



Audio Engineering Society Conference Paper

Presented at the Conference on
Headphone Technology
2019 August 27 – 29, San Francisco, CA, US

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Evaluation of Predicted Listening Effort for Active Noise Cancelling Headsets

Jan Reimes

HEAD acoustics GmbH, Germany

Correspondence should be addressed to Jan Reimes (telecom@head-acoustics.de)

ABSTRACT

Comparing and benchmarking different active noise cancellation (ANC) headsets objectively is a challenging task. No standardized testing methods for such devices are available so far and measurements derived from other fields of communication may not fit the domain of ANC.

While the pure noise reduction capability of an ANC device may be assessed with simple metrics, the usage in communication scenarios (receiving direction or external talker) still is cumbersome; on the one hand, speech quality and intelligibility of the desired talker has to be preserved while reaching utmost noise cancellation performance at the same time. Recently, the auditory and instrumental assessment of Listening Effort (LE) became more popular for such applications.

This contribution introduces a preliminary model for the prediction of LE, which is suitable for binaural recordings. Four commercially available ANC headsets were evaluated in typical applications. The analysis is based on speech signals, realistic noise conditions and binaural recordings obtained with a head and torso simulator (HATS). Results of the prediction model on this measurement series are presented and discussed.

1 Introduction

Environmental noise in daily life is increasing due to e.g. more traffic, traveling and crowded public places. Headphones providing noise cancellation functionality are a simple technical remedy for this issue. Due to decreasing prices and increasing signal processing power in embedded devices, this solution became quite popular recently.

Beside the reduction of ambient noise, also speech signals in communication scenarios may be attenuated as well. In most cases, this is an expected and even

desired behavior, if users do not want to be disturbed. However, even though some devices provide modes for passing speech-like signals, the voice of a second person is often attenuated or processed as well by the ANC functionality and conversation with other people becomes difficult. On the other hand, when using the device as a headset for telecommunication, the noise processing must not disturb or influence the downlink speech signal.

For the evaluation of both scenarios, numerous acoustic tests may be considered. Even though there are ongoing work items in standardization [1], so far no common

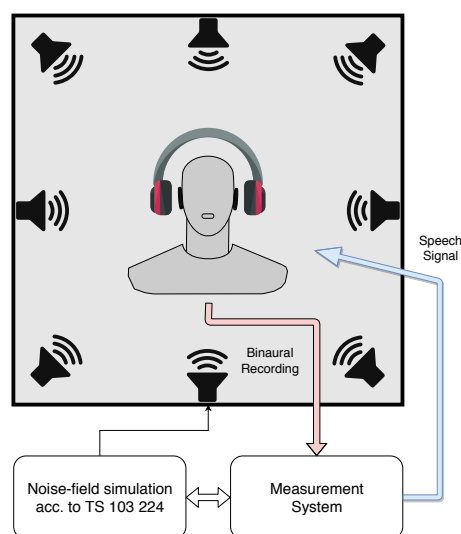


Fig. 1: Measurement setup for testing ANC

test procedures are available for e.g. benchmark or comparison of devices.

In order to investigate the impact on noise cancellation and speech transmission, several existing speech intelligibility metrics seem to be the most obvious way forward. However, in several studies [2, 3, 4] perceived listening effort was supposed to be a more suitable measure than intelligibility.

The upcoming specification ETSI TS 103 558 [5] aims at providing an auditory test design as well as an instrumental prediction model for LE. Even though still in a preliminary state, the algorithm is suitable for the use of binaural recordings, containing a mix of (residual) noise and processed speech signals.

Figure 1 illustrates the proposed acoustic setup for testing. The HATS according to ITU-T p.57 and p.58 [6, 7] is equipped with artificial ears of type 3.3. This ear type provides an anatomically correct pinna, which is valid for realistic mounting of in-ear/over-ear/on-ear headphones and headsets.

The artificial head is positioned in a measurement room equipped with a noise field simulation according to ETSI TS 103 224 [8]. The background noise system is equalized at eight microphone positions close / around the HATS and provides a highly accurate sound field simulation for lab testing. It is linked to the measurement system for triggered time-synchronous play and record of source signals.

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: Auditory scales for combined assessment

2 Auditory Test Methodology

Common auditory test procedures for speech in noisy environments are intelligibility tests. Here numerous test methods are generally available, e.g., matrix sentence tests as described for example in [9]. However, most of them are not described in freely accessible specifications, the corresponding audio material is not available and/or the focus of usage is mainly on audiology aspects, but not on telecommunication. Also the recently standardized auditory test procedure ITU-T P.807 [10] still has some drawbacks (e.g., rhyme test material is only available in English).

In addition, some intelligibility measures (like e.g., [11]) require separated speech and noise signals for analysis. Due to non-linear signal processing in telecommunication (and in particular for ANC applications), this precondition is almost never met.

An alternative approach for auditory evaluations is the perceived *Listening Effort* (LE). Here test subjects provide a self-assessment on a five-point categorical scale, similar to well-known speech quality testing methods. Several recent studies (e.g., [3, 4]) indicate that a wider range of signal-to-noise ratios (SNRs) can be evaluated regarding speech enhancement benefits, without reaching positive or negative saturation of intelligibility tests.

As opposed to speech intelligibility evaluations, this method has some advantages in practice. First, it is obvious that the speech material may be originated from

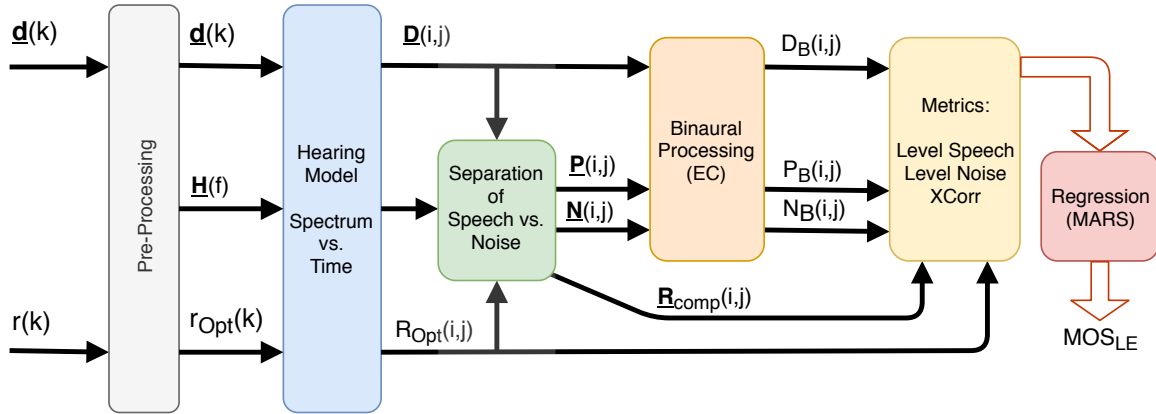


Fig. 2: Flow chart of prediction algorithm for Listening Effort (preliminary, acc. to ETSI 103 558 [5])

a much more limited corpus, because (within tolerable limits) repetitions of the stimulus do not corrupt or influence the test results. Another advantage is that this test design can be combined with additional scales, like e.g., speech quality or preferred loudness.

The usage of an optional *Speech Quality* (SQ) attribute supports test subjects in differentiating between the noise component (major impact on listening effort) of the signal and possible speech degradation (minor to medium impact on listening effort), which are included in a stimulus. Scales of both attributes were taken from ITU-T P.800 [12], as provided in Table 1.

In multiple other studies [2, 13, 14, 15], this approach was already proven to be successful. In each of these experiments, a mixture between the listening test methodologies ITU-T P.800 [12] and ITU-T P.835 [16] was used: The test subjects evaluated each presented sample twice. After the first presentation, a rating for LE was given. In a second trial, SQ was assessed.

3 Instrumental Prediction Model

3.1 Introduction

The model for predicted listening effort introduced in the following sections is currently under development. The work shown here represents an intermediate status and results are considered as preliminary. Thus, only a brief overview of the algorithmic parts is provided in this work.

Figure 2 provides an overview of the prediction algorithm, which consists of different stages. As described in section 1, binaural recordings are used as an

input and can be considered as a tuple of two single-channel signals (example given by equation 1). The clean speech reference $r(k)$ is a single-channel signal.

$$\mathbf{d}(k) = \langle d_L(k), d_R(k) \rangle \quad (1)$$

Similar to related speech quality metrics, an aurally-motivated transformation to the time-frequency domain of the signal is applied here (time index i , frequency bin j). These signal representations are denoted with capital letters $\mathbf{D}$ (egraded), $\mathbf{N}$ (oise), $\mathbf{P}$ (rocessed) and $\mathbf{R}$ (eference), which can again be considered as a tuple of spectra for the left and right ear (example given by equation 2).

$$\mathbf{D}(i, j) = \langle D_L(i, j), D_R(i, j) \rangle \quad (2)$$

3.2 Pre-Processing

The pre-processing stage aligns all input signals with regard to differences in temporal shifts and levels:

- The delay between degraded signals $\mathbf{d}(k)$ and the reference $r(k)$ is calculated by a cross-correlation analysis for both ears. The delay is selected from the ear providing the higher magnitude of the cross-correlation.
- The reference signal $r(k)$ is scaled to a fixed active speech level (ASL) of 79 dB_{SPL} ($r_{\text{Opt}}(k)$), which is assumed to be optimal regarding listening effort [17, 18].
- The transfer functions $\mathbf{H}(f)$ between degraded and reference signal are calculated by the H1 methodology (cross-power spectral density) in order to exclude non-correlated noise components.

3.3 Hearing Model Spectrum vs Time

The hearing model according to Sottek [19, 20] is calculated for the clean speech and processed signal. The transformation includes an auditory filter bank representation of the signal and a hearing-adequate envelope determination which leads to a frame resolution of about 8 ms. The frequency bandwidth $\Delta f(f_j)$ for the j -th frequency band is given by

$$\Delta f(f_j) = \Delta f(f_0) + c \cdot f_j. \quad (3)$$

The initial bandwidth $\Delta f(f_0)$ is set to 70 Hz and the factor c equals 0.14, which results in 27 bands up to 20 kHz. In contrast to other hearing-adequate frequency scales, the proposed method includes the whole fullband (FB) range.

This time-frequency representations is calculated for all pre-processed signals individually for left and right ear, resulting hearing model spectra vs time $\mathbf{D}(i, j)$ and $R_{\text{Opt}}(i, j)$.

3.4 Separation of Speech and Noise

For the calculation of metrics correlating with auditory assessed listening effort, it is useful to separate the degraded spectra $\mathbf{D}(i, j)$ into a (processed) speech ($\mathbf{P}(i, j)$) and a noise ($\mathbf{N}(i, j)$) component, as shown in equation 4.

$$\mathbf{D}(i, j) = \mathbf{P}(i, j) + \mathbf{N}(i, j) \quad (4)$$

For this purpose, the same principle as described in [21] is applied, in order to predict noisy speech quality and noise intrusiveness. A (pseudo-)Wiener filter is used here for the determination of $\mathbf{P}(i, j)$ as shown in equation 5.

$$\mathbf{P}(i, j) = \mathbf{D}(i, j) \cdot \mathbf{W}(i, j) \quad (5)$$

The Wiener gain $\mathbf{W}(i, j)$ is obtained by the interpolated magnitude of the transfer function according to equation 6 and a suitable noise estimation $\tilde{\mathbf{N}}(i, j)$.

$$\mathbf{W}(i, j) = \sqrt{\frac{\mathbf{R}_{\text{Comp}}(i, j)^2}{\mathbf{R}_{\text{Comp}}(i, j)^2 + \tilde{\mathbf{N}}(i, j)^2}} \quad (6)$$

$$\mathbf{R}_{\text{Comp}}(i, j) = R_{\text{Opt}}(i, j) \cdot \mathbf{H}(j) \quad (7)$$

For the determination of $\tilde{\mathbf{N}}(i, j)$, first a soft mask $\mathbf{M}(i, j)$ of active/inactive time-frequency bins is determined with $\mathbf{R}_{\text{Comp}}(i, j)$ (equation 7). The speech part classification algorithm according to [22] is carried out for

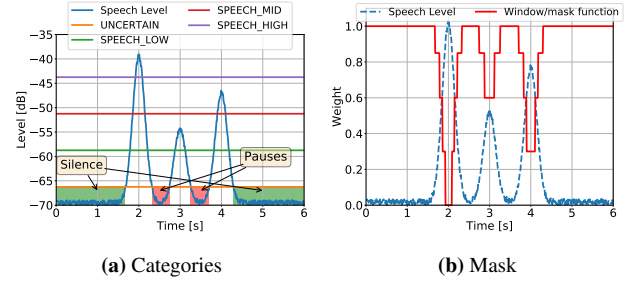


Fig. 3: Frame classification and mask generation

each frequency band. For silent and paused frames, a mask value of 1.0 is set, for high activity of speech to 0.0. Weights for medium, low and uncertain activity frames are calculated by linear interpolation. An example of this threshold-based method for one single frequency band is provided in Figure 3.

Then, the degraded spectra are multiplied by the masks, resulting in $\mathbf{M}_D$ as per equation 8. This step suppresses the active speech frames and can be considered as a windowed noise signal (time-frequency indices neglected for sake of clarity).

$$\mathbf{N} \cdot \mathbf{M} \approx \mathbf{D} \cdot \mathbf{M} = \mathbf{M}_D \quad (8)$$

With the Fourier operator $\mathcal{F}(\cdot)$, an estimate for the noise signal per frequency band can be described as a deconvolution problem $((x * y)^{-1})$ according to equations 9 to 11.

$$\mathbf{M}_D \leadsto \mathcal{F}(\mathbf{M}_D) = \mathcal{F}(\mathbf{N}) * \mathcal{F}(\mathbf{M}) \quad (9)$$

$$\Leftrightarrow \mathcal{F}(\mathbf{N}) = (\mathcal{F}(\mathbf{M}_D) * \mathcal{F}(\mathbf{M}))^{-1} \quad (10)$$

$$\Rightarrow \tilde{\mathbf{N}} = \mathcal{F}^{-1}((\mathcal{F}(\mathbf{M}_D) * \mathcal{F}(\mathbf{M}))^{-1}) \quad (11)$$

In general, multiple deconvolution algorithms would be possible here. For the current model, the algorithm described in e.g., [19] or [23] is used. The estimated noise spectra are directly used as the noise itself: $\mathbf{N}(i, j) \approx \tilde{\mathbf{N}}(i, j)$.

3.5 Binaural Processing

In order to address the capability of human hearing to improve SNR compared to monaural listening, a binaural processing stage is included in the prediction model. The spectral components for left and right ears are combined by a short-term equalization-cancellation (stec) according to [24]. This extension of the well-known model of Durlach [25] requires the

availability of the isolated speech and *masker* (noise-only) components. As a result of this stage, combined and enhanced hearing model spectra vs time are provided, like e.g., $D_B(i, j)$ for the degraded signal.

3.6 Metric Calculation

Based on the binaural spectra, several comparisons and single value metrics can be considered. A common method is to compare the degraded or processed signal to the reference, but also single-ended metrics without a comparison are possible. The development of metrics is still ongoing, for the current investigation, the following metrics were chosen.

3.6.1 Level-Based Metric

Since the hearing model spectra for processed speech and noise component are individually available, level metrics can be calculated in the frequency domain.

The active speech level S_{act} is calculated according to equation 12 on the processed signal. Based on the frame classification of the pre-processing, only active parts are taken into account. In addition, an A-weighted noise level N_A is calculated from the noise spectrum according to equation 13.

$$S_{act} = 10 \cdot \log\left(\sum_{j \in A} \sum_{i=0}^{27} P_B(i, j)^2\right) \quad (12)$$

$$N_A = 10 \cdot \log\left(\sum_{j=0}^{J-1} \sum_{i=0}^{27} (A(j) \cdot N_B(i, j))^2\right) \quad (13)$$

3.6.2 Correlation-Based Metric

Similar as the intelligibility metric introduced in [26, 27], a spectral cross-correlation is carried out in four different ways. Instead of the clipping procedure described in [26, 27], the threshold of hearing is applied [28] to the spectra. After this step, the non-linear loudness transformation according to [20, 19] is applied on all spectra (denoted as e.g., $D'_B(i, j)$ in the following). The normalization procedure introduced in [26, 27] is not necessary, since the data is already provided on an absolute loudness scale. Each band is divided into sub-frames of 320 ms ($N = 40$ frames, only active indices) with 50 % overlap.

The correlation metric $d(m, j)$ between two spectra X and Y is calculated per frequency band j and for the m -th sub-frame as shown in equation 14.

$$d(m, j) = \frac{\sum_n (X_j(n) - \frac{1}{N} \sum_l X_j(l)) (Y_j(n) - \frac{1}{N} \sum_l Y_j(l))}{\sqrt{\sum_n (X_j(n) - \frac{1}{N} \sum_l X_j(l))^2 \sum_n (Y_j(n) - \frac{1}{N} \sum_l Y_j(l))^2}} \quad (14)$$

Equation 15 provides the aggregation to an overall metric. With $X \in [D'_B, P'_B]$ and $Y \in [R'_{Opt}, R'_{B, comp}]$, four metrics can be derived by this analysis.

$$d_{X,Y} = \frac{1}{MJ} \sum_m \sum_j d(m, j) \quad (15)$$

3.7 Prediction of MOS

The metric stage provides multiple single values, which can be combined to an instrumental single listening effort score. Depending on the amount of training data, suitable and advanced machine learning techniques may be considered. Since only a few training databases are available now, *multivariate adaptive regression splines* (MARS) according to [29] are used for the current investigation. Parameters and coefficients of the regression are derived from data of similar studies related to the auditory assessment of listening effort [2, 13, 14, 15].

4 Application on ANC-Headsets

4.1 Introduction

Four commercially available devices under test (DUTs) were evaluated with the measurement as described in section 1. They are denoted as DUT A to D in the following clauses. DUT A, C and D represent over-ear headsets, while DUT B is an in-ear device.

Two noise types were used for the noise field simulation according to [8]:

- *Train Compartment*: This scenario is not part of any official noise database and was recorded in particular for this evaluation. On average, the level was determined to $\approx 64 \text{ dB}_{SPL}(A)$, measured close to the ears.
- *Airplane*: This scenario *Inside_Aircraft2_handset* was taken from the noise database of [8]. On average, the level was determined to $\approx 80 \text{ dB}_{SPL}(A)$, measured close to the ears.

For the evaluation of listening effort, speech signals can be inserted into the measurement setup by multiple ways. The type and calibration of speech playback depends on the specific use case, two of them are described in the following clauses. In all cases, background noise playback was triggered and synchronized with the start of the measurement, i.e. each device was exposed to exactly the same mixture of speech and noise.

As a source signal for all evaluations, a speech sequence consisting of 16 American English sentences according to Annex C of ETSI TS 103 281 [21] was used. The binaural signals were captured at the drum reference-point (DRP) including diffuse-field (DF) correction (cf. [7, 30]).

4.2 Communication Setup

A typical application of an ANC device is the usage as a headset, i.e. the device is connected to a mobile phone or computer for telecommunication purposes. Here the speech signal is inserted into receive (RCV) (or often also referred as *downlink*) direction. The measurement setup is illustrated in Figure 4a.

The investigated DUTs were connected via a Bluetooth™ reference gateway (cf. e.g., [31] or [32]). The connection was established with hands-free profile (HFP) and wideband (WB) mode, which allows an audio bandwidth up to 8 kHz.

The ASL according to ITU-T P.56 [33] under silent conditions was adjusted to a range of 73-79 dB_{SPL} per ear, which is considered as an optimum level regarding listening effort and quality [17, 18]. This adjustment via volume control was applied for each device separately. For all devices, ANC functionality was enabled and set to maximum cancellation operation mode (not applicable for all devices).

4.3 Talk-through Setup

When the listener is wearing an ANC device, speech signals originated from the surroundings (e.g., another talker or public address systems) may unintentionally be attenuated in the same way as the environmental noise. In order to evaluate the capabilities of devices to pass speech signals, but still canceling noise at the same time, the so-called *talk-through* (TT) mode is investigated more in detail. Manufacturers often have several and different designations for this functionality, which is often also referred as e.g., *transparency*, *ambient* or *voice pass-through* mode.

	Airplane	Train Compartment
Noise Level [dB _{SPL} (A)]	80.0	64.0
ASL Offset [dB]	9.1	4.8
ASL @ MRP [dB _{Pa}]	4.4	0.1

Table 2: Level adjustments for 2nd talker setup

	Airplane	Train Compartment
Noise Level [dB _{SPL} (A)]	80.0	64.0
ASL @ HATS [dB _{SPL}]	78.5	72.0
SNR [dB]	-1.5	8.0

Table 3: Level adjustments for loudspeaker setup

4.3.1 Second Talker

In order to simulate (the beginning of) a conversation, a second HATS was positioned at a distance of $d = 50$ cm to the left of the listening one, as shown in Figure 4b. The speech signals were played back over the artificial mouth of the second HATS, which can be considered as a highly directional loudspeaker. It was equalized as required per ITU-T P.581 [30] up to 20 kHz.

The speech signal was calibrated to obtain a reasonable level at the talker's side. The corresponding level settings are provided in Table 2. These increased playback levels (compared to the typical value for silence, -4.7 dB_{Pa}) address the Lombard effect [34] in noisy scenarios. The applied gains were determined as specified in [35]. Similar to the communication setup, ANC functionality was enabled and set to maximum cancellation operation mode for all devices.

4.3.2 Loudspeaker

In order to simulate e.g., an announcement over a public address system, a loudspeaker was positioned as shown in Figure 4c. This time, the speech is originated by a source located at height $h = 50$ cm above and 45° in front of the listener. Otherwise, the setup is identical to clause 4.3.1.

The loudspeaker output was calibrated to obtain a typical and reasonable level and SNR at the listener's ears (without wearing a device) for such scenarios. The corresponding level settings are provided in Table 3.

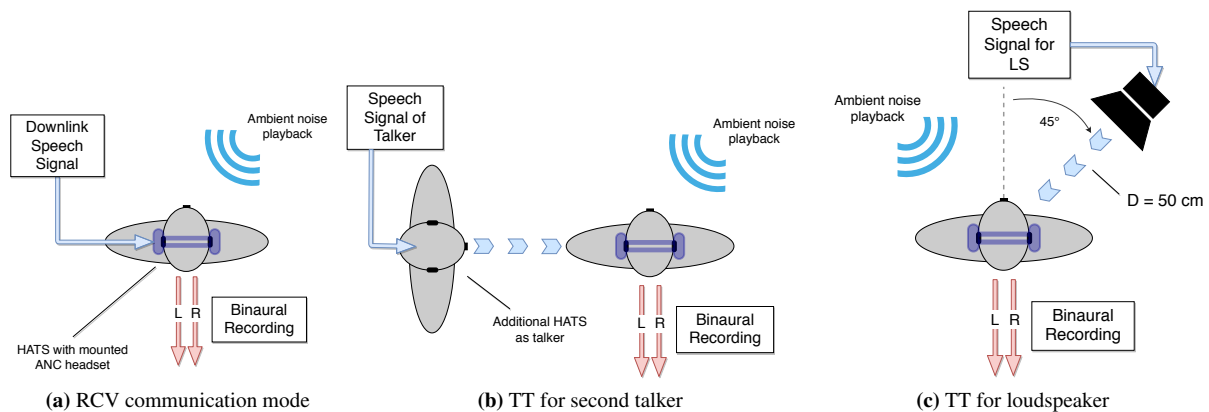


Fig. 4: Investigated ANC setups

As introduced in section 4.3, all evaluated devices provided at least one setting for improved talk-through mode, which was enabled in this particular setup.

5 Results

In the following, results for the three application modes as described in section 4 are presented and discussed. In order to analyze details of the prediction model, Table 4 and Table 5 provide the most impacting metrics, the speech and noise level as described in section 3.6.1. Additionally, the SNR (difference between speech and noise) as well as the final MOS_{LE} are included.

The spectral correlation metrics (section 3.6.2) only have slight impacts on the final MOS_{LE} , since disturbances like spectral shaping or audible artifacts can either be considered as constant across applications or are almost not present (at least for the current investigation). For the sake of simplicity, these metrics are neglected in Table 4 and Table 5 as well as in the following sections.

5.1 Communication Mode

The first column of Table 4 and 5 show the prediction results for the communication mode across the four DUTs. As expected, the predicted LE significantly decreases due to the increasing level of noise, when comparing results between *Train Compartment* and *Airplane* noise scenarios.

As expected, the speech level S_{act} seems not to be affected by the noise cancellation, the values remain stable within a range of ~ 74 - 80 dB_{SPL}.

On the other hand, the noise cancellation capabilities obviously seem to have more impact on MOS_{LE} . Even though DUT C provides quite low speech levels in both noise scenarios, this device outperforms the others with respect to residual noise level and thus, MOS_{LE} . For example, $MOS_{LE} = 2.8$ is assessed for the most demanding noise type *Airplane*, including about 7.4 dB less noise level than the DUT B. Also for the *Train Compartment* scenario, best overall performance regarding MOS_{LE} and noise level is obtained for this device.

In general, it would be expected that speech playback via headphone would result in even better values for MOS_{LE} . It has to be noted that many devices include amplification of the audio signal in the low frequency domain. By adjusting all devices to a common volume setting, this indirectly attenuates high-frequency components, which are important for intelligibility. Additional impacts of frequency response, limitation of bandwidth and coding may also reduce the similarity between perceived and optimum reference signal.

5.2 Talk-through - 2nd Talker

The second column of Table 4 and 5 show the prediction results for the talk-through mode with a second talker for the two noise scenarios across the four DUTs.

As expected, S_{act} decreases compared to the communication mode (in some cases more than 10 dB), due to active and passive attenuation of any sound field outside the headset. In general, the noise cancellation performance remains comparable across devices and noise types. However, in some cases an improved performance up to 9 dB can be observed (DUT D for *Train*

Mode Device	Communication				TT: 2 nd talker				TT: Loudspeaker			
	A	B	C	D	A	B	C	D	A	B	C	D
MOS _{LE}	1.8	2.2	2.8	2.3	1.9	2.3	2.7	2.1	2.0	2.2	1.9	2.2
SNR(A) [dB]	10.5	20.1	21.5	18.5	5.1	8.5	10.1	9.0	10.3	10.9	4.5	9.9
S _{act} [dB _{SPL}]	79.8	80.2	74.2	78.9	72.6	64.8	61.3	66.7	74.2	68.6	69.3	66.7
N _A [dB _{SPL} (A)]	69.4	60.1	52.7	60.4	67.5	56.3	51.3	57.6	63.8	57.6	64.8	56.7

Table 4: Results for *Airplanes* scenario in all application modes

Mode Device	Communication				TT: 2 nd talker				TT: Loudspeaker			
	A	B	C	D	A	B	C	D	A	B	C	D
MOS _{LE}	2.8	3.7	3.5	2.9	2.6	3.7	3.6	3.4	1.9	2.2	2.0	2.1
SNR(A) [dB]	25.0	36.7	32.0	25.1	12.3	23.2	16.8	18.2	5.5	8.6	-1.2	6.0
S _{act} [dB _{SPL}]	79.3	79.9	77.4	77.0	67.9	60.5	55.3	61.3	73.1	65.5	54.2	63.7
N _A [dB _{SPL} (A)]	54.3	43.2	45.4	51.9	55.6	37.3	38.5	43.2	67.6	56.9	55.4	57.7

Table 5: Results for *Train Compartments* scenario in all application modes

Compartment). This indicates that noise cancellation performance may be impaired by the simultaneous audio playback in communication mode.

This may also be explanation for similar MOS_{LE} results as in the previous mode, even though the setup is very different. This indicates that LE is less sensitive against the mere speech level. Especially for low SNR ranges, noise level reduction seems more important.

5.3 Talk-through - Loudspeaker

The third column of Table 4 and 5 show the prediction results for the talk through mode with an additional loudspeaker for the two noise scenarios across the four DUTs. In this setup, a dedicated talk-through mode was used. As expected, the noise cancellation performance decreases (much higher residual noise levels), while the speech level increases compared to the previous talk-through mode. The predicted listening effort seems not to benefit from this increase, results for MOS_{LE} are consistently lower. The concurrent gain of noise level is predominating here.

However, for some conditions, the prediction of MOS_{LE} may be investigated more in detail. Looking closer at e.g., noise type *Train Compartment* at DUT C, a negative SNR is provided by a quite low noise and speech level, both about 54-55 dB_{SPL}. Such conditions were not part of any training databases (c.f. section

3.7). On the other hand, this extrapolation leading to MOS_{LE} = 2.0 seems at least not to be unrealistic.

6 Discussion

Results shown in the previous clause indicate that the assessment of listening effort can mainly be explained by the (residual) noise and secondarily by the (mere) speech level.

From the user's perspective, it seems not sufficient to consider only SNR for the qualification of listening effort. It seems more important to investigate at which speech/noise level a certain SNR is obtained. Especially for the talk-through scenarios this becomes obvious, where a wide range of speech and noise levels are involved. However, in all three application modes, the noise cancellation performance seems to have more impact than an increase of speech level with a comparable amount.

In further work, it may be analyzed up to which degree the combination of speech and noise levels may already provide an sufficiently accurate prediction.

7 Summary

The work presented in this contribution investigated the performance of ANC headsets regarding several

aspects of speech processing. The auditory methodology of perceived listening effort was motivated and introduced. Based on this approach, a model for the instrumental assessment is currently being developed and a status of the preliminary prediction algorithm was briefly introduced.

A measurement series was conducted with four commercially available ANC headsets. Typical operational modes like communication and talk-through from different sources were considered. Realistic testing procedures based on binaural recordings obtained with an HATS, noise field simulation and real speech signals were introduced and allow a valid usage of the instrumental listening effort assessment.

Even though the prediction model was not yet tuned and validated for the application on ANC headsets, the proposed model provides at least plausible scores for the conducted measurement data. In particular, the trade-off between residual noise and sufficient speech level seems to be addressed correctly.

8 Outlook

The instrumental assessment of LE still is based on auditory test databases, which do not yet include ANC headsets. Even though the model already shows promising initial results, so far no comparison to suitable auditory tests are available. Especially the extrapolation to borderline cases needs further investigation, like e.g., the combination of very low speech and noise levels. The generation of listening test databases for training/tuning and validation is a necessary next step for the further work on the prediction model.

Another aspects to be considered in future work is the performance of headsets between activated and deactivated ANC functionality as well as comparisons to speech in noise without any headset. In such cases, an improvement in LE may provide even more useful hints than absolute values, which were regarded in the current work.

Updated work and results will be reported to ongoing work items in standardization. For example, ETSI TC STQ currently investigates several aspects on listening effort for applications in telecommunication [5] - including auditory and instrumental testing of ANC headsets. The same committee is also drafting a technical specification for the performance evaluation of ANC headsets [1]. Beside numerous measurements

focusing on individual aspects (like e.g., noise level attenuation in dB), listening effort assessment as described in this work would be a suitable tool for this upcoming specification.

References

- [1] *Test Methods and Performance Requirements for Active Noise Cancellation Headsets and other Ear-Worn Devices (draft; expected to be finalized 11/2019; may be shared upon request)*, ETSI TS 103 640 V0.0.2, 2019.
- [2] Reimes, J., "Listening Effort vs. Speech Intelligibility in Car Environments," in *Fortschritte der Akustik - DAGA 2015*, DEGA e.V., Berlin, 2015.
- [3] Pusch, A. et al., "Binaural listening effort in noise and reverberation," in *Fortschritte der Akustik - DAGA 2018*, volume 44, pp. 543–546, Berlin, Germany, 2018.
- [4] Rennie, J. and Kidd, G., "Binaural listening effort in noise and reverberation," in *Fortschritte der Akustik - DAGA 2018*, volume 44, pp. 615–616, Berlin, Germany, 2018.
- [5] *Methods for objective assessment of listening effort (draft; expected to be finalized 03/2020)*, ETSI TS 103 558 V0.0.4, 2019.
- [6] *Artificial ears*, ITU-T Recommendation P.57, 2011.
- [7] *Head and torso simulator for telephonometry*, ITU-T Recommendation P.58, 2013.
- [8] *A sound field reproduction method for terminal testing including a background noise database*, ETSI TS 103 224 V1.3.1, 2017.
- [9] Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Usler, V. N., Brand, T., and Wagener, K., "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International journal of audiology*, 54 Suppl 2, pp. 3–16, 2015.
- [10] *Subjective test methodology for assessing speech intelligibility*, ITU-T Recommendation P.807, 2016.
- [11] ANSI S3.5, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, 1997.

- [12] *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, 1996.
- [13] Reimes, J., "Auditory Evaluation of Receive-Side Speech Enhancement Algorithms," in *Fortschritte der Akustik - DAGA 2016*, DEGA e.V., Berlin, 2016.
- [14] Reimes, J. and Hans W. Gierlich, "Auditory Listening Effort Assessment for ICC systems," *Contribution to ITU-T SG12/Q4*, T17-SG12-C-0157, 2018, Geneva/CH.
- [15] Reimes, J. and Lüke, C., "Perceived Listening Effort for In-car Communication systems," in *ITG-Fachtagung Sprachkommunikation*, VDE Verlag, Oldenburg, Germany, 2018.
- [16] *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, ITU-T Recommendation P.835, 2003.
- [17] ITU-T, *Handbook on Telephonometry*, ITU, 1992, ISBN 92-61-04911-7.
- [18] ITU-T, *Practical Procedures for Subjective Testing*, ITU, 2011.
- [19] Sottek, R., *Modelle zur Signalverarbeitung im menschlichen Gehör*, Ph.D. thesis, RWTH Aachen, Aachen, 1993, aachen, Techn. Hochsch., Diss., 1993.
- [20] Sottek, R., "A Hearing Model Approach to Time-Varying Loudness," *Acta Acustica united with Acustica*, 102(4), p. 725–744, 2016, ISSN 1610-1928.
- [21] *Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*, ETSI TS 103 281 V1.2.1, 2018.
- [22] *Voice enhancement devices - Appendix II*, ITU-T Recommendation G.160 Amendment 2, 2011.
- [23] Aach, T. and Metzler, V., "Defect Interpolation in Digital Radiography - How Object-Oriented Transform Coding Helps," *SPIE*, 4332, 2001.
- [24] Wan, R., Durlach, N. I., and Colburn, H. S., "Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers," *The Journal of the Acoustical Society of America*, 136(2), pp. 768–776, 2014, doi:10.1121/1.4884767.
- [25] Durlach, N. I., "Binaural signal detection: Equalization and cancellation theory," *Foundations of Modern Auditory Theory*, Vol. 2, 1972.
- [26] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, Dallas, Texas, USA, 2010.
- [27] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), pp. 2125–2136, 2011, doi:10.1109/TASL.2011.2114881.
- [28] ISO/DIS 532-1, *Acoustics – Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions*, International Organization for Standardization, 2005.
- [29] Friedman, J. H., "Multivariate adaptive regression splines," *Ann. Statist.*, 1991.
- [30] *Use of head and torso simulator for hands-free and handset terminal testing*, ITU-T Recommendation P.581, 2014.
- [31] *Wideband hands-free communication in motor vehicles*, ITU-T Recommendation P.1110, 2017.
- [32] *Transmission requirements for narrowband wireless terminals (handset and headset) from a QoS perspective as perceived by the user*, ETSI TS 103 737 V1.3.1, 2018.
- [33] *Objective measurement of active speech level*, ITU-T Recommendation P.56, 2011.
- [34] Lombard, Étienne, "Le signe de l'élévation de la voix," *Annales des maladies de l'oreille, du larynx, du nez et du pharynx*, 37, pp. 101–119, 1911, ISSN 0150-9721.
- [35] *Narrowband hands-free communication in motor vehicles*, ITU-T Recommendation P.1100, 2017.