 doi:10.20944/preprints202211.0017.v1

Performance Comparison of TTS Models for Brazilian Portuguese to Establish a Baseline

Wilmer Lobato, Felipe Farias, William Cruz, Marcellus Amadeus

Alana AI Research, Brazil
{wilmer.lobato, felipe.farias, william.cruz, marcellus}@alana.ai

ABSTRACT

This paper compares the performance of three text-to-speech (TTS) models released from June 2021 to January 2022 in order to establish a baseline for Brazilian Portuguese. Those models were trained using dataset for Brazilian Portuguese. The experimental setup considers tts-portuguese dataset to fine-tune t he f ollowing T TS m odels: V ITS end-to-end model; glowtts and gradtts acoustic models both using hifigan vocoder. Performance metrics are arranged into objective and subjective metrics. As subjective metrics, the naturalness and intelligibility are measured based on the mean opinion score (MOS). Results shows that gradtts+hifigan model achieved naturalness of 4.07 MOS, close to performance of current commercial models.

Index Terms— text-to-speech; naturalness; intelligibility; Brazilian Portuguese

1. INTRODUCTION

Intelligent virtual assistants such as Amazon Alexa, Apple Siri, Microsoft Cortana and Google Assistant are machine-learning powered software capable to improve customer-machine interactions, making them more realistic and similar to human beings [1, 2]. They are composed of two complementary systems: automatic speech recognition (ASR), and text-to-speech (TTS). In case of a TTS model, it receives text (or phonemes) as input and produces synthesized speech as output with desired features like emotion, intonation, rhythm, etc [3].

A comprehensive TTS or end-to-end model is divided into two modules: acoustic model and neural vocoder. An acoustic model converts input text (or their equivalent phonemes) into mel-scalograms, which is a time-frequency representation. Some examples of this model are: dctts [4] and tacotron 2 [5]. On the other hand, neural vocoder converts a mel-scalogram into a synthesized speech. Some examples of this model are: griffin-lim [6], wavenet [7] and hifigan [8].

During the last years, TTS models have strongly improved their performance in terms of naturalness and intelligibility of synthesized speech. In [9], a survey focused on recent TTS models highlight a timeline of acoustic models, neural vocoders and end-to-end models released from the last five years (from 2016 to 2021). From here, the state of the art indicates that some TTS models such as VITS [10] (end-to-end model), glowtts [11] and gradtts [12] (acoustical models) both with hifigan vocoder [8], achieved very promissory results for English datasets like LJSpeech and VCTK.

Despite Brazilian Portuguese being the 6^{th} most spoken language worldwide (around 230 million native speakers), current TTS models are designed for English language [3]. In this way, and according to [1] and [3], Brazilian Portuguese is considered a low-resources language to develop TTS models due to the lack amount of high-quality speech corpora (with transcription) to train current deep learning models, becoming a great challenge for scientific community. However, some recent works as [13] and [3] expose the lack of TTS models trained with Brazilian Portuguese dataset as well as the lack of reported performance metrics to establish a baseline for comparison.

Thus, the aim of this work is to establish a baseline of TTS models for Brazilian Portuguese and compare their performance with the state of the art and commercial models. This comparison is based on widely known metrics such as naturalness and intelligibility of synthesized speech.

The remainder of this paper is organized as follows. Section 2, related works, describes the most relevant contributions on this topic. Section 3, experimental setup, describe all resources uses for model training, such as hardware specifications, datasets and implemented TTS models. Sections 4, performance evaluation, expose and compare the performance metrics obtained. Section 5, discussion, discourse about the obtained results and drawbacks. Finally, Section 6, conclusion, encloses the article and determines the fulfillment of the objectives and also defines some directions for future works.

2. RELATED WORKS

According to the current literature in TTS models for Brazilian Portuguese, [9] mention two main approaches on speech synthesis for Brazilian Portuguese: stochastic and neural. On the stochastic (or statistical parametric) approach, the speech is synthesized from a set of predicted acoustical features such as fundamental frequency, spectrum or cepstrum [9]. Exam-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0017.v1

ples are the Hidden Markov Models (HMM) implemented in [14], [15] and [1].

On the other hand, neural approach uses neural network based models instead parametric models in order to simplify acoustic features. The main advantages are a high voice quality in terms of naturalness and intelligibility, as well a few of human preprocessing. Thus, in the literature it was found two works. Firstly, in [13], it was trained a tacotron 2 [5] acoustic model with griffin-lim vocoder [6] using Common Voice dataset. Secondly, in [3], it was compared the performance between: a) dctts [4] acoustic model achieving a score of 3.03 MOS (mean opinion score), from 1-bad to 5-excellent; and b) tacotron 2 [5] acoustic model with wavenet [7] vocoder achieving 4.02 MOS, being the best reported performance for Brazilian Portuguese.

Despite these preliminary works, it is not clear which is the best open dataset for Brazilian Portuguese, how wide is the phonetic spectrum, and mainly which current TTS models have the better performance in terms of naturalness and intelligibility. In this regard this resarch paper contributes in order to outline a baseline of neural-based TTS models for Brazilian Portuguese.

3. EXPERIMENTAL SETUP

This section describes the resources used for all experiments and it is organized as follows: hardware and software environment, datasets, models, and model training.

3.1. Hardware and software environment

All experiments were run on a cloud instance with the following features: 8 vCPU with 61GB memory, 1 GPU Tesla V100 video card with 16GB memory, Linux operational system, using Python v3.9 and PyTorch v1.10.

3.2. Datasets

A high-quality TTS model such as [5, 10] needs to be trained by using a dataset with the following minimum features: 10 hours of recordings per speaker, sampling rate of 22.05kHz, 16 bits of resolution, and noise-controlled (NC). Despite Brazilian Portuguese is considered a low-resource language for TTS applications, the available open datasets used to establish the baseline are listed in Table 1, which describes the number of speakers (M - male, F - female), duration (in hours), if NC exists, and the number of bits. All datasets mentioned in Table 1 have a sampling rate of 22.05 kHz.

Although laps-bm [16] dataset has low duration, it has a phonetically-balanced text corpus, which is highly desired to train TTS models with wide phonetic coverage by reducing errors during speech synthesis. Constituição [16] dataset contain more audio recordings but a phonetically-unbalanced text corpus, and containing very long sentences (28s on average) which might difficult the model convergence. On the another hand, tts-portuguese [3] and globo [17] datasets have enough audio recordings, however, it is necessary to compare their phonetic richness for Brazilian Portuguese.

Table 1. List of audio corpora (with transciption) available for Brazilian Portuguese. NC is noise-controlled.

ref	dataset	speakers	duration	NC	bits
[16]	laps-bm	1M	1hrs	no	16
[16]	constituição	25M,10F	9hrs	no	16
[3]	tts-portuguese	1M	10hrs	no	32
[17]	globo	1M	20hrs	yes	16

In this way, we computed the percentage occurrence of each Brazilian Portuguese phoneme along tts-portuguese and globo text corpora. So, Figure 1 show the corresponding phonetic distribution using histograms for tts-portuguese (in purple bars) and globo (in pink bars). From this, we can verify that tts-portuguese has a broader phonetic coverage than globo dataset despite a few hours of recordings.

3.3. Models

As commented above, some acoustic models such as dctts [4] and tacotron 2 [5] were trained for Brazilian Portuguese datasets in [13] and [3]. In the last three years, recent TTS models were released such as described in [9] among acoustic models, neural vocoders and end-to-end models.

We have chosen the best three TTS models based on the following criteria: i) open-source; and ii) the highest reported MOS score for English datasets. According to this criteria, we decided to compare the following three models: a) glowtts acoustical model with hifigan vocoder, which achieved 4.01 MOS for LJSpeech dataset [11]; b) VITS end-to-end model which achieved 4.43 MOS for LJSpeech and 4.38 MOS for VCTK dataset [10]; and c) gradtts acoustical model with hifigan vocoder, which achieved 4.44 MOS for LJSpeech [12].

Glowtts [11] is a flow-based generative model that does not require external aligners between text and speech as preprocessing. To do this, glowtts combine the properties of flow and dynamic programming to search the most probable monotonic alignment.

VITS [10] is a parallel end-to-end TTS that use variational inference augmented with normalizing flows and adversarial training processes, which improves the expressive power of generative modeling.

Finally, gradtts [12] is a model with score-based decoder producing mel-spectrograms by gradually transforming noise predicted by encoder and aligned with text input by the monotonic alignment search.

3.4. Model training

In this section is presented the training process of the glowtts, VITS and gradtts models. In addition, in Figure 2 is illustrated the training curves of each model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0017.v1

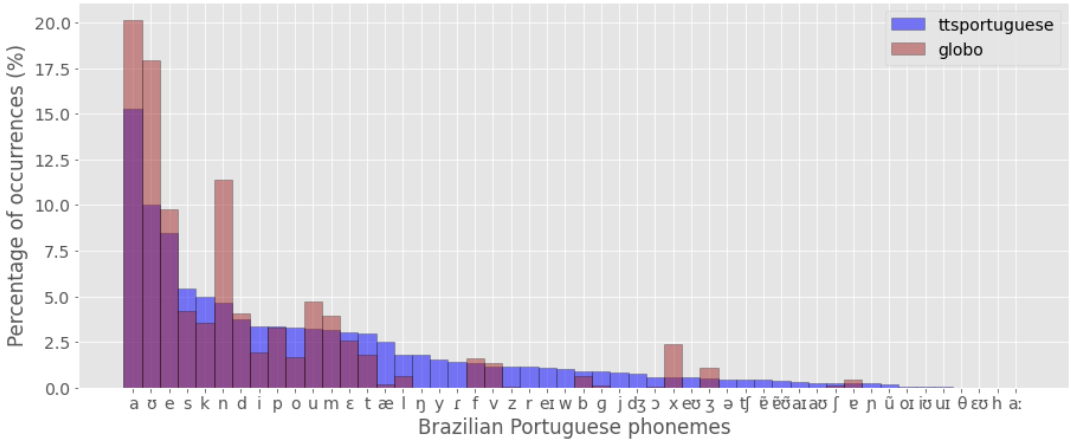


Fig. 1. Phonetic histogram obtained from tts-portuguese (in purple bars) and globo (in pink bars) text corpora.

- **glowtts+hifigan**: this experiment considered a pre-trained glowtts acoustic model with hifigan neural vocoder using tts-portuguese dataset during 300k steps and 15k steps, respectively. The fine-tuning process considered a synthetic dataset previously generated from a commercial TTS model divided into 70% for training, 15% for validation and 15% for tests. The fine-tuning process took 100k steps (approx. 2.3 days) for glowtts model. Figure 2a shows the glowtts average loss function during fine-tuning.

- **VITS**: this second experiment considered a pre-trained VITS end-to-end model using LJSpeech dataset (English) during 1M steps and fine-tuned using tts-portuguese dataset and phonemizer for Brazilian Portuguese. The dataset was divided into 70% for training, 15% for validation and 15% for tests, taking 128k steps (approx. 7.3 days) for fine-tuning. Figure 2b shows the VITS average loss function during fine-tuning.

- **gradtts+hifigan**: finally, the third experiment considered a pre-trained gradtts acoustic model with hifigan neural vocoder using LJSpeech dataset (English) during 2.5k epochs. The fine-tuning process considered tts-portuguese dataset during 1.5k epochs (270k steps, approx. 1.8 days) for gradtts model using the same sets for training, validation and tests than VITS model. Figure 2c shows the gradtts prior loss function during fine-tuning.

4. PERFORMANCE EVALUATION

This section describes the performance measure of each fine-tuned TTS model. This evaluation was divided into objective and subjective metrics.

4.1. Objective metrics

Objective metrics listed in Table 2 are divided into two aspects a: a) robustness: it quantifies how synthesized speech is correctly transcribed by listeners, such as character error rate (CER), word error rate (WER) and word information lost (WIL) [18, 19]; and b) latency: it measures the latency to synthesize speech, such as the real time factor (RTF) in [12].

Table 2. Objective metrics to compare TTS models based on robustness (WER and WIL) and latency (RTF).

models	WER	WIL	RTF
glowtts+hifigan	32.3%	47.2%	0.05478
VITS	27.2%	39.2%	0.04605
gradtts+hifigan	17.1%	28.8%	0.04484

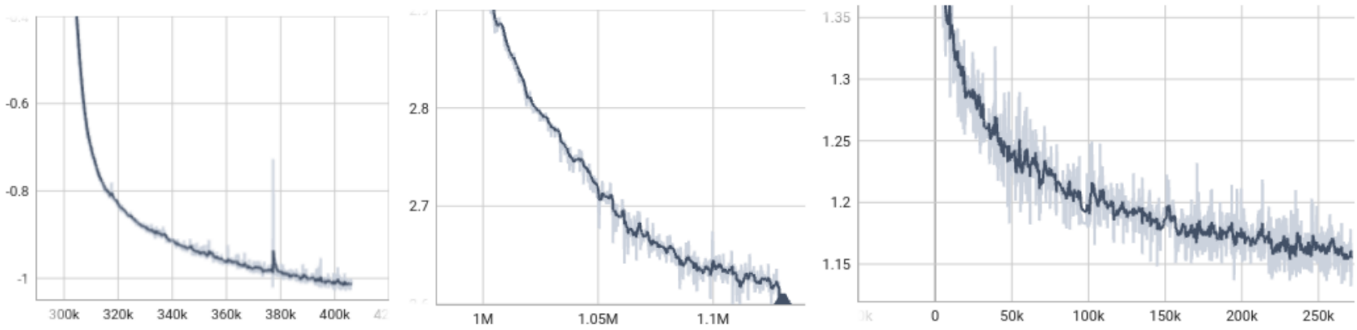
4.2. Subjective metrics

Subjective evaluation was performed using the mean opinion score (MOS), obtained from 5 hearing-healthy people through an online form[1]. For this experiment, a set of 30 sentences not used for training were randomly selected and synthesized for each model. After that, each volunteer assessed two important aspects of synthetic speech: naturalness and intelligibility.

Naturalness measures the degree of similarity of the synthesized speech with the speech produced by a human being. On the another hand, intelligibility is related to the degree of understanding of each heard sentence. Both naturalness and intelligibility are measured by scoring each synthesized audio file on a five-point scale as follows: 1-bad, 2-poor, 3-fair, 4-good and 5-excellent. Similar assessments were conducted in related works as [3]. In addition, the listener transcribed

[1]The authors would like to thanks Alana AI for the financial support and all collaborators that participated on the audio assessment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0017.v1



(a) glowtts (b) VITS (c) gradtts

Fig. 2. Convergence of training curves: average loss for a) glowtts (left); b) VITS (center); and prior loss for c) gradtts (right).

the heard sentence in order to compute the percentage of correct/mistaken words (i.e. the WER).

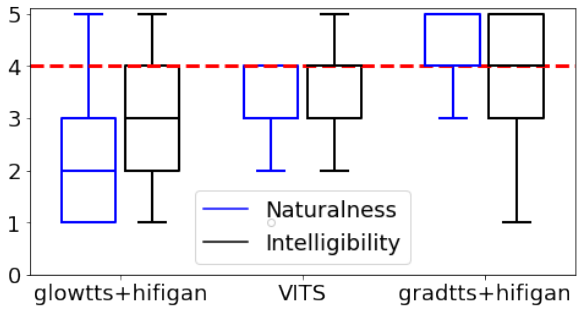


Fig. 3. Naturalness and intelligibility depicted as blue and black boxplots of the following TTS models: glowtts+hifigan [11], VITS [10] and gradtts+hifigan [12] models.

Figure 3 shows the boxplots for naturalness (blue) and intelligibility (black) of glowtts+hifigan, VITS and gradtts+hifigan models, representing the median value and quartiles. From Figure 3, we obtained the following naturalness metrics: a) glowtts+hifigan achieved 2.20 ± 0.98 MOS, b) VITS achieved 3.20 ± 0.62 MOS, and c) gradtts+hifigan achieved 4.07 ± 0.53 MOS. In terms of intelligibility, it is also possible to note a better performance by gradtts+hifigan. finally, a baseline of 4 MOS for speech naturalness was established through a red dashed line for Brazilian Portuguese.

5. DISCUSSION

Firstly, in Section 3.2, we perform a brief study about the features of open datasets for Brazilian Portuguese such as: constituição and laps-bm datasets [16], tts-portuguese [3] and globo [17]. From this, we concluded that tts-portuguese is the best open dataset to train a TTS model for Brazilian Portuguese mainly due to its phonetic coverage and duration of recording, sampling rate, bit resolution, etc.

Secondly, in Section 3.3, we selected three models: glowtts+hifigan [11], VITS [10] and gradtts+hifigan [12]. These models achieved more than 4.0 MOS of naturalness for English datasets. Third, in Section 3.4, these models were fine-tuned using tts-portuguese dataset. Figure 2a, 2b and 2c showed that all loss function converged, but VITS model took more than 3 times to converge than others.

Section 4.1, objective metrics demonstrated that gradtts+hifigan model has the highest robustness by achieving the lower WER and WIL. In addition, gradtts+hifigan also achieved the lowest latency measured trough the RTF, despite there are no significant statistical differences. In addition, in Section 4.2, subjective metrics measured the level of naturalness and intelligibility of synthesized speech based on the MOS scale. Results showed that gradtts+hifigan achieved the highest average naturalness of 4.01 MOS (with an upper bound of 5.0 MOS). This result is close to naturalness reported by commercial models. In addition, in terms of intelligibility, gradtts+hifigan achieved close to 4.0 MOS. It is important to emphasize that performance also depends on the quality of original recordings.

6. CONCLUSION

This work aimed to establish a baseline of text-to-speech models for Brazilian Portuguese, which is considered a low-resource language for this application. Along this study, we fine-tuned and compared the performance of glowtts+hifigan in [11], VITS in [10], and gradtts+hifigan in [12] using tts-portuguese dataset in [3]. Results showed that gradtts+shifigan model achieved the highest naturalness, intelligibility, robustness and latency. Obtained results are close to current commercial TTS models for Brazilian Portuguese. In this way, a minimum baseline of 4 MOS for naturalness was established.

As future work, the authors are interested in: a) fine-tune recent neural vocoders with better performance than hifigan; and b) propose a novel text corpus for Brazilian Portuguese with wide phonetic coverage and long duration.

7. REFERENCES

[1] Nelson Neto, Willian Rocha, and Gleidson Sousa, "An open-source rule-based syllabification tool for Brazilian Portuguese," *Journal of the Brazilian Computer Society*, vol. 21, no. 1, pp. 1–10, 2015.

[2] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee, "Alexa, Can i Trust You?," *Computer*, vol. 50, no. 9, pp. 100–104, 2017.

[3] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti, "SC-GlowTTS: an Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," *ArXiv*, apr 2021.

[4] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 4784–4788, 2018.

[5] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 4779–4783, 2018.

[6] Daniel W. Griffin and Jae S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *ArXiv preprint*, pp. 1–15, 2016.

[8] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[9] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A Survey on Neural Speech Synthesis," *ArXiv*, 2021.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning PMLR*, 2021.

[11] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, no. NeurIPS, 2020.

[12] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning PMLR*, 2021.

[13] Rodrigo Kobashikawa Rosa and Danilo Silva, "Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim," in *XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT 2021*, 2021, pp. 1–2.

[14] Igor Couto, Nelson Neto, Vincent Tadaiesky, Aldebaro Klautau, and Ranniery Maia, "An Open Source HMM-based Text-to-Speech System for Brazilian Portuguese," in *7th International Telecommunication Symposium (ITS 2010).*, 2010.

[15] Ericson Sarmento Costa, Anderson de Oliveira Monte, Nelson Neto, and Aldebaro Klautau, "Um Framework para Desenvolvimento de Sistemas TTS Personalizados no Português do Brasil," in *XXX Simpósio Brasileiro de Telecomuniações - SBrT 2012*, Brasília, DF, 2012.

[16] UFPA, "Audio corpora for Brazilian Portuguese," 2021.

[17] Pedro H. L. Leite, Edmundo Hoyle, Alvaro Antelo, Luis F. Kruszielski, and Luiz W. P. Biscainho, "A corpus of neutral voice speech in Brazilian Portuguese," in *Lecture Notes in Computer Science (LNBI)*, pp. 344–352. 2022.

[18] Andrew C. Morris, Viktoria Maier, and Phil Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," *8th International Conference on Spoken Language Processing, ICSLP 2004*, pp. 2765–2768, 2004.

[19] Phat Do, Matt Coler, Jelske Dijkstra, and Esther Klabbers, "A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2021, vol. 5, pp. 3306–3310, International Speech Communication Association.