OPEN ACCESS Freely available online
M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. Shirley, "Preferred Levels for Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech"
J. Audio Eng. Soc., vol. 67, no. 12, pp. 1003–1011, (2019 December.).
DOI: https://doi.org/10.17743/jaes.2019.0052

Preferred Levels for Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech

MATTEO TORCOLI,[1] *AES Member*, **ALEX FREKE-MORIN,**[1,2] *AES Associate Member*,
(matteo.torcoli@iis.fraunhofer.de) (afreke@aol.com)
JOUNI PAULUS[1,3], **CHRISTIAN SIMON,**[1] *AES Associate Member*, **AND**
(jouni.paulus@iis.fraunhofer.de) (christian.simon@iis.fraunhofer.de)
BEN SHIRLEY,[2] *AES Member*
(B.G.Shirley@salford.ac.uk)

[1] *Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2] *Acoustics Research Centre, University of Salford, UK*
[3] *International Audio Laboratories Erlangen, Germany, A joint institution of Universität Erlangen-Nürnberg and Fraunhofer IIS.*

In audio production, background ducking facilitates speech intelligibility while allowing the background to fulfill its purpose, e.g., to create ambience, set the mood, or convey semantic cues. Technical details for recommended ducking practices are not currently documented in the literature. Hence, we first analyzed common practices found in TV documentaries. Second, a listening test investigated the preferences of 22 normal-hearing participants on the Loudness Difference (LD) between commentary and background during ducking. Highly personal preferences were observed, highlighting the importance of object-based personalization. Statistically significant difference was found between non-expert and expert listeners. On average, non-experts preferred LDs that were 4 LU higher than the ones preferred by experts. A statistically significant difference was also found between Commentary over Music (CoM) and Commentary over Ambience (CoA). Based on the test results, we recommend at least 10 LU difference for CoM and at least 15 LU for CoA. Moreover, a computational method based on the Binaural Distortion-Weighted Glimpse Proportion (BiDWGP) was found to match the median preferred LD for each item with good accuracy (mean absolute error $= 1.97$ LU $\pm$ 2.50).

0 INTRODUCTION

One of the most common complaints to broadcasters is about the low intelligibility of the foreground speech (e.g., dialog and commentary) in TV programs due to the background (sometimes also referred to as *bed*) being too loud compared to the speech [1]. The background includes music and effects that are essential for the full understanding and enjoyment of the show. However, the background can energetically mask the speech [2], making it impossible or tiring to understand. In some cases, informational masking can also occur [3], e.g., during Voice-over-Voice (VoV) passages. An example for VoV is when a foreground voice translates or comments on a voice in the background.

Audio producers are challenged with the task of producing audio mixes of speech and background sounds with fully intelligible foreground speech, at least under favorable listening conditions. In order to do so, it is common practice to attenuate the level of the background during periods of foreground speech activity. This can be executed in various ways, which are all referred to as *ducking* in this paper. Techniques based on volume automation and side-chain compression or gating (i.e., when the level of an auxiliary or *side-chain* signal controls the amount of the attenuation of another signal) are what many audio engineers refer to as *ducking*. This paper uses the term *ducking* with a broader meaning, referring to any time-varying background attenuation with the aim of making the foreground speech clear.

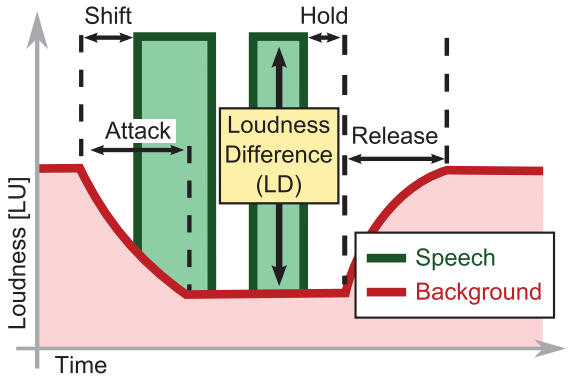


Fig. 1. Ducking of the background while foreground speech is active. This is controlled by parameters such as the Loudness Difference (LD) between speech and background during ducking and time constants (attack, shift, hold, release).

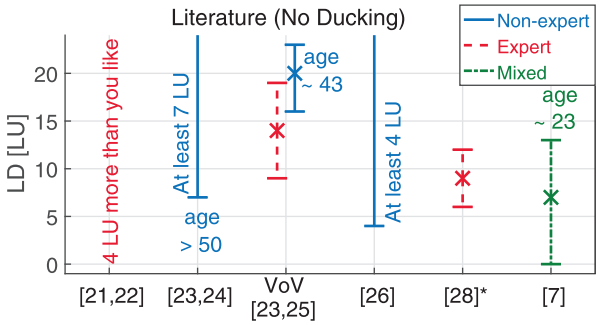


Fig. 2. Overview of the literature about desirable LDs in TV audio mixes, as detailed in Sec. 1.2. No work took into consideration ducking, but the LD was studied as a static level difference. A modified loudness measurement was used in [27] and it is unknown to what extent it is comparable with the other values.

In every case, a number of tunable processing parameters are involved.

Fig. 1 shows the main ducking parameters. These can be categorized into two groups: time constants and level parameters. The time constants can be described by typical compression parameters such as attack and release, complemented by a shift time, which describes the attack offset before speech is starting, and a hold time, which prevents *pumping* in short speech gaps. The main relative level parameter is the Loudness Difference (LD) between foreground speech and background. As a loudness measure we adopted the integrated loudness as per ITU BS.1770-4 [4], i.e., as per EBU R 128 [5]. This is measured in Loudness Units Full Scale (LUFS) if relative to digital full scale or in Loudness Units (LU) if related to another specified level, e.g., in the case of the LD. A useful property of the LU is that a level increase by 1 dB leads to a 1 LU increase.

Best practices for esthetically well-tuned ducking are not defined by mixing handbooks and broadcaster recommendations. Often the only recommendation given is that foreground speech has to be "comprehensible" and "clear."

This paper sheds some light on the ducking parameters by studying the LD. This is done by analyzing the practices found in a sample of TV documentaries and carrying out a preference test. The test also focuses on the preference differences between expert and non-expert listeners, as well as on the difference introduced by different types of background, such as music or ambience. Related works are also reviewed and compared with our findings.

1 LITERATURE REVIEW

1.1 The Perfect TV Audio Mix Is Subjective

The right balance between foreground speech and background was shown to depend on personal taste [6, 7], listener's hearing acuity [8–11], listening environment (e.g., environmental noise [12]), reproduction system [13], and listener's skill level in the content language [14].

Object-based audio systems, such as MPEG-H Audio, solve this problem by enabling the audience to personalize the relative level of foreground speech [15–18]. This is known as Dialogue Enhancement and has been shown to clearly increase the quality of the experience of the final user [7, 19, 20].

1.2 Desirable LD for the Default Mix

Even if object-based personalization is available, a default mix is needed. This should satisfy as many people as possible. Some works in the literature investigated the LD, even if considering a constant level of the background and not in the context of ducking. These works are reviewed in the following and their main results are visually summarized in Fig. 2.

In guidelines by the BBC [21, 22], general suggestions such as the following are given: *"Be aware of background noise"* and *"Take the music down. Our research showed that bringing music down slightly in the mix allowed people across the demographic to hear dialogue, including those with certain hearing loss. Once you're happy with your mix, try taking the music down 4 dB (one point on the PPM) and see if this impacts on your creative vision. The chances are it won't!"*

A set of guidelines focused on speech intelligibility published by German public broadcasters [23] suggests that the LD should be at least 7 LU and between 16 and 23 LU for VoV. These recommendations are based on the studies in [24] and [25]. In [24] a listening test is carried out where people could rate test items mixed with LD = 2, 7, 10 LU by means of a questionnaire. While participants under the age of 50 did not show a clear preference, participants over 50 preferred at least 7 LU for a comfortable listening experience. In [25] the focus is on VoV excerpts. These were mixed by audio engineers with LDs in the range 9–19 LU. Then, non-expert listeners between 23 and 58 years old (mean age 43) were asked to adjust the LD and set it to their preferred level by means of a slider, starting from the LD set by the audio engineers. The results showed two different groups of participants. One group favored LDs similar to initial ones (on average 14 LU), while a second group clearly preferred higher LDs (on average 20 LU). Different listening conditions were also tested, e.g., with/without

video, which did not lead to any significant difference in the results. Hence, an LD between 16 and 23 LU was suggested to meet the preference of the most critical group.

The Digital Production Partnership in the UK has produced technical requirements that recommend a minimum separation of 4 LU between dialog and background [26]. This figure was also reflected in Netflix guidelines as a recommended difference between dialog and effects, although this recommendation has been withdrawn in the most recent version of the Netflix guidelines [27], which make no such recommendation on relative loudness levels.

In a study by NHK [28] the balance between foreground speech and background was investigated using a variation of the ITU BS.1770 loudness with smaller time constants. It is unknown to what extent this is comparable to other loudness values. For documentary programs, the LD chosen by 12 mixing engineers was 9 ± 3 LU. A few samples of musical shows and sport programs were also considered. In these cases, LDs close to zero or even negative LDs were chosen by the same mixing engineers.

In [7] expert and non-expert listeners between 20 and 32 years old (median age 23) could set their preferred LD by means of the Adjustment/Satisfaction Test (A/ST). Among the test signals no VoV was present. Most of the preferred LDs were between 0 and 13 LU, with an average of 7 LU.

Finally, in [29], the preferred speech-to-background ratios, as well as the ratios that allow participants to only just understand everything in a sentence, were investigated and predicted by means of an objective intelligibility measure: the Binaural Distortion-Weighted Glimpse Proportion (BiDWGP). Results were given as energy ratios in dB instead of LU and could not be directly compared with the other reviewed works. Sec. 3.3 details a repeat of the prediction experiment on the new data collected in this study.

All of the research reviewed considered a static level of the background without ducking. However, ducking is widely used in real-world material. To the best of our knowledge, this work is the first study investigating desirable LDs during ducking. As a note, numerous publications have analyzed music mixing preferences, which are much better documented than those for broadcasting. The interested reader is referred to [30, 31] and references therein.

2 COMMONLY USED LDS DURING DUCKING

In this section, the intention was to gain an understanding of existing ducking techniques in TV by analyzing real-world content. Documentary programs were selected, as they are a common format in TV and tend to feature extensive use of ducking in the audio mix. In this type of content, the foreground speech is commonly a narrator while the background tends to belong to the depicted scene or create the mood of the scene.

Twelve documentaries broadcast in the UK, Germany, and France were considered. The documentaries were heterogeneous, comprising a range of production values. The full program files were normalized to equal integrated loudness. The LDs for 6 VoV, 12 Commentary over Music (CoM), and 6 Dialog over Music (DoM) excerpts were measured while the background was ducked. *Commentary* refers to speech recorded in a studio, e.g., in a sound booth, and features professionals who speak clearly. On the other hand, *(location) dialog* can be less clear than commentary, as it is recorded in a less controlled environment using portable recording equipment and can include location background sounds and non-professional speakers.

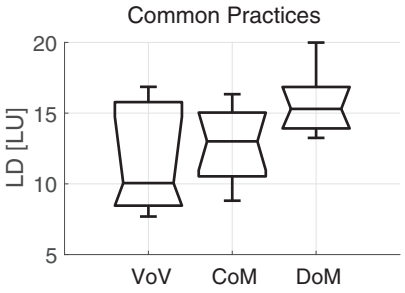


Fig. 3. Boxplot of the LDs during ducking found in 12 documentary programs broadcast in the UK, Germany, and France. Voice-over-Voice (VoV), Commentary-over-Music (CoM), and Dialog-over-Music (DoM) excerpts were considered.

VoV, CoM, and DoM excerpts were manually selected where ducking was employed and where the ducked background level was constant. For each excerpt, segments with active foreground voice and segments without foreground voice were isolated. The segments belonging to the same excerpt and voice activity class were concatenated. The LD of each excerpt was estimated as the difference between the integrated loudness of the concatenated mix segments when voice was active and the loudness of the background, i.e., the concatenated mix segments when voice was not active. For LDs above 5 LU, this estimation was found experimentally to give a mean absolute error of 0.6 LU (standard deviation 0.7) on similar synthesized excerpts.

Measured LDs are shown in Fig. 3 by boxplots.[1] Most of the values for VoV and CoM were within the 10–15 LU range, similar to the mixes created by professionals in [23, 25]. The lowest LDs were found for VoV, which was surprising considering the possible informational masking. However, the considered voice-overs were recorded by professionals (as in CoM), which is likely to enable comfortable listening in spite of the lower LDs. Higher LDs were found for DoM (14–17 LU). As location dialog can be less clear than commentary, audio engineers are likely to choose higher LD to compensate for this.

The usage of momentary and integrated loudness was also compared. When using momentary loudness, the values varied less than 1 LU with respect to integrated loudness. This can be explained by the short length and homogeneity of the analyzed excerpts.

Even though this research spanned different countries, topics, and production values, the considered selection of

[1] The ends of the boxes correspond to the 25/75% quantiles of the data; the central bar corresponds to the median. The vertical lines extending from the box (whiskers) indicate the minimum and maximum points within 1.5 IQR (interquartile range).

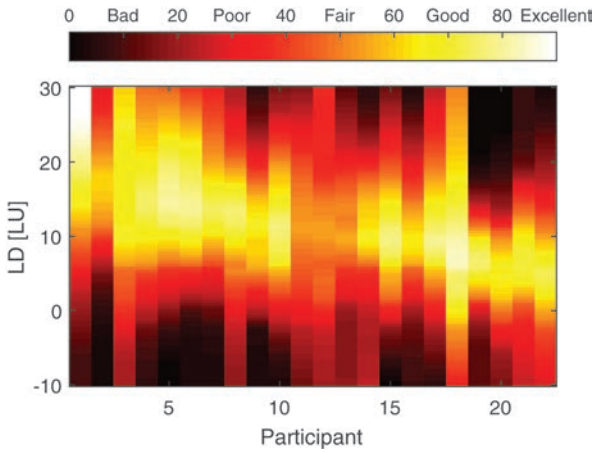


Fig. 4. Distribution of the mean ratings given by the participants to the different LDs proposed in the listening test. The rating is color-coded as explained by the upper bar.

programs is small and the values found should be corroborated in future research. Still, it is shown in the following that the found common practices are closely related to the results of the carried out listening test.

3 LISTENING TEST

This section describes the listening test carried out to determine the preferred LDs during ducking in the context of TV documentary programs. The setup is explained in Sec. 3.1, while the results of the listening test are analyzed in Sec. 3.2. A method to predict these results is proposed in Sec. 3.3. Then, Sec. 4 discusses how these results relate to the literature reviewed and the found common practices.

3.1 Listening Test Setup

Method: The test was a multiple stimuli test, where each stimulus (or condition) had a different LD. Participants rated their preference for each condition with a slider ranging 0–100 and labeled each 20-point range: bad, poor, fair, good, and excellent, as shown in the upper bar of Fig. 4, i.e., the same scale as in [32] was used.

Conditions: Nine conditions were presented on one test screen ordered based on the LD so as to simplify the task for the participant. As the conditions being assessed were only clearly distinct alterations in background level, randomization was considered to needlessly complicate the test for the participants. It is possible that some "centering bias" was present, but it was considered more critical that the test experiment was made simpler. The first condition featured no ducking and its initial LDs were randomly selected between –6 and 6 LU. The following conditions had the same background level when speech was not active but increasing LDs when speech was active, i.e., ducking was applied with increasing LDs. The last condition (with maximum LD) corresponded to 20–28 LU. The foreground speech had the same integrated loudness through all conditions.

Instructions: The following instructions were given to the participants: *"Imagine you are hearing the presented audio pieces while watching television in your living room. Your task is to rate these different mixes (of background sounds and speech) based on your overall preference. This means that you rate the mix(es) that you would personally rather hear with the highest score."*

Table 1. Ducking time constants used in the listening test. The *normal* constants were used for half of the items, while the *AD* constants were used for the other half. For the *normal* case, the full release was never to be heard in the test, so it is marked by (*).

Time Constant	Normal (ms)	AD (ms)
Attack	500	200
Release	500 (*)	200
Shift	500	200
Hold	1,500	0

Background Types: The test focused on commentary only: CoM as well as Commentary over Ambience (CoA). The minimum LDs for CoA were between –3 and 6 LU, i.e., these signals are examples of the special case in which ambience is particularly loud. Usually in broadcast material ambience has a naturally low level, resulting in high LDs. Here, the case in which ambience also has masking potential and ducking has to be applied was analyzed.

Test Items: The test involved 12 items (6 CoM and 6 CoA, each with the 9 different conditions). All items featured German commentary panned to the center over stereo backgrounds. The loudness range (LRA) for the commentary ranged between 3.4 and 5.2 LU (mean: 4.4 LU), while the mean LRA was 4 LU for the full mixes. Examples of ambience background were flowing water, street noise, car interior, and subway hall. Music backgrounds featured no lyrics and could be categorized as soft rock, ambience music, and orchestral music. No accompanying video was shown. Every item was 9 seconds long and had a sampling frequency of 48 kHz. A training item was presented before the 12 test items for a better understanding of the experiment. This item was not considered in the results. During the training phase, participants could modify the overall volume to a comfortable level.

Time Constants: Table 1 shows the ducking time constants applied in the test. Two sets were used. The *normal* set was used for half of the CoM items and half of the CoA items, while the *AD* set was used for the other half. The *AD* constants were faster and ducking characteristics resembled those commonly used for Audio Description (AD).

Participants: The test involved 11 expert listeners (between 21 and 43 years old, median age 26) and 11 non-expert listeners (between 23 and 59 years old, median age 25) without known hearing impairments. The expert listeners passed a listener-screening program, where they were verified to not have hearing impairments and to have high testing ability [33]. Five were professional audio engineers. All participants had German as first language, and they were reimbursed for taking part in the test.

Location: Two similar listening rooms with high-end studio monitors were used. The rooms were quiet and

Table 2. ANOVA of the preferred LDs: degrees of freedom (d.f.), effect size η^2 (given as percentage of the total variability), and p values (if lower than 0.05, we reject the null hypothesis). Significant effects are marked in bold.

Variable	d.f.	η^2 (%)	p
Participant	18	**31.7**	**0.00**
Item	10	**15.2**	**0.00**
Background	1	**14.3**	**0.00**
Age	1	**7.1**	**0.00**
Experience	1	**5.3**	**0.00**
Participant × Background	19	**4.3**	**0.00**
Item × Experience	10	1.3	0.24
Item × Age	10	0.9	0.58
Experience × Background	1	0.1	0.54
Age × Background	1	0.0	0.91
Age × Experience	1	0.0	0.98
Error	190	19.8	

acoustically treated so as to resemble low-reverberant living rooms.

3.2 Listening Test Results

Fig. 4 shows the distributions of the mean ratings given by the participants for the LDs proposed in the listening test. For each participant, bell-shaped distributions were observed with a more or less broad peak around the preferred LD. The size of the peak (the yellow area in Fig. 4) can be interpreted as an accepted range, where the LD is considered good even if it is not the preferred one. Participants had very different preferred LDs and they used the preference scale in different ways.

The main analysis of the results was carried out on the preferred LDs, i.e., by considering the condition(s) with the highest score for each participant and item. In a few cases, more than one condition was rated with the highest score. In this case, the mean LD was taken over the two or more preferred conditions. Focusing on the preferred LDs normalizes the differences in the use of the rating scale.

A nested ANOVA with five factors was carried out on the preferred LDs, as reported in Table 2. The factors are: *item*, *background* type (CoM or CoA), *participant*, participant *age*, and participant *experience* (i.e., if the participant is expert or non-expert). *Item* is nested inside *background* type. *Participant* is nested inside *age* as well as *experience*. *Item* and *background* are fixed factors, as they have fixed characteristics that can be used in a new experiment. *Participant*, *age*, and *experience* are random factors, as they are samples randomly taken from the relevant population, on which we would like to generalize.

Age describes two groups created by taking the participants' median age as threshold, which is 26 years. The first group comprises 14 participants with median age 24. The second group comprises 8 participants with median age 34.

The five factors are statistically significant. The factor *participant* alone explains 31.7% of the total variability in the data, confirming that personal taste plays a fundamental role. As expected, *item* and *background* are significant factors, accounting together for 29.5% of the total variability. *Age* explains 7.1% of the variability, although no hearing-impaired listeners were considered. *Experience* is also statistically significant, accounting for 5.3% of the variability. *Experience* is particularly interesting in our application, as TV mixes are created by experts to be consumed by non-experts. The only significant two-way interaction is between *participant* and *background*.

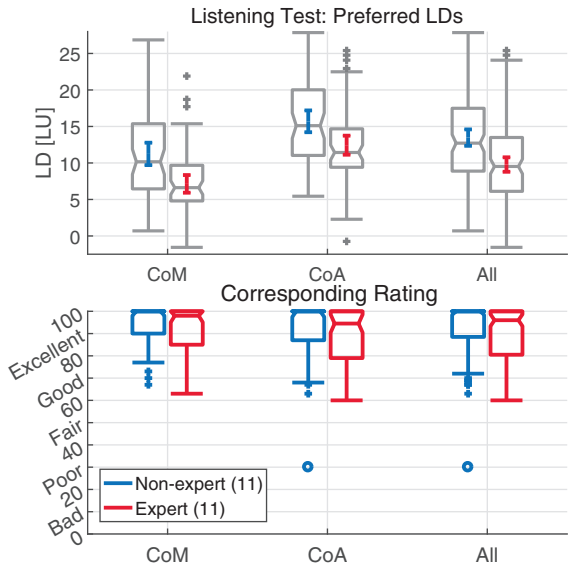


Fig. 5. The upper subplot depicts 95% confidence intervals overlapping boxplots of the preferred LDs by background type and experience. For each background type, non-experts are shown on the left and experts on the right. The lower subplot shows the ratings corresponding to the preferred LDs.

An ANOVA including the factor *time constants* (normal or AD) was also carried out. This factor was not found to explain significant variations in the preferred LDs. Hence, it was not included in the presented ANOVA or the following analysis of the results. Nevertheless, it is known from practical experience that ducking time constants are important for enjoyment and acceptance of an audio mix. Future research should study this aspect in more detail.

The upper subplot of Fig. 5 shows 95% confidence intervals and boxplots depicting the preferred LDs. The statistically significant difference between non-expert and expert listeners can be observed. Non-experts preferred higher LDs with interquartile values ranging from 6.5 to 15.4 LU for CoM and from 11 to 20 for CoA. On average, the measured difference in preferred LDs between experts and non-experts was equal to 4 LU, which supports BBC guidance of reducing music level by 4 dB [21, 22].

The lower subplot of Fig. 5 depicts the ratings corresponding to the preferred LDs. All the ratings lie in the "excellent" and "good" ranges with only one exception (rating 30). This exception is due to an expert participant who complained about the fact that the background did not have much to do with what was being said in that item.

In addition to analyzing preferred LDs, the range comprising the LDs rated above 60 points (i.e., in the excellent and good ranges) and no more than 10 points below the preferred LD was considered. This range is referred to as the

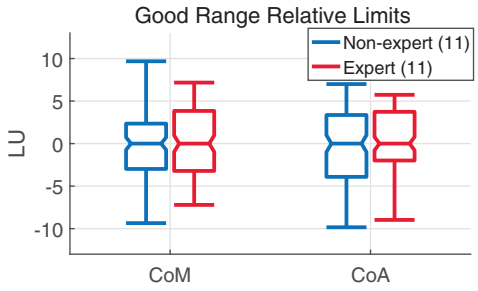


Fig. 6. Boxplots of the relative limit values of the *good range*. This range includes the LDs rated above 60 and no more than 10 points below the preferred LD for each item and participant.

good range. Its limit values (relative to the preferred LD for each item and participant) are shown in Fig. 6. They give a picture of the freedom that an audio engineer would have in order to meet the participant preference. The interquartile ranges suggest that a ± 3 LU around the preferred LD would still be well received.

Finally, Fig. 7 shows the preferences for each item. These are also matched with a computational method, as discussed in the following.

3.3 Automatic Prediction of the Per-Item LDs

The work [29] investigated the speech-to-background ratios (SBRs) that allow participants to only just understand everything in a sentence, and a ratio resulting in BiDWGP = 0.5 was shown to match the minimum level of full intelligibility. In a second test, participants were asked to select SBRs that resulted in *the speech content being intelligible enough and the background sound providing good atmosphere to the scene.* In other words, the participants completed a task similar to the task in our listening test, but the instructions were different (ours focused on the overall preference). The average selected SBRs were successfully matched with a computational method by adding an offset of 5.5 dB to the SBRs corresponding to BiDWGP = 0.5. The SBRs selected by the test participants ranged from –6 to 4 dB, i.e., they selected ratios where the speech has a level comparable or even lower than the level of the background. Even if these SBRs are not directly comparable with our results because of the measurement unit mismatch (SBRs in dB and LDs in LU), these SBRs seem to be clearly lower than the preferences recorded in our test and the previous works (Fig. 2). This is possibly explained by the different formulation of the task given to the participants. It is therefore not surprising that BiDWGP = 0.5 with an offset of 5.5 dB underestimates the preferred LDs found in our listening test.

Fig. 7 depicts the LDs corresponding to BiDWGP = 0.5 (shown by the purple solid line), which were computed considering only the signal portions for which the background was ducked. It can be clearly observed that all the preferred LDs are above BiDWGP = 0.5, suggesting that speech was fully intelligible in all cases. On our data, using an offset of 17.7 LU minimizes the mean absolute error (MAE) between the predictions and the median non-experts' preferences. The resulting MAE is 1.97 LU, standard deviation is 2.50 LU, and maximum error is 4.68 LU (observed for item M2).

4 DISCUSSION

Fig. 8 summarizes the LD values presented so far and discussed in the following, where our recommendations are also introduced. Our results show that non-expert listeners (such as consumers) prefer LDs that are, on average, 4 LU higher than the levels preferred by experts (a category that would include mix creators). This supports the rule of thumb suggested by the BBC to audio engineers on using LDs that are 4 LU higher than the ones they prefer [21, 22]. A similar difference was documented in [25].

In trying to standardize LD values, German public broadcasters recommend at least 7 LU in general and at least 16 LU for VoV [23]. This research, however, indicates that it is common practice to use similar LDs for VoV and CoM, i.e., between 10 and 15 LU. Higher LDs are found only for DoM (14–17 LU), probably due to the typically lower intelligibility of location dialog.

Preferred LDs as low as 0 LU were found for expert and non-expert listeners in [7], where the median age is 23. The particularly low age of the participants might explain the difference to other reviewed works. These lower values are not recommended for the default TV mix but can be made available via the personalization offered by object-based systems, such as MPEG-H Audio.

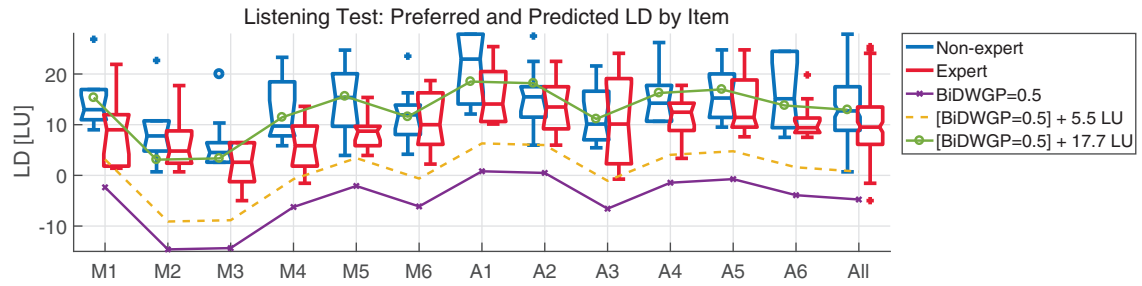


Fig. 7. Boxplots of the preferred LDs by item and experience. Items featuring CoM have IDs starting with *M*, while IDs start with *A* for CoA items. The median LDs preferred by non-experts are predicted based on the full intelligibility requirement suggested by BiDWGP =0.5. Using an offset of 17.7 LU minimizes the mean absolute error (MAE =1.97 LU, standard deviation =2.50 LU).

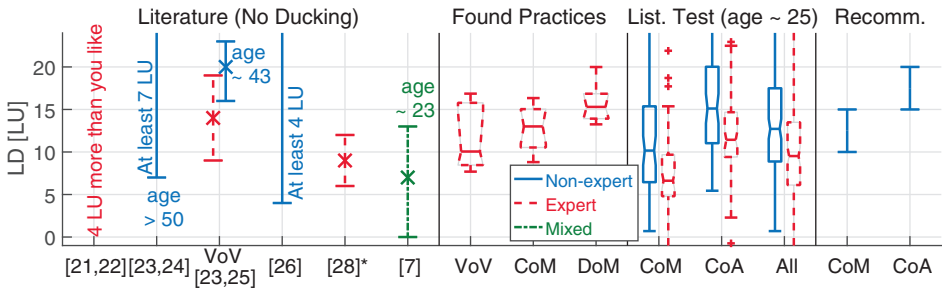


Fig. 8. Integrated Loudness Difference (LD) between foreground speech and background for the default mix in TV audio. Visual summary of the literature review (as in Fig. 2), common practices found in TV documentaries (as in Fig. 3), results from the preference listening test (as in Fig. 5), and final recommendations for Commentary over Music (CoM) and Commentary over Ambience (CoA).

A recommendation on the LD for the default mix has to take into consideration the preference of non-expert participants, as they are a closer representation of the audio mix consumers. Our listening test showed that non-experts prefer 6.5–15.4 LU for CoM and 11–20 for CoA. It also indicates that participants found a range of ±3 LU around the preferred LD acceptable. These are large ranges, but a recommendation can be made by suggesting *at least* the median values, i.e., trying to meet the preference of the half of the population that prefers higher LDs. Satisfying the higher half of the population as well would make the values closer to those that would be needed in suboptimal listening conditions or for hearing-impaired listeners.

Hence, as shown in the rightmost part of Fig. 8, we recommend at least 10 LU for CoM. Higher LDs need to be adopted for DoM. For CoA, at least 15 LU can be recommended. Esthetically pleasing upper limits can be considered 15 LU for CoM and 20 LU for CoA (i.e., the 75% quartile the non-expert participants' preferences). The range 10–15 LU for CoM was also observed to be commonly used in the real-world programs analyzed.

Finally, we found a good match between the intelligibility measure BiDWGP and the median preferred LD for each item. This shows that (BiDWGP = 0.5) + 17.7 LU is a promising computational method to predict the LD for one item. The offset 17.7 LU best fits the recorded LDs and it is more coherent with levels found in the literature. However, this deviates by 12.2 LU from the offset originally proposed in [29] and should be validated in future works.

The values discussed in this paper are for background signals with characteristics that do not vary largely over time. There may be cases and signals where the optimal LD could deviate significantly from the recommended values. In these cases, the know-how of an experienced audio engineer is irreplaceable.

5 CONCLUSION

Ducking is a technique that facilitates speech intelligibility while allowing the background to fulfill its purpose, e.g., to create ambience, set the mood, or convey semantic cues. Although it is extensively used in TV audio, optimal values for its parameters are not documented in the literature. This work filled this void by focusing on the integrated Loudness Difference (LD) between foreground speech and background during ducking. This was done by analyzing common practices found in a sample of TV documentaries and carrying out a listening test on the LDs preferred by 22 normal-hearing participants. Findings were compared with related works.

Results clearly showed that only the personalization offered by object-based audio technologies such as MPEG-H Audio can meet the preferences of all people. Nevertheless, a default mix is needed that satisfies as many people as possible. For a default mix that is esthetically pleasing and has clear speech, we recommend an LD of at least 10 LU for Commentary over Music (CoM) and 15 LU for Commentary over Ambience (CoA). Higher LDs should be used for location dialog.

Moreover, we found a good match between an objective intelligibility measure and the median preferred LD for each item. This seems to be a promising way to automate the creation of the default mix.

We believe that the know-how of an experienced audio engineer is irreplaceable, but it should be kept in mind that non-experts (such as the audio consumers) prefer LDs that are, on average, 4 LU higher than those preferred by expert listeners (such as the mix creators). This difference may be even higher when considering sub-optimal listening conditions. Future works should include participants with higher age and age-related hearing impairment. More items, background types, and reproduction configurations should also be studied. In addition, only the LD was investigated, which is the main parameter of ducking but not the only one. Other parameters (e.g., ducking time constants) should be analyzed in more detail in future works.

6 ACKNOWLEDGMENT

This paper is revised based on a paper presented at the 146th AES Convention in Dublin [34]. Many thanks to the members of the Convention Committee who were kind enough to award it Best Peer Reviewed Paper and to Bozena Kostek, who encouraged us to revise it for this journal. Special thanks also go to all who took part in the listening test, to Mariola Hatalski for taking care of them, and to Yan Tang for providing the code for BiDWGP.

7 REFERENCES

[1] M. Armstrong, "From Clean Audio to Object Based Broadcasting," http://www.bbc.co.uk/rd/publications/whitepaper324 (2016), accessed: 2019-08-09.

[2] J. Barker and X. Shao, "Energetic and Informational Masking Effects in an Audiovisual Speech Recognition System," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 3, pp. 446–458 (2009), doi:10.1109/TASL.2008.2011534.

[3] Y. Tang and T. J. Cox, "Improving Intelligibility Prediction Under Informational Masking Using an Auditory Saliency Model," presented at the *Int. Conf. on Digital Audio Effects (DAFx)* (2018).

[4] ITU-R Rec. BS.1770-4, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level" (2015).

[5] EBU Rec. R 128, "Loudness Normalisation and Permitted Maximum Level of Audio Signals" (2014).

[6] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement—Technology and Experiments," *EBU Tech. Rev.*, vol. Q2 (2012).

[7] M. Torcoli, J. Herre, H. Fuchs, J. Paulus, and C. Uhle, "The Adjustment/Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and Its Application to Dialogue Enhancement," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 524–538 (2018 June), doi:10.1109/TBC.2018.2832458.

[8] B. Shirley and P. Kendrick, "ITC Clean Audio Project," presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6027.

[9] B. Shirley and P. Kendrick, "The Clean Audio Project: Digital TV as Assistive Technology," *Technol. Disabil.*, vol. 18, no. 1, pp. 31–41 (2006).

[10] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *SMPTE Motion Imaging J.*, vol. 123, no. 5 (2014), doi:10.5594/j18429.

[11] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized Object-Based Audio for Hearing Impaired TV Viewers," *J. Audio Eng. Soc.*, vol. 65, pp. 293–303 (2017 Apr.), doi:10.17743/jaes.2017.0005.

[12] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Does Environmental Noise Influence Preference of Background-Foreground Audio Balance?" presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 9637.

[13] P. Mapp, "Intelligibility of Cinema & TV Sound Dialogue," presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 9632.

[14] M. Florentine, "Speech Perception in Noise by Fluent, Non-Native Listeners," *J. Acoust. Soc. Am.*, vol. 77, no. S1, pp. S106–S106 (1985), doi:10.1121/1.2022152.

[15] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Top. Sign. Process.*, vol. 9, no. 5, pp. 770–779 (2015), doi:10.1109/JSTSP.2015.2411578.

[16] F. Kuech, M. Kratschmer, B. Neugebauer, M. Meier, and F. Baumgarte, "Dynamic Range and Loudness Control in MPEG-H 3D Audio," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9465.

[17] ISO/IEC,"*Information Technology – High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 3: 3D Audio*," International Standard 23008-3:2018, Second Edition (2018).

[18] L. Ward and B. Shirley, "Personalization in Object-Based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners," *J. Audio Eng. Soc.*, vol. 67, pp. 584–597 (2019 Jul./Aug.), doi:10.17743/jaes.2019.0021.

[19] 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," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), convention paper 9842, doi:10.17743/aesconv.2017.978-1-942220-18-3.

[20] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch, and H. Fuchs, "Source Separation for Enabling Dialogue Enhancement in Object-Based Broadcast With MPEG-H," *J. Audio Eng. Soc., Spec. Issue on Object-Based Audio*, vol. 67, pp. 510–521 (2019 Jul./Aug.), doi:10.17743/jaes.2019.0032.

[21] BBC Editorial Guidelines, "*Hearing Impaired Audiences,"* http://downloads.bbc.co.uk/guidelines/editorialguidelines/pdfs/hearing-impaired.pdf (2011 Mar.), accessed: 2019-08-09.

[22] BBC Academy, "*Clear Sound: Best Practice Tips,"* https://www.bbc.co.uk/academy/en/articles/art20130702112135255 (2017 Aug.), accessed: 2019-08-09.

[23] ARD/ZDF, *Sprachverständlichkeit im Fernsehen, Empfehlungen für Programm und Technik (Intelligibility in Television, Recommendations for TV Program and Technique)*, https://www.irt.de/webarchiv/showdoc.php?z=NzE0MSMxMDA2MDE4I3BkZg== (2014), accessed: 2019-08-09.

[24] E. Ebert and E. Bodenseh, "Arbeitsergebnisse und die daraus resultierende Guideline (Speech Intelligibility in TV: A Guideline)," presented at the *28th Tonmeistertagung - VDT Int. Conv.* (2014).

[25] T. Liebl, S. Goossens, and G. Krump, "Verbesserung der Sprachverständlichkeit, speziell bei Voice-Over-Voice-Passagen (Improvement of Voice-Over-Voice Speech Intelligibility in Television Sound)," presented at the *28th Tonmeistertagung - VDT Int. Conv.* (2014).

[26] UK Digital Production Partnership (DPP), "Technical Specification for the Delivery of Television Programmes as As-11 Files v5.0," Sec. 2.2.1. Loudness Terms (2017).

[27] Netflix, "Netflix Sound Mix Specifications & Best Practices v1.1," https://partnerhelp.netflixstudios.com/hc/en-us/articles/360001794307-Netflix-Sound-Mix-Specifications-Best-Practices-v1-1, accessed: 2019-08-09.

[28] T. Komori, T. Takagi, K. Kurozumi, and K. Murakawa, "An Investigation of Audio Balance for

Elderly Listeners Using Loudness as the Main Parameter," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7629.

[29] Y. Tang, B. M. Fazenda, and T. J. Cox, "Automatic Speech-to-Background Ratio Selection to Maintain Speech Intelligibility in Broadcasts Using an Objective Intelligibility Metric," *Appl. Sci.*, vol. 8, no. 1, p. 59 (2018), doi:10.3390/app8010059.

[30] S. Fenton, "Automatic Mixing of Multitrack Material Using Modified Loudness Models," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), convention paper 10041.

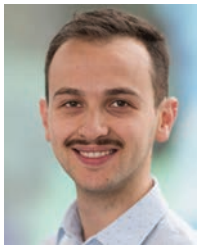
[31] B. De Man, J. Reiss, and R. Stables, "Ten Years of Automatic Mixing," presented at the *3rd Workshop on Intelligent Music Production* (2017).

[32] ITU-R Rec. BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems" (2015).

[33] N. Schinkel-Bielefeld, "Training Listeners for Multi-Channel Audio Quality Evaluation in MUSHRA With a Special Focus on Loop Setting," presented at the *8th Int. Conf. on Quality of Multimedia Experience (QoMEX)* (2016), doi:10.1109/QoMEX.2016.7498952.

[34] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. Shirley, "Background Ducking to Produce Esthetically Pleasing Audio for TV With Clear Speech," presented at the *146th Convention of the Audio Engineering Society* (2019 Mar.), convention paper 10175, doi:10.17743/aesconv.2019.978-1-942220-26-8.

THE AUTHORS


Matteo Torcoli


Alex Freke-Morin


Jouni Paulus


Christian Simon


Ben Shirley

Matteo Torcoli received his B.Sc. degree in computer engineering from Brescia University in 2011 and his M.Sc. degree in sound and music computer engineering from Politecnico di Milano in 2014, *cum laude*. He worked on his M.Sc. thesis on dereverberation for hearing aids at the International Audio Laboratories Erlangen. He then joined the Audio and Media Technologies division of Fraunhofer IIS, where he is currently working as an R&D engineer. His research focus is on applying digital signal processing and machine learning for developing more accessible and inclusive broadcasting and streaming services. In particular, he has been working on dialog enhancement, ways to enable it also without the original audio objects, and the subjective and objective evaluation of such an experience.

•

Alex Freke-Morin received his B.A Degree in television and radio production from the University of Salford specializing in audio production and broadcast in 2017. In 2018 he continued at the University of Salford for his M.Sc. in Audio Production with an interest in DSP and object-based audio. This led to a placement with Fraunhofer IIS (Erlangen, Germany); the success of the project (working with the authors of this paper) lead to the publishing of a conference paper, which won the AES Best Conference Paper Award in 2019.

•

Jouni Paulus received the M.Sc. (Eng.) and D.Sc. (Tech.) degrees in information technology from Tampere University of Technology (TUT) in 2002 and 2010, respectively. From 2002 to 2010 he was working as a researcher at the Department of Signal Processing at TUT with the topic of signal-based music content analysis. In 2010 he joined Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, as a research scientist, and as a member of the International Audio Laboratories Erlangen. Dr. Paulus has contributed to the development and standardization of MPEG-D SAOC and MPEG-H 3D Audio. His current research interests as a senior scientist at Fraunhofer IIS include object-based and spatial audio coding, informed and blind source separation, machine learning for audio applications, speech intelligibility enhancement, and subjective evaluation of the resulting audio processing algorithms.

•

Christian Simon received his Dipl.-Tonmeister in Audiovisual Media in 2010 from the Film University in Babelsberg, Germany and is a member of the SoundLab group at Fraunhofer IIS. He has 20 years of experience in audio production with a focus on mixing and dialog editing. With his award-wining startup Easy Listen, he was the first developer to realize a service for optimization of speech intelligibility for AV media in Germany. At present, his key focus is on Next Generation Audio, accessibility, and immersive mixing. Furthermore, he is a visiting lecturer at the Ansbach University of Applied Sciences.

•

Dr. Ben Shirley is a Senior Lecturer in audio technology at the Acoustics Research Centre, University of Salford, UK. He received his M.Sc. from Keele University in 2000 and his Ph.D. from the University of Salford in 2013. His doctoral thesis investigated methods for improving TV sound for people with hearing impairments. His research interests include audio broadcast, spatial audio, and audio-related accessibility solutions for people with sensory impairments.