Loudness differences for Voice-over-Voice audio in TV and streaming

Geary, D, Torcoli, M, Paulus, J, Simon, C, Straninger, D, Travaglini, A and Shirley, BG

10.17743/jaes.2020.0022

Loudness Differences for Voice-over-Voice Audio in TV and Streaming

Geary, David[1], Torcoli, Matteo[2], Paulus, Jouni[3], Simon, Christian[4], Travaglini, Allessandro[5], and Shirley, Ben[6]

[1]davidrichardgeary@icloud.com
[2]matteo.torcoli@iis.fraunhofer.de
[3]jouni.paulus@iis.fraunhofer.de
[4]christian.simon@iis.fraunhofer.de
[5]allessandro.travaglini@iis.fraunhofer.de, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[6]b.g.shirley@salford.ac.uk, University of Salford, Salford, UK

David Geary and Matteo Torcoli are co-first authors. This submitted paper was published in Journal of the Audio Engineering Society, Volume 68 Issue 11 pp. 810-818; November 2020. Publication Date: December 21, 2020, https://doi.org/10.17743/jaes.2020.0022

Voice-over-Voice (VoV) is a common mixing practice observed in news reports and documentaries, where a foreground voice is mixed on top of a background voice, e.g., to translate an interview. This is achieved by ducking the background voice so that the foreground voice is more intelligible, while still allowing the listener to perceive the presence and tone of the background voice. Currently, there is little published research on ducking practices for VoV and on technical details such as the Loudness Difference (LD) between foreground and background speech. This paper investigates the ducking practices of 9 expert audio engineers and the preferred LDs of 13 non-expert listeners of ages 57 years and older. Results highlight a clear difference between the LDs used by the experts and those preferred by the non-expert listeners. Experts tended towards LDs of 11.5-17 LU, while non-expert preferred a range of 20-30 LU. Based on these results, a minimum LD of 20 LU is recommended for VoV. High inter-subject variance due to personal preference was observed. This variance makes a substantial case for the introduction of personalization in broadcast and streaming. The audiovisual material used for the tests is provided at `https://www.audiolabs-erlangen.de/resources/2020-VoV-DB`.

1 INTRODUCTION

A common complaint from consumers in the media and broadcast sector is the lack of intelligibility of speech in TV, radio and film [1]. Often, the complaint is manifested as the speech *'not being loud enough'*. However, there are many variables involved which may impact speech intelligibility. This is especially problematic for consumers with hearing impairments [31]. The most common impairment is that of age-related hearing loss, which is the gradual degradation of hearing with age, and primarily affects the ability to hear high-frequency sounds, which are important for speech [10]. Furthermore, for many of these listeners, access to TV and other media can provide necessary social inclusion, with clear audio being rated as *very important* in a recent survey of hearing-impaired listeners [24]. For this reason, a practical solution to this problem is of great importance.

In recent years, swift technological advancements have led to significant developments in solving this issue. One of these is Object-Based Audio (OBA). OBA systems, such as MPEG-H Audio, solve this problem by enabling the audience to personalize the relative level of the foreground speech (or narrative dialog) with respect to the background [9]. This is known as Dialogue Enhancement and has been shown to clearly enhance quality of experience for the end-user [28, 27, 20, 23]. Even where object-based personalization is available, a default mix is required as a starting point, from which dialog enhancement can proceed. At present, there is no standard for the levels of foreground speech and background elements in broadcast; only rough guidelines exist, in which the foreground speech should be considered *clear'* [3]. As a result, there is a degree of variation in the sound mix from program to program. To address this, recent research has focused on looking at the relationship between the relative levels of speech and background sound. This is a complex issue and current evidence suggests that the appropriate relative levels between foreground and background are variable and depend on a number of factors including personal taste, listening environment (including playback system), program genre and content language [7, 17, 6]. Furthermore, different types of background sound (e.g., music, ambience, speech) may require different treatments [26, 25].

Voice-over-Voice (VoV) audio is a special case where both foreground and background audio contain speech. VoV cases are required in broadcast media for a number of applications. These include translation of interviews into foreign languages and spoken subtitles. VoV has been explored to only a limited extent in contrast to work covering Voice-over-Music and Voice-over-Ambience [26, 25].

In order to control the trade-off between clear speech yet retaining enjoyable background, a technique known as *ducking* is commonly implemented. This process involves a time-varying attenuation of the background audio when the foreground speech is present. This can be performed, for example, by manual volume control, automated sidechain attenuation or downward compression, or a combination of them. Sidechain compression is a tool commonly used by audio engineers in a music mixing context to *duck* selected audio using another audio source as a trigger [19]. Using sidechain compression in a VoV application, the foreground speech would act as a key input for attenuation or compression which then lowers the background audio level.

The parameter which dictates the relative levels between background and foreground audio is known as the Loudness Difference (LD), where an integrated loudness measurement, as defined by ITU BS.1770-4 [13], is adopted and is represented in Loudness Units (LU). The automated sidechain ducking features a number of time-based parameters which affect the envelope of the attenuation. Attack, hold and release controls affect the onset, sustain and release times respectively for the ducking process. While they do not affect the level of attenuation substantially, they are important as they affect the quality and enjoyment of the listening experience.

There is currently an increasing requirement for producing more accessible audio-visual content in broadcast media, such as TV and online services. This is partly due to the ageing demographic of TV audiences. With this in mind, the aim of this paper is to provide a comprehensive examination of how ducking is used, specifically in VoV scenarios. It has been previously stated that the most effective strategy for improving intelligibility of speech in broadcast and streaming is by increasing the LD or the speech to noise ratio (foreground to background), especially when object-based personalization options are provided [31]. Nevertheless, default sound mixes are still a requirement and need to be optimized for the maximum number of end users. On the basis of the experimental results obtained from the present work, recommendations for VoV audio are discussed and outlined. All the material used in the following experiments is available at: https://www.audiolabs-erlangen.de/resources/2020-VoV-DB.

2 LITERATURE REVIEW

The primary factor which affects speech intelligibility is background audio *masking* the foreground audio. This masking process occurs in two distinct classes.

The first of these, termed energetic masking (EM), is the *physical* loss of information due to the spectral overlap between a foreground source and a background masker signal. The extent of this masking is strongly dependent on the signal to noise ratio between the two signals. The second process known as informational masking (IM) is a higher-order cognitive phenomenon which is broadly described as the similarity between the target (foreground sound) and the masker (background sound) making it difficult for the listener to selectively attend to the correct source [2, 15]. VoV scenarios are complex as not only does EM occur but IM is also commonly present which makes it especially problematic for intelligibility [5]. Speech is an elaborate sound source that contains rapid frequency and amplitude modulation. Therefore, when two speech sources are played concurrently, it is fundamentally difficult to perceptually separate the two signals. These masking issues are challenging, particularly for older people [14, 8].

For VoV scenarios, voice characteristics are important. Using different sexes having differing vocal timbres for target and masking voices has been shown to result in higher intelligibility than using same-sex voices. Additionally, for the same sex voices condition, using different speakers for the target and mask is found to be more intelligible than if both the target and mask were the same speaker. Finally, signal to noise ratios have less influence on speech intelligibility with speech background masks as opposed to noise background masks, due to the presence of informational masking [4].

LD recommendations have been made previously for VoV of between 16-23 LU [22]. This recommendation was based on [16], where non-experts of a median age of 43 adjusted LD levels set by audio engineers to their preferred levels. Two groups were identified in the results, one which preferred lower LDs (mean of 14 LU) and one which preferred higher LDs (mean of 20 LU).

The aforementioned literature has focused on the static interaction between background and foreground audio. While this is important, ducking is commonly used for VoV content. In [26], the primary focus was looking at desirable LDs during ducking. Observational analysis was conducted on a selection of TV documentaries which featured VoV, Commentary-over-Music (CoM), Commentary-over-Ambience (CoA) and Dialog-over-Music (DoM). The range of LDs found amongst the documentaries were from 10-17 LU, with the highest LDs from DoM.

Interestingly the VoV samples [26] had the lowest observed LD values despite being most susceptible to both IM and EM [15]. Subjective testing was also conducted involving non-expert and expert listeners exploring desirable LDs where participants could rate different LD mixes depending on their preference. In this case, only CoM and CoA were investigated. A difference was identified between non-expert and expert listeners where the former preferred LDs on average 4 LU higher than the latter. Recommendations for ducking LD were made based on these findings. For CoM, an LD of at least 10 LU was suggested whereas an LD of 15 LU was advised as being suitable for CoA.

The evidence from the literature is that the LD between foreground and background audio is not a one-size-fits-all solution. The following experiments aim, firstly, to observe VoV mixing practices of experts in an attempt to verify and explain the LD values reported in [26]. Secondly, to discover LD preferences for VoV content of older individuals who suffer the most with speech intelligibility in broadcast and streaming.

3 MIXING TEST

This section outlines the design of the expert *mixing test* conceived to retrieve preferred VoV LD ducking levels as performed by professional audio engineers for various program mixes. The idea was to replicate real broadcast mixing conditions as closely as possible.

3.1 Test Format

Method: The test interface consisted of an AVID Pro Tools session, containing the item video and corresponding audio files. Subjects were instructed to mix the audio, as they would for a professional broadcast production, using only manual volume automation of the background speech. Once the tests were complete, the integrated LDs were measured by isolating and concatenating parts of the background where foreground speech was present. The NUGEN Audio VisLM Loudness Meter [29] was used to measure the integrated loudness across the isolated background region.

Test Items: The test involved 6 items, each consisting of a video and 2 mono audio files. The background audio files were a male English dialog recording and the foreground audio either a male German studio voice-over or a male German synthetic speech voice-over, recorded at a sampling rate of 48 kHz. The synthetic voice-over was generated with Google WaveNet. The reason for considering synthetic speech in addition to real speech is that this is sometimes used for spoken subtitles, so as a special case of VoV. The intention is to understand if there are differences in how engineers treat real and synthetic speech. There were three different English dialog recordings. These were recorded in different locations to create three background dialog items with different signal to noise ratios. The three location conditions were a dry, acoustically treated room, a reverberant hallway and a noisy outside location. These conditions were chosen, as they represented a range of broadcast content. Each foreground and background dialog clip was normalized to -24 integrated LUFS. All test material was recorded at Fraunhofer IIS.

Subjects: The subjects consisted of 9 individuals, each with broadcast mixing experience at a professional production level. Five of them were based at Fraunhofer IIS and four were external engineers. Eight of them were native German speakers. None of the subjects had any known hearing impairment.

Location: The internal tests took place in a treated studio environment at Fraunhofer IIS using Dynaudio BM6A studio monitors, in a stereo reproduction format. The external candidates were asked to use their own studio environments.

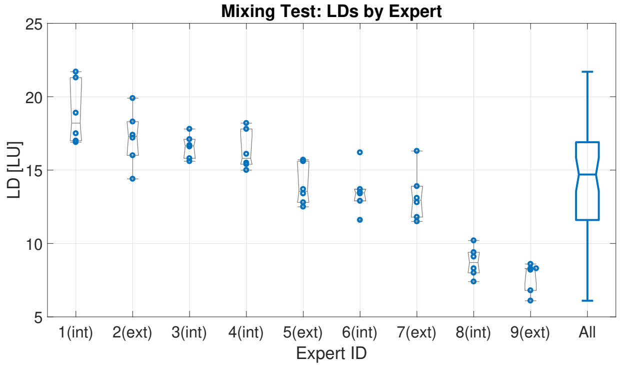

Figure 1: Boxplots and single data points depicting the LDs used by each subject for the 6 VoV test items, with the right-most boxplot showing all data points. Each subject is given a label of either (int) or (ext) which indicates whether they were an internal or an external engineer. The median LD for all subjects is 14.70 LU and half the per-subject medians lie in the 11.5-17 LU region.

Effect	d.f.	η^2	p
Subject	8	0.873	**0.00**
Background location	2	0.011	0.15
Voice-over type	1	0.000	0.94
Error	42	0.116	

Table 1: ANOVA test of the expert chosen LDs: degrees of freedom (d.f.), effect size η^2 (percentage of explained variability) and p-values (<0.05, reject the null hypothesis). Significant factors are outlined in bold.

3.2 Results and Discussion

The following results are all in relation to the measured LD, which is the difference in integrated loudness level between the foreground speech and the ducked background speech, and were compared with three factors: *Subject*, *Background location* (dry, reverberant, outside), and *Voice-over type* (natural, synthetic).

The data for all the subjects is shown in Figure 1. The median LD value is 14.70 LU while the interquartile range is from 11.5 to 17 LU, which is similar to the range observed by [26].

What is obvious from Figure 1 is that there is considerable inter-subject variability. The range in median values across all subjects is 9.95 LU which is high, despite half the per-subject medians lying in the 11.5 to 17 LU region. On the other hand, the intra-subject variability, or the range in LDs for individual subjects was quite small, an average of 3.74 LU. This suggests that each subject had a desired relative level for the foreground and background and achieved that level with consistency. Also, the intra-subject variability was similar across subjects (Levene's test p-value is 0.24). The inter-subject variability is a problem, as it explains the difference in sound mixes from program to program on TV, and highlights that appropriate LD levels are highly personal, even for expert sound engineers.

An ANOVA was run on the measured LDs to test the statistical significance of the three factors (Jarque-Bera normality test gave p-values greater than 0.5 for all considered groups). The 95% significance level (p = 0.05) was used. *Subject* was considered a random factor.

The results are presented in Table 1. Out of all factors, only *Subject* was statistically significant, accounting for 87.3% of the variance in the data. Neither *Location* nor *Voice-over type* were statistically significant. The overwhelming importance of the factor *Subject* can be explained by the different personal tastes of the audio engineers. The different mixing locations (and so different equipment, different setting, different overall volume, etc.) might also have had an impact. However, high inter-subject variability is observed also for the 5 engineers who performed the mixing task in the same location, labeled with *(int)* in Figure 1. This suggests that the importance of the mixing location is small compared to the effect of personal taste, but this conclusion should be corroborated on a broader sample size.

A further observation on this experiment is that the ducking envelopes seem to add to the variability of the LD values. Automation curves from different experts varied considerably. Some experts mixed the material in a static manner, with fast attack and release times while others had a more gentle approach with longer and more variable attack and release times. Further studies focusing on ducking time parameters are needed to explore this in detail.

4 LISTENING TEST

This section describes the VoV *multiple stimuli listening test* featuring non-experts between the ages of 57 and 75.

4.1 Test Format

Method: The test consisted of a multiple stimuli test where each condition had a different LD. Subjects rated their preference for each condition with a slider ranging from 0-100 and with labels at each 20 points for indicating the following ranges: bad, poor, fair, good and excellent, in accordance with a MUSHRA test [12]. This scale was translated and written in German, to make it understandable for all subjects. The translations were adopted from [30]. While there are a number of biases associated with the MUSHRA test methodology [32], using a derivative of it was considered most suitable for this test, especially considering the interface's ease of use by older individuals [30]. The test was completed in one sitting of approximately half an hour in length. A short questionnaire was completed after the test to provide feedback.

Instructions: The instructions for the test were written as follows, translated from the German original:

For each excerpt, different volume balances between the two voices are presented. Your task is to rate these different balances based on your overall preference. This means that you rate the balance that you would personally rather hear with the highest score. A chosen score below 40/fair is indicative of a balance you wouldnt accept from a broadcast program. You may rate multiple balances at the same score (including 100). For this test, the exact values of the slider scores are not important. The focus is more on the general ranking of each volume balance with respect to the *others. During the first item, please adjust the overall volume to a comfortable level for you, as you would do while watching television. Do not change this level after the first item.*

Test Items: The test featured 9 items, including two repeat items. These repeats were included to monitor any irregularities in the condition scoring of each subject. Each item was comprised of 15-second long segments of the speech material used in the expert mixing test. Therefore, three items were used from each location condition mentioned in Section 3. In total there were 8 LD conditions per item, ranging from 0 to 30 LU. In these conditions, the foreground speech remained at the same level and the background speech was ducked accordingly. The conditions as well as the items were presented in a random order. No accompanying video was shown in order to make the test easier for the participants and based on [16], where no significant difference was found for a similar task, with or without video. As time constants for the ducking, attack = 500 ms and release = 500 ms were used. All items had a sampling rate of 48 kHz.

Subjects: Thirteen subjects between the ages of 57 and 75 (median age 69) took part in the listening test. All were native German speakers, with varying levels of English language understanding, based on self-assessment. Each subject took an audiogram prior to starting the test and their hearing acuity was categorized as per the WHO audiometric descriptors [18]. This was achieved by averaging the hearing threshold level of the better ear, at 500 Hz, 1 kHz, 2 kHz and 4 kHz. Audiogram scores can be seen in Figure 2. The subjects were also post-screened based upon their preferred LD ratings on the repeat items. The preferred LD for each item is the the LD receiving the highest preference rating (or the mean in case multiple LDs shared the highest rating). The criterion was if the difference between the preferred LD for an item and its repeat exceeds 6 LU, the subject was excluded. According to this, one subject was excluded from the final analysis. Of the remaining subjects, six individuals had no impairment and six had a slight impairment. Interestingly, all individuals with no impairment were 69 years old or younger, while all individuals with slight impairment were older.

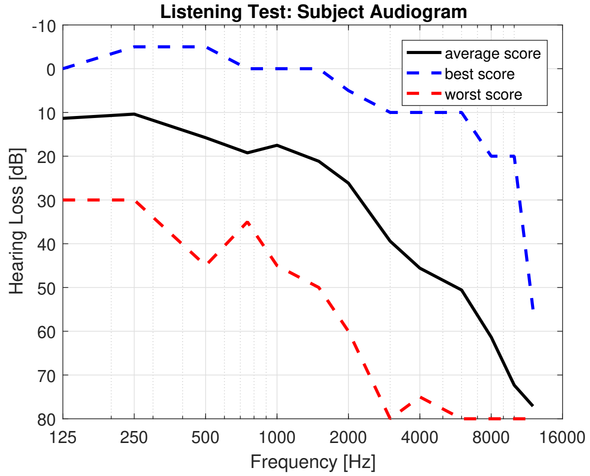

Figure 2: Average, best and worst audiogram scores between 125 Hz and 16 kHz for the listening test subjects. The increased hearing loss in the high frequencies is indicative of typical age-related hearing loss.

Location: An acoustically treated listening room adhering to ITU-R BS.1116 specification [11]. Dynaudio BM12A stereo monitor loudspeakers were used.

4.2 Results and Discussion

Figure 3 shows the distribution of the mean ratings given by the subjects for each LD presented in the listening test. Overall, each participant either displayed a positive linear trend with increasing LD, or a bell shaped distribution with the peak around the preferred LD, which is especially evident with subject 12. The size of the yellow/orange region (or a rating between 70 and 100) can be regarded as an accepted range, where the LD is considered satisfactory, even if not the preferred choice. There is a lot of variation between the participants and their use of the rating scale.

The main results are shown in Figure 4. For each subject, the preferred LD conditions for each item are plotted. The preferred LDs are the condition(s) with the highest score for each subject and item. In cases where more than one condition was given the highest score, the average LD was taken between the two conditions.

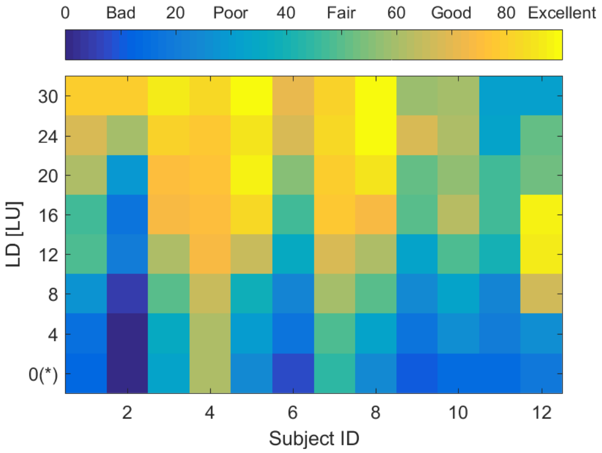

Figure 3: Distribution of the mean ratings given by the subjects to the different LDs in the listening test. The rating is color-coded and explained by the horizontal bar. 0(*) indicates the initial condition where no ducking is applied. The median LD for this condition is 0.35 LU, but it ranges between -0.15 and 3.5 LU depending on the item.

These results outline a strong preference for higher LDs for VoV content for older listeners. The interquartile range for all subjects ranges from 20 to 30 LU with a median value of 24 LU. Like the experts, high inter-subject variability is present. On an intra-subject scale, there is a degree of variability in preferred LD. The variance is larger than in the expert data, however this is to be expected, due to the experts having more experience in audio and critical listening.

Subjects 11 and 12 appear to be outliers in the data, preferring a range between 14 to 22 LU. An explanation for this is alluded to by subject 12 who stated in the questionnaire that they liked hearing the background dialog *"not too quiet and not too loud"*, and also mentioned they had never had previous problems understanding speech in TV. Moreover, they were the youngest subject in the test.

An ANOVA was run on the preferred LDs to check statistical significance. The factors tested were *Subject*, *Background location*, and *Hearing acuity* (no impairment, slight impairment). *Subject* is a random factor and it is nested inside *Hearing acuity*. The Jarque-Bera normality test gave p-values greater than 0.11 for all groups with the only exception of subject 3. This subject always preferred 30 LU (8 times) but once 27 LU was preferred. We consider this almost degenerate distribution as an acceptable deviation from the normality assumption. Moreover, we check the validity of the main conclusion from the ANOVA with a non-parametric test.

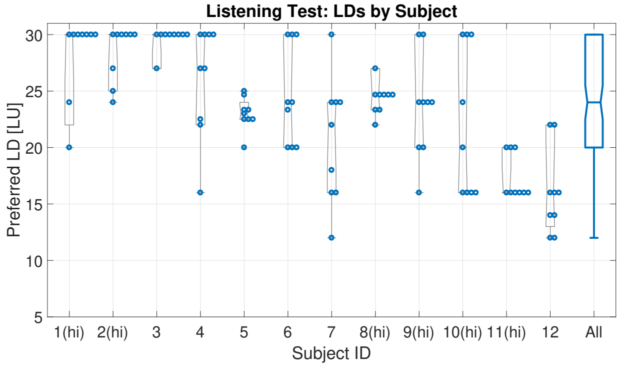

Figure 4: Boxplots and single data points showing the preferred LDs for all items per subject, with the rightmost boxplot showing all preferred scores. The subject ID is followed by (hi) if the subject is categorized as having slight hearing impairment.

Effect	d.f.	η^2	p
Subject	10	0.548	**0.00**
Background location	2	0.004	0.67
Hearing acuity	1	0.004	0.80
Error	94	0.444	

Table 2: ANOVA test of the preferred LDs: degrees of freedom (d.f.), effect size η^2 and p-values (<0.05, reject the null hypothesis). Significant factors are outlined in bold.

The ANOVA results presented in Table 2 show that *Subject* is the only significant factor, explaining 54.8% of the variability. The statistical significance of this factor is confirmed also by a Kruskal-Wallis test considering subjects as groups ($p = 0.00$). As in the expert mixing test, *Background location* was not statistically significant. Also *Hearing acuity* was not statistically significant. Both conclusions were confirmed by closely inspecting the data points. Especially the fact that hearing acuity was not statistically significant might be surprising at first. However, generally speaking, this does not imply that hearing acuity is not an important factor in the preference of LDs. Our ANOVA only shows that this factor is not statistically significant in this test. We believe that this can be explained by the homogeneous age range of the subjects, and by the fact that their hearing acuity was similarly close to the classification threshold, based on which the harsh division into no-impairment and slight-impairment was done. In fact, at least some degree of age-related hearing loss affect all non-expert subjects in this study, as indicated by the decline in hearing acuity after 2 kHz, shown in Figure 2. Including a control group of listeners without any hearing loss above 4 kHz might help in successfully rejecting the null hypothesis for this factor.

An analysis including *English level* was also carried out, but it was not found to be statistically significant, so it is excluded from the presented analysis. The variable itself could be considered somewhat unreliable as it was a self estimation. The real impact of understanding both sets of VoV languages, therefore, may still be untold.

5 COMPARATIVE DISCUSSION

A summary of the results from the experiments conducted in this study can be seen in Figure 5, alongside previous literature results and recommendations. Between the experts and non-experts there is an observable difference in preferred LD levels for VoV content. This result, not only corroborates the findings of [22] and [16] with regard to expert and non-expert preferences, but reveals a greater disparity between the two groups. While the expert LDs are in similar ranges between our results and the previous studies [22, 16], the non-expert LD preferences are a lot higher in our results compared with the prior. One clear difference to explain this is the difference in age (and possibly hearing acuity) of the participants between the experiments. Nevertheless, the subject is the most important determining factor, and is demonstrated by the high inter-subject variability. This again emphasizes the need for personalization options when it comes to VoV content.

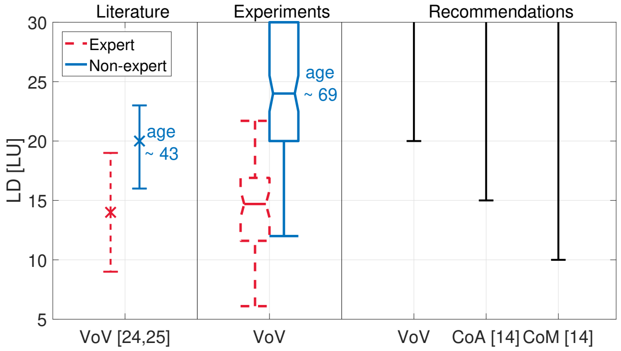

Figure 5: Preferred integrated Loudness Difference (LD) between foreground speech and background. Visual representation of Voice-over-Voice (VoV) preferences found in the literature, results obtained from this study, and resulting recommendations for VoV alongside Commentary-over-Ambience (CoA) and Commentary-over-Music (CoM) [26].

6 CONCLUSION

A common complaint from the public regarding broadcast content remains that foreground speech is often too quiet, or the background too loud. More simply, this is a speech intelligibility issue where some listeners struggle to understand speech under certain conditions. A test was designed and conducted to gain a more detailed evaluation of expert mixing practices. Nine sound engineers with experience in broadcast production were recruited and asked to mix 6 VoV passages. The findings showed that while most experts mixed the material with an LD in the 11.5-17 LU range, there was still significant inter-subject variability, ranging from 6 to 22 LU. However, on an intra-subject scale, variability tended to be small which highlights the mixers consistency for their personal preferred LD level. A nested ANOVA revealed that subject was the only statistically significant factor.

Ultimately, all content needs to be appropriate for the everyday consumer, of which the average age is increasing, especially for TV content. Due to this, an increasing portion of the audience of TV and streaming will have some form of hearing impairment [23, 21]. A subsequent listening test addressed to older non-experts was conducted. It consisted of a multiple stimuli listening test using the same VoV content as the previous expert test. Twelve subjects between the ages of 57 and 75 took part in the test containing 9 VoV items where they had to rate different LD conditions. The results showed a clear preference of LDs in the 20 to 30 LU range, almost 10 LU higher than those chosen by experts, with large inter and intra-subject variability. Variability is observed in the data of both experiments. This accentuates the problematic mismatch between the expert LDs and the preferences of older listeners, and highlights further that personal taste is a primary factor impacting preferences for listening to broadcast audio content. Consequently this emphasizes further the benefit of introducing personalization in broadcasting. This can be facilitated through object-based audio technology, such as MPEG-H Audio.

Nevertheless, a default mix is still required for providing a starting balance set to meet average preferences. A minimum LD value of 20 LU for VoV can be suggested, corresponding to the first quartile of the preferences of the non-expert subjects from this study. Non-expert subjects are considered for formulating this suggestion, as they represent the vast majority of the consumers of this type of audio signals. This recommendation can be seen alongside recommendations for CoA and CoM from [26] on the right-hand side of Figure 5, where a minimum LD of 15 LU is recommended for CoA and at least 10 LU is recommended for CoM.

Additional work is required to discover the importance of the time constants for ducking and how they impact the enjoyment of broadcast programs. Furthermore, both tests were conducted in a stereo reproduction format, as this remains the most common medium for media consumption. However with increasing access to multichannel and object-based audio reproduction, further work on ducking in these different formats is recommended.

7 ACKNOWLEDGEMENT

As with all studies, the research contributions extend further than that of just the authors. Many thanks to all the internal and external engineers who took part in the mixing test, as well as all those who took part in the listening test. Special thanks to Mariola Hatalski and the rest of the listening test team for organizing the experiment.

References

[1] M. Armstrong. *From Clean Audio to Object Based Broadcasting*. 2016.

[2] Jon Barker and Xu Shao. “Energetic and informational masking effects in an audiovisual speech recognition system”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17 (3) (2009), pp. 446–458.

[3] BBC Academy. *Sound Matters: Getting it right for our audience*. 2017.

[4] Douglas S. Brungart. “Informational and energetic masking effects in the perception of two simultaneous talkers”. In: *The Journal of the Acoustical Society of America* 109 (3) (2001), pp. 1101–1109.

[5] Raymond Carhart, Tom W. Tillman, and Elizabeth S. Greetis. “Perceptual masking in multiple sound backgrounds”. In: *The Journal of the Acoustical Society of America* 45 (3) (1969), pp. 694–703.

[6] Mary Florentine. “Speech perception in noise by fluent, nonnative listeners”. In: *The Journal of the Acoustical Society of America* 77 (S1) (1985), S106–S106.

[7] Harald Fuchs, S. Tuff, and C. Bustad. “Dialogue enhancementTechnology and experiments”. In: *EBU Technical review* 2 (2012), p. 1.

[8] Karen S. Helfer and Richard L. Freyman. “Aging and speech-on-speech masking”. In: *Ear and hearing* 29 (1) (2008), p. 87.

[9] Jürgen Herre et al. "MPEG-H 3D audioThe new standard for coding of immersive spatial audio". In: *IEEE Journal of selected topics in signal processing* 9 (5) (2015), pp. 770–779.

[10] Qi Huang and Jianguo Tang. "Age-related hearing loss or presbycusis". In: *European Archives of Oto-rhino-laryngology* 267 (8) (2010), pp. 1179–1191.

[11] ITU-R. "BS. 1116-3. Methods for the Subjective Assessment of Small Impairments in Audio Systems." In: *International Telecommunication Union - Radiocommunication Sector* (2015).

[12] ITU-R. "BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)". In: *International Telecommunications Union, Geneva* (2001).

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

[14] Tomoyasu Komori et al. "An Investigation of Audio Balance for Elderly Listeners using Loudness as the Main Parameter". In: *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[15] Marjorie R. Leek, Mary E. Brown, and Michael F. Dorman. "Informational masking and auditory attention". In: *Perception & psychophysics* 50 (3) (1991), pp. 205–214.

[16] 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)". In: *Proc. of the 28th Tonmeistertagung-VDT Int. Conv* (2014). (in German).

[17] Peter Mapp. "Intelligibility of Cinema & TV Sound Dialogue". In: *141st Audio Engineering Society Conv., Los Angeles*. Audio Engineering Society, 2016.

[18] Colin Mathers, Andrew Smith, and Marisol Concha. "Global burden of hearing loss in the year 2000". In: *Global burden of Disease* 18 (4) (2000), pp. 1–30.

[19] Bobby Owsinski. *The mixing engineer's handbook*. Nelson Education, 2013.

[20] Jouni Paulus et al. "Source Separation for Enabling Dialogue Enhancement in Object-based Broadcast with MPEG-H". In: *J. Audio Eng. Soc* 67 (7/8) (2019), pp. 510–521.

[21] Christian. Simon, Matteo. Torcoli, and Jouni Paulus. "MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming". In: *arXiv preprint arXiv:1909.11549* (2019).

[22] *Sprachverständlichkeit im Fernsehen, Empfehlungen für Programm und Technik (Intelligibility in Television, Recommendations for TV Program and Technique)*. (in German). ARD/ZDF. 2014.

[23] Davide Straninger. "Dialogue Enhancement in Object-based Audio – Evaluating the Benefit on People above 65". In: *arXiv preprint arXiv:2006.14282* (2020).

[24] Olaf Strelcyk and Gurjit Singh. "TV listening and hearing aids". In: *PLOS ONE* 13 (6) (2018).

[25] Matteo Torcoli et al. "Background ducking to produce esthetically pleasing audio for TV with clear speech". In: *Audio Engineering Society Convention 146*. Audio Engineering Society. 2019.

[26] Matteo Torcoli et al. "Preferred Levels for Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech". In: *Journal of the Audio Engineering Society* 67 (12) (2019), pp. 1003–1011.

[27] Matteo Torcoli et al. "The adjustment/satisfaction test (A/ST) for the evaluation of personalization in broadcast services and its application to dialogue enhancement". In: *IEEE*

Transactions on Broadcasting 64 (2) (2018), pp. 524–538.

[28] Matteo Torcoli et al. “The adjustment/satisfaction test (A/ST) for the subjective evaluation of dialogue enhancement”. In: *143rd Audio Engineering Society Conv., New York*. 2017.

[29] *VisLM | Nugen Audio*. Accessed: 2020-06-22.

[30] Christoph Völker and Rainer Huber. “Adaptions for the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants”. In: *DAGA, Mar* (2015).

[31] L. Ward and B. G. Shirley. “Personalization in object-based audio for accessibility: a review of advancements for hearing impaired listeners”. In: *Journal of the Audio Engineering Society* 67 (7/8) (2019), pp. 584–597.

[32] Slawomir Zielinski, Francis Rumsey, and Søren Bech. “On some biases encountered in modern audio quality listening tests-a review”. In: *Journal of the Audio Engineering Society* 56 (6) (2008), pp. 427–451.

Authors

David Geary is an audio technology Ph.D. student at the University of York, studying device orchestration in immersive audio experiences. He received his M.Sc. at the University of Salford in 2019 and worked on his M.Sc. thesis in collaboration with Fraunhofer IIS. His research interests include object-based audio and audio personalization and accessibility solutions.

Matteo Torcoli is an R&D engineer at the Audio and Media Technologies division of Fraunhofer IIS. He received his B.Sc. in computer engineering from Brescia University in 2011 and his M.Sc. in sound and music computer engineering from Politecnico di Milano in 2014 with highest honors. He worked on his M.Sc. thesis on dereverberation for hearing aids at the International Audio Laboratories Erlangen. His current focus is on audio signal processing and deep learning for developing more accessible and inclusive broadcasting/streaming services. He has been working on dialog enhancement, on ways to enable it also without the original audio objects, and on the subjective and objective evaluation of the resulting user experience.

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 include object-based and spatial audio coding, informed and blind source separation, machine learning for audio applications and speech intelligibility enhancement, and subjective evaluation of the resulting audio processing algorithms.

Christian Simon is a Tonmeister working as scientist and member of the Soundlab group at Fraunhofer IIS. He received his Diploma from Film University Potsdam-Babelsberg in 2011. He has 20 years of experience in audio production with a focus on mixing and dialogue editing. His current field of work is on Next Generation Audio and accessibility for AV media.

Davide Straninger received his B.A. degree in Multimedia and Communications specialising in audio and media technology in 2020 at the University of Applied Sciences in Ansbach, Germany. Since 2017 he has been working as a research assistant at Fraunhofer IIS in the SoundLab group. His Bachelor thesis investigated dialogue enhancement in object-based audio for people above 65. At present, he is studying Audiovisual Media with focus on Audio for a Master of Engineering at the Stuttgart Media University.

Alessandro Travaglini works as researcher at Fraunhofer IIS since 2019 where he joins the team developing algorithms and tools for MPEG-H authoring and processing. With more than 20 years background in sound design, audio production and post-production in broadcasting, he received his B.A Degree cum laude in Digital Music and Audio Technologies from Conservatorio Licino Refice in Frosinone in 2018 with a thesis on Next-Generation Audio and object-based technologies. He is a fellow member of several standard committees and international bodies including AES, EBU and CTA and has actively contributed to the definition of a few international recommendations with specific focus on loudness measurement and processing, including EBU R128, APEX-0415 and AGCOM-219/09/CSP.

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.