).

[b-Edjekouane 2] Edjekouane, Plapous, Quinquis, & Meunier (2015), *Subjective and Objective Measurements of Speech Loudness in Hands-Free Telephony – Toward an Extended Loudness Model for Telephonometry.* Audio Engineering Society Convention 139. Audio Engineering Society.

[b-Fiebig] A. Fiebig (2015), *Cognitive stimulus integration in the context of auditory sensations and sound perceptions.* PhD thesis, Berlin.

[b-Fletcher] H. Fletcher and R. H. Galt (1950), *The Perception of Speech and Its Relation to Telephony*. The Acoustical Society of America. 22(2).

[b-Florentine 1] Florentine, Buus, Poulsen (1996), *Temporal integration of loudness as a function of level*. The Journal of the Acoustical Society of America.
99(3), pp. 1633-1644.

[b-Florentine 2] Florentine, Epstein (2006), *To honor Stevens and repeal his law (for the auditory system)* In: Kornbrot DE, Msetfi RM, MacRae AW (eds), Fechner Day 2006. Proceedings of the 22nd Annual Meeting of the International Society for Psychophysics. St. Albans, England: ISP, 37-42

[b-github] Github, *Simple Compressor* (2019)
<https://github.com/music-dsp-collection/chunkware-simple-dynamics/tree/master/simpleSource>

[b-Glasberg 1] Glasberg, Moore (2006), *Prediction of absolute thresholds and equal-loudness contours using a modified loudness model*, The Journal of the Acoustical Society of America, 120(2), pp. 585-588.

[b-Glasberg 2] Glasberg, Moore (2002), *A model of loudness application to time-varying sounds*, J. Audio Eng. Soc, Vol. 50, No. 5, pp. 331-342.

[b-Glasberg 3] Glasberg (2016), *A loudness model for time-varying sounds incorporating binaural inhibition, Trends in Hearing*.

[b-Head acoustics] HEAD acoustics, Herzogenrath (Deutschland) (2014), *Calculation of loudness level and stationary and time varying sounds –Executable program for ISO_532-1*.

[b-Hots 1] V. J. Hots (2013), *Loudness of sounds with a subcritical bandwidth: A challenge to current loudness models?,* The Journal of the Acoustical Society of America, 134, pp. 334-339.

[b-Hots 2] V. J. Hots (2014), *Loudness of subcritical sounds as a function of bandwidth, centre frequency, and level.* The Journal of the Acoustical Society of America, 135(3), pp. 1313-1320.

[b-Moore 1] Moore, Glasberg (1996), *A revision of Zwicker's loudness model*, Acustica, Vol. 82, 335-345.

[b-Moore 2] Moore, Glasberg, Baer (1997), *A model for the prediction of*

thresholds, loudness and partial loudness, J. Audio Eng. Soc, Vol.45, No. 4, 224-240.

[b-Moore 3] Moore, Glasberg, (2007), *Modelling binaural loudness*, J. Audio Eng. Soc, Vol.121, No. 3, pp. 1604-1612.

[b-Pedersen] O. Z. Pedersen, (1987), *Loudness Rating and Other Measures of Loudness*, Brüel & Kjær.

[b-Rennies 1] Rennies, Verhey, Fastl (2010), *Comparison of loudness models for time-varying sounds*. Acta Acustica united with Acustica, 96(2), pp. 383-396.

[b-Rennies 2] Rennies, Verhey, Appell, Kollmeier (2013), *Loudness of complex time-varying sounds? A challenge for current loudness models*. In Proceedings of Meetings on Acoustics, Vol. 19, No. 1, p. 050189, Acoustical Society of America.

[b-Schlittenlacher] J. Schlittenlacher, T. Hashimoto, S. Kuwano and S. Namba (2014), *Overall loudness of short time-varying sounds*, Internoise, Melbourne, Australia.

[b-Stevens] Stevens (1957) *On the psychophysical law*. Psychological review 64(3), p.153.

[b-Widmann] Widmann, Lippold, Fastl (1998), *A Computer Program Simulating Post-Masking for Applications in Sound Analysis Systems*, In: Proceedings of NOISE-CON' 98, Ypsilanti Michigan USA, pp. 451-456.

[b-Zwicker] Zwicker, Fastl (1991), *Psychoacoustics – facts and models*, ISBN 3-540-52600-5.

[b-Zwicker 2] Zwicker (1958*). Über psychologische und methodische Grundlagen der Lautheit*. Acta Acustica united with Acustica, 8(Supplement 1), pp. 237-258.

[b-Zwicker 3] Zwicker, Fastl, Dallmayr (1984), *BASIC-program for calculating the loudness of sounds from their 1/3-oct. band spectra according to ISO 532B*, Acustica 55, pp. 63-67.

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