Audio Engineering Society

Convention Express Paper 84

Presented at the 154th Convention

2023 May 13-15, Espoo, Helsinki, Finland

This Express Paper was selected on the basis of a submitted synopsis that has been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This express paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Dialogue Enhancement with MPEG-H Audio: An Update on Technology and Adoption

Daniela Rieger, Christian Simon, Matteo Torcoli, and Harald Fuchs

Fraunhofer Institute for Integrated Circuits, Am Wolfsmantel 33, 91058 Erlangen, Germany

Correspondence should be addressed to Daniela Rieger (daniela.rieger@iis.fraunhofer.de)

ABSTRACT

Difficulties in following speech on TV due to loud background sounds are a common issue in broadcasting. Object-based audio (OBA) systems like MPEG-H Audio can solve this problem by providing a personalized speech level. Recently, international broadcasters have employed dialogue enhancement (DE) together with OBA, providing customization and improved accessibility to their audiences, e.g., during the football World Cup 2022. To also add customizable dialogues to material produced without OBA, deep neural networks (DNNs) can be applied to separate dialogues from the music and effects of the final audio mix. One of the technologies used for this is MPEG-H Dialog+, which has recently been adopted for the new "Clear Speech" service of the on-demand platform of the German public broadcaster ARD. This paper reviews the current state of DE, detailing real-world adoptions, with particular focus on the MPEG-H Audio system. The intention is to provide an up-to-date overview of successful implementations of DE solutions into production workflows as an example for further adoptions and developments.

1. Introduction

Difficulties in understanding speech on TV and the resulting complaints are a common problem that broadcasters have been dealing with for more than 30 years [1]. The urgency of these complaints has been confirmed by a study with over 2,000 participants carried out by Westdeutscher Rundfunk (WDR[1]) and Fraunhofer IIS in 2020 [2]: It showed that 68% of all participants, and 90% of the subjects aged 60 and above, often or very often had problems to understand speech on TV (see Figure 1). 83% of all study participants liked the possibility to switch to a dialogue-enhanced version, including those who do not normally struggle with speech intelligibility. This shows that the option to personalize sound is much appreciated. Using object-based audio (OBA), the MPEG-H Audio system can offer such options by providing an improved speech level that can also be adapted by the user. Additionally, the study highlights the need to provide dialogue enhancement (DE) for the large amount of existing content which contains only the final audio mix. In this case, deep-learning based solutions such as MPEG-H Dialog+ offer the possibility to automatically separate speech from the background and remix the content to a new speech-enhanced version that is easier to understand. In 2022, the German broadcaster ARD adopted MPEG-H Dialog+ for the on-demand segment of their new

[1] WDR is a constituent member of ARD, the joint organisation of Germany's regional public-service broadcasters.

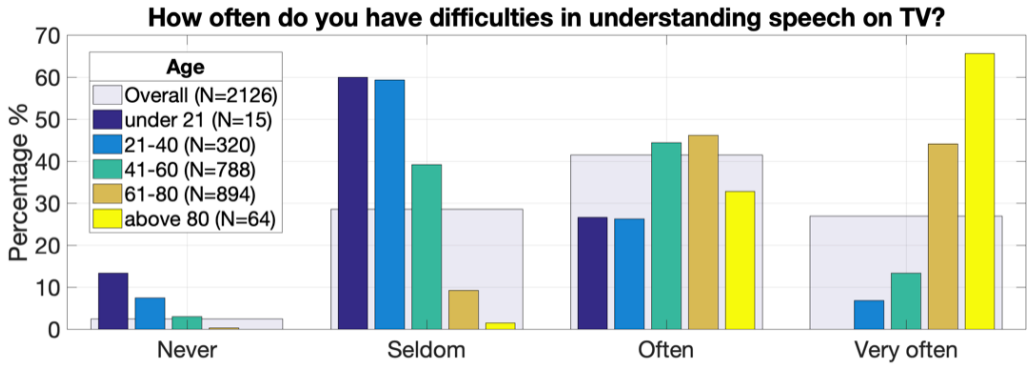

Figure 1. Survey by WDR and Fraunhofer IIS on understanding speech on TV carried out on a national scale in Germany [2].

accessibility service "Clear Speech" [3] (discussed in Sec. 4.2). Other European content providers are building similar DE services, for example the Swedish public broadcaster SVT [4].

This paper reviews the current state of object-based personalization and related technologies (Sec. 2.1). It details real-world adoptions of DE technologies with emphasis on the MPEG-H Audio system (Sec. 2.2). For that purpose, it examines cases where an object-based production is available as well as cases in which the objects are estimated by a deep neural network (DNN) (Sec. 3). It is meant to provide an overview of successful implementations of DE technology into existing production workflows in Brazil and South Korea (Sec. 4.1), as examples for further adoptions and developments.

2. Object-based Dialogue Enhancement

Sound mixing is an important asset for TV productions, as the composition of music, effects, and ambience not only serves as a background for the dialogue but is part of the storytelling. It is a complex task to produce content that, on the one hand, immerses viewers into the scene by providing a compelling mix, and, on the other hand, prevents loud background sounds from masking speech, allowing the audience to fully understand the dialogue without high listening effort. The topic of DE therefore has a long history, dating back more than 30 years.

2.1 Related Works

As early as 1991, BBC Research & Development published a study documenting regular complaints about speech that was difficult to understand – mostly due to loud background noises and music [1]. The goal of the study was to understand how much the background level would have to be lowered to significantly improve the intelligibility of speech on TV. However, the results were inconclusive; it only became clear later that the optimum volume difference between dialogue and music & effects (M&E) is highly personal. The study also pointed out that the broadcasting system at that time was not capable of transmitting an additional track with increased speech level [1].

In 2011, the BBC and Fraunhofer IIS jointly conducted a public field test during the Wimbledon Tennis Championships, in which viewers had the opportunity to personalize the dialogue level. The results showed that while one part of the subjects preferred to clearly enhance the dialogue level, the other part of the viewers chose to increase the ambience sounds [5]. The preferred loudness differences between dialogue and background were investigated in more detail in further studies that also showed a significant difference between expert and non-expert listeners [6].

Furthermore, in 2019, the BBC conducted a public trial together with the University of Salford, using a narrative importance approach [7]. That way, non-speech sounds are grouped together based on their importance to the narrative, preventing the loss of significant sounds due to global background attenuation [8]. Participants in the trial could not only enhance the dialogue, but also narratively important sounds. The trial was followed by a survey, showing that out of the 299 participants 73% rated the content more enjoyable or easier to understand when using the personalization feature.

2.2 MPEG-H Audio

MPEG-H Audio is a next generation audio system based on the open international standard ISO/IEC 23008-3, MPEG-H 3D Audio [9]. It is part of major broadcasting standards such as DVB [10], ATSC 3.0 [11], and SBTVD [12] [13]. MPEG-H Audio supports audio scenes consisting of object-based content as well as channel-based and scene-based content, or any combination of them, and thus enables an immersive audio experience via broadcast, streaming, and music services. Besides immersive sound, MPEG-H Audio offers personalization and "Universal Delivery", the latter allowing the optimal rendering of a production on all kinds of devices [14]. To enable these personalized sound experiences, MPEG-H Audio relies on metadata, which is produced (or "authored") and transmitted together with the audio content. A renderer provides an audio playback that is optimized for the individual user, end device, and surroundings [15].

Due to the object-based nature of an MPEG-H Audio scene, the individual components of the mix, such as dialogue and M&E, are not mixed together before transmission, but represented separately within the MPEG-H Audio stream. Thus, an MPEG-H Audio receiving device allows the individual adjustment of these components during rendering and playback. The personalization options are defined during production, e.g., the configuration of "presets", which are different representations of the audio scene based on metadata and user interaction [14]. Users could, for instance, select a preset with dialogue enhancement, where the relative level of M&E is attenuated when dialogue is active. Additionally, broadcasters can allow users to individually adjust the levels of dialogue and background objects in defined ranges, enabling them to set their preferred balance. To ensure consistent playback loudness, also in cases with user interaction, the content loudness is normalized during the rendering process based on metadata [14].

3. Deep-learning-based Dialogue Enhancement

Today, broadcasters and service providers still rely on a great amount of material created with non-object-based workflows, for which only the final audio mix is available, often in stereo format. For such cases, Dialogue Separation (DS) can be applied to separate speech and background elements (M&E) from the final audio mix and to create audio objects that users can interact with. DS can be used as a pre-processing step to enable OBA so that personalization is offered to the audience for both object-based and non-object-based productions.

3.1 Related Works

DS shares relevant characteristics with both speech enhancement and source separation. Both have been major research topics for decades, cf. e.g., [16]. Early works specifically addressing DS for TV developed signal processing strategies for extracting the dialogue from a final audio mix. These strategies exploited characteristics specific to dialogue in TV productions, e.g., the fact that dialogue is usually amplitude-panned in a stereo mix [17], typically located in the phantom center [18], or in any case a direct component correlated across channels [19], or a combination of these characteristics [20]. A more general approach was proposed in [21], where feature extraction is followed by a shallow neural network. This idea anticipated more recent works, in which deep neural networks (DNNs) are applied in the Short-Time-Fourier-Transform (STFT) domain [22] [23] [24] [25].

Recent advances in deep learning brought significant improvements in the quality of DS. While research in the field continues and further improvements are to be expected, some products delivering remarkable quality are already on the market. Besides the MPEG-H Dialog+ technology discussed further below, it is worth noting a selection of other post-production tools: IDC by Audionamix [26], RX Dialogue Isolate by iZotope [27], and Clarity Vx Pro by Waves [28].

3.2 MPEG-H Dialog+

MPEG-H Dialog+ is a DNN-based tool performing 1) DS, 2) automatic remixing, and 3) automatic metadata authoring. Dialog+ has been specifically developed to fill the gap between traditionally produced material available only as a final audio mix and object-based audio. The core DS solution contained in MPEG-H Dialog+ is under active development, but the version that is currently deployed in broadcasting and to service providers was described in [2], and updated with improvements similar to the ones proposed in [25]. It consists of a deep fully convolutional neural network, trained with original TV material for which the individual audio stems are available from production. The material

comprises various content types, such as nature documentaries, sports programs, and feature films. The audio stems are manually edited to exclude any parts where non-speech-sounds are present in the dialogue stem or speech is present in the M&E stem. This helps prevent faulty training where, for example, sounds are incorrectly identified as speech and separated. DS cannot always separate dialogue perfectly and some degree of coloration or distortions might be introduced. There are ongoing efforts to improve the quality of DS as well as to automatically control the final quality [29].

MPEG-H Dialog+ combines DS with automatic remixing, allowing global and time-varying background attenuation, or a combination of both. Users who prefer minimal background signal might benefit from global background attenuation, where the relative background level is attenuated by the same specified amount over the entire signal. Since globally attenuating the background might decrease ambience and sounds of narrative importance [8], time-varying background attenuation only lowers the background level when dialogue is present. The parameters for DS and remixing can be adjusted, and it is possible to combine global and time-varying attenuation [2]. Finally, the automatic authoring produces audio and metadata ready to be used in an object-based workflow. MPEG-H Dialog+ was shown to successfully enable personalization [2], and to reduce listening effort (LE), as discussed below.

3.3 Reducing Listening Effort (LE)

DE and the DS network of MPEG-H Dialog+ were evaluated in [30], where a multimodal evaluation of LE was carried out on subjects without hearing impairments and using TV excerpts. Pupil size was considered as a physiological indicator of LE together with self-reported LE and word recall rate. A common trend was observed across all measures of LE: It was shown that background music and effects cause significant LE in broadcast audio, even under ideal listening conditions. Decreasing the background level via DE consistently reduced the LE. This was shown to be the case both when the original audio objects were used for DE and when DE was enabled by the DS network of MPEG-H Dialog+. It was concluded that background music and effects can carry vital information and play an important role in engaging and entertaining the audience [31]. They do, however, come with a clear LE cost, requiring care at the production stage (cf. e.g., the recommendations in [32]), and the personalization functionalities provided by DE for users.

4. Dialogue Enhancement Production Workflows

In the following, we outline three current real-world use cases, where DE was implemented in broadcast and streaming workflows. The examples from Globo[2] (Brazil) and SBS[3] (South Korea) are live, object-based productions where speech and background sounds were provided as separate components. The described service by German public-service broadcaster ARD works with DS in a channel-based offline production environment. While DE for live transmission is normally faced with challenging latency requirements down to a maximum of a few frames, post-production workflows can typically be covered with offline processing, providing the option of more lookahead and processing time for DS and remixing.

4.1 Object-based Live Broadcast

The football World Cup 2022 in Qatar was a recent major event for which object-based DE was provided in several broadcasting and streaming services [33]. Both Globo and SBS implemented interactive, personalized MPEG-H Audio in their services. This enabled viewers to switch between different presets, including an option with attenuated background sounds for improved ease of listen. Additionally, due to the object-based production, advanced personalization options were possible: the balance between dialogue and background could be changed by users with a slider in the user interface, see Figure 2.

At the Globo facilities, two live broadcast services were set up during the world cup: One for the current SBTVD 2.5 broadcast standard with HE-AAC or AAC-LC and MPEG-H Audio [12], and one for SBTVD TV 3.0 (via DASH streaming), which is currently in the standardization phase and specifies MPEG-H Audio as the sole mandatory audio codec [34] [35]. Globo received the broadcast signals by the world cup's host broadcaster in their facilities in Rio de Janeiro. The local commentary by Globo sports channels was added to the live feed from Qatar.

[2] Globo is a Brazilian free-to-air television network, owned by media conglomerate Grupo Globo, and the largest commercial TV network in Latin America.

[3] Seoul Broadcasting System (SBS) is one of the leading South Korean television and radio broadcasters and the largest private broadcaster in South Korea.

To allow different presets and user personalization, the signal was enriched with MPEG-H metadata by an AMAU (Audio Monitoring and Authoring Unit) and passed along together with the video signals to the corresponding video encoders for simultaneous delivery over the TV 2.5 and TV 3.0 systems [36].

Figure 2. TV Globo football world cup broadcast with interactive DE user interface.

At SBS, the target application was the broadcasters' Android app, receiving a HTTP Live Streaming (HLS) signal. The four available presets for the users to choose from were "Basic", "Enhanced Dialogue", "Site", and "Dialogue Only". The live workflow also consisted of an AMAU-to-encoder-setup, fed by a mixing desk providing international feed and Korean commentary. In this case, the objects for the streaming service were produced in parallel to the regular master audio for the broadcast service.

4.2 DNN-based VoD Service

In contrast to the two OBA scenarios described above, the German public broadcaster ARD implemented a DE service called "Klare Sprache" ("Clear Speech") into their Video on Demand (VoD) service [37] using pure stereo audio. It is based on existing non-object-based content for which the separate audio stems are not available, but only the final audio mix.

The DS technology used for their VoD service is MPEG-H Dialog+, implemented in a centralized transcoding service. The MPEG-H Dialog+ processing in this use case is part of a containerized automatic workflow. It receives the audio mix from an archive or production. After DS processing and automatic remixing, stereo audio with enhanced dialogue is rendered to a file that is loudness-normalized according to EBU R128 [38]. The new audio mix is then muxed with video versions of different bitrates and encoded in parallel.

The created "Clear Speech" version is offered as alternative accessibility service in the VoD service "ARD Mediathek", see Figure 3. In this workflow, the produced MPEG-H Dialog+ output is a stereo rendering from the MPEG-H Audio scene authoring computed in the background. An optional ADM (Audio Definition Model) [39] output could be directly encoded and transmitted in case an OBA workflow is established.

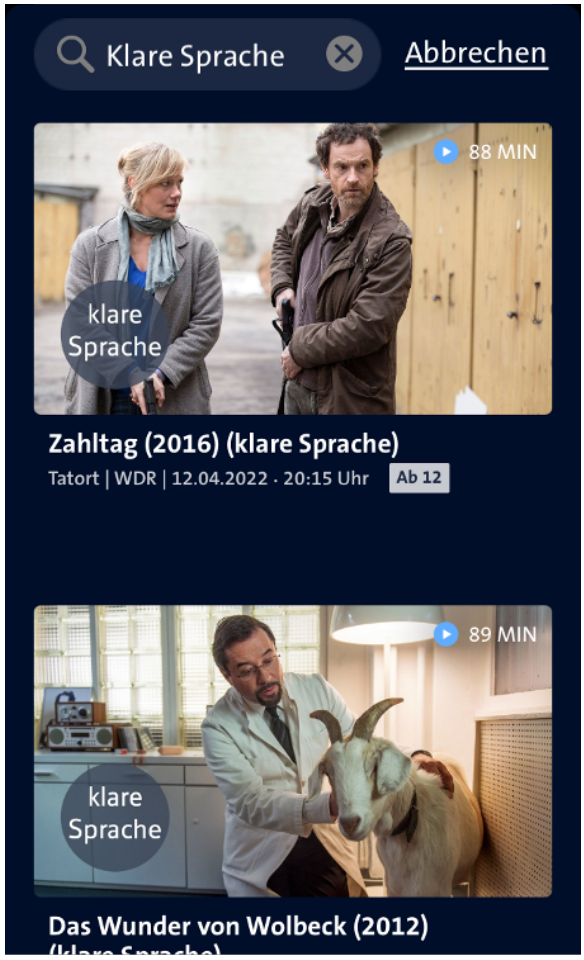

Figure 3. Movies with "Clear Speech" audio in the German VoD service "ARD Mediathek".

5. Conclusion

This paper reviewed the current state of Dialogue Enhancement (DE) in TV services, with emphasis on the MPEG-H Audio system. It showed real-world adoptions and implementations of DE technologies into current production and transmission workflows. DE can be a part of native object-based productions or enabled by Dialogue Separation (DS). While recent studies have shown that decreasing the relative level of the background via DE consistently reduces the listening effort, other studies have found that the optimal balance between dialogue and background is highly personal and situation-dependent. Only the personalization achieved through DE meets the needs and preferences of the audience in almost every situation.

Next to the relative balance of dialogue and background, other factors can make the TV content difficult to access [40]. Object-based audio (OBA) provides the tools to improve accessibility, e.g., by efficiently delivering multiple audio versions including audio descriptions, different languages and versions with simplified vocabulary or slower speech pace [40]. Employing DE and OBA means taking one concrete step towards improved accessibility and user satisfaction, as shown by the studies reviewed in this paper. Moreover, DNN-based DS can assist in driving the shift towards OBA, making existing content personalized and interactive.

6. Acknowledgement

Sincere thanks go to the Fraunhofer IIS Accessibility Solutions team and to Jennifer Karbach, Yannik Grewe, and Adrian Murtaza for their valuable support.

References

[1] C. Mathers, "A Study of Sound Balances for the Hard of Hearing," BBC White Paper, 1991.

[2] M. Torcoli, C. Simon, J. Paulus, D. Straninger, A. Riedel, V. Koch, S. Wits, D. Rieger, H. Fuchs, C. Uhle, S. Meltzer and A. Murtaza, “Dialog+ in Broadcasting: First Field Tests Using Deep-Learning-Based Dialogue Enhancement,” in *Technical Papers IBC 2021 (International Broadcasting Convention)*, Virtual, 2021.

[3] Verband Deutscher Tonmeister e.V., “KI in der Audioproduktion,” *vdt Magazin,* no. 2, 2023.

[4] A. Bidner, J. Lindberg, O. Lindman and K. Skorupska, “Deploying Enhanced Speech Feature Decreased Audio Complaints at SVT Play VOD Service,” in *9th Machine Intelligence and Digital Interaction (MIDI) Conference*, Poland, 2021.

[5] H. Fuchs, S. Tuff and C. Bustad, “Dialogue Enhancement - Technology and Experiments,” EBU Technical Review - Q2, 2012.

[6] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon and B. Shirley, “Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech,” in *AES Convention 146*, Dublin, Ireland, 2019.

[7] L. Ward, M. Paradis, B. Shirley, L. Russon, R. Moore and R. Davies, “Casualty Accessible and Enhanced (A&E) Audio: Trialling object-based accessible TV audio,” in *AES Convention 147*, New York, USA, 2019.

[8] L. Ward, “Improving Broadcast Accessibility for Hard of Hearing Individuals: using object-based audio personalisation and narrative importance,” PhD thesis, University of Salford, 2020.

[9] ISO/IEC 23008-3:2022, “High Efficiency Coding and Media Delivery in Heterogeneous Environments–Part 3: 3D Audio"".

[10] ETSI Standard TS 101 154 v2.3.1, “Specification for the Use of Video and Audio Coding in Broadcasting Applications Based on the MPEG-2 Trans- port Stream,” Feb. 2017.

[11] ATSC Standard 3.0 A/342:2021 Part 3, “MPEG-H System,” Mar. 2021.

[12] ABNT NBR 15602-2:2020, “Digital Terrestrial Television - Video Coding, Audio Coding and Multiplexing Part 2: Audio Coding,” 2020.

[13] TV 3.0 Project, “Terrestrial TV Evolution in Brazil,” [Online]. Available: https://forumsbtvd.org.br/tv3_0/#panel-phase2. [Accessed April 2023].

[14] Y. Grewe, P. Eibl, D. Rieger, M. Torcoli, C. Simon and U. Scuda, “MPEG-H Audio Production Workflows for a Next Generation Audio Experience in Broadcast, Streaming, and Music,” in *AES Convention 151*, Virtual, 2021.

[15] P. Eibl, Y. Grewe, D. Rieger and U. Scuda, “Production Tools for the MPEG-H Audio System,” in *AES Convention 151*, Virtual, 2021.

[16] E. Vincent, T. Virtanen and S. Gannot, Audio Source Separation and Speech Enhancement, John Wiley & Sons, 2018.

[17] A. S. Master, L. Lu, H. M. Lehtonen, H. Mundt, H. Purnhagen and D. Darcy, "Dialog Enhancement via Spatio-Level Filtering and Classification," in *AES Convention 149*, Virtual, 2020.

[18] J. T. Geiger, P. Grosche and Y. L. Parodi, "Dialogue Enhancement of Stereo Sound," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

[19] A. Craciun, C. Uhle and T. Bäckström, "An Evaluation of Stereo Speech Enhancement Methods for Different Audio-Visual Scenarios," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

[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," *Journal Audio Engineering Society (JAES),* vol. 67, no. 7/8, pp. 510-521, 2019.

[21] C. Uhle, O. Hellmuth and J. Weigel, "Speech Enhancement of Movie Sound," in *AES Convention 125*, San Francisco, USA, 2008.

[22] N. L. Westhausen, R. Huber, H. Baumgartner, R. Sinha, J. Rennies and B. T. Meyer, "Reduction of Subjective Listening Effort for TV Broadcast Signals With Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 3541-3550, 2021.

[23] J. Paulus and M. Torcoli, "Sampling Frequency Independent Dialogue Separation," in *European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022.

[24] D. Petermann, G. Wichern, Z.-Q. Wang and J. Le Roux, "The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks," in *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, Singapore, 2022.

[25] M. Torcoli and E. A. P. Habets, "Better Together: Dialogue Separation and Voice Activity Detection for Audio Personalization in TV," in *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, Rhodes, Greece, 2023.

[26] "IDC," Audionamix, [Online]. Available: https://audionamix.com/instant-dialogue-cleaner/. [Accessed March 2023].

[27] "RX Dialogue Isolate," iZotope, [Online]. Available: https://www.izotope.com/en/products/rx/features/dialogue-isolate.html. [Accessed March 2023].

[28] "Clarity Vx Pro," Waves, [Online]. Available: https://www.waves.com/plugins/clarity-vx-pro. [Accessed March 2023].

[29] M. Torcoli, J. Paulus, T. Kastner and C. Uhle, "Controlling the Remixing of Separated Dialogue with a Non-Intrusive Quality Estimate," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Virtual, 2021.

[30] M. Torcoli, T. Robotham and E. A. P. Habets, "Dialogue Enhancement and Listening Effort in Broadcast Audio: A Multimodal Evaluation," in *14th International Conference on Quality of Multimedia Experience (QoMEX)*, Lippstadt, Germany, 2022.

[31] L. Ward, B. G. Shirley, Y. Tang and W. J. Davies, "The Effect of Situation-Specific Non-Speech Acoustic Cues on the Intelligibility of Speech in Noise," in *INTERSPEECH*, 2017, Stockholm, Sweden.

[32] D. Geary, M. Torcoli, J. Paulus, C. Simon, D. Straninger, A. Travaglini and B. Shirley, "Loudness Differences for Voice-Over-Voice Audio in TV and Streaming," *Journal Audio Engineering Society (JAES),* vol. 68, no. 11, pp. 810-818, 2020.

[33] Fraunhofer IIS, "Football Fans Around the World Experience the Worldcup in Immersive and Personalized MPEG-H Audio," 2022. [Online]. Available: https://www.iis.fraunhofer.de/en/pr/2022/20221208_mpeg-h_audio_worldcup.html. [Accessed April 2023].

[34] Brazilian Digital Terrestrial Television System Forum; Brazilian Ministry of Communications, "Testing and Evaluation Report: TV 3.0 Project - Audio Coding," SBTVD Forum, Brazil, 2021.

[35] A. Murtaza, S. Meltzer, Y. Grewe, N. Faecks, M. Raulet and L. Gregory, "MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals," SET International Journal of Broadcast Engineering, 2021, doi: 10.18580/setijbe.2021.3.

[36] A. Murtaza, "Audio Advances Showcased During the FIFA World Cup 2022," *DVB Scene,* no. 61, p. 11, 2023.

[37] ARD Digital, "Tonspur Klare Sprache," ARD Digital, 2023. [Online]. Available: https://www.ard-digital.de/klaresprache. [Accessed March 2023].

[38] European Broadcasting Union (EBU), "Loudness Normalisation and Permitted Maximum Level of Audio Signals," EBU Recommendation 128, 2020.

[39] ITU-R BS.2076-2, "Audio Definition Model," 2019.

[40] C. Simon, M. Torcoli and J. Paulus, "MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming," 2019.