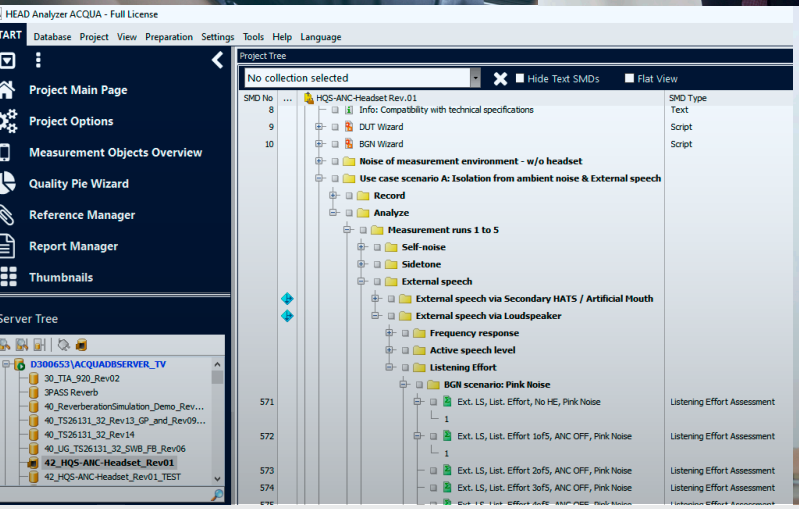
HEAD acoustics

APPLICATION NOTE


# ABLE

## Objective Assessment of Listening Effort in Telecommunications
## Background, Design and Application

# Application Note

## ABLE –
## Objective Assessment of Listening Effort in Telecommunications

Revision 0

## Legal notices

### Copyright



### Trademarks

HEAD acoustics® is a registered trademark of HEAD acoustics GmbH.

The Bluetooth® word mark and logos are registered trademarks owned by Bluetooth SIG, Inc. and any use of such marks by HEAD acoustics GmbH is under license. Other trademarks and trade names are those of their respective owners.

# 1 Introduction

The amount of environmental noise in daily life is increasing. People travel more frequently, traffic gets heavier and public places become more crowded. Everyone still expects to be able to use their audio and communication devices hassle-free. Smartphones providing near-end speech enhancement, headphones with active noise cancellation (ANC) and in-car communication (ICC) systems are the logical remedies, growing in popularity.

However, not every acoustic situation can be easily improved. Speech enhancement algorithms in smartphones cannot reduce local background noise. Other linear or non-linear signal processing may further degrade quality of transmitted speech.

In headphones, ANC reduces ambient noise including external speech signals. In most cases, this is expected and even desired if users do not want to be disturbed. However, certain situations require comprehensibility of external speech. Even though some devices provide modes for passing speech-like signals (‘talk-through’), ANC often attenuates or processes them as well. This makes conversation with other people difficult. When using the device as a headset for telecommunication, processing of ambient noise must not disturb or influence the downlink speech signal.

In-car communication systems must deal with the multitude of driving noises in the cabin. The continuous enhancement of soundproofing helps, but also increases absorption in the cabin and thus attenuates speech levels. Additionally, the ever-increasing size of cars expands the distance between occupants, making the issue worse. ICC systems facilitate conversation in the car, but only if their communication quality is decent.

In order to benchmark and/or compare systems and devices for their performance in these situations, several established speech intelligibility metrics seem to be the obvious way. However, in several studies [1] [2] [3], perceived listening effort proved to be a more suitable measure than speech intelligibility. This application note describes a new instrumental method for determining listening effort, which was recently standardized in ETSI TS 103 558 [4]. HEAD acoustics implemented this prediction algorithm as the ACQUA Option ABLE - Assessment of Binaural Listening Effort.

# 2 Background and Motivation

Speech intelligibility in telecommunication is mainly impacted by two factors: processing during transmission and near-end background noise. The near-end speech signal and the noise can be set in relation, which is typically described as signal-to-noise ratio (SNR). A low SNR leads to reduced speech intelligibility (SI) and an increased listening effort (LE), respectively. However, solely looking at the SNR value neglects several deciding factors affecting the perceived speech intelligibility and listening effort. Therefore, a more profound analysis is necessary.

SI and LE can be evaluated for all kinds of speech transmission. For telecommunication, typical systems and devices to be tested for near-end SI and/or LE are:

- In-car communication (ICC) systems
- Hands-free devices for conferences and in vehicles
- Smartphones and other smart devices (e.g. smart home)
- Headphones or headsets with active noise cancellation (ANC)

Listening tests are a viable way to assess speech intelligibility or listening effort, but they are costly and time-consuming. Predictive algorithms based on the same metrics are a much more efficient alternative. However, they are useful only if their prediction result reflects reality with acceptable accuracy.

## 2.1 Speech Intelligibility Prediction Metrics

Several established algorithms to assess speech intelligibility exist, presenting themselves as viable solutions for this task. The following sub-chapters outline the SI metrics most commonly used for telecommunication applications. There are other SI metrics from the academic field which are not in the scope of this document.

All metrics utilize one or more of the following signal types:

- Noise — Background noise only
- Clean speech — Speech before processing, without background noise
- Processed speech — Speech after processing, without background noise
- Degraded speech — Speech after processing, with background noise

### 2.1.1 Articulation Index (AI)

The Articulation Index (AI) is an outdated and basic method to assess intelligibility of speech. The calculation method was standardized in ANSI S3.5-1969 [5]. AI is based on averaged spectra of noise and speech, both idealized as stationary signals.

The algorithm processes single-channel, separate spectra of processed speech and noise, which is not always available in real measurement setups. Additionally, there is no comparison of the degraded speech with regard to the clean speech reference. For modern telecommunication devices, this method provides little information.

### 2.1.2 Speech Intelligibility Index (SII)

The Speech Intelligibility Index (SII) as standardized in ANSI S3.5-1997 [6] is the successor of the Articulation Index. The SII method represents an energy-based comparison of processed speech and noise. It is carried out over the two average 1/3-octave spectra of processed speech and the noise-only components. In addition, a simple masking model identifies how strongly the noise interferes with the speech signal in each spectrum.

Like the Articulation Index, SII requires separated noise and speech components, which are not always available in real measurement setups. Additionally, the lack of comparison between clean speech and degraded speech signals limits its informative value on intelligibility.

### 2.1.3 Speech Interference Level (SIL, PSIL)

The Speech Interference Level (SIL) metrics as described in, e.g. [7], solely analyzes the noise component to assess how strongly it interferes with speech intelligibility. To do this, the algorithm calculates the arithmetic mean of the (unweighted) SPL in octave bands considered to be relevant for intelligibility: 500 Hz, 1 kHz, 2 kHz and 4 kHz. The spectrum of each band is averaged, thus SIL operates only with stationary signals.

The Preferred Speech Interference Level (PSIL) is a speech-centered version of SIL, applying the same algorithm to only three octave bands (500 Hz, 1 kHz and 2 kHz). None of the SIL variants is standardized, producing varying results across different test setups and implementations. Additionally, they do not involve any speech signal into their rating.

### 2.1.4 Loudness / Roughness / Sharpness

Loudness is a psychoacoustic quantity that maps the human perception of the sound volume of acoustical signal to a linear scale. Loudness is based on calculations using signal processing that emulates the properties of human hearing. Several loudness calculation methods are available, each of which is specified in its own standard (e.g. ISO 532-1 [8]).

Depending on the nature of the source signal, each method produces loudness values varying considerably. This impedes comparability across different metrics. The appropriate calculation method must therefore be chosen according to the type of sound, as well as the objective of the examination.

Similar to SIL (see 2.1.3), loudness is often used to analyze only the noise component without taking any speech signal into account.

In a similar fashion, roughness and sharpness are metrics developed for purposes outside of speech signal analysis in telecommunication. They allow good assessment of the perceived roughness/sharpness of certain sounds, but are used only on the noise component of the signal.

### 2.1.5 STOI

The Short-time Objective Intelligibility (STOI) algorithm [9] [10] overcomes some drawbacks of the Speech Intelligibility Index (SII) method. STOI processes non-stationary signals on a short-time basis. Therefore, it is more suitable for speech signals. Also, the prediction algorithm of STOI incorporates degraded speech as well as the clean speech, which is desirable for most measurement setups.

STOI includes routines for automatic level adjustment of the reference signal in respect to the degraded signal (normalization step) and consideration of noise and distortions (clipping step). The correlation of these pre-processed 1/3-octave spectra are evaluated on active short-time frames of about 400 ms.

STOI is an improvement over previous speech intelligibility metrics as it can process real speech instead of averaged static spectra. On the other hand, STOI ignores the absolute and perceived level by automatically normalizing the reference signal level to the level of the degraded signal. Additionally, STOI has never left the academic R&D stage and therefore does not fulfill demands for standardized testing. Even though the method is widely used in the academic field, the usage for commercial purposes is restricted.

### 2.1.6 STI

The Speech Transmission Index (STI) method is part of many certification measurement procedures. It is standardized as IEC 60268-16 [11]. STI was originally developed for analyzing room acoustics, for which it works well. For non-linearly processed signals, its prediction quality is limited.

Based on an octave-band filter bank representation, modulation frequencies of each band are analyzed. The loss of modulation between the reference and the degraded signal is then determined for each octave band between 0.63 to 12.5 Hz. The final single value is then calculated as the average over time, modulation frequencies and octave bands.

STI generally uses modulated noise signals and therefore is not designed to process speech. Several modifications of this method evaluated the capability to handle real speech signals. All of these approaches originate from the domain of audiology, but none of these concepts has gone beyond R&D. Therefore, none has been validated and evaluated on e.g. publicly available listening test material.

### 2.1.7 ABC-MRT

ABC-MRT, the Articulation Band Correlation Modified Rhyme Test, emerged from extensive amounts of listening test data. The data originated from codec benchmarking for mission-critical voice transmission (e.g. emergency services) performed by 3GPP [12]. The main application of ABC-MRT is intelligibility on the sending side, focusing on noise reduction and coding technologies. How well ABC-MRT works for the receiving direction has not been adequately substantiated.

ABC-MRT can be seen as an advanced mixture between AI (see 2.1.1) and STOI (see 2.1.5). As a method, it is

a two-step process comprising of a prediction algorithm and subsequent analysis. It is important to understand that the output score produced by the prediction algorithm is *not* the final test result. First, it must be analyzed in the second step: an algorithm is used to “emulate” a modified rhyme test (MRT) by filling in for a real person. Like a real-life MRT, six words/samples are rated. The sample with the highest obtained algorithmic score is then chosen as the correct one. After a certain number of samples per condition, the real spoken words are compared to the algorithm’s selections. The obtained ratio is the final score for the condition.

For adequate prediction performance, it is recommended to use 1,200 samples/words per condition. One condition equals one specific volume setting with one specific background noise type. Even with short words and no pauses, 1,200 words roughly equals 50 minutes of source signal which needs to be measured. With sufficient computing power, the result analysis takes another 25 to 50 minutes. Thus, each single condition sums up to more than one hour of work.

Using less samples/words is not recommended to ensure reasonable prediction quality. Also, skipping the lengthy MRT procedure and simply using the score produced by the algorithm gives false results. Thus, testing according to ABC-MRT is at least a very time-consuming process.

### 2.1.8 PESQ and POLQA

The Perceptual Evaluation of Speech Quality is a method standardized in Recommendation ITU-T P.862 [13]. It is commercially available as PESQ. The method is used to predict speech *quality*, which must be differentiated from speech intelligibility and/or listening effort. It is applicable only at electrical interfaces, acoustic paths are in general excluded in the scope of ITU-T P.862, e.g. the short loudspeaker-to-ear path recurrent in telecommunication.

Even though the source code can be downloaded from the ITU-T website, it is not legal to use it without a valid license – even for academic purposes. Especially in the academic world, P.862 is often used in the context of noise reduction, speech enhancement and intelligibility enhancement in the presence of background noise. However, the standard explicitly rules out use cases containing any kind of noisy speech by stating that it will provide “inaccurate predictions” and therefore is “not intended to be used” for these purposes.

The Perceptual Objective Listening Quality Prediction, commercially available as POLQA, is the successor of P.862. In the latest version in force, P.863 [14], it is the first metric discussed in this chapter which is capable of audio analysis up to 20 kHz (full-band).

P.863 can only handle conditions with moderate background noise, thus with a high SNR. When SNR is low as to be expected in this context, the metric does not produce proper prediction results. Additionally, it fails already at slightly reverberant conditions/environments, which is also indicated in the scope of ITU-T P.863. Therefore, it generally is not suitable for hands-free scenarios like the ones at hand.

P.862 and P.863 both predict speech quality according to Recommendation ITU-T P.800 [15] with clearly laid-out listening test conditions. However, none of them are designed for reliably predicting speech intelligibility.

## 2.2 Conclusion on Speech Intelligibility Metrics

Most of the known metrics were developed either for purposes outside of telecommunication or for very specific scenarios which differ from near-end SI/LE testing as laid out at the beginning of this chapter. Therefore, they do not deliver comprehensive information when used in this context. Additionally, for the majority of metrics there is no standardization of measurement setups *and* listening tests, leading to results that are not comparable across different test setups.

None of the existing speech intelligibility metrics can process binaural signals, ignoring the enclosed spatial information that influences human hearing and thus speech intelligibility. Thus, binaural recordings made with artificial heads cannot be used. A possible workaround would be analyzing one signal at a time and combining the results, e.g. by summation or averaging. However, as none of the metrics for speech intelligibility clearly defines such a process, results would vary greatly, therefore this is not applicable.

An often overlooked fact is that a speech intelligibility algorithm's output does not translate directly to listening test results on a linear scale. Additionally, results obtained with different metrics are not directly comparable. This is understandable as they were originally developed for differing purposes. The Common Intelligibility Scale [16] shown in Fig. 1 shows the non-linear nature of various speech intelligibility tests and allows conversion of their results. However, if an speech intelligibility metric is used out of its scope, its prediction quality and therefore the validity of conversion via the CIS becomes uncertain.

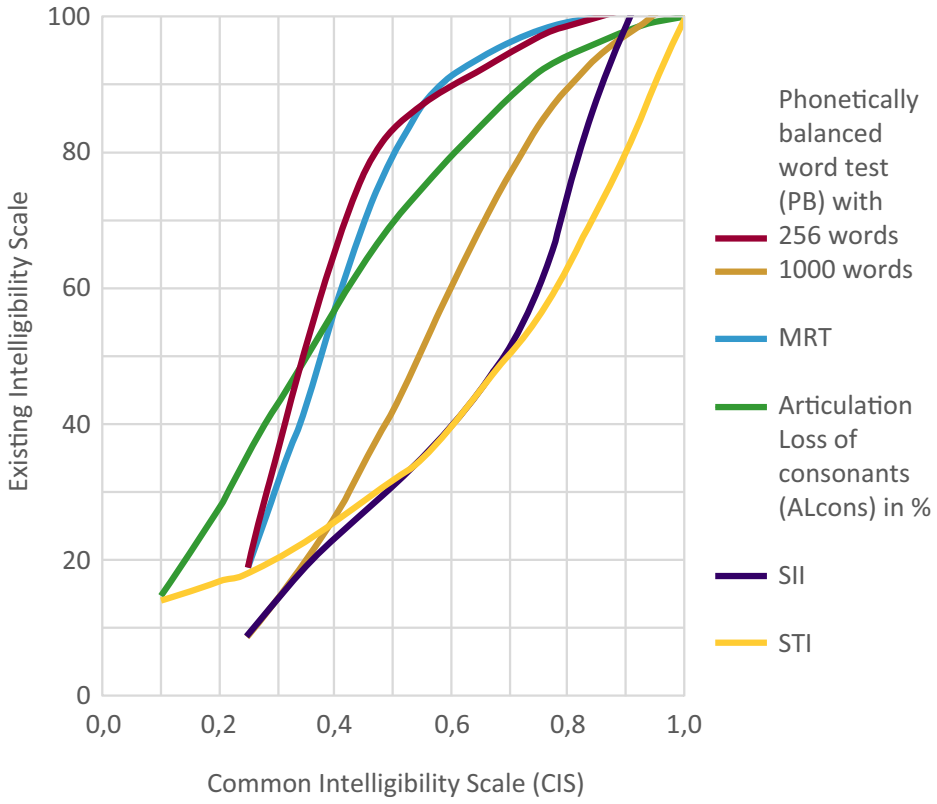



Fig. 1: Common Intelligibility Scale (CIS) for conversion of various intelligibility test metrics

## 2.3 Speech Intelligibility vs. Listening Effort

All perceptive tests of speech intelligibility are based on the participant's understanding of either singular syllables, words or whole sentences. At first glance this appears as a good measure for testing intelligibility. However, the specific characteristics of human hearing, combined with the desire to score well in tests, leads to improper scaling of test results.

Taking the common Modified Rhyme Test (see 2.1.7) as an example, a native speaker tends to get nearly all words correct beyond a medium SNR. Below a certain SNR, intelligibility becomes so poor that the same person hardly understands anything, thus scoring low. This leads to a compression of the ratio of intelligibility vs. SNR, which in turn reduces the informative value of intelligibility testing.

As an alternative, participants can be asked to give their assessment of the listening effort required to understand test sentences. By design, there is no aspiration to give the "right" answer. Instead, listening effort asks a person for their personal impression and then rate it on an Absolute Category Rating (ACR) scale similar to the well-known and established speech quality testing according to e.g., ITU-T P.800 [15].

A listening test performed to compare popular speech intelligibility metrics to listening effort produced results shown in Fig. 2. The test was conducted to assess the performance of an ICC system [1] while also comparing results of common speech intelligibility metrics with listening effort as described above.

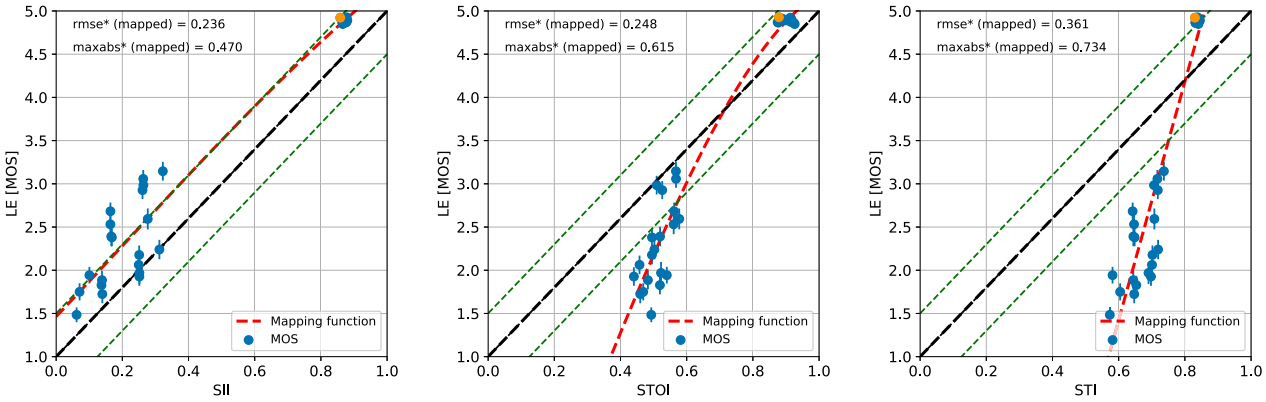



Fig. 2: Results of listening tests: MOS-LE vs. SII, STOI and STI respectively

Ideally, any of the speech intelligibility test metrics should have a linear relation to listening effort. While there is a general correlation between SI and LE, individual patterns emerge for each method.

The SII generally rates all not perfectly intelligible samples quite low on its scale, leaving a substantial gap of possible ratings unused. Additionally, various samples were rated with an identical SII value but varying LE values, hinting that SII does not differentiate between them where LE does. For example, several speech samples were rated with an SII around 0.25 while the MOS-LE ranges from 1.8 to 3.2 for the same samples.

The STOI shows noticeable compression of results. It groups many samples between 0.43 and 0.59 whereas the MOS-LE ranges from 1.4 to 3.3 for these samples. This narrow range of results makes interpretation via the STOI method difficult, e.g. when experimentally optimizing a speech transmission system.

STI displays a mixture of both behaviors. It shows compression of results similar to SII (but shifted to a higher value) while treating numerous speech samples as equal where LE differentiates between them.

In contrast, Listening effort shows a significantly larger spread in its test results, allowing a more precise analysis of the device under test (DUT). The increased gradation of results also enables a better evaluation of system improvements achieved through fine-tuning. Such small differences would not be noticeable on an SI scale.

The ITU-T Handbook [17] states that the design of listening tests and instrumental prediction method should be clearly coordinated with each other. None of the SI metrics achieves that, mostly lacking the description of listening test design. Various SI tests, mainly in the domain of audiology, are described in literature, but none of them currently are standardized. Additionally, necessary items often are not publicly available (e.g. audio sources or words/sentences in written form)

A major advantage of standardization is the definition of languages and speech bodies. As an exception, ITU-T P.807 [18] specifies an SI test, but only for American English. Translations are not available, and neither are freely available audio samples. Thus, even this well defined speech intelligibility test is usable only in very specific scenarios.

To allow meaningful and fully repeatable testing of listening effort with comparable results, in 2019 ETSI STQ specified listening test design and the suitable instrumental method in the specification ETSI TS 103 558 [4]. HEAD acoustics implemented this prediction algorithm as a software option for the Advanced Communication QUality Analysis system ACQUA: ABLE - Assessment of Binaural Listening Effort (ACOPT 37). ABLE is a unique method that processes binaural signals and therefore takes the influence of binaural hearing on listening effort into account. This allows a much more realistic assessment of human listening perception than the aforementioned monaural metrics for speech intelligibility.

Despite the outlined advantages of testing listening effort instead of speech intelligibility, ABLE has the advantage of speech material being available in different languages. Additionally, ABLE can process speech samples from other specifications like e.g. Recommendation ITU-T P.501 [19].

## 2.4 Auditory (Subjective) Test Databases

When evaluating Listening Effort, subjects provide a self-assessment on a five-point categorical scale similar to well-known speech quality testing methods. For evaluation of Listening Effort, the scale according to table 1 is used:

| Score | Listening Effort | Speech Quality |
|---|---|---|
| 5 | Complete relaxation possible, no effort required | Excellent |
| 4 | No appreciable effort required | Good |
| 3 | Attention necessary, moderate effort required | Fair |
| 2 | Considerable effort required | Poor |
| 1 | No meaning understood with any feasible effort | Bad |

Table 1: Overview of evaluation categories for listening effort MOS (MOS-LE)

The scale and the corresponding attributes are taken from ITU-T P.800 [15]. Besides the aforementioned benefits of Listening Effort testing, recent studies (e.g. [1]) indicate that speech enhancement benefits can be evaluated in a wider range of signal-to-noise ratios (SNRs) without reaching positive or negative saturation observed in intelligibility tests. Auditory tests can be conducted in groups in parallel.

# 3 Design of Listening Effort

## 3.1 The Instrumental Model ABLE

The instrumental model for Listening Effort standardized in ETSI TS 103 558 [4], and thus also its implementation as ABLE, describes two closely related operational modes of the algorithm – one with and one without a noise-only reference. The model below is the version without a noise-only reference, which is used most often as it can be applied in any use-case. For the model with a noise-only reference, please see [4].

The algorithm consists of different stages. Signals are either depicted as double-arrows with bold designations (binaural) or as single-arrows with regular designations (monaural). The different signal designations are:

- r / R for **r**eference signal (clean speech)
- N for **n**oise component
- P for **p**rocessed signal (processed speech without noise)
- d / D for **d**egraded signal (processed speech with noise)

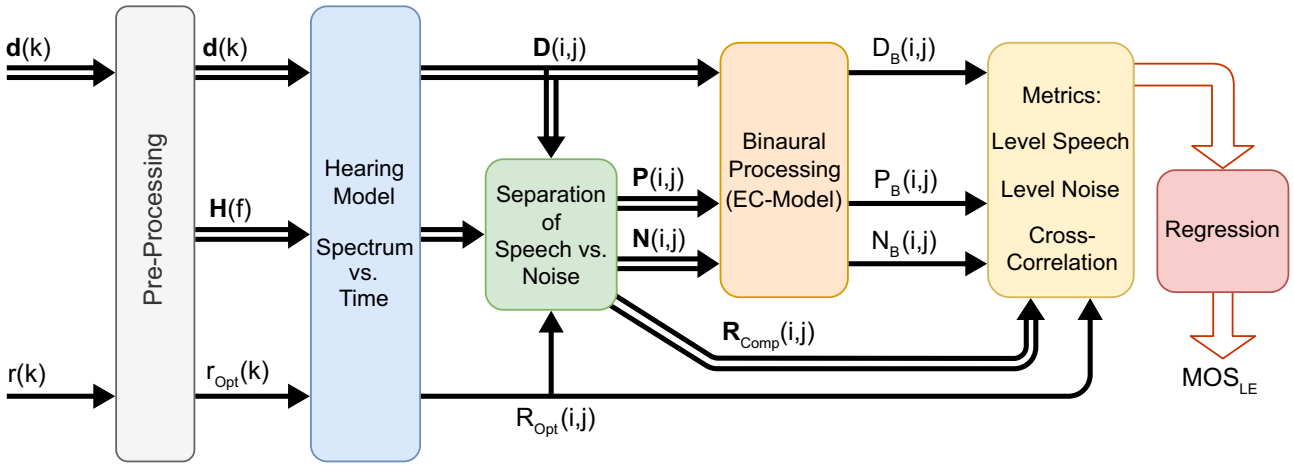



Fig. 3: Flow chart of prediction algorithm for Listening Effort without noise-only reference based on the chart in ETSI 103 558

### 3.1.1 Stages and Signals

ABLE is a binaural model, thus the input ideally are binaurally recorded scenarios. If not available, ABLE can also handle monaural signals. **d**(k) is the binaurally recorded degraded speech signal comprising the left and right channel, $d_l(k)$ and $d_r(k)$, respectively. r(k) is the monaural clean speech signal which is used as a reference.

- The *Pre-Processing* stage performs temporal and level alignment and calculates the binaural transfer function **H**(f) which describes the relation between degraded and reference signals in order to exclude non-correlated noise components
- The *Hearing Model* calculation performs an aurally adequate transformation using the HEAD acoustics hearing model [20] [21]. This results in separation into spectra (i) and time (j), only **H**(f) passes unaltered
- The *Separation of Speech & Noise* then separates the degraded spectra **D**(i, j) into (processed) speech **P**(i, j) and noise **N**(i, j)
- The *Binaural Processing* stage incorporates that the human ear is capable to improve the SNR when listening binaurally as compared to monaural listening (Equalization-Cancellation Model). This block requires the availability of the isolated speech and noise (masking) components which have been generated in the previous step. It puts out the “binaurally corrected” versions of all input signals, namely the degraded speech signals $D_B(i, j)$, the processed speech signal $P_B(i, j)$ and the noise component $N_B(i, j)$.
- The *Metrics* block combines the various signals generated by previous stages via level-based as well as correlation-based metrics. The results are aggregated to an overall metric
- From the metric block, various single values are available and combined by non-parametric r*egression analysis*. This results in a single Listening Effort score. Machine learning procedures are used here for the instrumental derivation of the Mean Opinion Score for listening effort (MOS-LE).

### 3.1.2 Auditory Databases for Training and Validation

As a basis of any instrumental prediction model, listening examples are required to train and validate the model. These listening examples need to cover the entire quality range of devices for which the model is to be developed. Furthermore, these listening examples need to include all impairments which may arise from signal processing and which are relevant for the user's subjective impression.

Different applications, ambient noises, devices, implementations and possible speech processing algorithms need to be taken into account when developing an objective test method. In order to predict all these scenarios reliably, numerous noisy speech samples have to be available. To generate these types of listening examples, advanced signal processing techniques for simulation as well as a large variety of devices should be available. In addition, a variety of realistic noise scenarios where these devices are used in, need to be available.

A set of test conditions are combined in a test database. Each test database contains a set of speech samples processed with different devices and/or algorithms combined with a certain noise type (or silence), ideally covering the complete range of Listening Effort and a set of reference conditions. Mapping of different test databases to each other is realized by mapping to the reference conditions. The purpose of auditory tests is twofold:

- So-called training databases are used to develop and train the objective model for best performance
- So-called validation databases are needed after model development is completed. The auditory test results of these databases were never seen before by the model. They provide information about the robustness and validity (generalization) of the objective model

More detailed information on training and validation databases can be found in ETSI TS 103 558 [4].

## 3.2 ABLE as an ACQUA Application

Beside a large number of devices under test, high quality test equipment is required to generate recordings for ABLE analysis and/or listening tests. A typical setup for generating such listening examples, which also is used for testing real devices, is shown in Figure 4.

Here, the example of testing an ANC headset in conjunction with a Head-and Torso Simulator (HATS) is illustrated. Prerecorded background noise is played back in a laboratory using a standardized sound-field generation system (see ETSI TS 103 224 [22]). This setup reproduces various background noises representative for the typical environmental noise situations with a high degree of accuracy around the device under test (DUT).

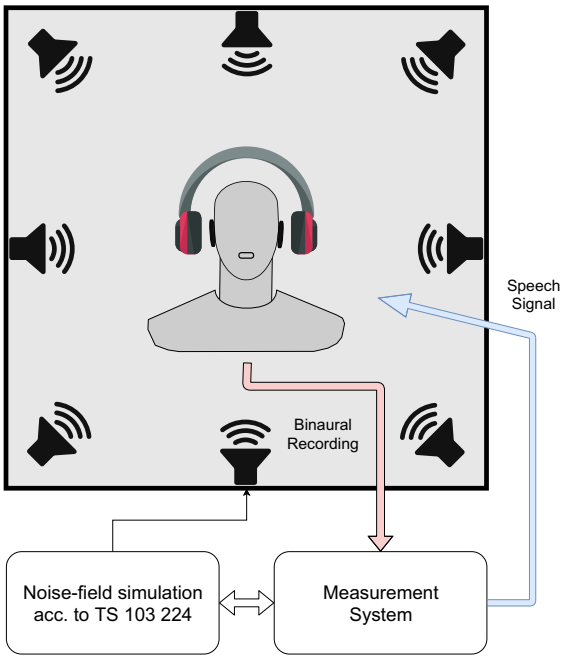



Fig. 4: Measurement setup for testing ANC

Based on the chapter 3.2, the following software and hardware components are generally required to test ABLE with ACQUA. Depending on the individual use case, additional hardware and/or software may be needed.

| Analysis Software | | Code |
|---|---|---|
| **ACQUA** | Advanced Communication QUality Analysis software | 6810 |
| **+ ACOPT 37** | Software Option ABLE | 6869 |

| Background Noise Simulation System | | Code |
|---|---|---|
| **3PASS *lab*** | 3-dimensional playback of acoustic sound scenarios - *lab* version | 6990 |
| **or** (depending on application) | | |
| **3PASS *flex*** | 3-dimensional playback of acoustic sound scenarios - *flex* version | 6995 |

| Head And Torso Simulator | | Code |
|---|---|---|
| **HMS II.3-33** | HEAD Measurement System with 3.3 pinnae, standard version | 1230 |
| **+ HIS L** | HEAD Impedance Simulator, left, for HMS II.3/4/5 | 1231 |
| **or** (depending on application) | | |
| **HMS II.3-LN** | HEAD Measurement System with 3.3 pinnae, low-noise version | 1230.3 |
| **+ HIS L-LN** | HEAD Impedance Simulator, left, low-noise version, for HMS II.3/4/5 | 1231.3 |

| Hardware Platform | | Code |
|---|---|---|
| ***lab*CORE** | Modular multi-channel hardware platform | 7770 |
| **+ *core*BUS** | *lab*CORE I/O mainboard | 7710 |
| **+ *core*IN-Mic4** | *lab*CORE microphone input board | 7730 |
| **+ *core*BEQ** | *lab*CORE Binaural EQualization, incl. filter set for one artificial head | 7740 |

A typical setup for testing listening effort with an ANC-capable Bluetooth® headset with ABLE is shown in figure 5. *lab*CORE transmits a clean speech signal to the headset via Bluetooth®. 3PASS *lab* simultaneously plays back typical environmental noises for this application. Via *lab*CORE, ACQUA receives the binaural degraded speech signals from the artificial ears of HMS. Clean speech and binaurally recorded degraded speech are sent to the prediction algorithm (see chapter 3.1.1) to calculate a MOS-LE for perceived listening effort.

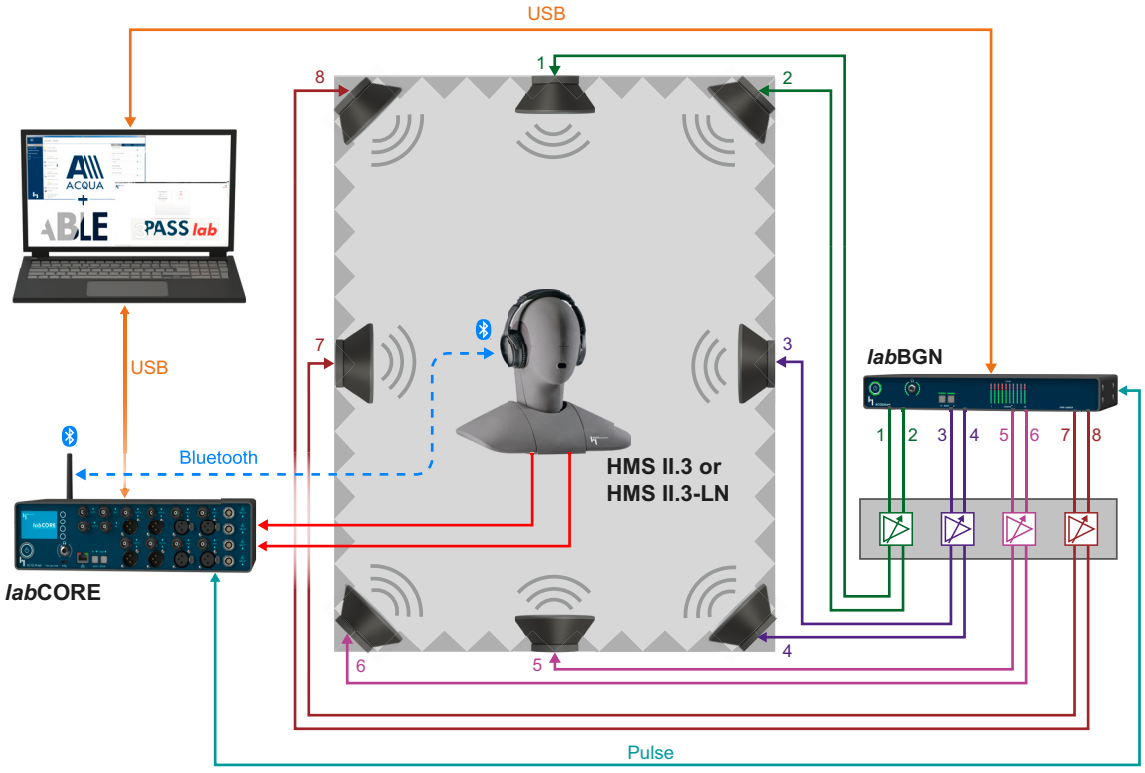



Fig. 5: Exemplary setup for testing listening effort with an ANC-capable Bluetooth® headset

ABLE can be added into ACQUA as one of many ACQUA Options, or in short ACOPTS. ABLE is ACOPT 37. Figure 6 shows a Single Measurement Descriptor (SMD) for a listening effort measurement with ABLE in ACQUA.

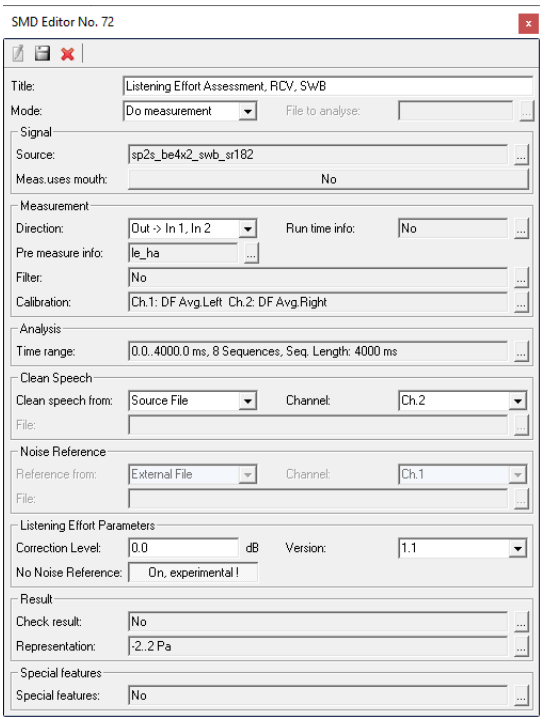



Fig. 6: ABLE measurement descriptor in ACQUA

# 4 Practical Application of ABLE

As laid out in chapter 2 of this document, assessing listening effort is an appropriate method to evaluate and analyze speech transmission for the receiving conversational partner. In this chapter, the instrumental analysis method ABLE is applied to real-world examples. The goal is to determine how ABLE fares in different fields of application that involve speech transmission.

Please note that in the following, a “better” or “improved” listening effort equals a lower actual effort for a person to listen to and understand speech. The the same manner, a “worse” or “worsened” listening effort equals a higher effort for a person to understand speech. Therefore, a *better* listening effort equals higher a MOS-LE, a *worse* listening effort equals a lower MOS-LE.

## 4.1 ANC Headsets

### 4.1.1 Scope & Setup

Active noise cancellation (ANC) is increasingly popular in consumer headsets. When combined with speech playback through the headset, the reduction of outside noise ideally improves listening effort by increasing the SNR at the ear.

To examine the properties of speech playback with ANC headsets in the presence of background noise, two representative consumer headsets with ANC were chosen. In a first step, their general noise cancellation performance is assessed. After establishing this baseline, both headsets are tested with speech playback for listening effort with ABLE to determine if and how their ANC performance influences listening effort.

Speech signals are sent to the headsets through a Bluetooth® connection. All measurements were performed with both headsets set to nominal volume, which equals a Receive Loudness Rating (RLR) of 2.0 dB. This depicts listening situations such as podcasts and radio broadcasts (speech in super-wideband) or telephony (speech in narrowband/super-wideband), all performed via Bluetooth®.

Measurements were performed in a test cabin equipped with a background noise simulation system according to TS 103 224: 3PASS lab. The system has been equalized with the symmetric microphone array MSA II to ensure accurate background noise simulation at both ears.

Recordings of the headsets were obtained with HMS II.3, which was equipped with artificial ears of Type 3.3 (ITU-T P.57 [23]). For an overview of further hardware and software for use with ABLE, please refer to chapter 3.3.

### 4.1.2 General Active Noise Cancellation Performance

The figures 7 and 8 show the passive isolation and active noise cancellation performance of the selected headsets in the presence of background noise. The reference, a pink noise signal, was played back by the background noise simulation system and recorded with the artificial ears of HMS II.3-33 without a headset. The below curves for passive noise isolation and active noise cancellation therefore represent the headsets’ performances as a directly readable decibel value.

In general, passive noise isolation of headsets is most effective at higher frequencies due to increased absorption. Active noise cancellation on the other hand is most effective at low frequencies – long wavelengths are easier to be matched and subsequently canceled than short wavelengths. Hence, the combination of passive noise isolation and active noise cancellation in a headset allows an effective reduction of background noise from a listener’s perspective.

To compare the ANC performance of the chosen headsets, three measurements at the right ear were performed for each headset:

1. Background noise (pink noise) without the headset
2. Background noise (pink noise) with the headset, ANC deactivated
3. Background noise (pink noise) with the headset, ANC active

The following three figures show the resulting spectra. The curves for passive noise isolation (black curve) and passive isolation + active noise cancellation (blue curve) show the response of the headset in respect to background noise. This type of presentation allows to directly read the headsets’ performances in decibels.

Headset A (Figure 7) shows effective passive isolation (black curve) only above 3 kHz. Below, and thus within the typical frequency range of human voice, passive isolation is poor. Additionally, between 180 Hz and 500 Hz background noise is amplified by acoustic resonances in the earcup cavity.

Active noise cancellation works most effectively between 70 Hz and 550 Hz. ANC performance is countered by the aforementioned undesirable acoustic amplification, which leads to reduced cancellation of background noise at the ear in that frequency range. On top of that, the ANC system amplifies background noise between 1.4 kHz and 6 kHz, countering the passive isolation performance to an extent. In combination, headset A shows poor total noise reduction performance.

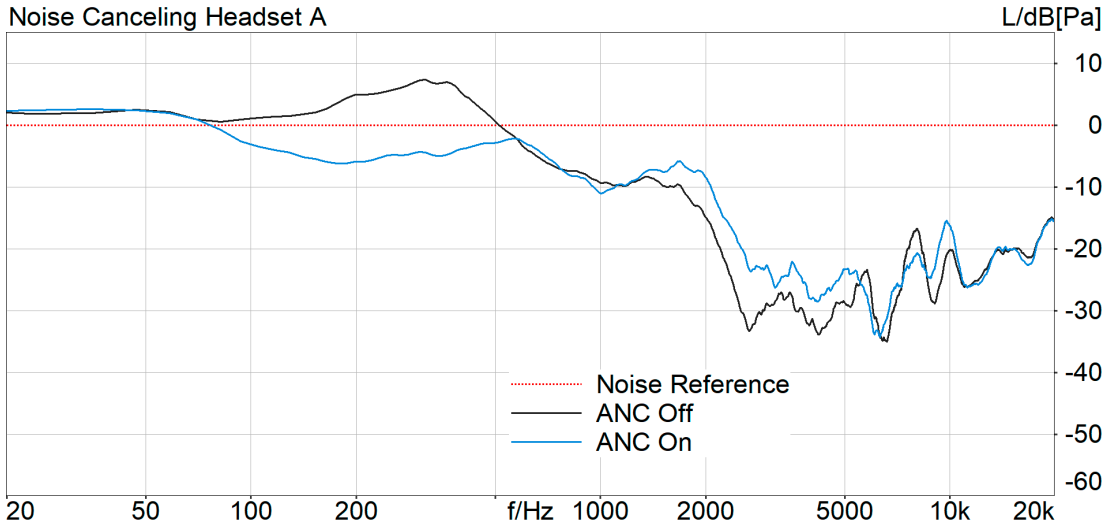



Fig. 7: Noise isolation (black) and total reduction (blue) performance of a headset with poor ANC (Headset A)

In contrast, headset B (Figure 8) shows an effective passive isolation performance, beginning beyond 300 Hz and improving further beyond 600 Hz. Active noise cancellation of headset B is very effective from 30 Hz to 700 Hz. ANC is well matched to the passive isolation, allowing for the effective reduction of background noise over most of the range of human hearing. Thus, headset B serves as an example for good noise cancellation performance.

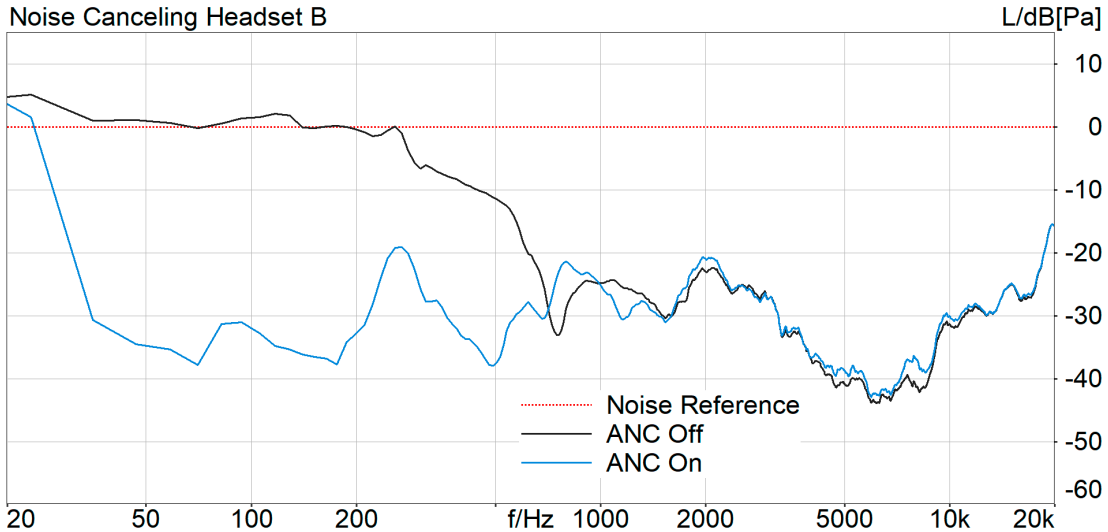



Fig. 8: Noise isolation (black) and total reduction performance (blue) of a headset with good ANC (Headset B)

Figure 9 displays the overall background noise reduction of both headsets in direct comparison. The curves represent the at-the-ear background noise reduction performance that users can expect.

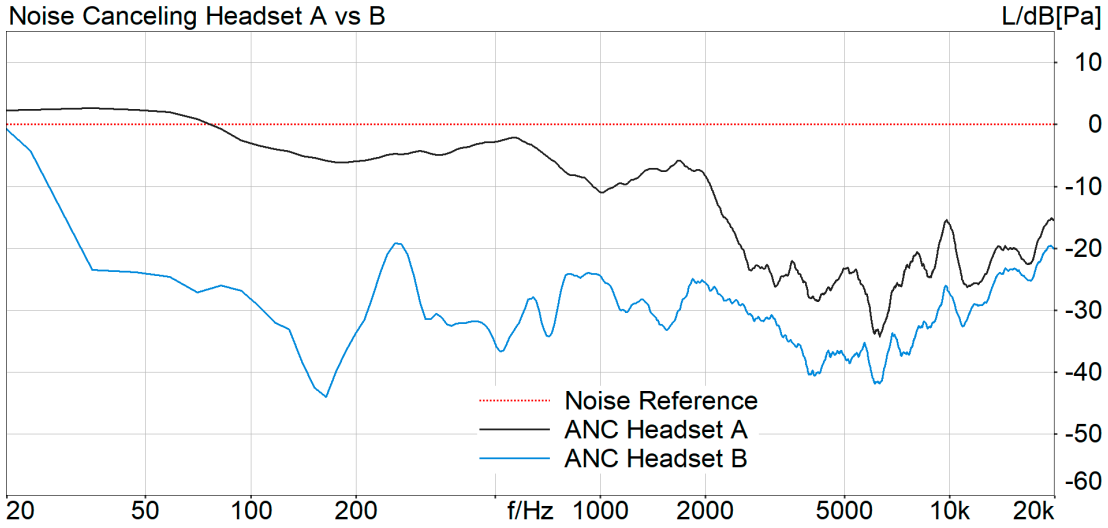



Fig. 9: Total noise reduction performance of both headsets in direct comparison

To better understand how this difference in performance influences speech playback and thus listening effort, the measurement data shown as curves in figure 9 are displayed as time data in figure 10.

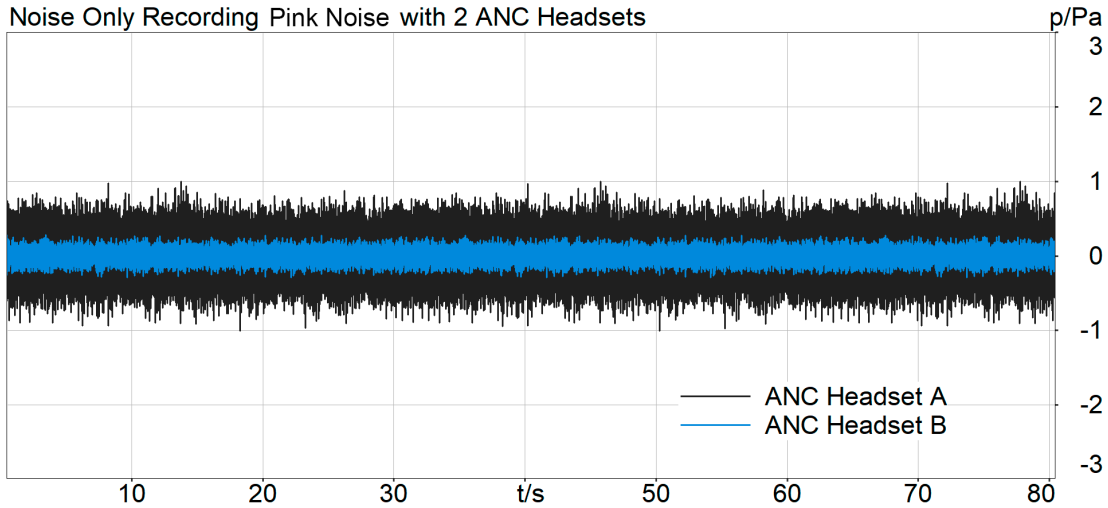



Fig. 10: Time data of both ANC headsets: background noise (pink noise) level at the right ear

The better cancellation performance of headset B (blue) is represented by the significantly lower level of noise at the ear in respect to headset A (black). Typically, ANC systems adapt well to stationary noise as their ANC systems require a finite amount of time to process their input signal and generate the inverse signal to cancel it out. As established previously, these systems usually are most efficient at low to mid frequencies, at which pink noise as well as many real world background noises, are most intense.

As a next step, a narrowband speech signal is played back through the headsets in the presence of the same background noise. Figure 11 shows the results as time data.

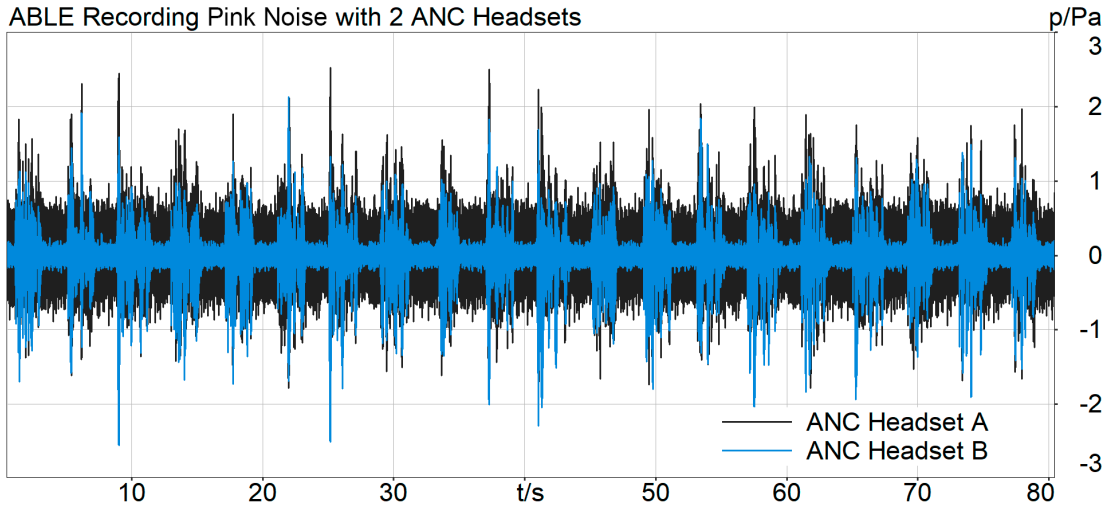



Fig. 11: Time data of both ANC headsets: background noise (pink noise) level at the right ear with added playback of narrowband speech through the headset

Due to its better noise cancellation performance, headset B (blue) achieves a significantly higher SNR at the ear than headset A (black) by reducing the noise component. However, this time data does not quantify by how much the better performance of headset B changes listening effort as perceived by the user. This is exactly what ABLE was developed for. The MOS values resulting from its analysis present a comprehensive and easily comparable evaluation of user perception. It is therefore expected that headset B requires less listening effort and thus reaches a higher MOS than headset A when analyzed with ABLE.

### 4.1.3 MOS-LE across various Noise Scenarios (NB)

Figure 12 shows the MOS-LE calculated by ABLE for both headsets. Background noise has been changed to four common noise scenarios. The speech signal again is narrowband (NB), measurement conditions match the conditions laid out in the preceding sub-chapters.

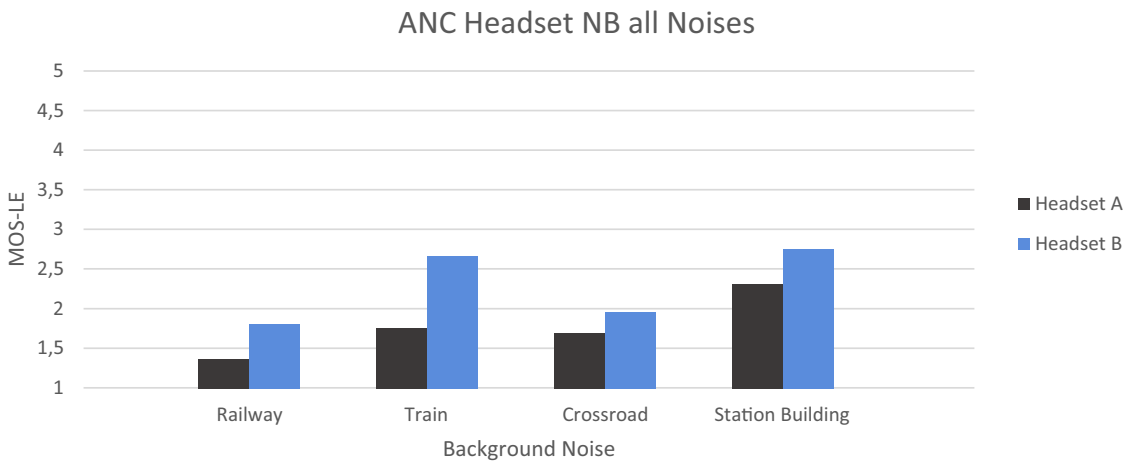



Fig. 12: MOS-LE of the two ANC headsets with narrowband speech for four background noise scenarios

As expected, the headset with superior noise cancellation performance – headset B – performs better than headset A across all scenarios. Highly time-variant noises with high volume levels (*'Railway'* & *'Crossroad')* lead to the lowest MOS-LE results as these soundscapes are more difficult for ANC systems to react to. Stationary signals like the *'Train'* scenario allow good ANC adaptation and thus reach a slightly higher MOS-LE. The more effective ANC of headset B becomes apparent in this scenario through its significantly higher MOS-LE of 2.66 as opposed to a MOS-LE of 1.75 for headset A.

The *'Station Building'* scenario, a mildly crowded railway station concourse, only has little surrounding noise and therefore leads to good scores for both headsets.

### 4.1.4 MOS-LE across various Noise Scenarios (NB vs. SWB)

The test described in the previous chapter is now repeated with speech playback in super-wideband (SWB) instead of narrowband (NB). As laid out in chapter 4.1.1, both headsets are set to a Receive Loudness Rating (RLR) of 2.0 dB. Figure 13 shows the results calculated by ABLE. The narrowband scores from the previous chapter are displayed for reference.

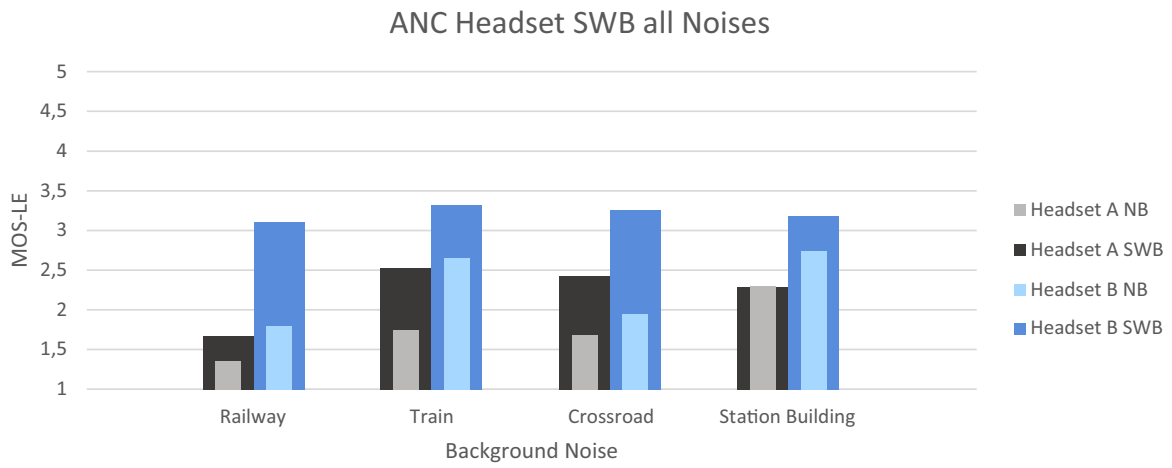



Fig. 13: MOS-LE of the two ANC headsets with super-wideband speech for four background noise scenarios

The enlarged frequency range of super-wideband speech can be reproduced very well by headsets. Their optimization for high quality reproduction of sound across the whole human range of hearing, of course, translates into the speech domain as well. Thus, the swap from narrowband to super-wideband generally increases performance of both headsets across all scenarios. There is only one exception: Headset A in the *'Station Building'* background noise scenario. In accordance with the super-wideband MOS-LE for *'Train'* and *'Crossroad'*, it appears that headset A reaches its peak performance at a MOS-LE between 2.3 and 2.5. Thus, there is no improvement over the narrowband MOS-LE for *'Station Building'*.

Headset B is able to substantially improve its performance across all four scenarios. Similar to headset A, it reaches peak performance, albeit in a higher MOS-LE range of between 3.1 and 3.4.

## 4.2 Mobile Phones

### 4.2.1 Scope & Setup

Mobile telephony is the most common application for which speech intelligibility and the related listening effort are important factors. Of course, background noise is a very common occurrence in everyday use of mobile phones. Additionally, the characteristics of codecs, issues in data transfer, the quality of microphones and loudspeakers as well as other factors all add their “individual signature” to the transmission chain.

To assess the impact of these factors on the quality of mobile telephony in a black box approach through ABLE, a common smartphone model was chosen. The following sub-chapters compare MOS-LE calculated with ABLE for different frequency ranges (narrowband vs. super-wideband), background noises, use cases (handset vs. handheld hands-free) and volume settings.

All measurements for the chapters 4.2.2 and 4.2.3 comparing MOS-LE between narrowband and super-wideband were performed at nominal volume. The volume comparison in chapter 4.2.4 was performed with narrowband speech. In handset mode, the left (free) ear signal consists of background noise, but no speech. This is taken into account by ABLE by its internal binaural processing.

Measurements were performed in a test cabin equipped with a background noise simulation system according to TS 103 224: 3PASS *lab*. The system has been equalized with the asymmetric microphone array MSA I to ensure accurate background noise simulation at the right ear.

Recordings of the handsets were obtained with HMS II.3, which was equipped with artificial ears of Type 3.3 (ITU-T P.57 [23]), and handset positioner HHP IV. For measurements, default mounting and the default/alternative handset position were used. For an overview of further hardware and software for use with ABLE, please refer to chapter 3.3.

The receive loudness ratings (RLR) play an important role in judging test results. They influence the measurement results and thus the MOS-LE. Of course, it is desirable to keep RLR deviations as small as possible. With the typically large steps between playback volume settings, variance can only be compensated to an extent. The following table shows the RLR that were used for the tests in chapter 4.2.

| | | **Volume Setting** | |
|---|---|---|---|
| | | Nominal (Nom) | Maximum (Max) |
| **Handset** | NB | 1.0 dB | -9.7 dB |
| | SWB | 0.8 dB | -8.0 dB |
| **Handheld Hands-free** | NB | 8.7 dB | 6.1 dB |
| | SWB | 9.8 dB | 7.1 dB |

Table 2: RLR values that were used for all measurements in this chapter.

### 4.2.2 MOS-LE in Handset Mode (Six Noise Scenarios, NB & SWB)

In this sub-chapter, the impact of the frequency range for speech transmission is examined. The chosen smartphone is used for a handset call in five different background noise environments. The call is performed in narrowband (NB) and super-wideband (SWB). Figure 14 shows the resulting MOS-LE calculated by ABLE.

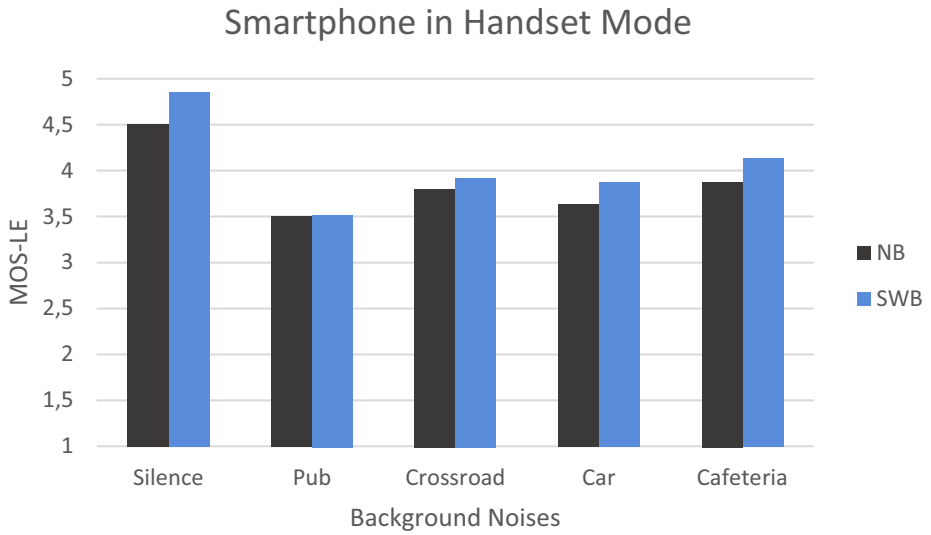



Fig. 14: MOS-LE of the smartphone in handset mode with narrowband (NB) and super-wideband (SWB) speech in the presence of background noise

The analysis shows a rather close relation between the MOS-LE for both frequency ranges, narrowband and super-wideband. The larger frequency range of SWB better depicts the real-life tonality of speech signals, but improves listening effort (higher MOS-LE) only to an extent when background noise is present.

This behavior is examined further with the use case *'Cafeteria'* as an example. Figure 15 shows absolute spectra of this scenario at the right ear in handset application. The curves for speech playback without background noise (black and blue) demonstrate the significantly wider frequency range of SWB. It is noteworthy that both speech signals reach a higher sound pressure level at the ear than the background noise without speech (red). The inversion of this level advantage below 200 Hz is not a deciding factor for listening effort as this is the very low end of the human vocal range.

Consequently, SWB produces an audibly better speech playback quality than NB, resulting in a slightly higher MOS-LE as shown in figure 14.

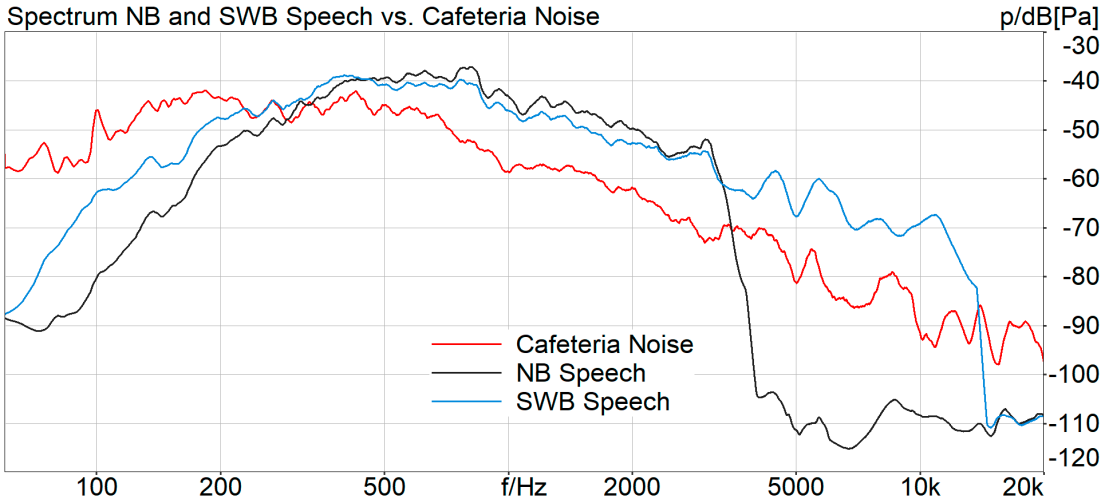



Fig. 15: Absolute spectra of the smartphone's handset loudspeaker with NB (black) and SWB (blue) speech in the presence of the background noise *'Cafeteria'* (red)

### 4.2.3 MOS-LE in Handheld Hands-Free Mode (Five Noise Scenarios, NB & SWB)

In this clause, testing as laid out in chapter 4.2.2. is repeated with the smartphone in handheld hands-free mode. Figure 16 shows the respective MOS-LE results for the same five background noise scenarios.

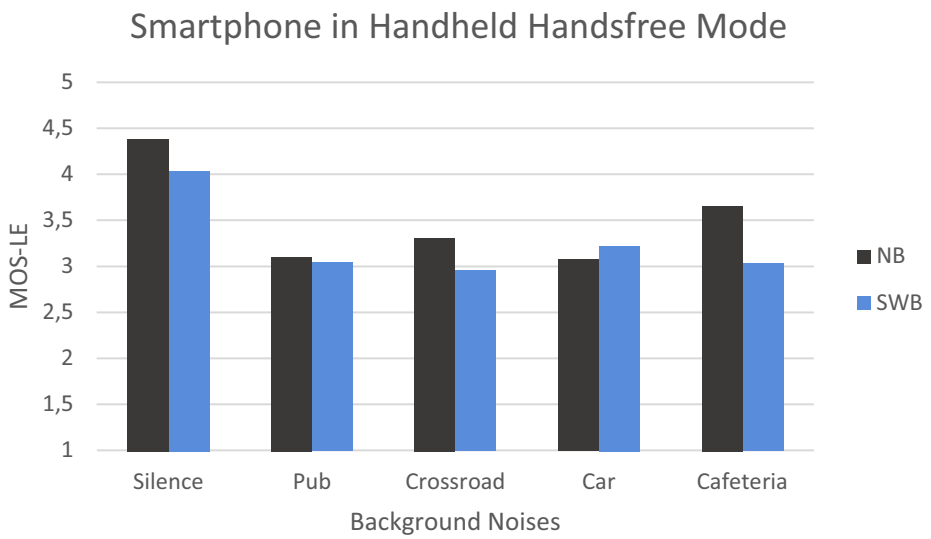



Fig. 16: MOS-LE of the smartphone in handheld hands-free mode with narrowband (NB) and super-wideband (SWB) speech in the presence of background noise

Again, silence scores high in respect to all noisy environments. In contrast to handset mode however, super-wideband does not necessarily reach a higher MOS-LE than narrowband. The reasons for this are twofold:

1. The receive loudness rating (RLR) is 1.2 dB higher in favor of NB (see table 2)
2. The overall sound pressure level of speech generally is low in respect to background noise

The nature of hands-free application makes the smartphone's loudspeaker struggle to reach a good SNR at the ear. The loudspeaker is aimed away from the ear, the phone is positioned at a distance to the listener. Surrounding noise therefore becomes highly intrusive in respect to speech playback. Additionally, the hands-free loudspeaker has physical limits in terms of frequency reproduction range and maximum volume. Aiming it away from the listener additionally attenuates the higher frequencies of the human vocal range due to the loudspeaker's inherent directivity.

As a result, the speech-centered narrowband is able to counter the challenges of this situation better than super-wideband. The additional low- and high-frequency components of SWB are either not reproduced at the ear at all or tie up capacities from the frequency range in which the hands-free loudspeaker works most efficiently: the narrowband. Please also see figure 17 for further information.

The only exception in terms of MOS-LE results is the car scenario, in which the driving noise of a car at 130 km/h is simulated via 3PASS *flex*. This noise is stationary, low frequencies are very pronounced. Narrowband speech therefore is masked effectively. Super-wideband on the other hand is able to "escape" the masking driving noise through its high frequency components, effectively improving listening effort. This leads to SWB having a slightly higher MOS-LE in this use case. Highly dynamic noises with a wider range of frequencies like *'Pub'* and *'Cafeteria'* on the other hand give NB speech a clear advantage.

To analyze these results, the use case *'Cafeteria'* is examined in more detail in as an example. Figure 17 shows the absolute spectra of the background noise scenario *'Cafeteria'* (red), the speech signal for narrowband (black) and super-wideband (blue) at the right ear for handheld hands-free application.

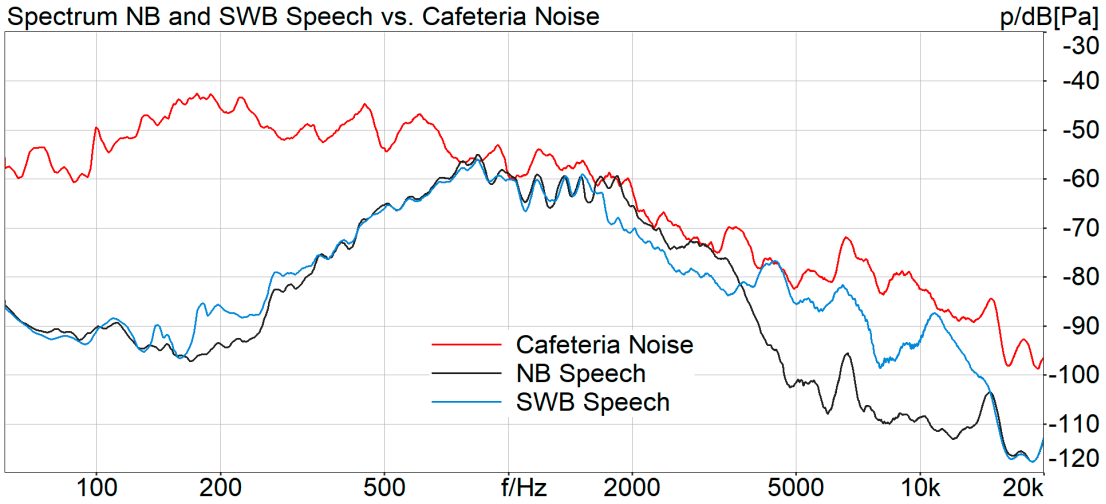



Fig. 17: Absolute spectra of the smartphone's hands-free loudspeaker with NB (black) and SWB (blue) speech in the presence of the background noise *'Cafeteria'* (red)

In this use case, the extension to low frequencies of speech playback in SWB is negligible at the ear. The extension to high frequencies is present, but superimposed at the ear by background noise. Additionally, the switch to SWB reduces speech output level in respect to NB by an average of 6 dB between 1.8 kHz and 3.7 kHz. This, combined with the small RLR advantage, leads to narrowband being equal or advantageous for hands-free application in the majority of environments.

### 4.2.4 MOS-LE Volume Comparison in Handset Mode (Volume Comparison)

In this sub-chapter, the calculated MOS-LE for nominal (Nom) and maximum (Max) volume of the smartphone in handset mode and compared. The background noise scenarios are identical with the previous chapters 4.2.2 and 4.2.3. All measurements are performed in narrowband. Figure 18 visualizes the MOS-LE results.

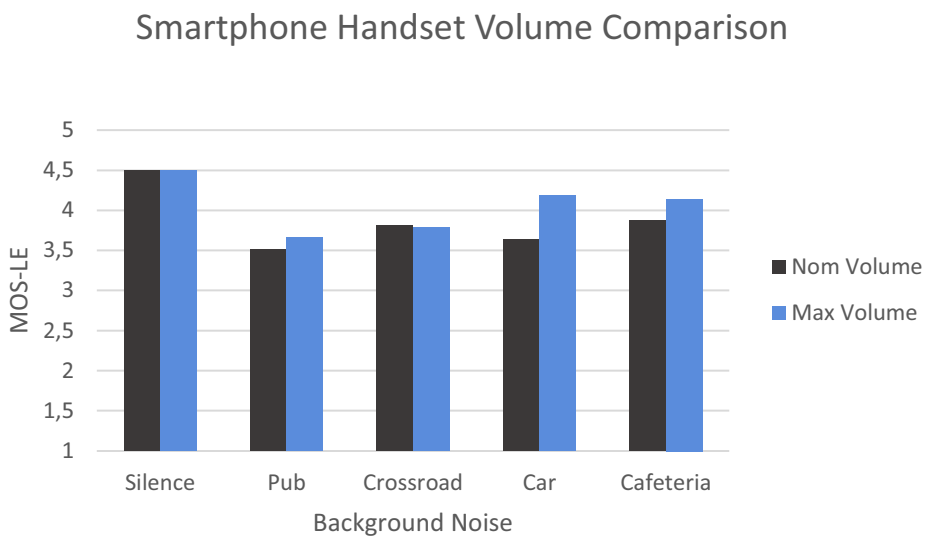



Fig. 18: MOS-LE of a mobile phone in handset mode at nominal and maximum volume in the presence of background noise

In handset mode, the decibel advantage for the maximum volume setting is very substantial: + 10.7 dB for NB, + 8.8 dB for SWB. As a result, the latter is advantageous in most environments. However, the benefit of a higher playback volume is countered by deteriorated playback quality. At maximum volume, the phone's loudspeaker and amplifier operate at their physical limits, which adds significant amounts of distortion to speech playback.

Consequently, from a viewpoint of listening effort, maximum volume is not necessarily better than nominal volume. This is visible in the MOS-LE results for *'Silence'* and *'Crossroad'*.

Without interfering background noise, the advantage of higher volume and loss of quality balance out. The highly time-variant *'Crossroad'* scenario on the other hand is so intrusive that single test sentences are incomprehensible regardless of playback volume. In the remaining three scenarios, maximum volume gains an advantage by improving the SNR at the ear.

### 4.2.5 MOS-LE Volume Comparison in Handheld Hands-Free Mode (Volume Comparison)

This sub-chapter repeats testing as described in the previous sub-chapter, but with the phone being in handheld hands-free mode instead of handset mode. Figure 19 shows the associated MOS-LE.

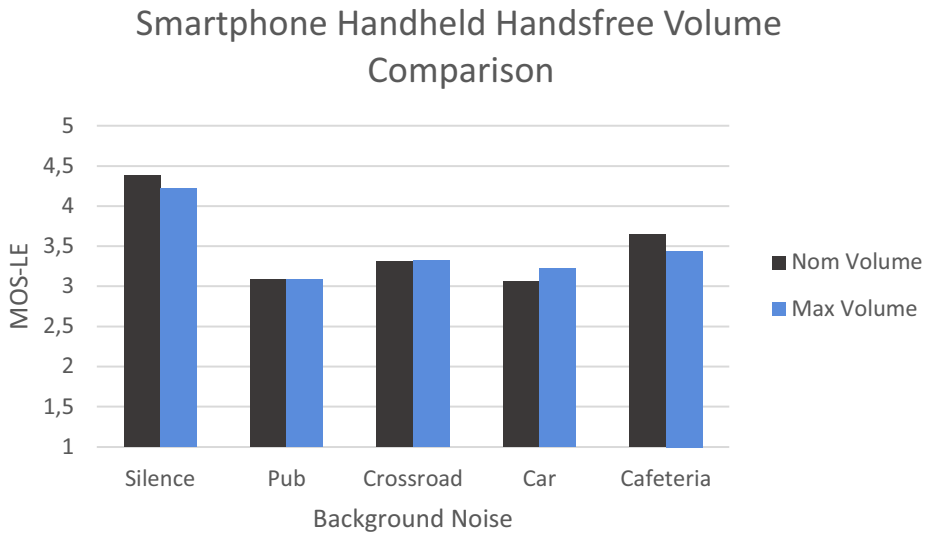



Fig. 19: MOS-LE of a mobile phone in handheld hands-free mode at nominal and maximum volume in the presence of background noise

As opposed to handset mode, the RLR between nominal and maximum volume is much smaller than in handset application: + 2.6 dB for NB, + 2.7 dB for SWB (see table 2). Additionally, the natural physical limits of loudspeaker and amplifier are much more dominant in far-from-the-ear applications (also see chapter 4.2.3). Consequently, possibilities to improve listening effort through raising the playback volume are miniscule. As laid out in the preceding chapter, setting the phone to maximum volume is generally accompanied by significantly higher playback distortion. Again, this effect is pronounced in hands-free application.

The combination of these effects leads to an equal or worsened listening effort and thus lower MOS-LE for maximum volume. Only one background noise scenario reaches a higher MOS-LE for maximum volume: the *'Car'*. This background noise is intrusive, but also time-invariant. This allows human hearing to separate speech from noise more efficiently than time-variant noise such as '*Pub'*, *'Crossroad'* and *'Cafeteria'*. In *'Car'*, the additional speech playback distortion in hands-free application is masked by driving noise. Thus, the small advantage in playback volume is able to counteract the additional distortion in this use-case and reach a slightly higher MOS-LE.

## 4.3 ICC systems

### 4.3.1 Scope & Setup

ICC systems aim to make conversation in vehicles easier by capturing typically the driver's voice and playing it back into the cabin in real time. To judge how well an ICC system fulfills this duty, assessing listening effort for vehicle passengers is a viable method.

In the following, a representative ICC system of a car was chosen. Measurements were performed in a stationary vehicle equipped with the background noise simulation system 3PASS *flex*. Talker and listener were simulated with two HMS II.3-33, one in the drivers seat and one in the rear behind the passenger seat. For additional hardware and software used for these measurements, please refer to chapter 3.3.

### 4.3.2 MOS-LE - ICC-System (on/off, 60/120 km/h)

The acoustic situation at hand – two conversational partners at a distance and not facing each other in a vehicle cabin with driving noise – "naturally impedes" good conversational quality. This is exactly why ICC systems can be very beneficial.

Figure 20 compares MOS-LE with the ICC system on and off, both calculated by ABLE for 60 km/h and 120 km/h.

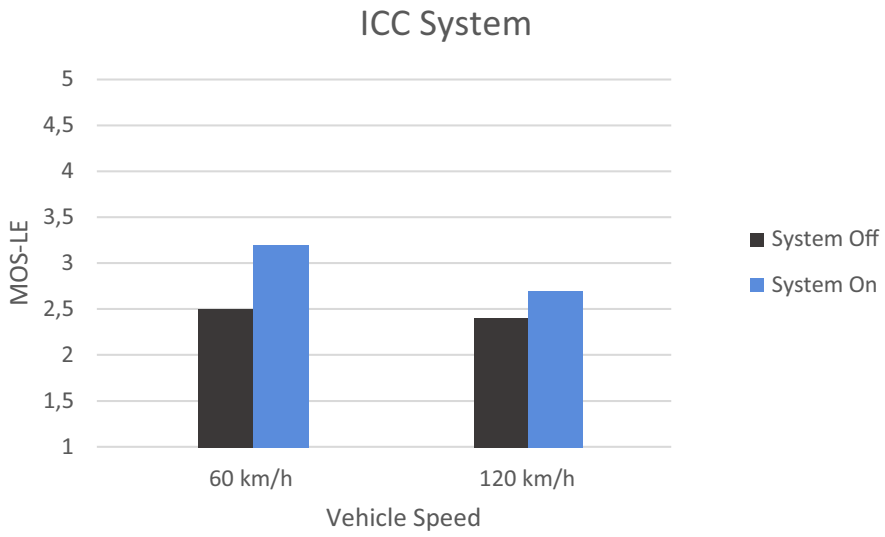



Fig. 20 MOS-LE comparison with deactivated vs. active ICC system at 60 km/h and 120 km/h

The MOS-LE results show two noteworthy patterns for this communication scenario:

1. Using the ICC system increases the MOS-LE at both speeds, proving the effectiveness of the system under test.
2. At 120 km/h, listening effort is generally worse than at 60 km/h due to increased driving noise. Additionally, the jump in MOS-LE with the ICC system is smaller than at 60 km/h.

Beyond low speeds, driving noise becomes the dominant noise component in the cabin. With high levels of driving noise, even clean speech played back through the ICC system at high volume levels can only improve MOS-LE within a limited range. As a consequence, small increases of the MOS-LE become more relevant with increasing vehicle speed.

This knowledge allows selective improvement of the ICC system, e.g. by increasing amplification and/or shaping the output's frequency response in relation to vehicle speed. Therefore, calculating the MOS-LE with ABLE allows manufacturers to test general performance of their ICC system as well as tweaking it for best performance in arbitrary driving situations without the need to perform actual driving.

Tests for functionality and quality of ICC systems and components are defined in Recommendation ITU-T P.1150 [24] issued in January of 2020. The recommendation incorporates listening effort as standardized in ETSI TS 103 558 [4]. Thus, ABLE can be used to test listening effort in this application.

# 5 Summary

This application note outlined the task of analyzing near-end voice communication quality in the presence of background noise. Existing SI-based metrics were examined regarding their qualification for this challenge. As none of them were explicitly developed for this task, inconsistent test results are inevitable. Additionally, no SI metric provides a standardized listening test design *and* the corresponding instrumental testing method at the same time. Therefore, reliable testing is impossible.

To solve this dilemma, the committee ETSI TC STQ developed a metric for near-end voice communication quality analysis, which was published as ETSI TS 103 558 (2019-11) [4]. It introduces perceived listening effort as a better suited approach to analyze near-end voice communication quality. Detailed analyses have shown that the results calculated via the described instrumental model match results obtained in listening tests very well. The ‘Assessment of Binaural Listening Effort’, or in short ‘ABLE’, is the implementation of the instrumental model by HEAD acoustics as a software option for ACQUA.

This document explained the operating principle of ABLE and its application to various different fields in telecommunication. Analyses of measurements showed that the method produces very plausible test results – the calculated MOS-LE match the expectations across all scenarios.

ABLE uses noisy speech signals as perceived by the user as an input. The highly accurate background noise simulation of 3PASS complements the measurement setup. Measurements with ABLE are fully repeatable and thus ensure consistent testing. Both input signals – binaurally recorded degraded and clean speech – are readily available or easy to obtain for any use case. The use of the established Mean Opinion Score (MOS, [15]) allows quick evaluation and comparison of results. The MOS-LE cumulates all information contained in the input signals and combines it with a sophisticated model of human auditory perception to obtain a single numerical value.

It can be concluded that ABLE is a simple, but universal, efficient and yet conclusive method for comprehensive analysis of near-end voice communication quality in telecommunication.

# 6 References


[1] J. Reimes, “Listening Effort vs. Speech Intelligibility in Car Environments,” in *Fortschritte der Akustik - DAGA 2015*, DEGA e.V., Berlin, 2015.

[2] J. Reimes, “Instrumental assessment of near-end perceived listening effort,” in *Perceptual Quality of Systems Workshop*, Berlin: DEGA e.V., 2016.

[3] J. Reimes and C. Lüke, “Perceived Listening Effort for In-car Communication systems,” in *ITG Fachtagung Sprachkommunikation*, VDE Verlag, Oldenburg, Germany, 2018.

[4] ETSI TS 103 558: *Speech and multimedia Transmission Quality (STQ); Methods for objective assessment of listening effort* V1.1.1 (2019-11).

[5] ANSI S3.5-1969. (R1986): *Methods for the calculation of the articulation index,* New York (NY): American National Standards Institute, 1969.

[6] ANSI S3.5-1997: *Methods for Calculation of the Speech Intelligibility Index*, 1997.

[7] *Speech interference level as a predictor of face-to-face communication in noise*, The Journal of the Acoustical Society of America 63, 581, 1978.

[8] ISO/DIS 532-1, *Acoustics – Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under freefield and diuse-field listening conditions*, International Organization for Standardization, 2005.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Dallas, Texas, USA), pp. 4214–4217, Mar 2010.

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,” *IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 7, pp. 2125–2136*, 2011.

[11] IEC TR 60268-16:2012-05, Sound system equipment - Part 16: *Objective rating of speech intelligibility by speech transmission index (IEC 60268-16:2011)*, 2012-05.

[12] https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/TSGS4_86/Docs/S4-151231.zip

[13] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 02/2001

[14] ITU-T Recommendation P.863, *Perceptual objective listening quality prediction*, 03/2018

[15] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, 08/1996.

[16] Barnett PW, Knight RD, “The common intelligibility scale,” *Proceedings of the Institute of Acoustics*, 17(7): 201–206, 1994

[17] ITU-T, *Handbook on Telephonometry*, ITU, 1992, ISBN 92-61-04911-7.

[18] ITU-T Recommendation P.807, *Subjective test methodology for assessing speech intelligibility*, 2016.

[19] ITU-T Recommendation P.501, *Test signals for use in telephonometry*, Jan. 2012.

[20] R. Sottek: “Modelle zur Signalverarbeitung im menschlichen Gehör,” *Ph.D. thesis*, RWTH Aachen, Techn. Hochsch., Diss., 1993.

[21] R. Sottek: “A Hearing Model Approach to Time-Varying Loudness,” *Acta Acustica united with Acustica*, 102(4), p. 725–744, 2016, ISSN 16101928.

[22] ETSI TS 103 224: *A sound field reproduction method for terminal testing including a background noise database* V1.4.1 (08-2019).

[23] ITU-T Recommendation P.57, *Artificial ears*, 12/2011.

[24] ITU-T Recommendation P.1150, *In-car communication audio specification*, 01/2020.