AI FOR AUDIODESCRIPTION: A NATURAL VOICE FOR ACCESSIBILITY

Pedro H. L. Leite[1], Luiz F. Kruszielski[2], Marcelo Lemmer[3] and Edmundo Hoyle[4]

[1]Federal University of Rio de Janeiro (UFRJ) and [1,2,3,4] Globo, Brazil

ABSTRACT

This study describes the development and implementation of an AI-based natural voice synthesis and automated mixing workflow for audio description (AD) in Brazilian television drama content, with a real demonstration case of success. With home-built AI technologies and automation algorithms, our approach advances a scalable and cost-effective solution capable of producing high-quality AD in real broadcast and VOD environments with lower costs and fast paced schedules, overcoming real barriers of modern audiovisual content production. AI architectures for Text-to-Speech (TTS) systems were employed to synthesize speech, ensuring natural and transparent narration. An automated mixing process was also specifically designed to integrate the AD narrations with original content seamlessly, maintaining speech intelligibility even in high dynamic range sound scenarios. The system was field-tested over the air and in VOD, where it demonstrated robust performance and received positive feedback from both general audiences and AD specialists.

INTRODUCTION

The evolving landscape of media consumption underscores the crucial need for inclusivity, particularly for those with visual impairments. Audio description (AD) plays an indispensable role in making media accessible, providing a verbal representation of visual content that allows visually impaired individuals to experience films, television, and live performances in meaningful ways. As described by Audio Description provides narration of the visual elements - action, costumes, settings, and the like - of theatre, television/film, museum exhibitions, and other events. The technique allows patrons who are blind or have low vision the opportunity to experience arts events more completely - the visual is made verbal. AD is a kind of literary art form, a type of poetry. Using words that are succinct, vivid, and imaginative, describers try to convey the visual image to people who are blind or have low vision” (J. Snyder, (1)). However, traditional methods of producing audio descriptions are fraught with challenges, including high production costs and significant time demands, which have historically limited the accessibility and timeliness of such services. According to the 2010 data from the Brazilian Institute of Geography and Statistics (IBGE) (MEC, (2)), there are approximately 6.5 million people in Brazil with significant or severe visual impairments. This statistic is supported by findings from the 2019 National Health Survey (PNS) (IBGE, (3)), which indicates that 3.4% of the population, or around 3.978 million people, experience some form of visual impairment. It is crucial to recognize that audio description benefits not only those who are completely blind but also those with partial and severe vision loss. Additionally, other groups, including individuals with intellectual disabilities and learning disorders, can greatly benefit from audio description as it serves as an alternative sensory channel that aids in quicker and more effective comprehension of visual content.

In Brazil, television and locally produced drama content hold a central place in the cultural landscape, (Keske and Scherer, (4)) with telenovelas—a genre of daily dramas—being particularly significant. These shows are watched by approximately 33% of the population every day (Hibou, (5)), illustrating their widespread appeal and cultural importance. In this landscape, audio description production is not merely a technical enhancement but a formative social advancement. It embodies a commitment to cultural inclusivity, ensuring that all individuals, regardless of visual ability, have equal access to cultural narratives and entertainment.

This research, focuses on the implementation and production of an artificial intelligence (AI) natural synthetic voice and automated workflow for audio description, tested in real broadcast environments and aimed particularly at enhancing the accessibility of Brazilian television content. In the second chapter, the challenges to implement this system will be discussed. In the third chapter we will talk about the technical aspects of the implementation. In the fourth chapter we will discuss in detail the deployment of our first AI audio description broadcast and its reception by the public. In the last chapter, we present our conclusions and plans for future work.

CHALLENGES IN PRODUCING AUDIO DESCRIPTION FOR DRAMA

There is significant demand for AD in drama content for Brazilian television. In 2008 an association of visual impaired people wrote an open letter demanding drama content to be produced with AD. However, the availability of such content is still scarce. Motta (L. Motta. (6)) highlights this issue, stating: "For a long time, I have been talking about the lack of audio description for blind people in cinemas, theatres, and on television, but always in a theoretical way, because initiatives like this are still so rare that we can hardly get a taste of them." Since 2022, Globo has been producing drama series with AD using the traditional method.

Some technical challenges contribute to the scarcity of AD. Firstly, different from live AD, where the narrator does a description of the events and transmits it in real time, the dynamic nature of drama content, where scenes can shift dramatically from one to another, requires audio describers to adapt quickly to describe diverse settings and introduce new characters effectively. This task demands a high level of precision and creativity to ensure that the narrative is accessible and engaging for visually impaired viewers. It necessitates a script that is meticulously written for the recording, calculating the actions, scenery, characters, and objects to be narrated in the available time slot that avoids overlapping with dialogues or other important sounds. This is a complex process and requires that the content is in its final cut, with no further edits possible.

Moreover, the production process for television drama often involves "open works," where scripts are not finalized until close to broadcast time and this could be a script altered prior to shooting, based on audience reception of the on-air drama, or more immediate and closer to exhibition post-production edits, caused by other external factors such as scheduling constraints or commercial break necessities. This fluidity complicates the timing and synchronization of audio descriptions, as any changes in the scene layout require corresponding adjustments in the AD script and the AD audio mix.

In a regular AD production method, first, the AD script-writer receives the audio-visual finished material, which is carefully watched and analysed, and then a narration script is written. Once an AD script is finished, it is recorded by a narrator with a clear pleasant voice with a neutral tone – the emotion of the scene must be in the listener, not in the narrator – in a professional studio. Further minor adjustments to the AD script can be made at this stage to fit the content properly. Then, this recording is edited and mixed into the content of the original audio track. If the narration does not fit properly into the mix, a re-dub may also be requested. In the mixing process, the sound technician must make sure that all the AD narration is clear and maintains a constant level. Also, in this process the narration should affect the original sound as little as possible, and the narration does not make the original sound inaudible. The background sound, such as music, sound ambiance, or effects, is equally important to the perception of the visually impaired. The final material is then checked by a consultant, which necessarily has a visual deficiency, and only then can be finally exported to the final material. These steps are often time-consuming and involve a tight schedule between staff members (narrator, scriptwriter, audio editor) and the available technical resources (recording and mixing room) (KruszieIski et. al, (7)).

AI FOR AUDIO DESCRIPTION

The recent developments in synthetic voice using AI plays a key role in allowing the possibility of using a scalable voice for AD, avoiding the need of recording procedures for each content. At the start of our project, recent algorithms had obtained excellent results for high resource languages such as English and Mandarin, but we could observe that models developed for Brazilian Portuguese were still behind in quality and reliability, due to specific data and development limitations. To address this problem, we started our work by creating a background environment to enable the development of our AD software, generating language-aware data and algorithms for artificial voice systems.

Technical Implementation

For audio description, the technique used to synthesize voice is Text-to-Speech (TTS), where the text is the input to a computer algorithm that generates speech as output. Modern TTS systems are created using machine learning architectures, specifically deep learning. These structures can capture semantic and linguistic relationships between the words in the text and can generate pronunciation and intonation consistent with the textual intent. Thus, the data need for TTS applications is in general for text/speech aligned files containing the transcriptions and the corresponding speech, within 10-30s excerpts. The recordings must comprise natural human-like readings of the text, with the intonation intended for the final use case. For the AD case, a neutral voice is desirable, having in mind that the narrations should be emotionally transparent.

In that sense, we generated two datasets for TTS applications in Brazilian Portuguese: one containing male voice speech (Leite et. al, (8)) and another female (Leite et. al, (9)), both recorded with professional equipment and in a controlled environment and with neutral speech emotion. The intention was to create base models and algorithms that could serve as research hot starts for the final use cases, leveraging the robustness of pronunciation and general audio cleanliness. With these two datasets, we could create high quality TTS systems in Brazilian Portuguese for those voices, as shown in the previous works (8) and (9). This was achieved using the following machine learning design: The training process involves an initial stage with a neutral voice using the data available in our base datasets.

After obtaining a model with a completely neutral voice, there is a second stage related to transfer learning for the target voice, which has a smaller amount of recorded hours available and is the voice that users are used to hearing in our AD delivery. In this final stage, fine-tuning of timbre (acoustic properties of the voice) and prosody (rhythm, stress and intonation of the speech) is performed to grasp the individualities of the voice that will carry out the audio description. The learning architectures used were those described in works Shen et. al (10) and Yang et. al (11), Tacotron2 and Multiband-MelGAN, respectively, using an open-source code implementation. The Tacotron2 model uses convolutional (Goodfellow et. al, (12)) and bidirectional recurrent networks (Goodfellow et. al, (13)) (which use future and past context in sequences) with attention mechanisms to decode the text and relate it to psychoacoustic characteristics, which are implicitly modelled by the network weights. These characteristics are subsequently transformed into a mel-spectrogram by convolutional layers and linear projections. Since mel-spectrograms do not contain any phase information and have a bandwidth narrower than desired, a subsequent step of generating the waveform in time is still necessary, where the second model used in this work, described below, comes into play.

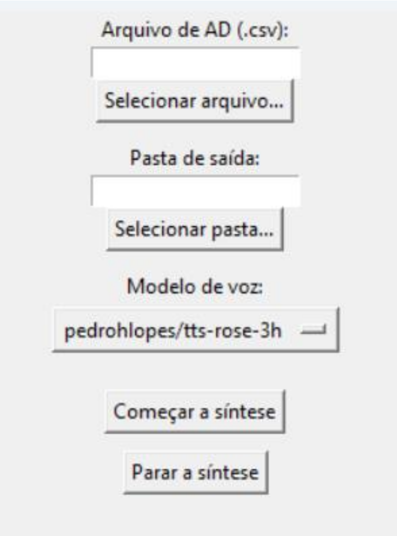
The neural vocoder Multiband-MelGAN is a machine learning model based on generative networks whose intuition is to leverage the characteristics of generating faithful samples of Generative Adversarial Networks (GANs) (Goodfellow et. al, (14)) to reconstruct waveform time-amplitude signals from mel spectrograms. Multiband-MelGAN is an extension to the original MelGAN (Kumar et. al, (15)), which uses convolutional networks in its generator and a multi-scale audio discriminator, using down-sampling techniques. In the Multiband case, processing is done in sub bands, creating and joining signals in several individual frequency bands, instead of a single full-band signal as in the case of simple MelGAN. This division in processing may help the network learn which parameters are important for each frequency band, consistent with the predictions of psychoacoustic models that show that our auditory apparatus excites frequency bands differently and that our perception is also heterogeneous in this sense.

To increase the reliability of our TTS pipeline, we also trained a single-stage end-to-end model, VITS (Kim et. al, (16)), which also comprises a GAN for synthetic waveform generation, but its text/speech language modules rely on variational autoencoders (Kingma et. al, (17)) and normalizing flows (Papamakarios et. al, (18)) to encode and decode relevant features. With two models, we could assure that some of the flaws of the first model were covered by switching to this new architecture, and vice-versa.

Automatic Mixing and Interfaces

With these architectures rightfully trained over the audio descriptors' voice, we could transform the description script directly into audio waveforms containing the corresponding speech with adequate timbre and prosody, using our pipelines.

Furthermore, we also generated an intelligent automatic mixing process, which takes the speech files and inserts them into the original content, without hard blocking important sound cues that were already present. The final goal of the automatic mixing system was then to keep the AD narrator voice intelligible even in adverse dynamic sound level environments. It should be possible to hear sound effects and ambiences in soft sound level scenes, and still have the AD voice at an intelligible volume in moments where high intensity level music is occurring.



Modelo de voz: pedrohlopes/tts-rose-3h

Figure 1 – Batch audio description generation interface.

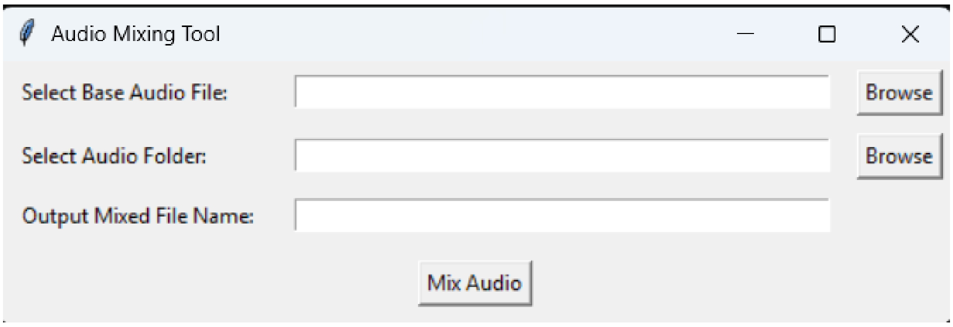


Figure 2 – Mixing interface.

The automatic mixing works as follows: The input speech excerpts and the insertion timecodes are passed in through a csv file and the general audio loudness is calculated and stored before the process starts; After that, we reduce the local audio loudness with parametric equalization filters chosen to increase intelligibility of speech, applied from right before until right after the end of the speech insertion times; The full mixed content is then stored in a new audio file and is ready to be used in the VOD platforms. These pipelines were all processed using *python programming language* classical libraries such as *scipy* and *numpy.* All algorithms regarding audio processing for these parametrical equalization and mixing were built in-house, without using pre-developed libraries, to guarantee performance and sound reliability.

Then, our systems were stress tested and put to work programmatically within our command line interfaces. To get that into the production workflow and to enable easy usage for content producers, we proceeded to design two user interfaces: one for synthetizing the script texts into speech and another to automatically mix the generated audio with the original content. As we can see in figures 1 and 2, the users could now use the interface to choose the csv containing the text to be synthetized and the timecodes, the desired TTS model and just hit a button to generate all the content. If something goes wrong, the content producers could just change a line in the csv file to change the timecodes or the scripted content, saving time and being cost-efficient. With the audio files validated, the speech is also automatically mixed into the product using a single button.

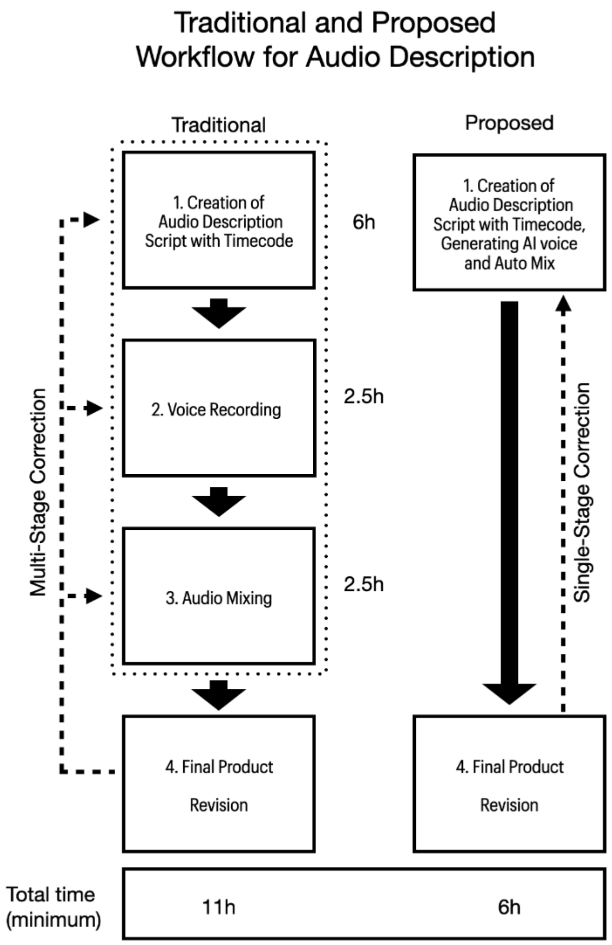

Traditional and Proposed
Workflow for Audio Description
Traditional
Proposed
1. Creation of Audio Description Script with Timecode
6h
2. Voice Recording
2.5h
3. Audio Mixing
2.5h
4. Final Product Revision
Multi-Stage Correction
1. Creation of Audio Description Script with Timecode, Generating AI voice and Auto Mix
Single-Stage Correction
4. Final Product Revision
Total time (minimum)
11h
6h

Figure 3 – Classical and novel workflow comparison.

Novel Workflow in Production

In figure 3, we can see the comparison between the traditional workflow and our novel AD workflow that was put into production with this work. In the recording stage, the narration process would be replaced by generating a synthetic voice. It is important to maintain the same requirements at this stage as the traditional method - a neutral, clear, pleasant voice with optimal recording quality. One of the biggest challenges and the most important aspect at this stage to be successful with a synthetic voice, is that at no point can this voice be perceived as artificial - the viewer must interpret it as a natural, non-robotic voice with the absence of audible artifacts. A failure in this aspect could lead the viewer to a break in the immersion of the storytelling. The perception of something "strange" in the voice can draw the viewer's attention to the voice itself and not to the story being told. This would also occur with failure in the way of speaking words correctly. As our current models have met these criteria in production, we could also successfully validate the quality of the proposed architectural designs. Also, an advantage that synthetic voice can bring to this matter is vocal continuity, as the voice remains the same regardless of content production volume or content duration. This allows the audio description to be made by different people and even simultaneously, avoiding changes in the voice that the viewer is used to hearing.

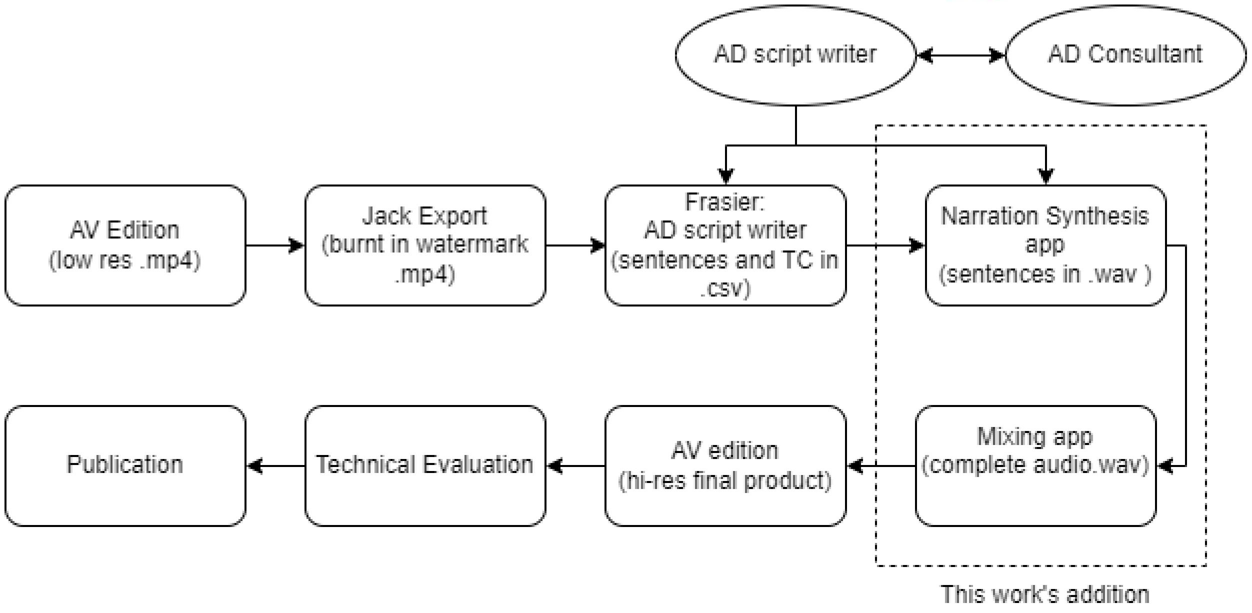


Figure 4 – Detailed audio description workflow with the inclusion of AI and automations derived from this work.

Regarding the mixing process, it was also important to maintain the characteristics required in the traditional workflow. A drama mix generally has extensive dynamic sound level variations, which usually are a significant challenge in creating a mixing system and requires extra attention for the mixing professionals. As our system was specifically designed to deal with such environments and stress tested in real use case scenarios, we also obtained successful results when putting the automatic mixing into production, with all current VOD drama AD produced content done directly with it and zero overlaps. Also, to avoid overlap of in content dialog with AD, the scriptwriter can validate the AD voice duration as the script is created, directly making the necessary changes to the text, fitting the speech excerpts within the dialog gaps on the go. This would significantly reduce the time needed to produce an AD, considering that stages 2 and 3 of figure 3 happen instantly and thus the AD scriptwriter has direct access to the final mix. It also allows different people to work on the same project, such as two AD scriptwriters doing different sections of the same chapter, enabling an even greater reduction in production time. Another advantage that emerges from the current flow is regarding eventual changes and corrections. Changes at an advanced stage of the pipeline required backward corrections and could provoke duplicated work at all previous phases, which can be quite complex and time-consuming. When the adjustment is made with a single-stage setup, the result and the testing of this adjustment is immediate, transforming a multi-stage process with several people into a single interaction. This also allows easier small corrections, where tasks such as dividing a program into different blocks with no text change could be made directly by the video or audio editor.

In figure 4, we show in detail our novel workflow that is currently in production for our drama series: the audiovisual product is edited and formatted to a final cut version, that is delivered to an internal platform (namely Jack), that safely exports the content with a burnt watermark; From there, the audio description scriptwriter builds the sentences to be synthetized and tags the corresponding timecodes using Frasier video-to-voice platform, exporting a .csv file containing AD cue times and texts; Then, our systems come into play by taking the exported .csv, synthetizing all narrations and putting them into place with the automatic mixing; During this last process, the AD consultant reviews the final product and necessary changes can be directly made to the script on the go; Finally, the product is exported back to our sites to generate the high-resolution version, that is evaluated by quality assurance technicians to find minor quality problems; When the product passes this last step, it is ready to be delivered to the consumers.

CONTENT PRODUCTION AND PUBLIC RECEPTION

The first pilot project created with the proposed new workflow of AD was the episode from series “Histórias (Im)possíveis: Levante” - (Im)possible stories: Uprising. It is a 45-minute episode, that aired on November 20, 2023. It had a total audience of about 11 million people. The audio description was available as a separate stream that could be accessed through the secondary audio program (SAP) remote button. The content was then posted in the GloboPlay VOD service (Globoplay, (20)). The program was presented to a group of audio description specialists – audio description writers, narrators, and consultants, both with normal and impaired vision, to ask about the overall quality of the audio description. Without the previous knowledge that the voice was generated with AI, they complemented the quality of the audio description as a whole and had no complaints or any comments on the voice quality or audio mixing. This result took place even though the community of professional audio describers exhibits considerable resistance to the integration of artificial intelligence-generated voices in audio description. There is a prevalent scepticism among audio describers regarding the capacity of AI to provide an equivalent or enhanced user experience, fearing that the use of synthetic voices might compromise the quality and effectiveness of audio descriptions for visually impaired audiences. From this first pilot to the publication of this article, all new drama series produced and created by Globo used this same workflow, on a total of more than 50 episodes. The series “Justiça 2” and “Encantados 2” were also broadcast on Television and released on VOD, and no comments about the voice quality or mixing were observed – both from general public and audio description specialists. This outcome has been highly encouraging for the progress of this research project.

CONCLUSION AND FUTURE WORKS

This paper has presented a comprehensive examination of an innovative approach to audio description (AD) for Brazilian television, using advanced artificial intelligence (AI) technologies to generate synthetic speech and automate mixing processes. This AI-driven method not only addresses the critical need for inclusivity in media consumption but also offers a scalable solution that can significantly reduce production times and costs for accessibility content production, while assuring the high quality needed for television broadcasting and VOD streaming.

The research outlined in the referenced documents highlights the successful implementation of an AI system capable of generating natural-sounding, emotionally resonant synthetic voices for the purpose of audio description. The system was tested in real-world broadcasting scenarios, specifically in the drama series "Histórias Impossíveis - Levante", reaching millions of Brazilian homes. The AI-generated voice was indistinguishable from human narration to the general audience, a testament to the technological advancements in speech synthesis and natural language processing employed by the research team. Following this success, the technology has been applied to over 50 series episodes, demonstrating its robustness and reliability in a production environment. The implemented AI workflow effectively streamlined the audio description process, which traditionally involves significant time-consuming processes including scriptwriting, narration and sound engineering. By automating these aspects, the AI system reduced the dependency on extensive and inefficient pipelines, which lead to increased production times and higher costs.

For the visually impaired community, the enhanced accessibility offered by reliable and timely audio description significantly enriches their television viewing experience and fosters greater cultural inclusion. The high quality of the AI-generated audio description ensures that all viewers, regardless of visual ability, can enjoy and fully understand the broadcast content.

Future work directions and final remarks

While the results are promising, the scope for further research is vast. Future studies could explore the possible customization of the voices by the end user, extending this technology to regional dialects and accents, different ages and cultures. There is also an opportunity to expand the application of this technology to other formats of media beyond television and OTT, such as live theatre, public events, and educational resources, all of which could benefit from the automated audio description process. Research into user interface design can also ensure that these technologies are accessible and easy to use for content producers around the world.

Importantly, plans are underway to implement this technology in a daily telenovela show, a format that has historically been challenging to accommodate with traditional AD methods due to its rapid production turnaround and daily broadcast schedule (tight deadlines and high production volume). This expansion may represent a significant breakthrough in making daily serialized content accessible to visually impaired audiences, showcasing the adaptability and potential of AI to enhance media accessibility at scale that was once unreachable.

We note that the successful implementation of AI in the production of audio description for Brazilian television represents a significant milestone in the field of media accessibility. This work should not only demonstrate the feasibility of such technologies but also highlight the positive impact they can have on society. As we look forward to further advancements, it is imperative that we continue to focus on ethical considerations and the potential for these technologies to improve the lives of all viewers, particularly those with individuals with special accessibility needs. By doing so, we can ensure that the future of broadcasting and media production is inclusive, innovative, and responsible, meeting the needs of more diverse audiences.

REFERENCES

1. J. Snyder, 2022. Fundamentals of audio Description. The Audio Description Project. August 2022. Available at: https://adp.acb.org/adi/ADA%20Fundamentals.doc.pdf
2. MEC (Brazilian ministry of education), 2018. Data reafirma os direitos das pessoas com deficiência visual. Portal MEC. Available at: http://portal.mec.gov.br/component/tags/tag/deficiencia-visual
3. Agência de notícias do IBGE (Brazilian statistic and Geography News Agency) 2019: país tem 17,3 milhões de pessoas com algum tipo de deficiência. PNS - Pesquisa Nacional de Saúde. Available at: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/31445-pns-2019-pais-tem-17-3-milhoesde-pessoas-com-algum-tipo-de-deficiencia.
4. H. I. G. Keske, M. M. Scherer, 2013. A Telenovela Brasileira e a Cultura de Massa: Uma Relação Muito Além do Zapping. Polem!ca, v.12 n.2, June 2013, pp. 239–255. H

5. Hibou. 2023. Audiência: O Que o Brasileiro Assiste e Porque. Available at: https://lehibou.com.br/wp-content/uploads/2023/10/23HB_AUDIENCIA01.pdf
6. Motta, L. M. 2011. A Audiodescrição na Escola: Abrindo Caminhos Para Leitura de Mundos. Ver com Palavras. Available at: http://www.vercompalavras.com.br/pdf/a-audiodescricao-na-escola.pdf
7. Kruszielski, L., Leite, P., Bravo, P., et al, 2023. The Use of Artificial Intelligence Enabling Scalable Audio Description on Brazilian Television: A Workflow Proposal. International Journal of Broadcast Engineering, v. 9. August 2023, pp. 39–44.
8. P. H. L. Leite, E. Hoyle, Á. Antelo, L. F. Kruszielski, and L. W. P. Biscainho, 2022. A corpus of neutral voice speech in Brazilian Portuguese. Computational Processing of the Portuguese Language, May 2022, pp. 344–352.
9. P. H. L. Leite, E. Hoyle, Á. Antelo, L. F. Kruszielski, and L. W. P. Biscainho, 2023. Neutral TTS Female Voice Corpus in Brazilian Portuguese. XLI Brazilian Symposium on Telecommunications and Signal Processing - SBrT 2023, October 2023.
10. J. Shen, R. Pang, R. J. Weiss, et al., 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). April 2018, pp. 4779–4783.
11. G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, 2021. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. 2021 IEEE Spoken Laguage Technology Workshop (SLT). January 2021, pp. 492–498.
12. I. Goodfellow, Y. Bengio e A. Courville, 2016. Convolutional Networks. Deep Learning Book, 1st ed, pp. 326–366.
13. I. Goodfellow, Y. Bengio e A. Courville, 2016. Sequence Modeling: Recurrent and Recursive Nets. Deep Learning Book, 1st ed, pp. 367–413.
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al, 2014. Generative Adversarial Nets. Advances in Neural Information Processing Systems, v. 27. December 2014, p. 67.
15. K. Kumar, R. Kumar, T. de Boissiere et al., 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. Advances in Neural Information Processing Systems, vol. 32. December 2019.
16. Kim, J., Kong, J. and Son, J., 2021, July. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. International Conference on Machine Learning. July 2021, pp. 5530-5540.
17. Kingma, D. P., Welling, M., et al, 2019. An introduction to variational autoencoders. Foundations and Trends in Machine Learning, v. 12, n. 4. pp. 307–392.
18. Papamakarios, G., Nalisnick, E., Rezende, D. J., et al, 2021. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research, v. 22, n. 1. January 2021, pp. 1–64.
19. Video to Voice Frazier software: Available at: https://www.videotovoice.com/
20. Globoplay, 2023. Historias (Im)possíveis: Levante, November 2023, Available at https://globoplay.globo.com/v/12082658/

ACKNOWLEDGEMENTS

The author would like to thank former members of our development team, Álvaro Antelo and Pedro Serrado without whom this work would never reach its current state of maturity. Also, to the continuous and effective orientation of Professor Luiz W. P. Biscainho from the Audio Processing Group (GPA) of the Signals, Multimedia and Telecommunications Laboratory (SMT) at UFRJ, which significantly raised the technical coherence levels of our algorithms and brought notable academic experience to the group.