The Adjustment / Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and its Application to Dialogue Enhancement

Matteo Torcoli, Jürgen Herre, *Senior Member, IEEE,* Harald Fuchs, *Member, IEEE,* Jouni Paulus, and Christian Uhle

***Abstract*—Media consumption in broadcasting is heading towards high degrees of content personalization also in audio thanks to next-generation audio systems. It is thus crucial to assess the benefit of personalized media delivery. To this end, the Adjustment / Satisfaction Test (A/ST) was recently proposed. This is a perceptual test where subjects interact with a user-adjustable system and their adjustments and the resulting satisfaction levels are studied. Two configurations of this test paradigm are implemented and compared for the evaluation of Dialogue Enhancement (DE). This is an advanced broadcast service which enables the personalization of the relative level of the dialog and the background sounds. The test configuration closer to the final application is found to provide less noisy data and to be more conclusive about the Quality of Experience. For this configuration, DE is tested both in the case in which the original audio objects are readily available and in the case in which they are estimated by blind source separation. The results show that personalization is extensively used, resulting in increased user satisfaction, in both cases.**

***Index Terms*—Adjustment / Satisfaction Test (A/ST), Audio Systems, Advanced Broadcast Services, Blind Source Separation (BSS), Dialogue Enhancement (DE), Digital Audio Broadcasting, MPEG-H, Next-Generation Audio (NGA), Perceptual Evaluation, Personalization, Quality of Experience (QoE), User Satisfaction.**

I. Introduction

ULTRA High-Definition Television (UHD TV) is being deployed around the world, offering many advantages to TV users, such as advanced interactivity and a highly personalized experience. This is possible also for the broadcast audio thanks to Next-Generation Audio (NGA) systems such as MPEG-H Audio (see Section II).

Thus it is fundamental for content producers, system engineers, and broadcasters to understand which personalization services are desired by the users and how these services improve the Quality of Experience (QoE).

On an abstract level, given a user-adjustable system and assuming that the available personalization is designed so as to improve the users QoE, we want to answer the following research questions RQ1 and RQ2.

RQ1) To what extent is the personalization used by the subjects? Is it used at all or are the subjects satisfied with a given default setting?

RQ2) How much is the QoE increased by the available personalization?

In order to investigate RQ1 and RQ2, the work in [1] recently proposed the Adjustment / Satisfaction Test (A/ST), where the user satisfaction is used as a measure of the QoE. The current paper expands upon [1], giving a more complete review of related technologies and literature, new subjective data, a deeper interpretation of the results, and discussing a variant of the test design.

The rationale of the A/ST is to provide subjects with the possibility of adjusting a prototype of the final application (the so-called adjustment phase). Hence, the subjects are asked to rate the difference between the given default setting and the one preferred during the adjustment phase.

Even if designed with audio in mind, the rationale is more general and can potentially find its application also with other media. The focus of this paper is on the application to Dialogue Enhancement (DE), i.e., a system where end users can personalize the level ratio between dialog and background. In the context of DE, the term *dialog* refers to all types of foreground speech, including monologs, narrations, and news reading. All the other sound sources are referred to as *background*, as they mostly consist of background music, sound effects, and ambient noise.

DE can be implemented with object-based audio. This requires that the audio objects are separately available on the encoder side. Still, the audio is often only available as mono, stereo, or a 5.1 mix, especially for archive content or low-budget productions. In these cases, methods for decomposing the mixture signals into separate signal components are needed to open the way for DE. Decomposition strategies that can be adapted to DE are numerous in the literature on speech enhancement and Blind Source Separation (BSS) [2]–[8]. These techniques are not able to perfectly reconstruct the original objects, and artifacts, distortions, or changes in timbre may be introduced. These may influence the way in which

Manuscript accepted March 31, 2018 for publication in the IEEE Transactions on Broadcasting, vol. 64, no. 2, pp. 524-538, June 2018, Digital Object Identifier 10.1109/TBC.2018.2832458.

M. Torcoli, J. Herre, H. Fuchs, J. Paulus, and C. Uhle are with Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany. Correspondence should be addressed to Matteo Torcoli (matteo.torcoli@iis.fraunhofer.de).

J. Herre, J. Paulus, and C. Uhle are also with International Audio Laboratories Erlangen, a joint institution of Universität Erlangen-Nürnberg and Fraunhofer IIS.

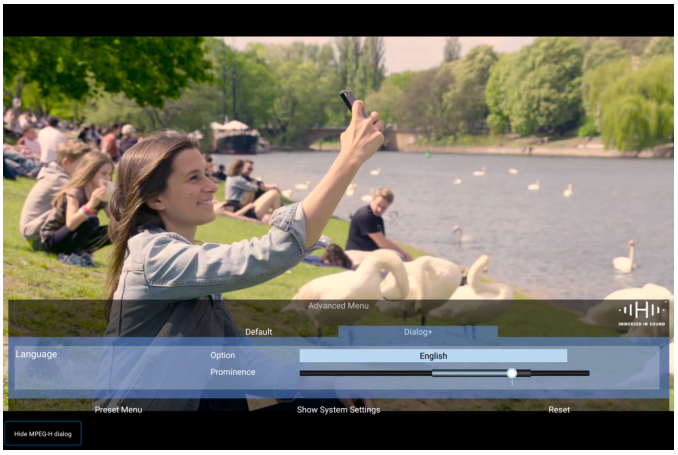

Fig. 1. Example of a user interface for Dialogue Enhancement in MPEG-H.

the users interact with DE and the resulting QoE. Thanks to the A/ST, we are able to study not only the QoE for DE applications where the original audio objects are available, but also the case where BSS is employed and the differences between the two cases.

The remaining part of this paper begins with Section II, which gives an introduction to MPEG-H Audio and the broad range of personalization possibilities. Section III proceeds to review works related to the evaluation of personalization in audio. Subsequent to this, the A/ST is described in detail in Section IV; its application to DE and the collected subjective data are discussed in Sections V – VII. Conclusions are given in Section VIII.

II. Next-Generation Audio Systems

A. MPEG-H Audio

NGA systems such as MPEG-H Audio [9], [10] offer true immersive sound and advanced user interactivity features. Their object-based concept of delivering separate audio elements with metadata within one audio stream enables new ways of personalization and universal delivery.

MPEG-H Audio provides a complete integrated audio solution for delivering the best possible audio experience, taking into account the final reproduction system and the listening environment. To achieve this, it includes rendering and downmixing functionality, together with advanced Loudness and Dynamic Range Control (DRC) tools. MPEG-H Audio enables immersive sound, i.e., the sound can come from all directions, including above or below the listener, using any combination of the three well-established audio formats: Channel-based, Object-based, and Higher-Order Ambisonics (HOA).

MPEG-H Audio [11]–[13] is adopted by the Digital Video Broadcasting Project (DVB) [14] and the Advanced Television Systems Committee (ATSC) standard ATSC 3.0 [15] and is selected by the Telecommunications Technology Association (TTA) in South Korea as the sole audio codec for the terrestrial UHDTV broadcasting system [16]. These broadcast specifications refer to the MPEG-H 3D Audio Low Complexity Profile Level 3 that allows up to 16 audio elements (channels, objects or HOA signals) to be decoded simultaneously.

B. Interactivity and Personalization

MPEG-H Audio enables viewers to interact with the content in new ways and personalize it to their preference. The MPEG-H Audio metadata carries all the information needed for personalization, such as attenuating or increasing the level of objects, muting them, or changing their spatial position. The metadata also contains information to control and restrict the personalization options, including setting the limits within which the user can interact with the content.

Object-based audio delivery together with metadata is the basis for enabling features like DE. The dialog signal is encoded and delivered as a separate audio element within the audio bitstream. In the most basic example, all other audio content is mixed into a second audio element, i.e., the background signal, sometimes also referred to as “channel bed”.

In the receiving device, the dialog element can be boosted for better intelligibility. This can be done automatically in the device based on user preferences, or the user can adjust the dialog enhancement manually via a user interface. An example user interface is shown in Fig. 1. The level range (i.e., the minimum and maximum levels) are carried in the metadata. This maximum value for the lower and upper end of the scale can be set differently for different content. Additionally, the MPEG-H Audio system automatically compensates for loudness changes that result from user interaction (e.g., dialog boost) to keep the overall loudness on the same level. This ensures a constant loudness level not only over programs but also after user interactions.

Another example for personalization is an advanced Video Descriptive Service (VDS, also known as Audio Description, AD). For this use case MPEG-H Audio allows to carry the video description object in several languages for user selection and additionally enables the user to spatially move the video description object to a user selected position (e.g., to the left or right), enabling a spatial separation of main dialog and video description. Further examples in sports events are different commentaries that the user can select from, like one commentary for the home team and another one for the away team, or additional audio elements such as the team radio in car racing.

III. Related Works

The origins of the method of adjustment are reviewed in Section III-A. The application of this method to the investigation of subjective preferences in music mixing is discussed in Section III-B. Section III-C reports on works about preferences in TV mixes, while works evaluating DE systems are reviewed in Section III-D. Final comments are made in Section III-E.

A. The Method of Adjustment

The story of perceptual experiments via the method of adjustment starts when Gustav Fechner, the founder of psychophysics, borrowed the method from astronomy [17]. Thanks to this, Fechner was able to formulate Weber’s law and introduce the concept of Just-Noticeable Difference (JND) [18]. In an experiment with the classic form of the

method of adjustment, subjects are asked to adjust a physical value in such a way that a perceptual quantity matches a given reference stimulus. For instance, in [19], subjects adjust the level of amplitude modulated sounds until they sound as loud as a reference unmodulated sound. Many other experiments of this kind are mentioned in [20]. The method of adjustment greatly contributed to our understanding of perception of sound (and so to the development of advanced audio technologies such as transparent lossy coding), but it was applied to other fields as well, such as the perception of light and weight, to name a few.

This method was also used without an explicitly given reference. Instead, subjects are asked to match their subjective internal reference so that the experimenter can elicit an abstract threshold, e.g., for the acceptability of sound isolation for music recording [21] or for the acceptability of ambient noise [22].

Personal preferences can also be studied in a similar way by asking subjects to adjust a control parameter and set it so as to match the preferred value, given a specific task. This approach was used in many recent works, especially in the field of music production and consumption as discussed in the following.

B. Preferences in Music Mixes

Understanding the preferences of audio engineers while mixing music is the topic of much research with the goal of designing Automatic Mixing systems [23], [24], and references therein. These works aim to understand the "optimum" properties of a music mix, e.g., track level balancing, spatial characteristics (panning), equalization, dynamic range compression, and artificial reverberation. The dataset presented in [23] is produced by having numerous audio engineers mix the given music material with a limited, yet rich set of software tools. This can be seen as a generalized adjustment method, where many control parameters are to be adjusted at the same time. After this, the audio engineers are asked to rate and comment on the versions created by their own and by other engineers. Expert audio engineers preferences are consistent among different educational institutions [23] and exhibit low variation over time [25].

Other works have focused on the preferences of music consumers, e.g., by studying the aesthetically preferred level of reverberation: a detailed literature review can be found in [26], where works using methods of adjustment are reviewed, e.g., [27].

In [28], the relative level of the vocals and the instruments preferred by cochlear implant users is studied with a method of adjustment using the original tracks, followed by pairwise comparisons of a few fixed vocals-to-instruments ratios. High variability among users is shown. In general, cochlear implant users prefer vocals-to-instruments ratios up to 12 dB higher than people with normal hearing. In [29], a similar study is carried out, this time using BSS. This is shown to be successfully applied in order to create modified music mixes, which can be better enjoyed by cochlear implant users, even though the original tracks are not available.

Moreover, the works in [28], [29] are useful to understand that personalization has a potential benefit also in music consumption, by helping people enjoy music even if in disadvantageous listening conditions.

C. Preferences in TV Audio Mixes

While several aspects of music production have been studied in literature, little research focuses on preferences in TV audio mixes. The works in [30], [31] carry out attempts to understand the optimum level of the dialog in TV material. In [30], subjects are asked to rate via a questionnaire the level of listening fatigue and the pleasantness of TV excerpts where the dialog was mixed with a loudness difference of 2, 7, or 10 Loudness Units (LU)[1] with the background. The experimenters attempt to show that the tonmeisters' best practice of mixing with a loudness difference between dialog and background of 7-10 LU is somehow valid. In [31], the focus is on voice-over-voice passages, e.g., when an external voice translates an interview in a foreign language. In order to investigate this, subjects are asked to adjust the voice-over-voice ratio and to set it to their preferred level by means of a slider. The given recommendation is 16-23 LU level difference between the voices of the two talkers. However, high variation among items and listeners is reported, e.g., the disagreement between groups of people can be as big as 10 LU for the same test item.

The optimum level ratio between foreground speech and all other sound sources is indeed a choice that depends on personal and contextual factors. The best understood ones are:

- Listener's hearing acuity [34]–[36];
- Listening environment, e.g., environmental noise [37];
- Reproduction system, e.g., type of TV set [38];
- Listener's mother tongue and content language [39];
- Personal taste [1], [40].

It follows that a unique *one-size-fits-all* mix can hardly satisfy the needs of the audience in all cases. This is also indicated by the increasing number of complaints about the difficulty in understanding what is said in the broadcast material, with too loud background sounds being the major cause of them [41]. This issue is addressed by DE applications.

D. Evaluation of Dialogue Enhancement Systems

DE or similar applications are evaluated in [5]–[7], [34]–[37], [40], [42]. A set of objective measures for DE was proposed in [42], focusing on the case in which BSS is used to estimate dialog and background and on the distortions that this can introduce. Even if objective measures can be of great help during the development phase, they cannot answer the research questions RQ1 and RQ2.

User adjustment is the cornerstone of the subjective tests employed in [36], [37], [40]. These works analyze:

- The preferred level of commentary over court ambience during sport events [40], showing a bivariate distribution;
- The preferred level of dialog over other audio objects categories with hearing impaired listeners [36];

[1]LU is herein meant as per BS.1770 [32] and so as in EBU R 128 [33].

- The preferred level of dialog over typical TV backgrounds with listeners in a noisy environment [37].

These works focus on RQ1, but do not address RQ2 as they do not quantify the impact of the personalized levels, e.g., on a satisfaction scale.

On the other hand, DE systems are subjectively assessed by having listeners assess pre-determined mixes with fixed dialog-to-background ratios, i.e., without addressing RQ1 in [5]–[7], [34]. BSS is also considered in [5]–[7]. The listeners are asked to rate the different mixes and to base their judgement on various criteria such as overall sound quality and speech clarity. A blind forced-choice AB comparison is adopted in [6], [34], while in [5], [7] direct scaling is used, i.e., the subjects are asked to convert the sensation formed from the comparison of multiple stimuli into a sensory magnitude and report it on a scale. Finally, DE is shown to significantly improve speech intelligibility for hearing-impaired subjects in [35], where three different pre-determined mixes are compared.

E. Comments

The method of adjustment was successfully applied in order to study personal preferences and it seems like the natural choice for answering our research question RQ1. However, our main focus is not on finding the optimum values of one or more parameters, such as in the works reviewed in Section III-B. Instead, we use the method of adjustment to study the QoE of users provided with the possibility of personalization, as pointed out by RQ2. For this reason, in the design of the A/ST, an adjustment phase is complemented by a phase focusing on user satisfaction, which is a novelty with respect to the related works. Moreover, the reviewed works adopt heterogeneous tests, some of them are designed ad hoc, and none of them employs one of the numerous tests standardized or recommended by international organizations. Guidance through these standards is given in [43]. Neither were we able to find a suitable method among these standards to address our research questions. These considerations motivated us to design the A/ST.

IV. The Adjustment / Satisfaction Test

We now introduce the A/ST for the subjective evaluation of user-adjustable systems.

Let S be a system that can be personalized via the set of parameters p that is controlled by the user, e.g., via rotating knobs, a remote control, or similar devices. Let us evaluate S via the A/ST. An introductory phase (Phase 0) is followed by adjustment (Phase 1) and satisfaction assessment (Phase 2).

A. Phase 0: Explaining Envisioned Usage

First, the envisioned usage scenario and the goal of the personalization are described to the participants. These concepts have to be very clear to the users, as they define their expectations and thus their satisfaction. In fact, QoE is defined as resulting from the fulfillment of user expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state [44].

Hence, it is also important that the test environment reproduces the main characteristics of the envisioned usage environment and that the test material is representative of the application.

In this introductory phase, it is also explained how to operate the interface. In order to minimize the risk of a poor comprehension of the task, written instructions are given to the participants, they operate the test with a training item, and any doubt that may rise is verbally clarified.

B. Phase 1: Adjustment

The adjustment phase addresses RQ1. During this phase, the test participants interact in real time with S by adjusting p. The goal is to find the preferred p according to the criteria introduced in Phase 0.

As an example, Fig. 2 shows the user interface that we use. No visual feedback of the current status of p is provided, except an indicator that the end of the allowed range has been reached. Also, the adjustment steps are not perceivable while operating the knob(s) by which the user controls p.

During the personalization, it is possible to compare the adjusted settings with a default setting p_0 by instantaneously switching between the two versions using a button. This p_0 is also included in the range of p. The possibility of comparing against p_0 is important for two reasons. First, it prevents the frustration a user may experience for small adjustment steps. Second, default values for p can help undecided users: if the user likes p_0, she/he is encouraged to find p_0 or similar values in the available range of p; if the user does not like it, she/he is stimulated to find a different settings for p.

Changing the settings of p produces physically different outputs. If the physical differences are perceptually relevant, they may or may not result in differences in terms of user satisfaction. Phase 1 studies if and how p is adjusted. This is complemented by assessing if and how the adjustment of p affects the user QoE, as done by Phase 2.

C. Phase 2: Satisfaction Assessment

Phase 2 aims to assess the user satisfaction resulting from the adjustment of p, i.e., investigates RQ2. The participants are asked to rate the difference in satisfaction between p_0 and the chosen p by means of a provided labeled scale. The Comparison Category Rating scale is used for this purpose [45]. The points and labels of this scale are displayed by the user interface shown in Fig. 3.

This test design provides a post-screening criterion: the satisfaction experienced with the chosen p cannot be worse than the one with p_0. If p_0 is preferable, this should be selected in Phase 1. Hence, satisfaction levels lower than "The same as" reveal low reliability of the participant or the task being misunderstood. Subjects that select "Worse" or "Much worse" are excluded from the analysis of the results. We decide to accept "Slightly worse" because even if the selected p violates the test assumptions, it is likely to be close to what the participant actually prefers (more on this in Section VII-B).

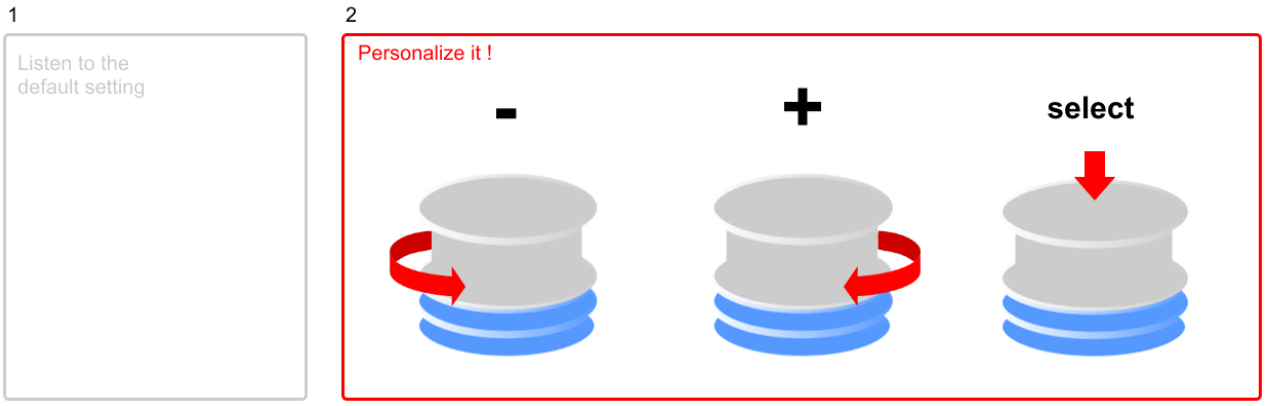

Fig. 2. A user interface for the adjustment phase of the A/ST.

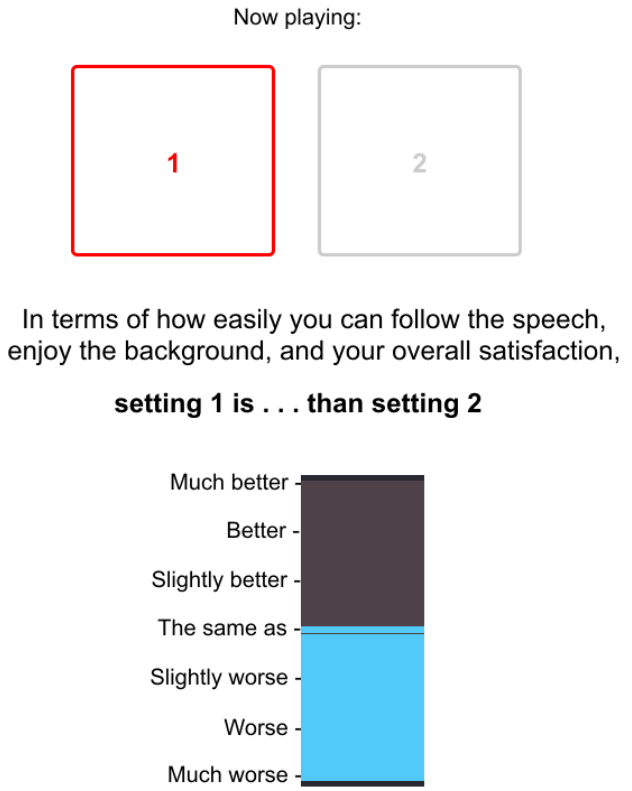

Fig. 3. A user interface for the satisfaction assessment phase of the A/ST.

D. Experience Configuration and Unlabelled Configuration

Two different configurations of the test are implemented and compared; they are referred to as Experience Configuration and Unlabelled Configuration. The *Experience Configuration* has the following features:

- For each test item, Phase 2 takes place right after Phase 1 and they are repeated iteratively;
- During Phase 2, the test participant is explicitly informed by the graphical interface about which is the default setting and which is the setting that he/she selected in the previous phase.

On the other hand, the *Unlabelled Configuration* has the following features:

- Phase 2 takes place after Phase 1 has been completed for all the test items (with a small break between the phases);
- During Phase 2, the true labels for the default setting and the selected setting are hidden below the labels "1" and "2" and their assignment is randomized for each new comparison.

As the results will suggest, the procedure of the Experience Configuration clearly biases the participants towards a "correct" or more positive answer in favor of the setting that he/she selected. Yet, this configuration is very close to the final application, as these biases are present also in it, e.g., Fig. 1. Hence, the overall experience is evaluated by this configuration, taking into consideration also the satisfaction that can come from having the chance to personalize the system, which is a relevant part of the user experience. In [1], only this configuration is employed.

The goal of the Unlabelled Configuration is to study what happens when these biases are not present. To this end, the satisfaction assessment phase is separated in time and consists of a blind comparison. Hence, the focus is moved from the overall QoE to the actual adjustment preferences.

V. The A/ST for Dialogue Enhancement

The goal of the adjustment of a DE system is to find an enjoyable mix, where the dialog can be easily followed. To this end, p consists of one control parameter, which adjusts the Loudness Difference (LD) between the loudness of the dialog and that of the background, where the loudness is calculated as per BS.1770 [32] and measured in Loudness Units (LU). All outputs have equal integrated loudness.

In the following, the standard DE system that has access to the original objects (OO) is referred to as S_{OO}. The Experience Configuration of the A/ST is also applied for the assessment of a DE system that estimates the audio objects from their stereo mixture by a BSS algorithm. This is referred to as S_{BSS} and may introduce distortions such as artifacts or changes in timbre for strong modifications of the LD. The default mixes (p_0) are used as inputs to BSS.

The subjects are first asked to imagine being home and watching television for a long time. During Phase 0, they are asked to adjust the overall volume.

The core text of the instructions is as follows: *"You are going to listen to audio items that contain speech that may be difficult or tiring to understand. If this is the case, you want to change the audio so that you can easily follow the speech, yet keeping the rest of the content (e.g., background music) enjoyable. To this end, you can adjust the relative level of the speech by means of the provided knob. Please note that the speech adjustment process may cause a degradation in quality. If this happens, please find the best compromise between the level of the speech that you would like and a sound quality that you would accept in television. The graphic interface (Fig. 2) shows visual feedback (not shown in Fig. 2) while you are operating the knob: a blue frame around one of the turn knob icons indicates that the audio is changed according to the direction of rotation; a red frame around the icon indicates that the audio cannot be modified further in that direction. You can switch between the personalized setting and the default setting by pressing 1 and 2 on the keyboard. When you find the parameter setting that allows you to follow the speech easily, yet keeping the rest of the content enjoyable, please select it by pressing the knob from the top. During a second phase, the window in Fig. 3 will display a question about your satisfaction with the selected setting. Also here, you can compare the selected setting with the default setting.*

In the case of the Unlabelled Configuration, the instructions also explain the following: *"You can switch between the two settings by pressing 1 and 2. However, in this second phase, the audio settings are randomly assigned to 1 and 2 for each comparison. So, please read carefully the displayed question (you always rate 1 with respect to 2) and listen carefully."* On the other hand, in the Experience Configuration, the labels "1" and "2" in Fig. 3 are replaced by the explicit labels "selected setting" and "default setting".

Note that the participants are asked to jointly consider the ease of listening to the dialog and the enjoyment of the background. There are cases where these two goals diverge [34]. In these cases, the preferred trade-off has to be found.

VI. Independent variables of the A/STs

We run two A/STs in two different sessions, once in the Experience Configuration and once in the Unlabelled Configuration. For both of them we use same listening environment, test items, default mixes, and type of listeners. These shared independent variables are explained in Section VI-A. Other independent variables (e.g., the number of participants and the adjustment range) are different for the two sessions, as described in Sec. VI-B and VI-C.

A. Shared independent variables

Listening environment. The experiment is carried out in a listening room that resembles a quiet low-reverberant living room. Other listening environments could be simulated in future works. Stereo signals are reproduced over two Genelec 8250A studio monitors, which are positioned approximately at the height of the listener's head, 2 meters away from her/him, and $\pm 30^\circ$ from her/his looking direction. The user interface is displayed on a TV positioned between the loudspeakers. The participants sit on a chair with fixed position, and the knob and the keyboard controlling the interface are on a little table nearby.

Item	LD_0 [LU]	Type of Background
AR2_de	1.03	Cheering crowd
AR4_en	2.99	Instrumental jazz-rock music
AR1_de	4.03	Train station hall noise
TV3_en	-0.74	Instrumental classical music
AR3_de	6.00	Rain and distant thunders
Mean	2.66	

TABLE I
LD_0 VALUES CORRESPONDING TO p_0 AND TYPES OF BACKGROUND FOR THE MIXES FOR WHICH THE ORIGINAL AUDIO OBJECTS ARE AVAILABLE.

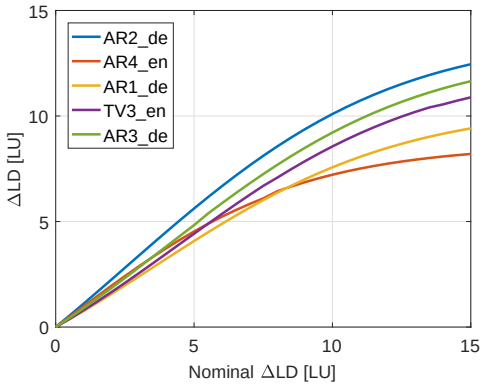

Fig. 4. When S_{BSS} is used, the nominal ΔLD roughly estimates the Loudness Difference (LD) between the dialog and the background. In this plot, the relationship between the nominal ΔLD and the actual ΔLD is shown for the mixes for which the original audio objects are available.

Test items. As test items we use material that was broadcast in Germany or in the UK as well as artificially created mixes. We employ 13 test signals. One of them is used as the training item and it will not be shown in the results. Five of the remaining 12 items are presented twice, once with S_{OO} and once with S_{BSS}. It follows that S_{OO} is tested on 12 items, while S_{BSS} only on 5 of them. The repetitions of one item are not presented one after the other, but interleaved with other items. Sampling frequency is 48 kHz. The length of the items varies between 8 and 17 seconds and the playback loops over the entire duration until the subject decides to proceed to the next item. The stereo backgrounds of the items comprise music (classical, ambient, jazz-rock, and pop) and environmental recordings (rain, sea waves, cheering crowd, train station hall, and construction site). The dialog is panned to the center and features German and English language, male and female speakers. The accompanying video for this material is not shown, as its quality can influence the perception of audio quality [46].

The item names are composed as follows. The name starts with "TV" for the real broadcast content, while it starts with "AR" for the artificially created mixes. A numerical ID and an underscore follow. Finally, the language of the content is indicated by "en" for English or by "de" for German.

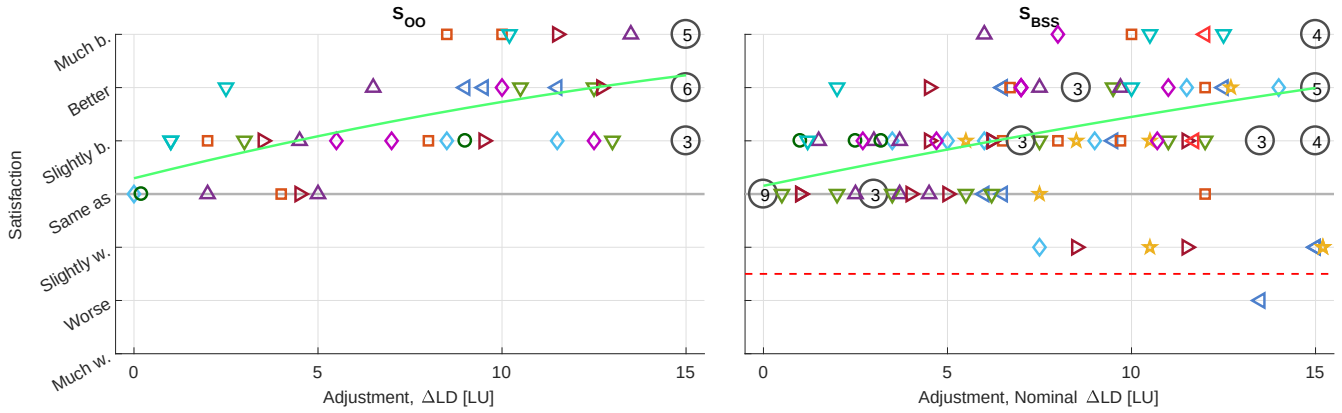

Fig. 5. Experience Configuration. Satisfaction levels as a function of the selected adjustment while using S_{OO} (left plot) and S_{BSS} (right plot). Each of the 11 listeners is represented by a unique combination of symbol and color. The red dashed line depicts the post-screening threshold. Dark gray circles are used when many different subjects overlap; the number of overlapping listeners is printed inside the circle. The first-order polynomial optimally fitting the data in the least-squares sense is depicted in light green. All data points after post-screening are used for the line fitting.

Default mixes (p_0). The original broadcast signals are used as default mixes corresponding to p_0. For some of these signals, the broadcasters received complaints from the audience because of the loud background. Other real-world broadcast signals were selected so to have a similarly low LD. The artificially created mixes were produced imitating these cases. Table I reports the default LD (LD_0) for the artificially created mixes and for the real broadcast item for which the original audio objects are available. The default mixes are also used as staring point of the adjustment in Phase 1. All items are normalized to have equal integrated loudness [32] both in their default and adjusted versions.

An additional signal (AR5_de) with $LD_0 \gg 0$ LU is also presented in the test. AR5_de consists of the same dialog and background signals in AR2_de, but they are mixed with a LD of 18 LU.

In the following, the difference between a modified LD and the LD_0 is referred to as ΔLD.

When S_{BSS} is employed, only a rough estimate of the ΔLD is available, referred to as *nominal* ΔLD. The actual ΔLD depends on the performance of the BSS, which is item-dependent. We can calculate the actual ΔLD, also in the case with S_{BSS}, only if the original audio objects are available. If this is the case, the mix modified by S_{BSS} is decomposed into dialog component, background component, and artifacts with the help of the BSS Eval toolbox [47]. The LD is then calculated between the dialog and the background components. Fig. 4 shows the relationship between nominal ΔLD and the actual ΔLD for the mixes for which the original audio objects are available but unknown to S_{BSS} for evaluation purposes.

Subjects. All the participants have normal hearing and are voluntary, remunerated, non-expert, initiated[2], and (mostly German) university students of different disciplines. Based on interviews, we can divide them into two groups. Subjects that claimed to be passionate about Hi-Fi, music, or audio/video production are referred to as *Hi-Fi lovers*. On the other hand, we adopt the term *naive listeners* for participants that do not claim any particular interest in audio besides regularly using the main platforms for music or film streaming.

The number of subjects for the two test sessions vary. Six listeners participated in both sessions with an interval of 7 months between them.

[2]A person who has already taken part in a sensory test is referred to as *initiated*.

B. Session with Experience Configuration

Subjects. The session involves 11 participants, between 19 and 32 years old (median age is 25). Six of them are *Hi-Fi lovers* and the other five are *naive listeners*.

Available range for p. Values of ΔLD ranging from 0 to $+15$ LU are used for S_{OO}, with steps of 0.5 LU. The same range is used for the nominal ΔLD of S_{BSS}.

Implementation. The test software is implemented in Max/MSP. It is made available for general non-commercial use at `https://www.audiolabs-erlangen.de/resources/2017-AES-AST`.

C. Session with Unlabelled Configuration

Subjects. This session involves 21 participants, between 20 and 32 years old (median age is 23). Seven of them are *naive listeners*.

Available range for p. Values of ΔLD ranging from -10 to $+20$ LU are used for S_{OO}, with steps of 0.8 LU.

Intermediate satisfaction levels. Four additional unlabeled levels are present between each labeled satisfaction level.

VII. Results

The results obtained with the Experience Configuration are given in Section VII-A; Section VII-B elaborates on the effect of the available adjustment range; Section VII-C reports on the test run with the Unlabelled Configuration.

A. Session with Experience Configuration

Fig. 5 depicts the selected level of satisfaction as a function of the adjustment levels. Here, each listener is represented by a different combination of symbol and color. The dashed

red line depicts the post-screening threshold as introduced in Section IV-C. One naive listener is excluded from the analysis of the results, as she/he selected "Worse" once in the satisfaction assessment phase for S_{BSS}. The detailed results for the remaining ten subjects are given in the Appendix (Figs. 13 and 14). Fig. 5 also shows the first-order polynomials optimally fitting the data in the least-squares sense. The data points after post-screening are used for the line fitting. The resulting lines are here defined only in the studied range ΔLD $= [0, +15]$ LU. Assuming that "Much worse" corresponds to a satisfaction level of -3 and that each satisfaction label is separated by 1 satisfaction unit, the resulting line for S_{OO} is the following: $satisfaction = 0.123\Delta\text{LD} + 0.427$. While for S_{BSS}: $satisfaction = 0.121\Delta\text{LD} + 0.207$. Interestingly enough, a small gain in satisfaction is present even for $\Delta\text{LD} = 0$ LU. This is probably due to the bias introduced by the limited allowed range of adjustment, see Section VII-B.

Fig. 6 depicts the mean of the listeners' adjustments and satisfaction levels for S_{BSS}, together with box plots[3]. A clear correlation between the mean levels of ΔLD and satisfaction can be observed (Pearson's $r = 0.81$), meaning that the adjustment has a noticeable and positive effect.

Furthermore, Fig. 7 compares the selections for the items presented with both S_{OO} and S_{BSS}. Also in this case, we can observe a clear correlation in the adjustments selected with S_{OO} and S_{BSS} ($r = 0.94$) and in the satisfaction they provide ($r = 0.996$). Still, lower levels of ΔLD are preferred for S_{BSS}, on average 3 LU lower than S_{OO}, resulting in lower satisfaction. As confirmed by interviewing the participants, this is due to the fact that the subjects have to trade-off between the desired ΔLD (selected while operating S_{OO}) and the distortions, which S_{BSS} introduces for high values of ΔLD.

Fig. 8 compares the adjustments for the items presented for both S_{OO} and S_{BSS}, expressing them in absolute LD. The first and third quantile for the mean item adjusted via S_{OO} cover the range 9-15.5 LU. Yet, high variation among items is observed, ranging from 3 to 21 LU. Moreover, the limited allowed range of adjustment has an effect on these values, as detailed in Section VII-B.

Throughout Figs. 5 – 8, high subjective variance is evident. This shows that subjects have very different preferences for the relative levels of dialog and background. It follows that a unique *one-size-fits-all* mix would hardly satisfy all listeners. Hence, the personalization offered by DE is desired, even by subjects with normal hearing in quiet and controlled listening conditions. Personal taste is likely to be the main reason behind this discovery.

In conclusion, personalization is extensively used and translates into clearly increased satisfaction not only for the item processed by S_{OO}, but also for the items processed by S_{BSS}. This suggests that both DE systems offer a useful service, despite the artifacts or change in timbre potentially introduced by S_{BSS}.

[3]A box plot is a compact way of visualizing the distribution of data points. Here the box is depicted vertically and hourglass-shaped. Its lower end corresponds to the first quartile $Q1$, the central bar corresponds to the median (in this paper always in black), and the upper end corresponds to the third quartile $Q3$. Hence, the hight of the box corresponds to the Interquantile Range $IQR = Q3 - Q1$. Vertical lines (often referred to as whiskers) extend from the box indicating the variability outside the upper and lower quartiles; they are concluded with horizontal bars positioned at the maximum or minimum point within $1.5IQR$. Points outside the whiskers range are displayed with a cross if they are between 1.5 and 3 times the IQR and with a circle if they are outside 3 times the IQR.

B. On the effect of the available range

The results presented for the Experience Configuration show clusters of adjustments around the maximum of the available range, i.e., $\Delta\text{LD} = +15$ LU. This can be observed in Fig. 5 and in the asymmetry of many of the box plots in Figs. 6 – 8 (and more in detail in the Appendix, Fig. 13 and 14). This phenomenon is a consequence of the limited provided range $[0, +15]$ LU and especially due to the fact that one extreme of the range, i.e., $\Delta\text{LD} = 0$ LU corresponds with p_0, i.e., the starting point of the adjustment. To prove this statement, we had six listeners repeat the adjustment phase after 7 months. They could operate S_{OO} in exactly the same conditions but with a larger range available allowing also decreases of the level of the dialog: $\Delta\text{LD} = [-10, +20]$ LU is the range used with steps of 0.8 LU. The starting point of the adjustment is still the same p_0, i.e., $\Delta\text{LD} = 0$ LU.

Fig. 9 compares the preferred adjustment levels for the two cases. With $\Delta\text{LD} = [-10, +20]$ LU, the vast majority of the selections lies far from the extremes and $\Delta\text{LD} = +15$ LU is not selected in any case.

When the range $\Delta\text{LD} = [0, +15]$ LU is allowed, there is a clear tendency towards higher levels of ΔLD. This tendency is then *clipped* at the maximum $\Delta\text{LD} = +15$ LU. A possible explanation for this is that listeners generally like the fact that they can personalize the mix, even if it was explained that an unmodified mix, i.e., $\Delta\text{LD} = 0$ LU was a legitimate choice. Hence, they prefer to modify the mix, even if only with small adjustments. Any of these adjustments is directed towards positive ΔLD values, as they are the only available. On the other hand, this effect tends to be averaged out if negative ΔLD values are also available. In fact, small adjustments not caused by a strong preference but by the liking of the chance of personalization can now go towards both negative and positive ΔLD values.

Allowing only positive ΔLD can be closer to a final application where the content-provider allows only dialog enhancement and not its suppression, but it introduces a bias towards higher levels of adjustment. On the other hand, having also negative ΔLD in the test can tell us more about the relative levels of dialog and background actually preferred by the listeners.

C. Session with Unlabelled Configuration

As the Unlabelled Configuration intends to study the level preferences more closely, negative values of ΔLD are made available during it. The initial point for the adjustment remains p_0. In this session we also allow listeners to select unlabeled intermediate satisfaction levels. The aim is to reduce the *quantization noise* that we can observe in the results of the satisfaction assessment during the Experience Configuration, Fig. 5.

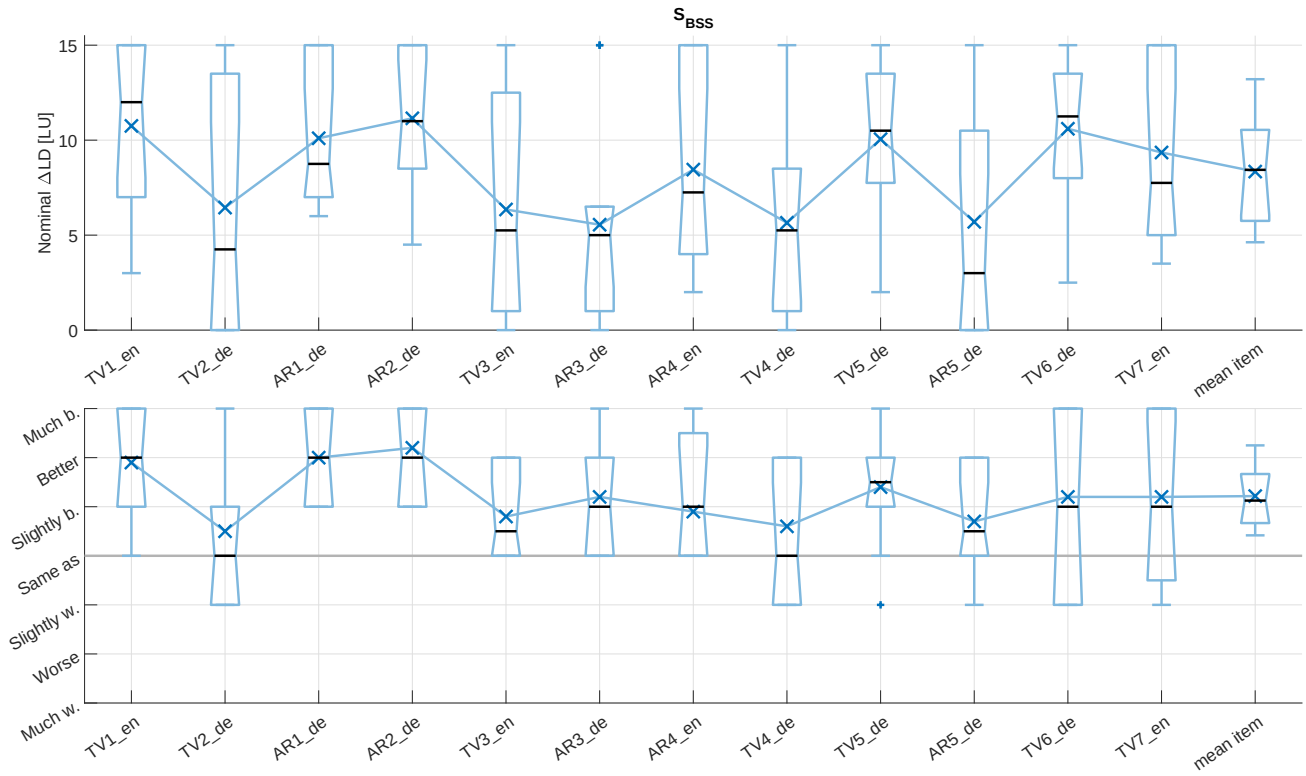

Fig. 6. Experience Configuration. Mean selections (main blue lines with crosses) and box plots for the preferred nominal ΔLD (upper plot) and resulting satisfaction levels (lower plot). Data recorded for S_{BSS}, i.e., Dialogue Enhancement employs blind source separation. The ordering of the items from left to right reflects the order in which they were presented in the test.

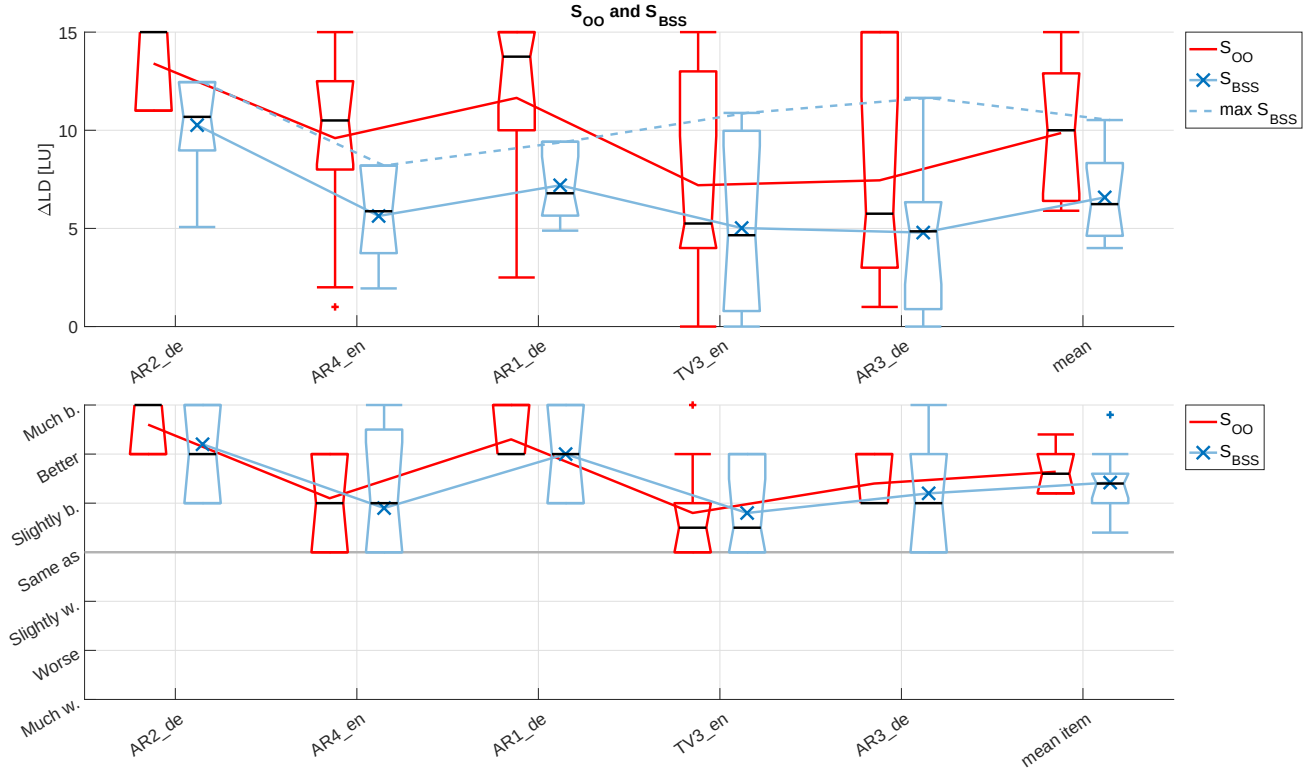

Fig. 7. Experience Configuration. Mean selections and box plots for the preferred ΔLD (upper plot) and resulting satisfaction levels (lower plot). S_{OO} (main red lines without crosses) is compared with S_{BSS} (main blue lines with crosses). The maximum ΔLD achievable by S_{BSS} is also depicted (dashed line).

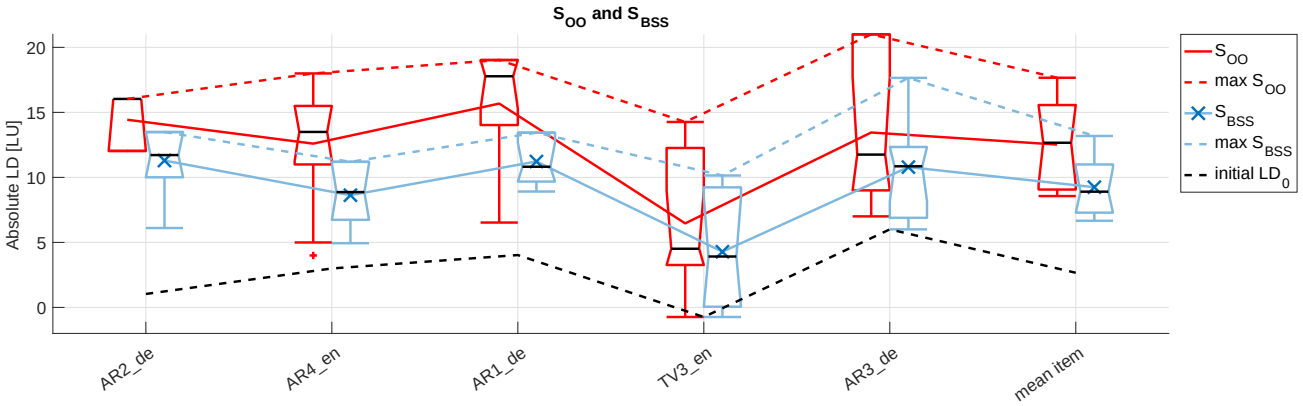

Fig. 8. Experience Configuration, 10 subjects. Mean selections and box plots for the adjustment levels expressed in terms of absolute LD. S_{OO} (main red lines without crosses) is compared with S_{BSS} (main blue lines with crosses). Maximum and minimum available absolute LDs are also shown (dashed lines).

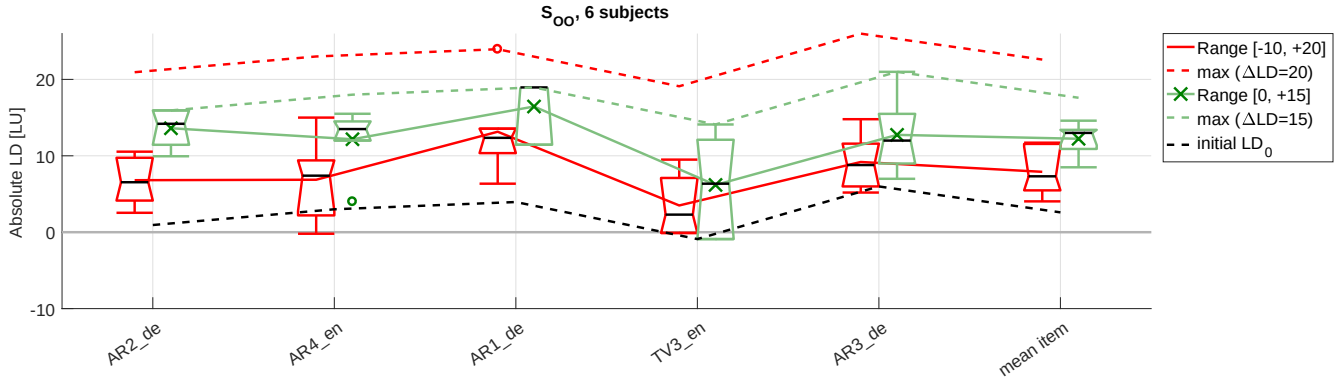

Fig. 9. Experience Configuration. Mean selections (main solid lines) and box plots for the adjustment levels preferred by the same six listeners with S_{OO} providing different adjustment ranges.

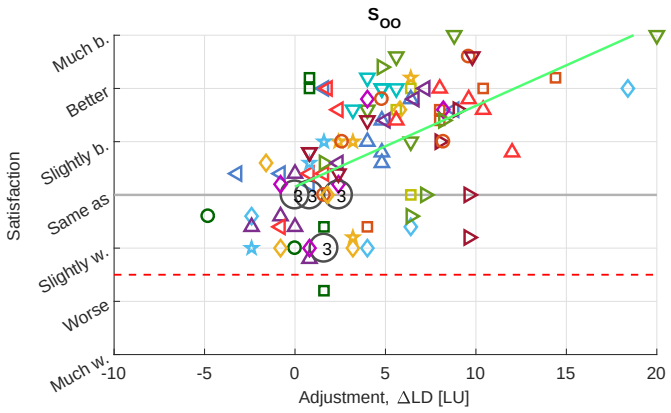

Fig. 10. Unlabelled Configuration. Satisfaction levels as a function of the selected adjustment while operating S_{OO}. Each of the 21 listeners is represented by a unique combination of symbol and color. The red dashed line depicts the post-screening threshold. Dark gray circles are used when many different subjects overlap; the number of overlapping listeners is printed inside the circle. The first-order polynomial optimally fitting the data in the least-squares sense is depicted in light green.

Fig. 10 depicts the selected level of satisfaction as a function of the adjustment levels. The results for the satisfaction assessment phase are presented so that the shown level would answer the question "The selected setting is ... than the default setting", although the signals were not always rated in this order and the listener was not explicitly informed about which signal corresponded to the selected or to the default setting. Again, one naive listener is excluded from the analysis of the results according to our post-screening criterion. The detailed results for the remaining twenty subjects are given in the Appendix (Fig. 15).

The first-order polynomial optimally fitting the data in the least-squares sense is also shown in Fig. 10: $satisfaction = 0.152\Delta\text{LD} + 0.153$. This does not hold for $\Delta\text{LD} < 0$, as we do not have enough data points to fit a meaningful trend there.

A first difference with the Experience Configuration is that satisfaction levels below "The same as" can be observed even when no source separation artifacts are present. This is particularly the case for small adjustments, i.e., $\Delta\text{LD} = [-3, +3]$ LU, resulting in satisfaction levels between "Slightly worse" and "Slightly better". A possible explanation for this comprises the following factors, which also emerged while interviewing the listeners:

- We focus on common users and we do not select the listeners to be *reliable*, e.g., as defined in [48], i.e., able to repeat themselves on the set of evaluated stimuli. If the adjustment is perceivable but small, the listeners might

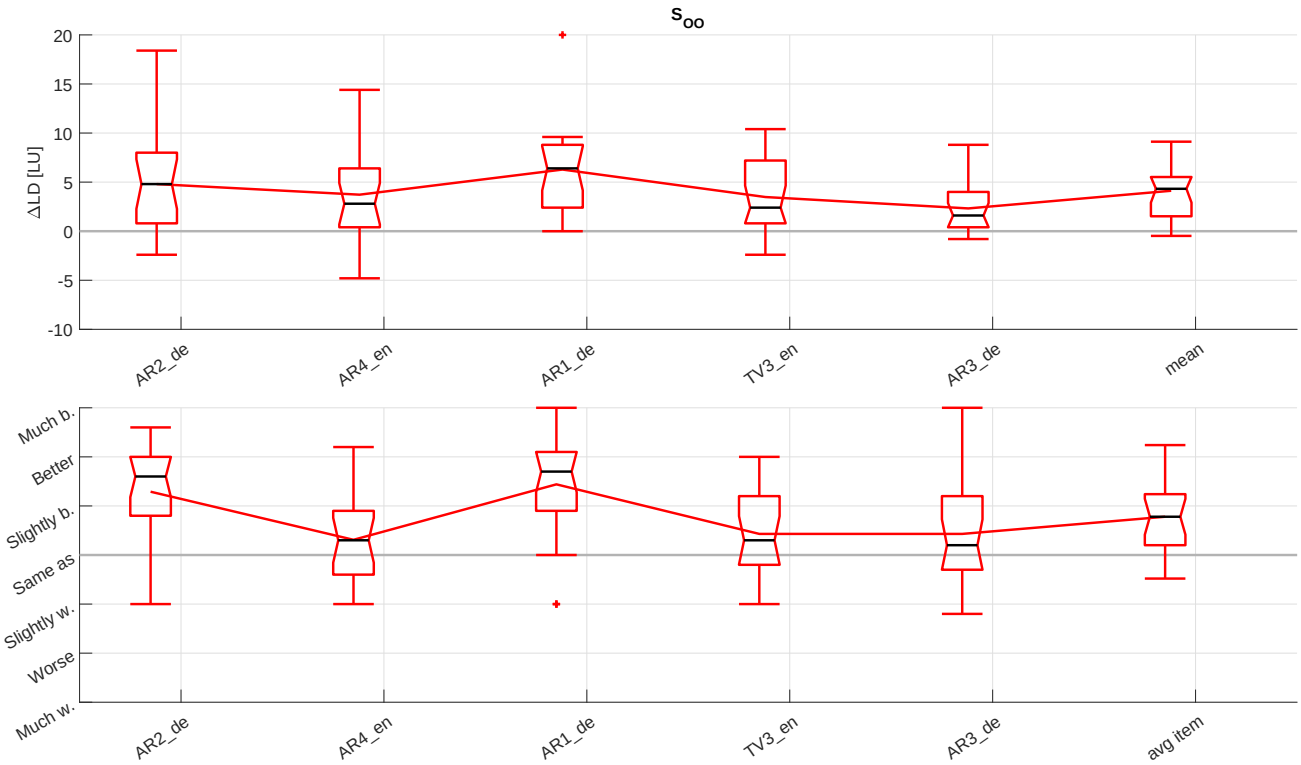

Fig. 11. Unlabelled Configuration, 20 subjects. Mean selections (main solid lines) and box plots for the preferred ΔLD and resulting satisfaction levels with S_{OO}. Available adjustment range is $[-10, 20]$.

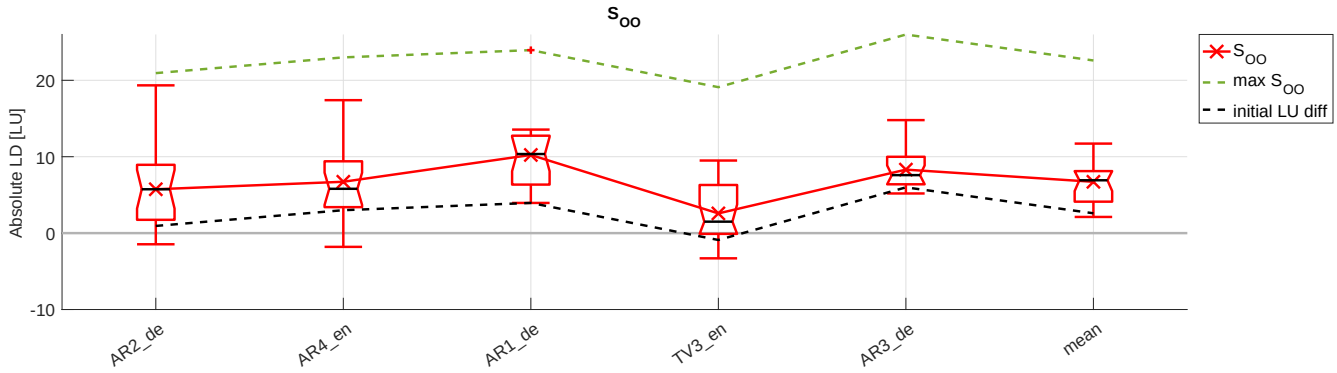

Fig. 12. Unlabelled Configuration, 20 subjects. Mean selections (main solid red line) and box plots for the adjustment levels expressed in terms of absolute LD with S_{OO}. The maximum and minimum available absolute LD are also depicted (dashed lines).

not have a strong preference, causing an almost random assessment between "Slightly worse" and "Slightly better" in a blind assessment. This might be emphasized if some time passes between the adjustment and the blind satisfaction assessment: in our case about 20 minutes passed on average.

- There is no difference that the listener can actually hear, e.g., $\Delta\text{LD} \approx 0$ LU. Nevertheless, the subject thinks there should be one and that "The same as" is not a useful answer and he/she randomly *slightly* prefers one.
- The subject misunderstands the given task and thinks that different criteria apply during the adjustment and during the assessment phase.

This can be considered measurement noise and its effect can be averaged out if enough data points are collected: a small adjustment randomly rated as "Slightly better" corresponds to a small adjustment randomly rated as "Slightly worse" giving a zero adjustment on average rated as "The same as".

Fig. 11 depicts the average selections and box plots. Similarly to the Experience Configuration, high correlation can be noticed between the average adjustment and the average satisfaction ($r = 0.87$).

Finally, high variance among subjects and items can be observed. Fig. 12 shows the absolute LDs. The LDs for the first and third quartile are 4 and 8 LU. However, these vary between 0 and 13 LU if single items are considered and between -3

and 24 LU if single subjects are considered.

Also from the data collected through this configuration we can conclude that DE is a useful broadcast service, as personalization is extensively used and results in increased satisfaction.

VIII. Conclusion

The Adjustment / Satisfaction Test (A/ST) was recently proposed for the assessment of user-adjustable systems. In this paper, two different configurations of this test paradigm were successfully applied to the evaluation of Dialogue Enhancement (DE) for broadcasting. Thanks to the A/ST, it was possible to note that the personalization offered by DE is extensively used and results in increased user satisfaction. This confirms that there is a clear benefit introduced by DE for the Quality of Experience (QoE) in broadcast services.

Throughout both configurations it was possible to observe a very high variance among listeners and items in terms of the preferred relative levels of dialog and background. This can be explained as a consequence of personal taste, as a relatively homogeneous group of normal hearing subjects took part in the tests in quiet and controlled listening conditions.

The Unlabelled Configuration of the A/ST gave us a better understanding of the range of the preferred Loudness Differences (LDs) between the dialog and the background. Most of the preferred LDs lie between 0 and 13 LU and the ones for a mean item are between 4 and 8 LU.

Still, the Experience Configuration of the A/ST is less noisy and more conclusive about the final QoE. In this configuration we also tested a DE system, where the original objects are not available, but they are estimated via Blind Source Separation (BSS). Also in this case, the personalization enabled by BSS was extensively used and resulted in increased satisfaction, albeit with a gap of about 3 LU with respect to the standard DE system.

In future works we intend to involve a larger number of users with a wider age interval, possibly including listeners with age-related hearing loss. In order to easily reach out to more listeners, a web-based version of the A/ST is currently under development. Furthermore, in order to better understand the preferred relative levels of dialog and background, a broader range of signals should be included in future.

Even if designed with audio in mind, the rationale of the A/ST is more general and can potentially find application also with other media in broadcasting.

Appendix
Detailed Results

For the adjustment phase as well as for the satisfaction assessment phase, the collected data points lie in a space with three dimensions: items, listeners, and selections (adjustment or satisfaction level). The complete post-screened data for the Experience Configuration is shown by Fig. 13 (adjustment phase) and Fig. 14 (satisfaction assessment phase), where a symbol and a color are fixed for each item. Similarly, the data collected during the Unlabelled Configuration is shown by Fig. 15. The left-hand plots of these figures show the projection of the data space onto the item plane, while the right-hand plots show the same data projected onto the listener plane. Figs. 13 and 14 also show a comparison between the case with S_{BSS} and the one with S_{OO}, please refer to the plot titles. Bigger markers are used in case data points overlap. The size of the markers is proportional to the number of overlapping points, which is printed inside the marker. In the plots on the right, dark gray circles are used when many different markers overlap. The ordering of the items from left to right reflects the order in which they were presented in the test. The ten listeners of the Experience Configuration are represented by numerical labels: from 1 to 4 for naive listeners and from 5 to 10 for Hi-Fi lovers. The twenty listeners of the Unlabelled Configuration correspond to the labels from 1 to 6 for naive listeners and from 7 to 20 for Hi-Fi lovers.

The reader should not be overwhelmed if these figures appear complex at first sight: a large amount of information is indeed displayed. The most evident fact is, however, that there is a very high variance among listeners and items in terms of preferred ΔLD.

An indication of the coherent behavior of the participants throughout the test is given by the ΔLD selected for AR2_de and AR5_de, i.e., the items created by mixing with different LD_0 the same dialog and background objects. Almost all the listeners select a ΔLD for AR2_de (LD_0=0.97 LU) that is significantly bigger than the one that they select for AR5_de (LD_0=18 LU).

The low number of naive listeners makes it impossible to analyze the differences between naive listeners and Hi-Fi lovers in details. Still, the observed trend would suggest that naive listeners select higher levels of dialog than the Hi-Fi lovers, even if with high personal variations. This should be investigated in future.

Acknowledgment

The authors would like to thank Joshua D. Reiss, Oliver Hellmuth, Sascha Disch, Andreas Silzle, and Matthew Jefferson for the fruitful discussions. Special thanks also go to all who took part in the tests and to Ornela Pali for taking care of them.

References

[1] M. Torcoli, J. Herre, J. Paulus, C. Uhle, H. Fuchs, and O. Hellmuth, "The Adjustment/Satisfaction Test (A/ST) for the Subjective Evaluation of Dialogue Enhancement," in *Proceedings of the 143rd Audio Engineering Society Convention, New York*, 2017.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[3] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[4] S. Makino, *Audio Source Separation*. Springer, 2018.

[5] C. Uhle, O. Hellmuth, and J. Weigel, "Speech Enhancement of Movie Sound," in *Proceedings of the 125th Audio Engineering Society Convention, San Francisco*, 2008.

[6] J. Geiger, P. Grosche, and Y. Parodi, "Dialogue Enhancement of Stereo Sound," in *Proceedings of the 23rd IEEE European Signal Processing Conference, Nice*, 2015.

[7] A. Craciun, C. Uhle, and T. Bäckström, "An Evaluation of Stereo Speech Enhancement Methods for Different Audio-Visual Scenarios," in *Proceedings of the 23rd IEEE European Signal Processing Conference, Nice*, 2015.

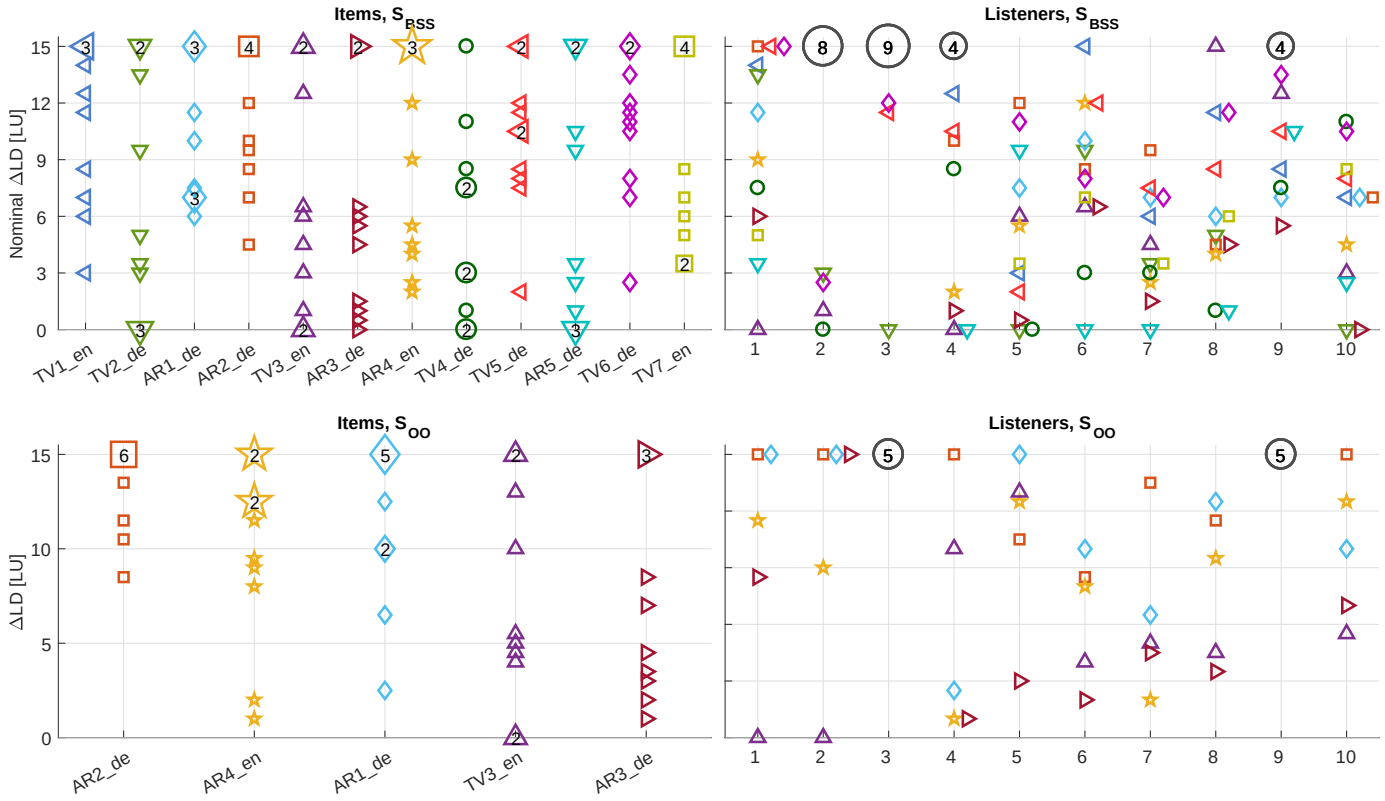

Fig. 13. Appendix. Experience Configuration: adjustment phase. The preferred levels of the dialog-to-background ratio (ΔLD) while operating S_{BSS} (upper plots) and S_{OO} (lower plots) are projected onto the item plane (plots on the left) and onto the listener plane (plots on the right). A symbol and a color are fixed for each item.

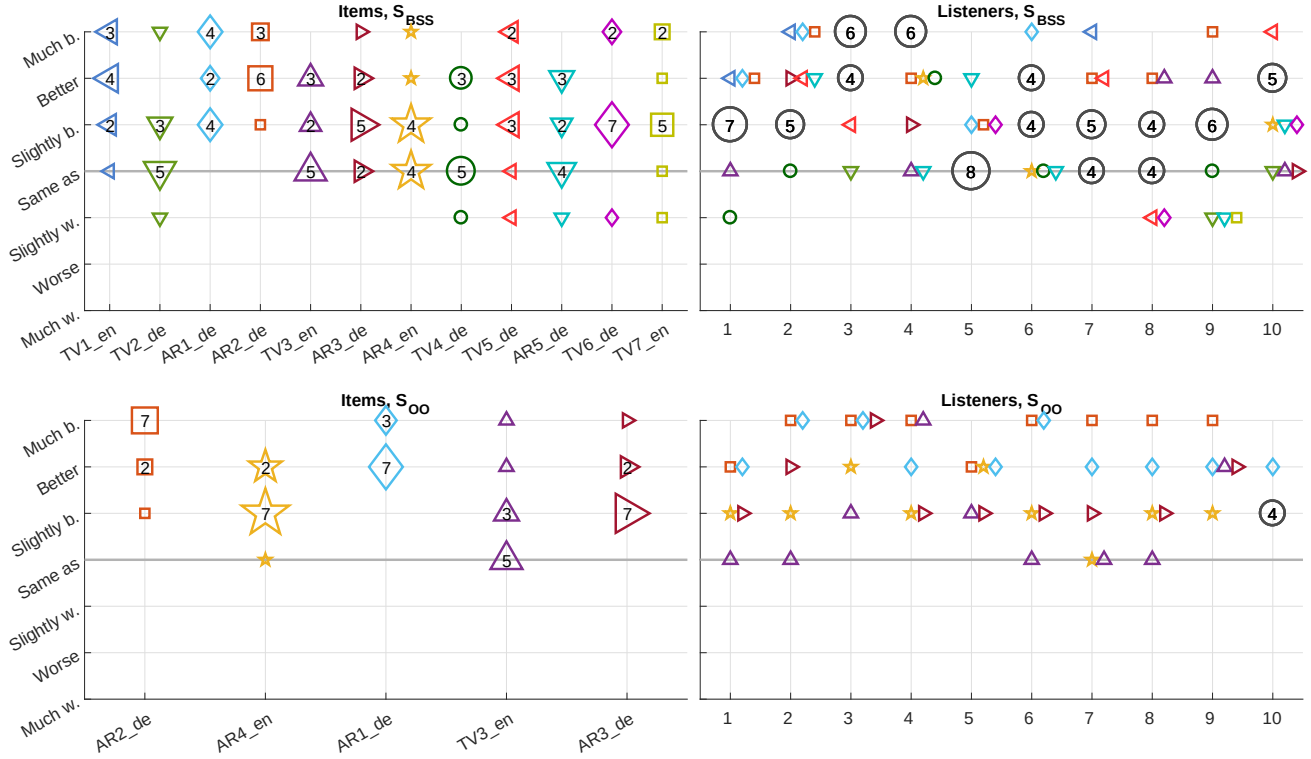

Fig. 14. Appendix. Experience Configuration: satisfaction assessment phase. Satisfaction levels related to the adjustment of S_{BSS} (upper plots) and S_{OO} (lower plots) projected onto the item plane (plots on the left) and onto the listener plane (plots on the right).A symbol and a color are fixed for each item.

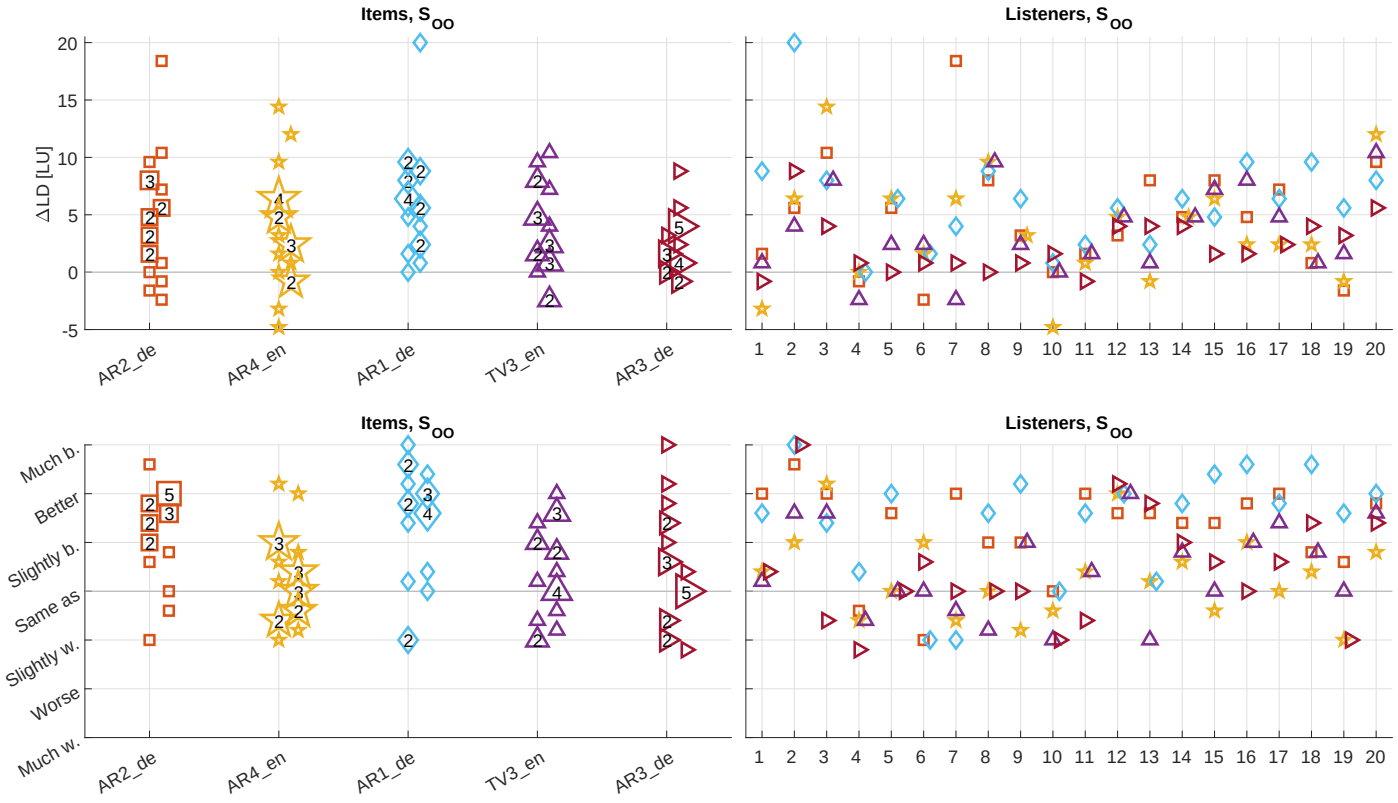

Fig. 15. Appendix. Unlabelled Configuration. Preferred adjustment levels (upper plots) and resulting satisfaction levels (lower plots) for S_{OO} projected onto the item plane (plots on the left) and onto the listener plane (plots on the right). A symbol and a color are fixed for each item.

[8] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *arXiv:1708.07524*, 2017.

[9] *Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio*, International Organization for Standardization (ISO), Geneve, 2015, Standard ISO/IEC 23008-3:2015.

[10] *Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio - Amendment 3: Audio Phase 2*, International Organization for Standardization (ISO), Geneve, 2016.

[11] R. L. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza *et al.*, "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Trans. Broadcasting*, vol. 63, no. 1, pp. 202–236, 2017.

[12] R. L. Bleidt, H. Thoma, W. Fiesel, S. Kraegeloh, H. Fuchs, R. Zeh, J. DeFilippis, and S. M. Weiss, "Building the World's Most Complex TV Network: A Test Bed for Broadcasting Immersive and Interactive Audio," *SMPTE Motion Imaging Journal*, vol. 126, no. 5, pp. 26–34, July 2017.

[13] A. Murtaza, H. Fuchs, and S. Meltzer, "MPEG-H TV Audio System for Cable Applications," *Journal of Digital Video Subcommittee*, vol. 2, no. 1, pp. 30–52, July 2017.

[14] ETSI TS 101 154 v2.3.1, *Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream*, Digital Video Broadcasting (DVB), February 2017.

[15] Advanced Television Systems Committee (ATSC), "ATSC 3.0 Standards: A/342 Part 3, MPEG-H System," February 2017.

[16] TTAK-KO-07.0127R1, *TTA - Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1*, Telecommunications Technology Association of Korea (TTAK), December 2016.

[17] B. Cardozo, "Adjusting the Method of Adjustment: SD vs DL," *Journal of the Acoustical Society of America*, vol. 37, no. 5, pp. 786–792, 1965.

[18] G. Fechner, *Elemente der Psychophysik Elements of Psychophysics*. Breitkopf und Härtel, 1860.

[19] B. Moore, D. Vickers, T. Baer, and S. Launer, "Factors Affecting the Loudness of Modulated Sounds," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2757–2772, 1999.

[20] B. Moore, B. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[21] F. Hubach and B. Edwards, "Empirical Determination of Sound Isolation Requirements for Recording Studio Isolation Booths," in *Proceedings of the 93rd Audio Engineering Society Convention, San Francisco*, 1992.

[22] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Determining the Threshold of Acceptability for an Interfering Audio Programme," in *Audio Engineering Society Convention 132*, 2012.

[23] B. De Man and J. Reiss, "The Mix Evaluation Dataset," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[24] B. De Man, J. Reiss, and R. Stables, "Ten Years of Automatic Mixing," in *Proceedings of the 3rd Workshop on Intelligent Music Production, Salford*, 2017.

[25] R. King, B. Leonard, and G. Sikora, "Variance in Level Preference of Balance Engineers: A Study of Mixing Preference and Variance over Time," in *Proceedings of the 129th Audio Engineering Society Convention, San Francisco*, 2010.

[26] B. De Man, K. McNally, and J. Reiss, "Perceptual Evaluation and Analysis of Reverberation in Multitrack Music Production," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 108–116, 2017.

[27] J. Paulus, C. Uhle, J. Herre, and M. Höpfel, "A Study on the Preferred Level of Late Reverberation in Speech and Music," in *Proceedings of the 60th Audio Engineering Society International Conference (DREAMS), Leuven*, 2016.

[28] W. Buyens, B. van Dijk, M. Moonen, and J. Wouters, "Music Mixing Preferences of Cochlear Implant Recipients: A Pilot Study," *International Journal of Audiology*, vol. 53, no. 5, pp. 294–301, 2014.

[29] J. Pons, J. Janer, T. Rode, and W. Nogueira, "Remixing Music Using Source Separation Algorithms to Improve the Musical Experience of Cochlear Implant Users," *Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4338–4349, 2016.

[30] E. Hildebrandt, "Sprachverständlichkeit im Fernsehen *(Intelligibility in Television)*," Master's thesis, Universität für Musik und darstellende Kunst Wien, 2014.

[31] T. Liebl, G. S., and K. G., "Verbesserung der Sprachverständlichkeit, speziell bei Voice-Over-Voice-Passagen *(Improvement of Voice-Over-Voice Speech Intelligibility in Television Sound)*," in *Proceedings of the 28th Tonmeistertagung - VDT International Convention, Köln*, 2014.

[32] ITU-R Recommendation BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," 2015.

[33] EBU Recommendation R 128, "Loudness Normalisation and Permitted Maximum Level of Audio Signals," 2014.

[34] B. Shirley and P. Kendrick, "ITC Clean Audio Project," in *Proceedings of the 116th Audio Engineering Society Convention, Berlin*, 2004.

[35] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *SMPTE Motion Imaging Journal*, vol. 123, no. 5, 2014.

[36] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized Object-Based Audio for Hearing Impaired TV Viewers," *Journal of the Audio Engineering Society*, vol. 65, no. 4, pp. 293–303, 2017.

[37] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Does Environmental Noise Influence Preference of Background-Foreground Audio Balance?" in *Proceedings of the 141st Audio Engineering Society Convention, Los Angeles*, 2016.

[38] P. Mapp, "Intelligibility of Cinema & TV Sound Dialogue," in *Proceedings of the 141st Audio Engineering Society Convention, Los Angeles*, 2016.

[39] M. Florentine, "Speech Perception in Noise by Fluent, Non-native Listeners," *Journal of the Acoustical Society of America*, vol. 77, no. S1, pp. S106–S106, 1985.

[40] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement - technology and experiments," *EBU Technical Review*, vol. Q2, 2012.

[41] M. Armstrong, "Audio Processing and Speech Intelligibility: a literature review," in *BBC Research & Development White Paper, WHP190*, 2011.

[42] M. Torcoli and C. Uhle, "On the Effect of Artificial Distortions on Objective Performance Measures for Dialog Enhancement," in *Proceedings of the 141st Audio Engineering Society Convention, Los Angeles*, 2016.

[43] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2006.

[44] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi *et al.*, "Qualinet White Paper on Definitions of Quality of Experience," *Output from the Fifth Qualinet meeting, Novi Sad*, 2013.

[45] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union - Telecommunication Standardization Sector (ITU-T), 1996.

[46] J. Beerends and F. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[47] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[48] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge - a Measure of Assessor Expertise in Audio Quality Evaluations," in *Proceedings of the 38th Audio Engineering Society International Conference: Sound Quality Evaluation, Piteå*, 2010.