Acoustic Analysis of Prosodic Features in Natural versus Synthesized Speech Samples from YourTTS and SYNTACC Models

Julio Cesar Galdino[1], **Gustavo Evangelista Araújo**[1], **Arnaldo Candido Junior**[2], **Miguel Oliveira Jr.**[3], **Moacir Antonelli Ponti**[1], **Sandra Maria Aluísio**[1]

[1]Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo (USP), São Carlos, SP, Brazil

[2]Universidade Estadual Paulista (UNESP) Rio Preto, SP, Brazil

[3]Universidade Federal de Alagoas (UFAL) Maceió, AL, Brazil

{juliogaldino,gustavo_evangelista}@usp.br, arnaldo.candido@unesp.br

miguel@fale.ufal.br, {moacir, sandra}@icmc.usp.br

Abstract. *This study presents an acoustic analysis of prosodic features in both natural and synthesized speech samples, using two state-of-the-art speech synthesis models: YourTTS and SYNTACC. By analyzing spontaneous speech data, the duration of intonational units and syllables produced by these models was compared. The findings reveal that both models generate speech with significantly shorter and less variable durations of intonational units and syllables compared to natural speech. These results highlight the differences in syllable duration and speech rate between synthesized and natural speech, emphasizing the need for more refined prosodic metrics to accurately assess the quality of synthesized speech.*

Index Terms: Acoustic Analysis of Prosodic Features, Speech Synthesis Models Evaluation, Portuguese language, Spontaneous Speech

1. Introduction

Speech synthesis, a branch of Natural Language Processing (NLP) [Caseli and Nunes 2024], focuses on converting written text into natural-sounding speech. Also known as Text-to-Speech (TTS), this technology has evolved significantly from early mechanical synthesizers to modern systems capable of producing highly realistic voices. Recent advances in Deep Learning have greatly contributed to the development of these systems. Recurrent and convolutional neural networks, as well as Transformer-based models like the one introduced by [Li et al. 2019], have been important in generating high-quality speech. Additionally, flow-based generative models, such as those by [Kingma et al. 2016] and [Hoogeboom et al. 2019], have provided more flexibility in controlling the prosodic features of synthetic speech.

Prominent models from the early 2020s, such as VITS [Kim et al. 2021], YourTTS [Casanova et al. 2022b], and SYNTACC [Nguyen et al. 2023], combine variational inference and adversarial learning techniques to generate high-quality, customizable voices. These architectures enable speech synthesis in multiple languages and dialects, and also offer greater control over aspects such as emotion and style.

The quality of speech synthesis systems applied to resource-rich languages has drawn attention to the possibility of applying these models to low-resource languages. Until mid-2019, there were no datasets with a significant amount of high-quality audio hours available to train deep learning-based speech synthesis models for Brazilian Portuguese (BP) [Caseli and Nunes 2024]. Even with limited data [Casanova et al. 2022a], models have been used for BP over the years, such as YourTTS [Casanova et al. 2022b], which incorporates a multilingual text encoder into the VITS architecture while maintaining speaker identity.

Conventional multi-speaker text-to-speech (TTS) synthesis systems generated speech with a certain level of naturalness, but they were not capable of generating speech in different accents, which motivated the development of the *Synthesizing Multi-Accent Speech By Weight Factorization* (SYNTACC) architecture. SYNTACC was designed to evaluate the ability to generate speech that preserves the international linguistic variants of English. However, the potential to synthesize speech with sensitivity to international accents also raises questions about its applications for regional linguistic variants of a given language. For example, BP is spoken in a country of continental dimensions, with 26 states, and has fewer speech resources compared to English. Therefore, it is important to assess whether the quality of speech synthesis is still maintained in these contexts.

While these TTS models are increasingly capable of producing fluent speech that closely resembles human speech at the phonetic and lexical levels, capturing the full variability of speech beyond these levels remains a significant challenge. Research suggests that these models often struggle to accurately reflect the diversity of speech contexts [Chan and Kuang 2024], resulting in inadequate prosody representation. This shortcoming is critical, as prosody is essential for generating natural-sounding speech.

The quality of TTS models is often evaluated using subjective methods, remarkably the Mean Opinion Scores (MOS), which assigns an overall numerical value from 1 (bad) to 5 (excellent) to a given model. However, it fails to capture the complexity of prosody, particularly nuances like intonation, rhythm, and emphasis, all crucial for the perception of naturalness. On the other hand, auxiliary objective measures such as the Speaker Embedding Cosine Similarity (SECS), only provides the likeliness of the generated speech with respect to those of the natural one for a given speaker, and do not provide a clear indication of what makes speech "natural" and do not clearly indicate which linguistic factors contribute to these judgments. Such limitations highlights the need to develop more sophisticated metrics capable of capturing the diversity and complexity of human speech [Chan and Kuang 2024]; [Chiang et al. 2023]; [Le Maguer et al. 2024].

Considering the limitations of objective and subjective evaluations mentioned earlier, we believe progress in the field of speech synthesis has been constrained by the absence of metrics that assess prosodic attributes such as fundamental frequency, segmental duration, and intensity. Recent studies have proposed new approaches to overcome these limitations by developing more rigorous metrics capable of quantifying prosodic features such as lexical prominence and phrasal structure [Chan and Kuang 2024]; [Galdino 2023]. By deeply exploring the acoustic properties of the synthesized speech signal, these metrics provide a more objective and comparative assessment between natural speech and different synthesis models.

The main aim of this study was to carry out an objective acoustic analysis of speech duration, comparing natural speech samples with those generated by two state-of-the-art text-to-speech (TTS) models: YourTTS and SYNTACC. This study introduces an evaluation method widely recognized in Linguistics, highlighting the importance of detailed prosodic analysis when assessing the quality of synthesized speech. The results offer a foundation for future research, helping to improve TTS models by making their output more natural and expressive. The key contributions of this work are: (1) Comprehensive evaluation of prosodic features in state-of-the-art TTS models for BP; (2) The processing and analysis of a BP dataset; and (3) A detailed analysis of the duration of various linguistic units, including intonational units by utterance type, different syllable types (nuclear and non-nuclear), and pauses, alongside a measurement of speech rate.

2. Data Description and TTS Models Selected

2.1. Revised Sample of the Museu da Pessoa (MuPe) Corpus

Given the significant lack of datasets for evaluating regional linguistic varieties [Matos et al. 2024], we used a revised sample from the Museu da Pessoa (MuPe) Corpus, a virtual and collaborative museum of life stories containing metadata for variations of Brazilian accents (city and state of origin), created by the Tarsila Project[1]. The MuPe life stories audios were automatically transcribed by WhisperX [Bain et al. 2023] using Whisper's large-v2 model [Radford et al. 2023] and diarization via pyannote-audio[2]. To maximize transcription accuracy, linguists performed a manual review, correcting errors in the automatic transcription and validating the results. The final dataset includes detailed information about each audio segment, such as identification, duration, transcription, and demographic information of the speaker. Additionally, annotation labels were included to indicate the presence of noise, abrupt cuts, or other relevant audio characteristics.

The dataset used in this study is referred to as post-processed MuPe-v1, and it is a sample with 87,076 segments and 90.25 hours of speech, providing a reasonable amount of data for research in natural language processing and speech synthesis for BP (Table 1). It features a diversity of 86 speakers from 11 birthplaces in Brazil (states) and two foreign countries (Germany and Portugal) from which the speakers immigrated to Brazil, and a rich vocabulary with 18,116 unique recognized words. The linguistic varieties are not evenly represented in the dataset, that is, the sample is not balanced regarding regional linguistic varieties. The largest representation is from the São Paulo variety, with about 62 hours of audio, followed by Minas Gerais (approximately 6.6 hours) and Pernambuco (approximately 6.4 hours) (Figure 1).

2.2. TTS Models Preprocessing and Training

The MuPe corpus sample was preprocessed (Section 2.2.1) and used to train the SYNTACC and YourTTS speech synthesis models, described in Sections 2.2.2 and 2.2.3.

[1] https://sites.google.com/view/tarsila-c4ai/home

[2] https://github.com/pyannote/pyannote-audio

Table 1. Overview of the post-processed MuPe-v1 dataset

Corpus Data	Values
Language	Brazilian Portuguese
Regional linguistic variations	11
International linguistic variations	2
Speakers	86
Speech style	Spontaneous speech
Text genre	Bibliographic interview
Segments	87,076
Total hours	90.25
Average segment length	3.74 ± 3.5 seconds
Recognized vocabulary size	18,116 unique words
Unrecognized vocabulary size	9,463 unique words

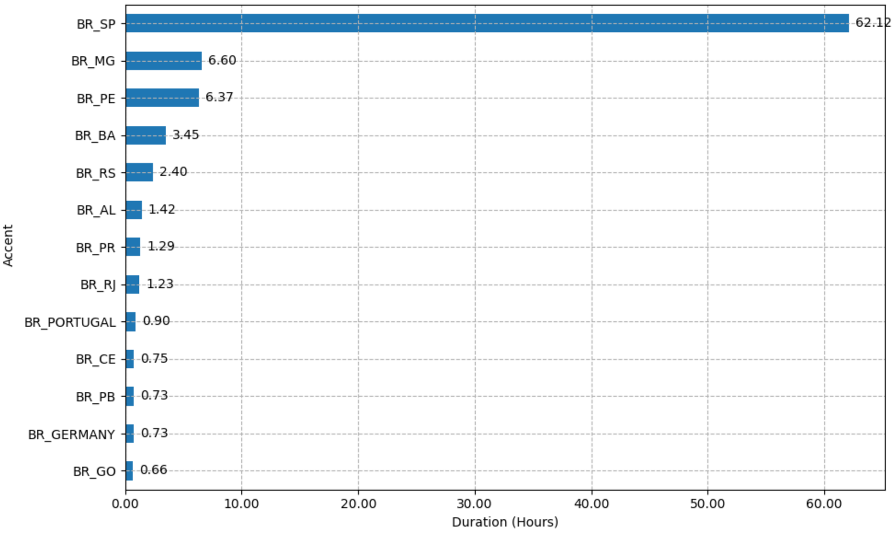

Figure 1. Distribution of audios by place of birth

2.2.1. Preprocessing and Training Settings

Preprocessing included the selection of segments and metadata, table joining, quality analysis of transcriptions, and data cleaning. Typographical errors were corrected, values were normalized, and unnecessary markers were eliminated. For training, weight factorization was applied to optimize training in low-resource scenarios. Both models were trained with a diverse test set, using audio resampled to 16kHz and continuous process monitoring. The "AdamW" optimizer and the "ExponentialLR" scheduler were used to adjust learning rates. For the accent transfer test, 10 utterances from each of the 13 linguistic varieties were selected (Table 1). A script was used to calculate the duration distribution of each accent and classify the utterances into three categories: short, medium, and long. From each category, 10 utterances were randomly selected, following the methodology described by [Nguyen et al. 2023]. Thus, for each linguistic variety, a group of 30 utterances of different durations was generated.

2.2.2. YourTTS: Zero-Shot Multi-Speaker TTS

Similarly to its predecessor VITS, YourTTS uses a Transformer-based encoding-decoding architecture, where the encoder receives the text sequence as input and generates an intermediate representation, which is subsequently processed by the decoder to generate the mel spectrogram, a frequency over time representation that is reconstructed into audio by a vocoder.

YourTTS model is illustrated for the training (left) and inference (right) stages in Figure 2. The *variational autoencoder* (VAE), identified as the *posterior encoder*, receives a linear spectrogram and speaker embeddings during training to predict a latent variable z. This variable is used both as input to the vocoder and in the flow-based decoder, which conditions z and the speaker embeddings to a probabilistic distribution. The *Monotonic Alignment Search* (MAS) aligns the output of the flow decoder with that of the text encoder. The *stochastic duration predictor* diversifies the rhythm in text synthesis by receiving the duration from MAS and the speaker and language embeddings. During inference the flow-based decoder is inverted to output z to the vocoder, so that the posterior encoder can be removed. This model implements: (i) concatenation of 4 trainable language embeddings to each input character, allowing multilingual training; (ii) the Speaker Embedding module, which adapts the model to the voice of different speakers, maintaining the consistency of vocal identity during synthesis, even in a multilingual environment.

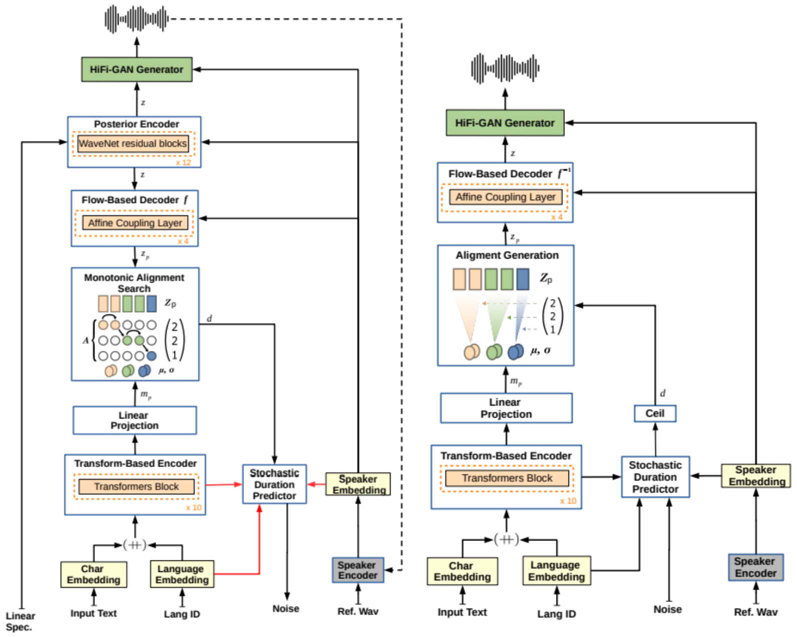

Figure 2. YourTTS diagram representing (left) training procedure and (right) inference procedure [Casanova et al. 2022b].

2.2.3. SYNTACC: Synthesizing Multi-Accent Speech By Weight Factorization

The SYNTACC model [Nguyen et al. 2023] is based on YourTTS architecture, but focuses on speech synthesis with multiple accents via a weight factorization technique. This approach divides the model's weights into shared and accent-specific components, optimizing training in low-resource scenarios. This enables speech synthesis to be adaptable and applicable in multicultural contexts such as BP. It allows explicit control of accents by partially freezing the weights assigned to them, thus enabling speech to be synthesized more specifically with the desired accent.

3. Acoustic Analysis

For the acoustic analysis, utterances from the São Paulo variety were pre-selected due to their representativeness in the dataset and the wider diversity of speakers, which allows for a more robust and generalizable analysis. To ensure data quality and reliability, specific criteria were applied for the final selection of utterances. Only utterances that met the following conditions were included:

1. **Contains at least one Intonational Unit (IU) with complete meaning:** Ensures that intonational properties essential for prosodic analysis are captured.
2. **Is a neutral statement:** Enhances the generalizability of the analysis by avoiding bias introduced by specific content.
3. **Is free of truncations:** Prevents the distortion of intonational features that could result from incomplete utterances.
4. **Has no speech overlaps:** Avoids complications in segmentation, ensuring accurate analysis.
5. **Is free of laughter, coughing, or other extra-linguistic noises:** Prevents interference with prosodic analysis caused by non-linguistic sounds.

A total of 18 utterances were selected for acoustic analysis based on these criteria. These utterances were then annotated using the Praat application [Boersma and Weenink 2024], with segmentation performed at various levels:

1. **IU (Intonational Unit)**: Utterances were segmented into IUs with complete meaning, numbered sequentially. For example, IU0 refers to the first intonational unit, while IU1, IU2, IU3, etc., indicate subsequent units in sequence.
2. **IU+**: This segmentation considers IUs marked by intonational contours and their characteristic nuclei. The nuclear intonational unit (nIU) contains the main information of the utterance. Units following the nIU are numbered as IU+1, IU+2, etc., and those preceding it are labeled as IU-1, IU-2, and so on.
3. **Sil (Orthographic syllable)**: Utterances were segmented into orthographic syllables.
4. **Sil+ (Syllables and pauses)**: This segmentation includes detailed information about different syllable categories, essential for intonational description, along with information about pauses present in the utterances. The categories include:
   - **sa**: The first syllable of an IU+;
   - **spa**: The syllable immediately following sa;
   - **snm**: The last stressed syllable of an IU+ that does not correspond to the end of a complete IU;
   - **sprnm**: The syllable preceding the nuclear syllable in snm;
   - **spsnm**: The syllable following the nuclear syllable in snm, if present;
   - **snf**: The last stressed syllable of an IU+ that corresponds to the end of a complete IU;

- **sprnf**: The syllable preceding the nuclear syllable in snf;
- **spsnf**: The syllable following the nuclear syllable in snf, if present;
- **su**: The last syllable of an IU+, if it exists, categorized as su if there is only one syllable after the core;
- **s**: Any syllable that does not fall into the categories above;
- **ps**: Silent pause;
- **pp**: Filled pause.

5. **SS (Syllables per IU)**: Utterances were segmented into IUs, indicating the number of syllables in each IU, with reference to IU+.

The original utterances (ground-truth, GT) selected for analysis comprise 45 intonational units, 21 of which are nuclear. The average duration of each intonational unit is 1,692 milliseconds, and the total duration of all analyzed utterances, including pauses, is 72,548 milliseconds. The synthesized versions of these utterances showed varied distributions of intonational units and durations, which will be addressed in the results section. Acoustic analysis was carried out using the Praat software, with the duration, minimum, maximum, and average fundamental frequency, as well as intensity, automatically extracted through the AnalyseTier script [Hirst 2012]. Statistical analyses were performed using R software [R Core Team 2024].

4. Results and Discussion

In this study, Analysis of Variance (ANOVA) was used to assess multiple variables, followed by the Tukey test for post hoc comparisons. Statistical significance was determined at a p-value threshold of 0.05.

Table 2 presents the descriptive statistics of the durations of IU by type of utterance, indicating that GT tends to have longer and more variable units, whereas SYNTACC exhibits units with more consistent and relatively shorter durations. Descriptive statistics grouped by utterance type reveal considerable variation in the average duration of units across different groups. There were statistically significant differences in the average durations among the different types of utterances. The differences in durations between SYNTACC and GT (difference of 498.57 ms) and between YourTTS and GT (difference of 435.93 ms) were statistically significant. In contrast, the difference between YourTTS and SYNTACC is not statistically significant. These findings indicate that intonation unit durations vary significantly among types of utterances, with the GT type having the longest average duration and variability, while SYNTACC presents units with the shortest duration and least variability.

Table 2. Average, minimum and maximum duration of IU by type of utterance

Type	Count	Average Duration (ms)	Standard deviation (ms)	Minimum duration (ms)	Maximum duration (dm)
GT	41	1692	876	306	3971
SYNTACC	39	1194	522	306	2737
YourTTS	36	1256	535	380	2832

Thus, speech generated by synthesizers can result in comprehension failures or perception problems, because the shortening of intonation units and the lack of variation in syllables can compromise the rhythmic organization of speech [Cagliari 1992].

Table 3. Speech Rate of Intonational Units

Utterance Type	Average (Syllables per Second)	Standard deviation	Number of Occurrences
GT	6.67	2.31	41
SYNTACC	9.35	1.49	39
YourTTS	9.64	1.24	36

In terms of speech rate, expressed in syllables per second (Table 3), the results revealed significant differences among the three types of utterances. A statistically significant difference in mean syllables per second was observed between the utterance types and among pairs of types. When comparing GT with SYNTACC, it was found that the mean syllables per second is significantly lower in GT (6.67) than in SYNTACC (9.35). Similarly, GT has a significantly lower mean compared to YourTTS (9.64). Between SYNTACC and YourTTS, the difference is smaller but still significant. These results indicate that natural speech (GT) has a lower speech rate compared to the speech synthesized by the SYNTACC and YourTTS models. Although the difference between the two synthesis models is small, YourTTS exhibits a slightly higher rate. It is important to note that, in general, the speech rate for adult speakers of BP in spontaneous speech situations is approximately between 4 and 6 syllables per second [Gonçalves 2017]. The production of faster speech can delete small segments or even unstressed syllables, which impairs intelligibility [Kent and Read 2002].

The descriptive results reveal significant differences in syllable duration between types of utterances (Table 4). Natural utterances (GT) tend to have longer average durations for various syllables, especially the nuclear ones (“snf” and “snm”), compared to those synthesized by the SYNTACC and YourTTS models. Specifically, the average syllable duration in GT expressions was 164 ms, while it was 108 ms in SYNTACC and 104 ms in YourTTS, the latter being the shortest. This analysis highlights that synthesis models generally produce syllables with shorter and more consistent durations, as evidenced by the lower standard deviations compared to natural utterances. This phenomenon can be attributed to the intrinsic characteristics of synthesis models, which often prioritize regularity in syllable duration. We observed a significant overall difference between the groups, indicating that the average syllable durations differ significantly between the types of utterances. Post-hoc tests confirmed the significance of all pairwise statement comparisons. Specifically, the syllable duration of GT statements was significantly longer in both SYNTACC (estimated difference of 56.44 ms) and YourTTS (estimated difference of 60.25 ms). Furthermore, the difference between SYNTACC and YourTTS, although smaller, was also significant (estimated difference of 3.81 ms).

The statistical analysis results on syllable duration by syllable type, differentiated by utterance type, revealed significant differences in the means. Both syllable type and utterance type have significant effects on syllable duration. Additionally, the interaction between syllable type and utterance type was also significant, indicating that the relationship between syllable type and duration varies according to the utterance type. The results of the post-hoc tests, conducted to evaluate specific differences between the mean syllable durations, showed that the mean duration of “GT s” syllables was significantly higher compared to other categories, such as SYNTACC “s” and YourTTS “s”. Furthermore, the comparison between “GT s” and “GT sa” also revealed a significant difference.

Table 4. Average duration and standard deviation of each syllable type by utterance type

Utterance type	Syllable type	Average duration (ms)	Standard Deviation (ms)	Count
GT	s	138.0	76.4	245
GT	sa	118.0	57.0	41
GT	snf	255.0	74.4	11
GT	snm	289.0	128.0	30
GT	spa	155.0	90.7	38
GT	sprnf	132.0	40.6	11
GT	sprnm	152.0	69.6	30
GT	spsnm	67.0	NA	1
GT	su	156.0	55.9	25
SYNTACC	s	97.5	36.5	251
SYNTACC	sa	91.9	35.0	39
SYNTACC	snf	178.0	47.8	10
SYNTACC	snm	152.0	58.2	31
SYNTACC	spa	93.6	38.7	35
SYNTACC	sprnf	117.0	49.3	10
SYNTACC	sprnm	112.0	41.5	28
SYNTACC	spsnm	37.0	NA	1
SYNTACC	su	131.0	47.7	26
YourTTS	s	96.3	36.2	267
YourTTS	sa	92.0	31.5	36
YourTTS	snf	178.0	66.2	13
YourTTS	snm	132.0	45.4	27
YourTTS	spa	102.0	32.3	33
YourTTS	sprnf	110.0	38.0	12
YourTTS	sprnm	107.0	44.4	23
YourTTS	spsnm	47.0	NA	1
YourTTS	su	128.0	51.6	23

In general terms, the average durations varied considerably between different types of syllables and utterances. For instance, the average durations for the "GT" type across different categories showed significant variations, with the longest being for "snm" (289.3 ms) and the shortest for "spsnm" (67.0 ms). This indicates that, regardless of the type of syllable, there is substantial diversity in durations, influenced by the type of utterance. The analysis of the duration of nuclear and non-nuclear syllables revealed significant differences in both natural and synthesized speech data (Table 5).

Table 5. Mean values and standard deviation of nuclear and non-nuclear syllables

Utterance type	Syllable type	Count	Average duration (ms)	Standard deviation (ms)
GT	Non-nuclear	391	139.0	74.0
GT	Nuclear	41	280.0	117.0
SYNTACC	Non-nuclear	390	100.0	39.2
SYNTACC	Nuclear	41	158.0	56.4
YourTTS	Non-nuclear	395	99.1	37.8
YourTTS	Nuclear	40	147.0	56.5

For nuclear syllables, the average duration is considerably longer compared to non-nuclear syllables across all types of utterance. This difference is statistically significant across all types of utterances. Within prosodic boundaries, the duration of syllables occurs consistently and systematically in BP, expressing word stress and rhythm [Ferreira 2014]. Therefore, it is essential that synthesis also maintains this regularity, making it more similar to human speech.

Natural speech (GT) predominantly features silent pauses (11) compared to filled pauses (4) (Table 6). However, synthesis models, particularly YourTTS, exhibit limitations in accurately simulating silent pauses.

Table 6. Means and standard deviations of pause durations

Utterance type	Pause type	Count	Average (ms)	Standard deviation (ms)
GT	Filled (pp)	4	574	177
GT	Silent (ps)	11	483	436
SYNTACC	Filled (pp)	3	282	88.9
SYNTACC	Silent (ps)	4	222	54.7
YourTTS	Filled (pp)	3	203	92.5
YourTTS	Silent (ps)	0	N/A	N/A

There are no statistically significant differences in the mean durations of filled and silent pauses when discriminated by type of utterance. Therefore, the observed differences in mean durations are not statistically significant, suggesting that both filled and silent pauses have comparable durations across utterance types. Although the results are not significant, the pause (silent or filled) marks a prosodic boundary [Raso et al. 2020], playing a role in speech demarcation. In general, the results of the duration analysis show that natural speech (GT) is characterized by greater variability and average duration of IU, syllables, and pauses compared to speech synthesized by the SYNTACC and YourTTS models. Specifically, GT has longer average durations and a lower speech rate, aligning with the expected rates for spontaneous speech. In contrast, the synthesis models produce faster speech with shorter and more consistent syllable durations, with the YourTTS model having a slightly higher speech rate. Nuclear syllables are consistently longer than non-nuclear ones in all types of utterances, and the difficulty of the synthesis models in replicating silent pauses suggests the need for adjustments to improve the naturalness of synthesized speech. In the field of speech synthesis, models already allow for duration control at the utterance level. An example is the Non-Attentive Tacotron, a version of Tacotron 2 that replaces the attention mechanism with a duration predictor, trained on a 354-hour dataset. These models are trained with a large amount of publicly available data, which BP does not yet have [Caseli and Nunes 2024].

5. Conclusions and Future Work

In this study, we employed two recent speech synthesis models, YourTTS and SYNTACC, to conduct an acoustic analysis of prosodic features in samples of both natural and synthesized speech. Our findings revealed that both models synthesize speech with shorter duration and less variation, while natural speech has a slower rate. We emphasize the need for comprehensive acoustic analyses, such as the one presented here, to effectively assess the quality of synthesized speech.

Regarding the study of prosody, we only addressed the aspect of duration in our work. We intend to analyze the fundamental frequency (F0) using Praat scripts [Jadoul et al. 2018] in future work. Additionally, recent studies have proposed models that incorporate prosodic features to improve the quality of synthetic speech. For example, [Raitio et al. 2022] created a fast non-autoregressive parallel neural TTS front-end architecture with hierarchical prosody modeling and control using intuitive prosodic features such as pitch, phone duration, speech energy, and spectral tilt, which are easy to calculate from audio. Thus, this work will be an excellent candidate for evaluation with BP data, to be compared with SYNTACC and YourTTS, described here.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support of São Paulo Research Foundation (FAPESP grant #2019/07665-4) and IBM Corporation. It was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of Oct 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44. We also thank the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) Academic Excellence Program (PROEX), n.o 88887.920395/2023-00.

References

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, pages 4489–4493.

Boersma, P. and Weenink, D. (2024). Praat: doing phonetics by computer [computer program]. *http://www. praat. org/.*

Cagliari, L. C. (1992). Prosódia: algumas funções dos supra-segmentos. *Cadernos de estudos linguísticos*, 23:137–151.

Casanova, E., Junior, A. C., Shulby, C., Oliveira, F. S. d., Teixeira, J. P., Ponti, M. A., and Aluísio, S. (2022a). Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese. *Language Resources and Evaluation*, 56(3):1043–1055.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. (2022b). Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.

Chan, C. and Kuang, J. (2024). Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. In *Proc. Speech Prosody 2024*, pages 27–31.

Chiang, C., Huang, W., and Lee, H. (2023). Why we should report the details in subjective evaluation of TTS more rigorously. In Harte, N., Carson-Berndsen, J., and Jones, G., editors, *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 5551–5555. ISCA.

Ferreira, L. P. (2014). A duração como correlato acústico do acento de palavra no português brasileiro e no espanhol: desafios para o ensino de suprassegmentais e preparação de material didático. *Signum: Estudos da Linguagem*, 17(1):74–101.

Galdino, J. C. (2023). Em 200 metros, vire à esquerda: a entoação dos comandos de GPS. Master's thesis, Universidade Federal de Alagoas.

Gonçalves, C. S. (2017). Taxa de elocução e taxa de articulação em corpus utilizado na perícia de comparação de locutores. *Letras de Hoje*, 52:15–25.

Hirst, D. (2012). Analyse tier praat script.

Hoogeboom, E., Van Den Berg, R., and Welling, M. (2019). Emerging convolutions for generative normalizing flows. In *International conference on machine learning*, pages 2771–2780. PMLR.

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Kent, R. and Read, C. (2002). *The Acoustic Analysis of Speech*. Singular/Thomson Learning.

Kim, J., Kong, J., and Son, J. (2021). Vits: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. ICML*, pages 5530–5540.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Le Maguer, S., King, S., and Harte, N. (2024). The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Li, N., Liu, S., Liu, Y., Zhao, S., and Liu, M. (2019). Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6706–6713.

Matos, A., Araújo, G., Junior, A. C., and Ponti, M. (2024). Accent classification is challenging but pre-training helps: a case study with novel brazilian portuguese datasets. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 364–373.

Nguyen, T.-N., Pham, N.-Q., and Waibel, A. (2023). Syntacc: Synthesizing multi-accent speech by weight factorization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28492–28518. PMLR.

Raitio, T., Li, J., and Seshadri, S. (2022). Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7587–7591. IEEE.

Raso, T., Teixeira, B., and Barbosa, P. (2020). Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences*, 9:105–128.