The Use of Artificial Intelligence Enabling Scalable Audio Description on Brazilian Television: A Workflow Proposal

Luiz F. Kruszielski
Pedro H. L. Leite
Pedro Bravo
Marcelo Lemmer
Edmundo Hoyle

Kruszielski, Luiz F.; Leite, Pedro H. L.; Bravo, Pedro; Lemmer, Marcelo and Hoyle, Edmundo; 2023. The Use of Artificial Intelligence Enabling Scalable Audio Description on Brazilian Television: A Workflow Proposal. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2023.3. Web Link: https://dx.doi.org/10.18580/setijbe.2023.3

COPYRIGHT This work is made available under the Creative Commons - 4.0 International License. Reproduction in whole or in part is permitted provided the source is acknowledged.

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING - SET IJBE V.9, 2023, Article 3, 5 p
2023 SET - Brazilian Society of Television Engineering / ISSN (Print): 2446-9246 / ISSN (Online): 2446-9432

The Use of Artificial Intelligence Enabling Scalable Audio Description on Brazilian Television: A Workflow Proposal

Luiz F. Kruszielski, Pedro H. L. Leite, Pedro Bravo, Marcelo Lemmer, Edmundo Hoyle

***Abstract*—Recently, Artificial Intelligence (AI) technologies have been gaining ground in various areas of knowledge, significantly impacting many academic and business spheres. One application that can benefit from AI is the inclusion of people with disabilities in audiovisual content, where the scaling capacity of certain processes can bring new accessibility opportunities. In this work, we show what a traditional workflow of an audio description for dramaturgy audiovisual content looks like, and from there, we propose a new workflow for generating audio description audios for visually impaired people using synthetic voice created with Artificial Intelligence models. The proposed workflow simplifies and considerably reduces production time and costs, besides allowing the generation of audios on a larger scale compared to a traditional workflow, enabling a broader reach of the target audience. It also allows multiple people to work simultaneously on the same project while preserving sound identity through the synthetic voice and standardized mixing. With this proposal, we believe that accessibility on Brazilian television can be expanded to serve a much larger audience.**

***Index Terms*—Audio Description, Artificial Intelligence, Voice Synthesis, Accessibility,**

I. Introduction

Audio description(AD) is the most important tool for enabling inclusion of people with visual impairments, but its implementation is not always feasible due to the complexities of the workflow and scalability. In AD, elements and actions determined as important in a scene are narrated by a voice, providing the possibility of understanding the plot and narrative context without the need for visual presence.

According to Joel Snyder, "Audio Description provides narration of the visual elements - action, costumes, settings, and the like - of theater, 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."[1] Most audio-visual content does not have AD, and this deprives the entire population with visual impairment or low vision of this content. The cultural implications caused by this type of impediment are enormous. This restriction not only prevents the delivery of entertainment content but also enables cultural integration with the general public, as we understand that much of the television content is part of the daily life of Brazilians. Enabling the consumption of this content by a part of the population that, otherwise, would not have access, creates a common culture, and this is a way to create inclusion.

In this article, we are presenting the workflow being implemented in the production of audio description for series and daily drama series (novela) of Grupo Globo's programming. This process incorporates synthetic voice generation and automatic mixing techniques. We believe that the proposed workflow has the capacity to reach a larger number of people with visual impairments, providing access to this resource that is often unavailable. The work will be divided into five distinct sections. Initially, we will address the current panorama of audio description and the challenges it presents. In the second section, we will detail the technique employed to create synthetic voice. The third section will address the conventional audio description production process, in addition to introducing the workflow proposed by this study. In the fourth section, we will share the challenges and the advantages we identified when using the new workflow. Finally, in the last section, we will present our conclusions.

A. Current Scenario of Audio Description:

Data from IBGE in 2010 indicate that Brazil has a population of about 6.5 million people with high or severe visual impairment [2]. This data is corroborated by the National Health Survey (PNS) of 2019[3], which shows that 3.4% (3.978 million) of the population has visual impairment. It is important to emphasize that audio description is not only important for people with total vision loss, as those with partial and severe loss can also benefit from this technique. Other groups, such as people with intellectual disabilities and learning disorders, can also take advantage of audio description "...as it is a second sensory channel to be used for faster comprehension of visual information [4]."

Audio description for open television began in 1982 by the American network PBS, where it was simultaneously transmitted on television and the audio description was transmitted on FM Radio [4]. In Brazil, except for some film

L. F. Kruszielski is with Grupo Globo, Rio de Janeiro, RJ, Brazil (e-mail: luiz.fk@g.globo).
Pedro H. L. Leite. is with Grupo Globo, Rio de Janeiro, RJ, Brazil (e-mail: pedro.hleite@g.globo).
Edmundo Hoyle is with the Grupo Globo, Rio de Janeiro, RJ, Brazil (e-mail: edmundo.hoyle@g.globo).
Marcelo Lemmer is with the Grupo Globo, Rio de Janeiro, RJ, Brazil (e-mail: marcelo.edward@g.globo).

This open access article is distributed under a Creative Commons Attribution (CC-BY) license.
https://set.org.br/ijbe doi: 10.18580/setijbe.2023.3 Web Link: https://dx.doi.org/10.18580/setijbe.2023.3

content, the production of audio description for dramaturgical content is very scarce. In recent years, this production has started to grow, but it is still far from covering most of the television content. This is partly due to the dynamics that exist in the process of producing television dramaturgical content. In the Globo group, most of the content produced daily by dramaturgy is made in the "open work" format, that is, the daily drama is not recorded with all its chapters finalized and is written according to the repercussion of the chapters that are being broadcast. This means that the delivery of episodes is very close to the exhibition, which can significantly complicate the construction process of AD since it can only be executed with the completed material. The public demand for this technology in this type of dramaturgy is not new. In 2005, for example, a group of visually impaired people wrote an open letter requesting that the soap opera América, which featured a visually impaired character, be produced with audio description [4]."

By the no. 188 ordinance from the Brazilian agency of telecommunications, Anatel, (2010) [5], it is mandatory for all open television networks to broadcast a minimum of 20 hours per week of content with AD. Partly, the material produced today for AD is performed in auditorium programs, being live audio description. In this process, a person narrates in real-time what is happening in the program. This type of use of AD is relatively simpler than when applied to dramaturgy content and eventual mistakes tends to be admitted. This is because it is in a relatively controlled environment where the variation of the content displayed on the screen is significantly less complex than in drama. The program takes place entirely in the same environment, with a group of participants and pre-defined agendas without major changes. In dramaturgy, the environment where a scene takes place can be very different from the next, requiring the audio describer, that is, the one who is visualizing and narrating in loco, to verbally present that scenario in each situation, or an introduction of various characters. Due to the complexity, to create audio description for daily dramaturgy, it is necessary that the content is locked, that is, there is no more alteration in the editing or sound design of it, so that AD script can be made and subsequently narrated. More about this workflow is detailed in section III. One of the difficulties in creating AD content for daily dramaturgy is working with open works, that is, works that do not have the script written from start to finish, and can be modified according to the response from the audience. This condition means that the time between what is produced and what is broadcast can be greatly reduced. Another complicating factor is that the audiovisual product can be changed according to external factors, such as the availability of the broadcast schedule on the day or the sale of commercial breaks, and which occur even closer to the broadcast. An automated tool that allows anyone to edit audio description content is extremely useful and necessary for this type of situation in order to have AD.

II. Use of Synthetic Voice to Generate Audio Description:

The recent developments in synthetic voice using AI plays a key role in allowing the possibility of using an scalable voice for AD, avoiding the need of recording procedures for each content. In this section we describe the method proposed for creating a voice that sounds natural and is not perceived as machine generated content. Also, the neutral intonation required for AD fits the current data availability for voice synthesis, specially in Brazilian Portuguese [6].

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.

To build these structures, some training steps are required, during which pairs of text and corresponding audio signals are presented to the models in batches. The training process involves an initial stage with a neutral voice using the data available in [6]. After obtaining a model with a neutral voice, there is a second stage related to transfer learning for the target voice, which has a smaller amount of recorded hours available. In this final stage, fine-tuning of timbre (acoustic properties of the voice) and prosody (rhythm, stress and intonation of the speech) is performed for the individualities of the voice that will carry out the audio description. The learning architectures used were those described in works [7-8] (Tacotron2 and Multiband-MelGAN, respectively) using an open-source code implementation[1]. The Tacotron2 model, whose block diagram can be seen in Figure 1, uses convolutional [9] and bidirectional LSTM recurrent networks (which use future and past context in sequences) [10] with attention mechanisms to decode the text and relate it to psychoacoustic characteristics, which are implicitly modeled 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 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) [11] to reconstruct waveform time-amplitude signals from mel spectrograms. Multiband-MelGAN is an extension to the original MelGAN [12], which uses convolutional networks in its generator and a multi-scale audio discriminator, using downsampling techniques. In the Multiband case, processing is done in subbands, 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

[1] https://github.com/TensorSpeech/TensorFlowTTS

This open access article is distributed under a Creative Commons Attribution (CC-BY) license.
https://set.org.br/ijbe doi: 10.18580/setijbe.2023.3 Web Link: https://dx.doi.org/10.18580/setijbe.2023.3

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.

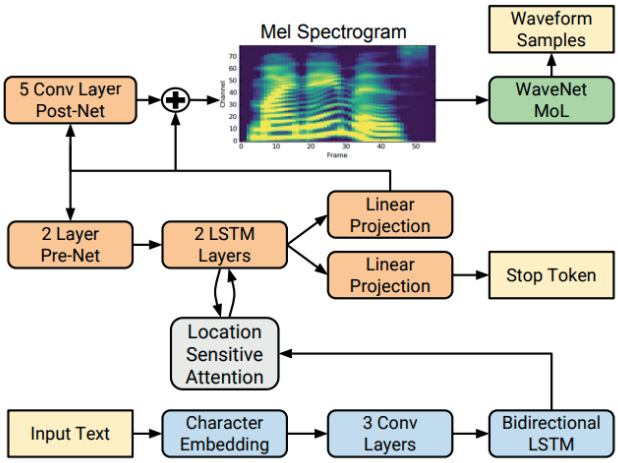


Figure 1. Block Diagram of Tacotron 2 [6].

III. Audio Description workflow:

The AD technical process can be described in four stages (Fig.2):

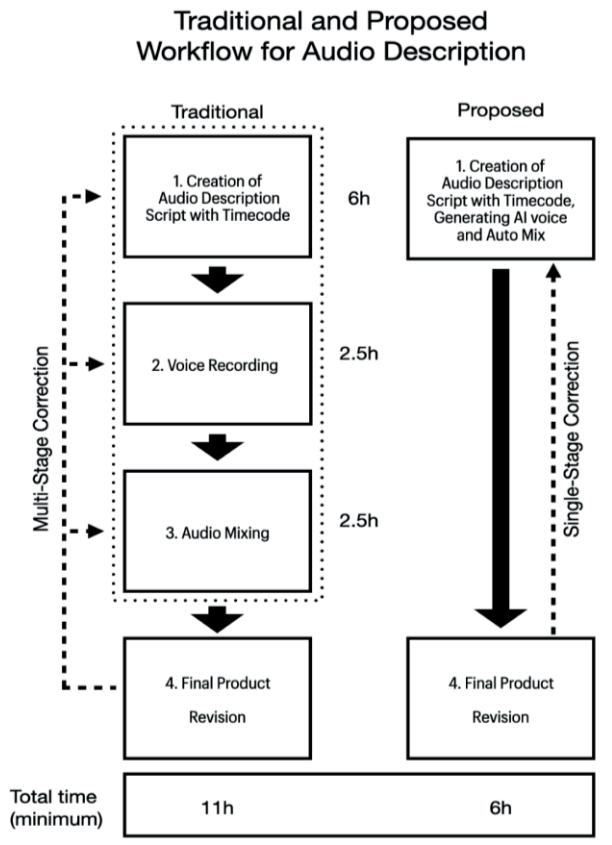


Figure 2. Traditional and proposed workflow for Audio description. The two processes are showed parallel in a vertical disposition. The dotted line arrow shows how the process of the revision can make influence in the different stages and cause a cascade correction effect.

1- Script Creation: A document containing the information to be narrated in the audio description is created by an audio describer, based on the finished audiovisual content. In this stage, to create the speech that narrates the scene, the audio describer already considers the space between existing voices in the audiovisual content, so that the narration does not overlap any dialogue. Then, a narration script is created from the approximate timecode already marked during the scriptwriting process. The minimum estimated time to adequately complete this stage for a 40-minute program is about 6 hours.

2- Audio Recording: The narrator must have a clear and pleasant voice and must be able to read the script text naturally and neutrally. It is also necessary to have high-quality recording equipment, minimally compatible with a broadcast standard. This includes a recording booth with adequate acoustics and quality microphones. For content with multiple episodes, it is important to use the same voice to maintain narrative unity. The voice is perceived by the viewer as a character, and changing it generates a strangeness similar to changing an actor for the same role during a program. The minimum estimated time to adequately complete this stage for a 40-minute program is about two and a half hours.

3- Audio Mixing: The audio recorded by the narrator is mixed with the original video. The audio editor must adjust the volume of the description audio so that it does not interfere with the original video sound. This primarily involves avoiding overlap between the audio description narration and dialogues originally contained in the video. For this, the audio editor can often use the artifice of speeding up the narration to fit it into a specific time frame. If the editor feels that the result was not satisfactory, they can request a re-recording or a text modification to fit that time. Ambient sounds and sound effects are also important for understanding the scene, as they are an integral part of the actions and help the viewer understand what type of environment a scene occurs in. The minimum estimated time to adequately complete this stage for a 40-minute program is about two and a half hours.

4- Evaluation of the Final Content: A team of technical evaluators verifies, among other things, the quality of the audios and the mix. It is worth mentioning that at all stages, the participation of a consultant is extremely important for a good understanding of the product. This consultant necessarily needs to be someone with visual impairment and usually not only accompanies the final product but can also follow and interfere in other stages of the process. All the times shown here do not consider eventual changes and corrections in the previous stages, where it is often necessary to redo the entire process to change a section.

The time estimation for each stage was based on the experience of producing AD of drama content made by Grupo Globo. This time refers to an estimate of the minimum necessary to deliver a product in suitable conditions. However, this value can increase significantly according to the complexity of each work and unforeseen events that may occur in a production.

The proposed process would modify the traditional workflow by altering stages 2 and 3. It would be executed by the scriptwriter themselves during stage 1, where the script would be inputted into a program, and the voice generation

This open access article is distributed under a Creative Commons Attribution (CC-BY) license.
https://set.org.br/ijbe doi: 10.18580/setijbe.2023.5 Web Link: https://dx.doi.org/10.18580/setijbe.2023.1

and mixing would be immediately done, which can be tested and validated by the scriptwriter.

IV. Challenges and Advantages of Using the New Workflow

In stage 2, recording the narration would be replaced by generating synthetic voice. It is important to maintain the same requirements at this stage of 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. With the model presented in session 2, we believe we have enough quality to meet these criteria. An advantage that synthetic voice can bring at this stage 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, avoiding possible changes in the voice that the viewer is used to hear.

Regarding the mixing process in stage 3, it is also important to maintain the characteristics required in the traditional workflow. A drama mix have an extensive dynamic sound level variation, that is a significant challenge in creating a mixing system. An automatic mixing system was created that kept the AD narrator voice intelligible in varied sound dynamic levels environments. In this system, it is 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. The AD narration was placed at a volume similar to the dialogue level, and the program could interpret the dynamics of current events to act only where and when necessary. In order to avoid overlap of in content dialog with AD, the scriptwriter validates the AD voice duration as the script is created, making the necessary changes to the text, fitting the speech excerpts within the dialog gaps.

This would significantly reduce the time needed to produce an AD, considering that stages 2 and 3 happen instantly. 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 the time it takes to create an AD.

Another advantage that emerges from the proposed 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.

V. Conclusion:

In this work, we evidenced how a significant portion of the population can benefit from audio description. We discussed about some of the existing difficulties for large-scale implementation of it in a television production flow. We mapped the technical stages necessary to perform an audio description. For each stage, the necessary requirements were raised to perform quality audio description. From this understanding, we propose a method that reduces implementation time, complexity, and therefore the production cost. In this way, the use of artificial intelligence becomes an important tool that can speed up and simplify the AD process. The time to make a quality audio description often conflicts with the agility needed in a production line of television drama content. We believe that the model proposed here can lead to a more comprehensive and efficient implementation of audio description. The demand for this type of content exists and is not always met. Our proposal can not only help fulfill a relevant social role for the media production but also enable the integration of a significant portion of the population that still has no independent access to some kinds of content (such as daily drama). With this, we can deliver programs to a more diverse public and thus potentially increase the quality of life and social inclusion of millions of people.

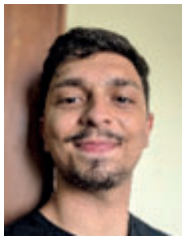
References

1. J. Snyder, (2022, Aug) “Fundamentals of audio Description” Available: https://adp.acb.org/adi/ADA%20Fundamentals.doc.pdf
2. MEC (Brazilian ministry of education) “Data reafirma os direitos das pessoas com deficiência visual ” Available: http://portal.mec.gov.br/component/tags/tag/deficiencia-visual
3. Agência de notícias do IBGE (Brazilian statistic and Geography News Agency) “PNS 2019: país tem 17,3 milhões de pessoas com algum tipo de deficiência” Available: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/31445-pns-2019-pais-tem-17-3-milhoes-de-pessoas-com-algum-tipo-de-deficiencia
4. L.M. Villa, P. Romeu. (2010). *Audiodescrição : transformando imagens em palavras.* (1st ed.) [Online]. Available: https://www.ufrgs.br/comacesso/wp-content/uploads/2019/01/Audiodescric%CC%A7a%CC%83o-Transformando-Imagens-em-Palavras.pdf
5. Anatel: Portaria 188 de 2010 - *Dispõe sobre audiodescrição e estabelece novos prazos de implementação*. Available: https://informacoes.anatel.gov.br/legislacao/normas-do-mc/443-portaria-18.x
6. P. H. L. Leite, E. Hoyle, Á. Antelo, L. F. Kruszielski, and L. W. P. Biscainho, “A corpus of neutral voice speech in Brazilian Portuguese,” in *Computational Processing of the Portuguese Language*, Fortaleza, 2022, pp. 344–352.
7. J. Shen, R. Pang, R. J. Weiss, et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, 2018, pp. 4779–4783.
8. G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, “Multi-band melgan: Faster waveform generation for high-quality text-to-speech,” *2021 IEEE Spoken Laguage Technology Workshop (SLT)*, Shenzen, 2020 pp. 492–498.
9. I. Goodfellow, Y. Bengio e A. Courville, “Convolutional Networks” in *Deep Learning Book,* 1st ed. Cambridge, 2016.

This open access article is distributed under a Creative Commons Attribution (CC-BY) license.
https://set.org.br/ijbe doi: 10.18580/setijbe.2023.5 Web Link: https://dx.doi.org/10.18580/setijbe.2023.1

10. I. Goodfellow, Y. Bengio e A. Courville, “Sequence Modeling: Recurrent and Recursive Nets,” in *Deep Learning Book,* 1st ed. Cambridge, 2016.
11. I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative Adversarial Nets,” in Advances in Neural Information Processing Systems,vol. 27, Montreal, 2014.
12. K. Kumar, R. Kumar, T. de Boissiere et al., “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” em *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, 2019.

Luiz F. Kruszielski is graduated in Music-Sound Production at UFPR (Curitiba, Brazil 2004) and have a master (Tokyo, Japan 2011) and a doctor degree in Sound and the Environment at Tokyo University of the Arts (Tokyo, Japan 2013). He worked as a professional sound designer since 2003, and from 2013, he works at Globo TV Network (Rio de Janeiro, Brazil) as a researcher for sound technologies, later becoming a Sound Producer, where he was the technical responsible of the sound for more than 10 drama series and telenovelas for a total of more than 400 episodes. He currently work as an Innovation Specialist in the same institution.



Pedro H. L. Leite is graduated in Electronics and Computer Engineering (BSc.) at UFRJ (Federal University of Rio de Janeiro, 2021). He currently works as an innovation researcher at Grupo Globo and is a masters student at the Audio Processing Group at the Signals, Multimedia and Telecommunications lab in UFRJ (GPA/SMT-UFRJ). His main research interests are audio/speech processing and artificial intelligence.

Pedro Bravo is an undergraduate student in Control and Automation Engineering at UFRJ. Engaged in research in the field of Biomedical Engineering, focusing on electromechanical systems and biological signals for myoelectric prosthetics. Currently working in the Innovation department at Globo, focusing on projects related to programming and machine learning.

Edmundo Hoyle received his BSc. in Physics at National University of Trujillo (Peru) in 2004 and his DSc at the Federal University of Rio de Janeiro (Brasil) with specialization in Image Processing in 2013. Currently, he works as researcher at Grupo Globo and his main research interests are image processing, computer vision and Artificial Intelligence.

Marcelo Lemmer graduated in Music Education from UFRJ, worked as a music producer and since 2017 has been working as an audio description consultant for theater plays, short films, live musical events, and series. Currently, works in the area of audio/audio description production at Rede Globo.

Received in 2023-06-08 I Approved in 2023-07-08

This open access article is distributed under a Creative Commons Attribution (CC-BY) license.
https://set.org.br/ijbe doi: 10.18580/setijbe.2023.3 Web Link: https://dx.doi.org/10.18580/setijbe.2023.3