# Certas Palavras: A 1980s-90s Brazilian Radio Corpus to Test TTS Models in Noisy Multi-Speaker Dialogues

**Gustavo Evangelista Araújo[1], Moacir Ponti[1], Arnaldo Candido Junior[2], Sidney Leal[1,4], Edresson Casanova[5], Renato Moraes Silva[1], Miguel Oliveira Jr.[3], Adriana Barbosa Santos[2], Gustavo Wadas Lopes[1], Sandra Aluísio[1]**

[1]University of São Paulo (USP), [2]São Paulo State University (UNESP)
[3]Federal University of Alagoas (UFAL)
[4]Venturus - Centro de Inovação Tecnológica, Campinas, SP, Brazil
[5]NVIDIA Corporation, São Paulo, SP, Brazil
**Correspondence:** gustavo_evangelista@usp.br

## Abstract

Robust text-to-speech (TTS) systems must be trained on speech that mirrors the variability and imperfections of spontaneous dialogues. However, TTS systems trained on existing Brazilian Portuguese datasets are typically limited to clean, scripted, or studio-recorded speech. Certas Palavras (CP) bridges this gap with 70 hours of spontaneous, multi-speaker dialogues from a Brazilian radio program broadcast in the 1980s–1990s. The extensive manual annotation process captures conversational dynamics, including orality markers, filled pauses, and hesitations. For the analog medium, we annotated non-verbal phenomena as musical interference, noise and segmental corrections, describing a challenging acoustic environment for synthesis. Baseline YourTTS and F5-TTS models were trained on a 9-hour single-speaker subset corresponding to one of the main program hosts. Objective evaluation shows that the synthesized speech remains intelligible, with moderate WER and CER. In contrast, subjective evaluation reveals a clear gap in perceived naturalness, with lower MOS scores and higher inter-rater variability compared to ground-truth audio. Together, these properties make the dataset a strong benchmark for TTS robustness.

## 1 Introduction

Text-to-speech (TTS) systems have reached remarkable levels of naturalness and intelligibility in recent years (Casanova et al., 2024, 2022b; Chen et al., 2025b). However, their performance remains highly dependent on the quality and homogeneity of the training data (Tan et al., 2021; Xie et al., 2025). Most Brazilian Portuguese corpora used for synthesis are based on clean, read, and studio-recorded speech (Pratap et al., 2020; Oliveira et al., 2023; Casanova et al., 2022a), resulting in models that perform well in controlled scenarios but degrade when exposed to spontaneous, multi-speaker, or noisy environments (Leal et al., 2025). Recent multilingual research has shown that training with diverse and imperfect data improves the robustness and adaptability of neural TTS models (Lee et al., 2023), particularly in conversational or emotionally variable contexts. However, there remains a scarcity of Brazilian Portuguese resources that capture these real-world variations.

To address this gap, we introduce *Certas Palavras*[1], a 70-hour archival corpus drawn from a Brazilian radio program broadcasted between the 1980s and 1990s. The dataset captures spontaneous, unscripted conversations between two main hosts, rotating celebrity reporters, and a wide variety of guests, from several states of Brazil. Unlike studio-based corpora, these recordings exhibit analog tape artifacts such as background hiss and compression, along with frequent interruptions, overlapping turns, and non-speech elements, such as music, jingles, and ambient effects.

With the release of this dataset, we provide the following contributions:

- A **70-hour multi-speaker corpus** of spontaneous, noisy, and stylistically rich broadcast dialogue **with demographic information** for ASR, expressive TTS and regional accent tasks: **190 speakers** (136 Male | 54 Female);
- **Human-verified annotations** for diarization and speaker roles (host, guests, and co-hosts), transcriptions and annotations of speech and non-speech interference;
- **Subjective** evaluation of YourTTS and **objective** evaluation of YourTTS (Casanova et al., 2022b) and F5-TTS (Chen et al., 2025b) models, both fine-tuned with *Certas Palavras*.

[1]https://huggingface.co/datasets/nilc-nlp/certas_palavras



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 819–829
April 13-16, 2026 

| Corpus | Hours | Speech Genre | Style | #Spk | License |
|---|---|---|---|---|---|
| **English corpora** | | | | | |
| LJSpeech | 25 | Audiobook reading | read | 1 | CC0 1.0 |
| VCTK | 44 | Read prompts | read | 109 | ODC-By v1.0 |
| LibriTTS | 586 | Audiobook reading | read | 2,456 | CC-BY 4.0 |
| HiFiTTS-2 (22 kHz) | 36.7k | Audiobook reading | read | 5,013 | CC0 1.0 |
| HiFiTTS-2 (44 kHz) | 31.7k | Audiobook reading | read | 4,631 | CC0 1.0 |
| DailyTalk | 20 | Conversational dialogues | spontaneous | 2 | CC BY-SA 4.0 |
| **Brazilian Portuguese corpora** | | | | | |
| Multilingual LibriSpeech (MLS) | 130 | Read audiobooks | read | 54 | CC BY |
| BRSD v2 | 158 | Read prompts | read | 775 | – |
| Multilingual TEDx (PT) | 170 | Stage talks | prepared | – | CC BY-NC-ND 4.0 |
| CORAA ASR 1.1 | 290 | Interviews, readings, conversations and talks | spontaneous, prepared,read | 1,689 | CC BY-NC-ND 4.0 |
| Common Voice 17 (PT) | 175 | Crowdsourced prompts | read | 3,453 | CC0 |
| NURC-SP Audio Corpus | 239 | Sociolinguistic interviews | spontaneous | 401 | CC BY-NC-ND 4.0 |
| MuPe Life Stories | 365 | Biographical interviews | spontaneous | 289 | CC BY-NC-ND 4.0 |
| CML-TTS (PT slice) | ∼69 | Read | prepared | 48 | CC-BY 4.0 |
| **Certas Palavras (ours)** | **70** | **Conversational dialogues** | **spontaneous** | **190** | **CC BY-NC 4.0** |

Table 1: Comparison of Brazilian Portuguese and English speech corpora relevant to this study.

## 2 Related Work

Recent progress in text-to-speech synthesis has been driven by small and high-quality corpora (see Table 1), such as LJSpeech (Ito and Johnson, 2017) and VCTK (Yamagishi et al., 2019), for applications such as voice assistants. Research needs have shifted towards the creation of large ones, such as LibriTTS (Zen et al., 2019), which was derived from audiobooks and features 585 hours of spoken audio from 2,456 speakers, 1,185 of whom are female and 1,271 male, which have supported the development of highly natural zero-shot multi-speaker TTS to replicate the voice of a speaker not seen during training (Casanova et al., 2022b, 2024; Chen et al., 2025a). Recently, HiFiTTS-2 (Langman et al., 2025) was launched to overcome two problems large datasets for training TTS have: (i) non-commercial license because its underlying sources have varying licenses and copyrights; (ii) mixed-bandwidth data, as several datasets are often downsampled to 16 kHz to generate transcripts of their audios. HiFiTTS-2 authors remove data from LibriVox speakers who have explicitly requested that their recordings not be used for machine learning applications, to generate this large-scale speech dataset designed for high-bandwidth speech synthesis, with 2 subsets (22.05 kHz subset and 44.1 kHz subset). However, these resources consist primarily of clean, read speech recorded in controlled studio conditions, offering limited diversity in speaker interaction, background conditions, and expressive variation. In contrast, corpora such as DailyTalk (Lee et al., 2023) have introduced conversational dialogues designed to evaluate prosody, turn-taking, and speaker consistency, demonstrating the importance of spontaneous and multi-speaker data for training robust synthesis systems.

For Brazilian Portuguese, the available corpora remain comparatively narrow in both acoustic and stylistic scope. Datasets such as the Portuguese subset of Multilingual LibriSpeech (Pratap et al., 2020), BRSD v2 (Macedo Quintanilha et al., 2020), and Common Voice 17 (Ardila et al., 2020) provide extensive read material collected in consistent recording setups. Other initiatives, including Multilingual TEDx (PT) (Salesky et al., 2021) and CORAA ASR 1.1 (Candido Junior et al., 2022), add a mixture of read, prepared and spontaneous speech but remain acoustically homogeneous relative to real conversational scenarios. NURC-SP Audio Corpus (Lima et al., 2025) and MuPe Life Stories (Leal et al., 2025) extend the range of spontaneous speech available in Brazilian Portuguese, introducing sociolinguistic and biographical interviews that highlight narrative expressiveness and demographic and accent diversity. Nevertheless, none of these corpora capture the acoustic complexity and broadcast realism present in archival media with conversational dialogue speaking genre.

## 3 Certas Palavras: A radio program of books and ideas

*Certas Palavras* was a Brazilian radio program conceived and hosted by Jorge Marques de Vasconcellos and Claudiney José Ferreira. The program focused on books, ideas, and the Brazilian publishing market, establishing itself as a reference point in the literary and artistic spheres during the 1980s

and 1990s. Its independent production model, centered on interviews, debates, and cultural profiles, attracted prominent figures from both national and international artistic circles.

The CEDAE (from the Portuguese *Centro de Documentação Cultural "Alexandre Eulalio"*) currently preserves the Programa Certas Palavras archive, composed of recordings of interviews and reports broadcasted until the mid-1990s, along with additional historical documentation. Since 2015, this collection has been digitized. In a partnership with the Center for Artificial Intelligence(C4AI) at the University of São Paulo (USP), it was possible to digitalize the remaining tapes and cassettes, ensuring broader access to this unique cultural heritage.

*Certas Palavras* fills the gap of previous works by combining spontaneous multi-speaker dialogue, overlapping turns, and environmental noise typical of analog recordings, creating a new category of "broadcast conversational" corpus for Brazilian Portuguese. The dataset complements clean, read resources such as Common Voice and BRSD, and spontaneous but structured ones like MuPe Life Stories, by introducing variability in background conditions, speaking style, and high speaker interaction.

Together, these contributions position *Certas Palavras* as a unique benchmark for evaluating TTS models under realistic and adverse acoustic and stylistic conditions, a complementary resource for advancing TTS research in Brazilian Portuguese.

Specifically, it can enable (i) assessment of generalization, expressiveness, and robustness of Brazilian Portuguese TTS systems in noisy, dynamic, and conversational settings; (ii) evaluation protocols for speaker diarization (iii) construction and evaluation of regional accent tasks, leveraging speaker diversity and accent/dialect variation; (iv) complementary analysis of ASR robustness; and (v) testing of prompt-driven audio separation/segmentation models such as Meta's SAM Audio (Shi et al., 2025).

Beyond expanding linguistic and stylistic coverage for Brazilian Portuguese, this corpus contributes to the growing focus on robustness evaluation in TTS research (Chen et al., 2025a; Lee et al., 2023; Casanova et al., 2022b). It encourages a shift from purely naturalness-oriented evaluation toward assessing the ability of synthesis systems to maintain intelligibility, timbral consistency, and expressive stability under diverse input conditions.

### 3.1 Audio Processing and Annotation

The transcription process of *Certas Palavras* began with an automatic pipeline using WhisperX (Bain et al., 2023), which provided high-quality transcriptions and precise timestamp alignment through the large-v2 Whisper model[2], combined with Voice Activity Detection (VAD) and speaker diarization from Pyannote-Audio (Bredin, 2023; Plaquet and Bredin, 2023).

Audio recordings were standardized to a 16kHz sampling rate, and to mono channel format to ensure temporal and acoustic consistency. From November 2023 to May 2024, the automatically generated transcripts were subsequently manually verified and enriched by a team of 12 annotators, who corrected automatic transcription errors, and added real speaker identity labels to the generic ID label attributed by Pyannote-Audio.

Fleiss's kappa was used to assess the reliability of agreement between the 12 raters, on two Certas Palavras radio programs (`004_CP700` with 300 classifying segments and `001_CP344` with 362 classifying segments) for the diarization task. The kappa values were 0.809 and 0.962, respectively. Each episode/program was further annotated with guest and reporter names, demographic information, and broadcast dates, using metadata obtained from the archival collection of Carmem Silvia P. Teixeira (1997).

The final transcripts followed a structured set of 11 annotation guidelines addressing both verbal and non-verbal phenomena. These rules covered (i) orality markers, (ii) filled pauses, (iii) repeated hesitations, (iv) number and letter transcription, (v) acronyms, (vi) foreign terms, (vii) punctuation and capitalization, (viii) expressive or emotional sounds (e.g., laughter), (ix) misunderstood words, (x) segmentation corrections, and (xi) sentences with music or sound effects (SFX). The same transcription protocol was adopted in the MuPe Life Stories corpus (Leal et al., 2025) and NURC-SP Audio Corpus (Lima et al., 2025), ensuring methodological consistency and comparability between datasets in terms of how spontaneous, overlapping, and emotionally expressive speech is represented.

This multi-stage procedure ensured that the dataset preserved the spontaneous character and interactional richness of the original radio dialogues while providing high-quality, time-aligned data for speech synthesis and analysis.

[2] `https://github.com/openai/whisper`

### 3.2 Corpus

As shown in Table 2, *Certas Palavras* differs markedly from clean Brazilian Portuguese corpora.

| feature | Total |
|---|---|
| # speakers | 190 |
| # words | 689,574 |
| Total duration (h) | 70.98 |
| # radio program segments | 42,486 |
| Mean duration/segments (s) | 5.95 |
| Mean words/episode | 4257 |
| Mean turns/episode | 55 |
| # sentences with unintelligible words (i) | 972 |
| # sentences with filled pauses (ii) | 8968 |
| # sentences with emotional content (viii) | 594 |
| # sentences with segmental errors (x) | 609 |
| # sentences with music/SFX (xi) | 1807 |
| Speakers with state_of_birth: SP | 61 (32,8 %) |
| Speakers with state_of_birth: RJ | 32 (17,2 %) |
| Speakers with state_of_birth: MG | 13 (7,00 %) |
| Speakers with others state_of_birth | 23 (12,1 %) |
| Speakers with no birth info | 61 (32,8 %) |
| Foreign-borns | 13 (0,68 %) |

Table 2: Detailed statistics of *Certas Palavras.*

The number of dialogue turns and sentence-level segments is considerably higher than in comparable datasets such as MuPe Life Stories or CORAA ASR 1.1, reflecting the inherently interactive nature of broadcast exchanges. The number of sentences with artifact annotations shows how extensively the dataset captures real acoustic disturbances that comes from speaking genre, style and media.

In terms of speaker composition, the dataset comprises a total of 190 distinct speakers, reflecting the conversational structure of the program. This set includes two primary hosts, Claudiney Ferreira and Jorge Vasconcellos, as well as one recurring co-host, Ivan Lessa.In addition, the corpus contains contributions from 187 guest speakers, of whom 133 are male and 54 are female. Some demographic metadata examples can be found in Appendix A.

The corpus includes speaker birthplace metadata for a substantial subset of the speakers, these metadata therefore provide partial but explicit geographic coverage, which can support controlled analyses of regional variation and related speaker-background effects. However, we do not claim that the present distribution constitutes a balanced accent inventory; rather, these fields should be understood as auxiliary metadata that may enable future accent-oriented studies.

This distribution mirrors the guest dynamics of the original broadcast and contributes to substantial inter-speaker variability in voice characteristics, speaking styles, and interactional roles. The presence of stable hosts alongside a large and diverse pool of guests reinforces the dialogic nature of the dataset while providing a clear distinction between the types of hosts frequently represented.

## 4 Methods and Materials

We release the selection scripts, speaker lists, and fixed random seeds for sampling, along with CSV manifests defining utterance boundaries and labels for each split, following dataset card recommendations. A public Docker image with identical environment specifications were released alongside the dataset and training scripts on a GitHub Page[3].

### 4.1 YourTTS Model

The YourTTS model (Casanova et al., 2022b) adopts a Transformer-based encoder-decoder architecture in which the encoder processes the input text sequence and generates an intermediate representation that is then decoded into a mel-spectrogram, later converted into audio by a vocoder.

The model stands out for its intermediate components, such as the posterior encoder implemented as a variational autoencoder (VAE), which enhances end-to-end learning by capturing variations and uncertainties more effectively. This module receives a linear spectrogram and speaker embeddings to predict a latent variable, used by both the vocoder and the flow-based decoder, which conditions this latent variable and the speaker embeddings on a probabilistic distribution to improve realism and variability in the synthesized speech. The monotonic alignment search (MAS) aligns the outputs of the text encoder and the flow-based decoder, while the stochastic duration predictor introduces controlled rhythm variability based on MAS outputs and speaker and language embeddings. YourTTS also introduces innovations such as multilingual support through the concatenation of four trainable language embeddings at each input character and a speaker embedding module capable of adapting to new voices from short audio samples while maintaining vocal identity across languages.

The model was initially trained on a very small Brazilian Portuguese dataset, demonstrating remarkable robustness and generalization capacity despite the limited data, and later extended to larger multilingual datasets such as the CML dataset

[3] https://github.com/GustavoEvangelistaAraujo/CertasPalavras/

(Oliveira et al., 2023), which significantly expanded its coverage of speakers and languages.

These characteristics make YourTTS an appealing baseline, particularly for scenarios like the Certas Palavras dataset, where speaker diversity and limited samples per speaker are key challenges.

### 4.2 F5-TTS Model

The F5-TTS model (Chen et al., 2025b) is a text-to-speech model designed to generate speech by gradually transforming random noise into a structured audio representation.

F5-TTS implements a non-autoregressive TTS system via Conditional Flow Matching with Optimal Transport, regressing vector fields to transform Gaussian noise into log-mel spectrograms through ODE integration, enabling parallel generation across all frames without sequential autoregression. To better connect textual information with acoustic features, character sequences are tokenized, padded with filler tokens to match the speech length, and refined via ConvNeXt V2 blocks before channel-wise concatenation with masked speech and noisy input. The joint representation feeds a Diffusion Transformer, employing RoPE self-attention, convolutional positional embeds, and flow matching loss for implicit text-speech alignment during text-guided infilling training; Classifier-Free Guidance enhances conditioning through unconditional dropout.

During inference, optimized sampling strategies prioritize the most important transformation steps, enabling faster synthesis while maintaining naturalness and prosodic detail. It applies Euler ODE solver with Sway Sampling, followed by Vocos vocoding, while preserving zero-shot prosody from 100K-hour multilingual pretraining.

In general, F5-TTS stands out for its context-dependent prosody modeling, with a prominent DiT backbone to deal with noisy data, enabling natural rhythm, intonation, and duration assignment without rigid phoneme predictors, a complementary DiT model for the experiments.

### 4.3 Training Environments

All experiments were conducted using the same computational environment to ensure reproducibility across models and dataset splits. Training was carried out on a dedicated server. The system comprises $8\times$ NVIDIA A100 80GB Tensor Core GPUs (700GB total), six NVIDIA NVSwitches, and petaFLOPS-class AI throughput ($\sim$10 petaOPS INT8).

Each training run employed mixed-precision optimization (`fp16`) and the AdamW optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and a constant learning rate of $2 \times 10^{-4}$. Training lasted approximately 400 k steps for full datasets, using a batch size of 32. All experiments were performed with fixed random seeds to facilitate reproducibility, and logs were managed with Weights & Biases[4] for experiment tracking. Hyperparameters (optimizer, LR, batch size, steps) are shared across regimes to isolate the effect of data composition.

### 4.4 Training Split

To ensure a controlled and interpretable evaluation, we restricted our analyses to a single speaker: the program's main host, Claudiney Ferreira, who is also the most represented voice in the dataset, with a 9-hour subset. This choice minimizes variability caused by speaker imbalance, as TTS models typically perform better on voices with more training data. Using the host, therefore prevents performance differences from being confounded with data scarcity. The host appears consistently throughout the recordings, across diverse acoustic environments and interactional contexts, providing the model with a stable and statistically rich representation of timbre and prosody. This makes it easier to isolate the effects of *Certas Palavras*' acoustic challenges without additional noise introduced by underrepresented speakers. To support transparency and reproducibility, the exact list of training segments and corresponding metadata are publicly available in the project's GitHub[3].

### 4.5 Subjective metrics

The Mean Opinion Score (MOS) is the average of the rating-scale scores assigned by listeners, a type of Absolute Category Rating (ACR) (Ribeiro et al., 2011). MOS has emerged as the most widely used descriptor of perceived media quality. It has been broadly adopted in the field of speech quality assessment and has since been extended to other modalities such as audio, images, video, and audio-visual content, being applied in contexts ranging from controlled laboratory tests to in-service monitoring (ITU - T, 1996).

For MOS computation, human evaluators listen to the synthesized and natural audio samples and

[4] https://wandb.ai/site/

assign a score from 1 to 5, where the final value corresponds to the average of all evaluators' ratings, as shown in Table 3. The translated version for Portuguese experiment and further information were also added in the GitHub repository[3].

| MOS | Quality | Naturalness |
|---|---|---|
| 5 | Excellent | Completely natural |
| 4 | Good | Mostly natural |
| 3 | Fair | Equaly natural and unnatural |
| 2 | Poor | Mostly unnatural |
| 1 | Bad | Completely unnatural |

Table 3: Description of the used MOS reference for the experiments for naturalness evaluation.

### 4.6 Objective metrics

The Word Error Rate (WER) is a standard metric used to quantify the intelligibility of synthesized or transcribed speech by comparing the predicted word sequence to a reference transcription. WER measures the rate of substitution ($S$), deletion ($D$), and insertion ($I$) errors relative to the total number of words in the reference ($N$). It is defined in Equation 1.

$$\text{WER} = \frac{S_{\text{word}} + D_{\text{word}} + I_{\text{word}}}{N_{\text{words}}} \tag{1}$$

The Character Error Rate (CER) provides a finer-grained measure of intelligibility by evaluating errors at the character level. CER follows the same formulation as WER, but replaces words with characters, enabling more sensitivity to minor transcription differences such as orthographic errors or phonetic approximations. This definition is given in Equation 2.

$$\text{CER} = \frac{S_{\text{char}} + D_{\text{char}} + I_{\text{char}}}{N_{\text{char}}}, \tag{2}$$

As a complementary metric for the subjective evaluations, we compute Kendall's coefficient of concordance ($W$) to quantify the degree of agreement among annotators. Kendall's $W$ measures the consistency of rankings assigned by multiple raters to the same set of stimuli, with values ranging from 0 (no agreement beyond chance) to 1 (complete agreement). In this work, $W$ serves as an indicator of inter-annotator reliability, enabling the assessment of whether perceived differences between stimuli are consistently judged across participants. This formulation is defined in Equation 3.

$$W = \frac{12 \sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2}{m^2 \left(n^3 - n\right)} \tag{3}$$

### 4.7 Evaluation process

The division of audio lists for evaluation was carried out considering two types of audio: synthesized and natural. These audio samples were organized into two lists: List A and List B, which were assigned to participants in a balanced manner. Thirty natural audio samples were placed in List A, while their corresponding synthesized versions were allocated to List B. Likewise, thirty synthesized audio samples were included in List A, with their corresponding natural versions assigned to List B.

Regarding participant selection, prior work recommends a proportion of at least ten expert listeners for every twenty non-experts in subjective speech evaluations (Loizou, 2011). In this study, we intentionally exceeded this minimum by recruiting a total of 27 expert annotators, in order to strengthen the reliability of expert-level judgments and to mitigate potential effects of participant dropout. These experts were distributed across two evaluation lists, with 15 participants in list A and 12 in list B. The resulting imbalance between the lists reflects natural attrition during the experimental process, as some participants withdrew before completing all evaluation stages, rather than deviations from the intended sampling design. Both experiments followed the same allocation strategy, ensuring consistent exposure conditions and preserving diversity in evaluative perspectives for comparative analysis between expert and non-expert assessments of the audio samples.

To reduce potential sources of bias in the subjective evaluation, several measures were implemented based on previous related work (Araújo et al., 2024): (i) participants never listened to the paired natural and synthetic versions of the same utterance, preventing direct comparison effects, (ii) all audio samples were presented in a randomized order, minimizing sequence effects, learning effects, and listener fatigue, (iii) each participant evaluated only a subset of 60 audio samples, ensuring a manageable cognitive load throughout the experiment, (iv) all selected samples were short utterances exhibiting clear intonational unity, allowing listeners to fully perceive the prosodic characteristics of each stimulus while keeping the evalua-

| Model | MOS ↑ | Kendall's W interval |
|---|---|---|
| Ground truth audio | 4.56 ($\sigma$ 0.77, $\pm$ 0.03) | 0.079 |
| YourTTS-CML→CP-ft$_{Claudiney}$ | 3.28 ($\sigma$ 1.77, $\pm$ 0.04) | 0.278 |

Table 4: Subjective evaluation of *Certas Palavras* models. MOS measure perceptual quality.

| Model | Original | Synthetic | t-statistic (two-sample) | p-value |
|---|---|---|---|---|
| *WER* ↓ | | | | |
| YourTTS-CML→CP-ft$_{Claudiney}$ | 13.70 | 23.86 | -6.1082 | 5.2423e-09 (<0.01) |
| F5TTS→CP-ft$_{Claudiney}$ | 13.70 | 18.12 | -3.0251 | 0.0028 (<0.01) |
| *CER* ↓ | | | | |
| YourTTS-CML→CP-ft$_{Claudiney}$ | 7.45 | 11.79 | -3.4708 | 0.0006 (<0.01) |
| F5TTS→CP-ft$_{Claudiney}$ | 7.45 | 10.49 | -2.5096 | 0.0129 (<0.05) |

Table 5: Objective evaluation of TTS models. WER/CER intelligibility measure via ASR back-evaluation.

tion consistent across conditions, (v) audio pairs in which the natural version contained external noise artifacts, such as music or sound effects, were excluded to avoid providing unintended perceptual cues to the evaluators, (vi) natural speech phenomena such as hesitations, expressive sounds, and disfluencies were preserved.

## 5 Results

For clarity, we define the model train splits criteria as follows: **YourTTS-CML→CP-ft**$_{Claudiney}$ refers to the YourTTS model pre-trained in the CML dataset and fine-tuned on the single speaker split of the Certas Palavras corpus, while **F5TTS→CP-ft**$_{Claudiney}$ denotes the F5TTS model fine-tuned on the same single-speaker split.

Due to the higher operational cost of large-scale human listening tests, the subjective MOS evaluation was conducted only for the YourTTS sythesis, which served as the primary reference point for perceived naturalness on *Certas Palavras*. In contrast, objective intelligibility metrics (WER and CER) were extended to include an additional synthesis baseline (F5), enabling broader, cost-effective comparisons between models under the same transcription and scoring protocol.

As a subjective evaluation method, a naturalness test was performed with MOS. As shown in Table 4, the ground-truth natural audio received a high MOS of 4.56 $\sigma$ 0.77, reflecting the expected perceptual quality of the original broadcast material after digitization. In contrast, the synthesized speech produced by the YTT-CML→CP-ft$_{Claudiney}$ model obtained a MOS of 3.28 $\sigma$ 1.77. Although lower than the natural reference, this score indicates that the fine-tuned model was nevertheless capable of generating speech that listeners perceived as moderately natural despite the challenging acoustic characteristics present in the *Certas Palavras* dataset.

The standard deviation for the ground-truth audio (0.77) is relatively high, which is consistent with the heterogeneous nature of the archival material. Because the original recordings include tape noise, compression artifacts, and variable background conditions, listeners naturally diverge more in their quality judgments compared to clean studio datasets. The larger standard deviation observed for the synthetic samples (1.17 vs. 0.77 for ground truth) suggests greater variability in listener judgments, likely due to inconsistencies in timbre, prosody stability, and how the model handles background artifacts present in the *Certas Palavras* training data.

As objective evaluation method, an intelligibility test was performed using a recent ASR model (Leal et al., 2025), which was trained on spontaneous speech. The WER and CER metrics, calculated from the synthesized speech, are reported in Table 5.

The objective results contextualize the subjective findings. As shown in Table 5, both models preserve intelligibility despite being trained on a relatively small subset of *Certas Palavras* characterized by spontaneous dialogue and broadcast noise. For the **YourTTS-CML→CP-ft**$_{Claudiney}$ model, WER increases from 13.70 when evaluated on original text to 23.86 on synthesized speech, indicating a degradation that is expected when moving from natural recordings to generated audio under adverse acoustic conditions. A similar trend is observed for **F5TTS→CP-ft**$_{Claudiney}$, although with a smaller relative increase, reaching 18.12 WER on synthesized speech.

Comparing the two systems, F5-TTS consistently achieves lower WER and CER values on

synthesized audio, suggesting stronger preservation of linguistic content and improved robustness to noise and spontaneous prosodic variation. This behavior aligns with its non-autoregressive flow-matching architecture, which is designed to handle long-range dependencies and acoustic variability more effectively.

## 6 Discussion

As *Certas Palavras* is considered a challenging dataset for Brazilian Portuguese TTS, some limitations should be considered when interpreting the results and reusing the corpus.

All audio was standardized to a 16 kHz sampling rate. While this choice is compatible with common ASR pipelines and with the original archival quality of the recordings, it imposes an upper bound on the frequency content available to TTS models and may limit the reproduction of very fine-grained timbral and prosodic cues compared to higher sampling rates (e.g., 22.05 kHz or 24 kHz). Consequently, the dataset particularly well suited for studies focusing on robustness, but less ideal for purely high-fidelity synthesis. However, because all audio was manually reviewed and all sound artifacts (e.g., hiss, compression, jingles, music bleed, laughter) were consistently labeled, the dataset can indeed support fine-grained robustness evaluation.

Our experimental evaluation is intentionally limited in scope. We report results for a small set of architectures (YourTTS and F5-TTS) and primarily focus on the main host target voice in fine-tuning experiments. Objective intelligibility is measured with a single ASR system trained in spontaneous Brazilian Portuguese speech.

Future work will include the release of a new version of the dataset with a higher sampling rate (48kHz), enabling high-bandwidth model training and improved spectral modeling. In addition, expanded stress-test studies will be conducted to analyze the impact of specific annotated artifacts on speech synthesis, along with broader cross-model comparisons to better understand how adverse acoustic conditions affect both intelligibility and perceptual quality in neural TTS systems and will also replicate the current training protocol using full-precision optimization (`fp32`).

## 7 Conclusion

*Certas Palavras* introduces a challenging benchmark for text-to-speech systems, combining 70 hours of spontaneous, multi-speaker Brazilian Portuguese dialogue recorded under real broadcast conditions. Its detailed annotations of overlap, speech and non-speech interferences, disfluencies, and speaker imbalance enable controlled robustness experiments that are rarely addressed in existing corpora. The subjective evaluation reflects the acoustic complexity of the dataset: although natural recordings received high MOS scores, listener variability was substantial, and synthesized speech exhibited a clear reduction in perceived naturalness.

The objective evaluation provides a complementary perspective. WER and CER values indicate that the synthesized speech remains intelligible, even under adverse acoustic and interactional conditions. This result is particularly noteworthy given the limited size and heterogeneous, noisy nature of the training data, and suggests that exposure to *Certas Palavras* promotes increased robustness in TTS systems. By releasing the dataset along with fixed evaluation protocols, we aim to ensure that both objective (WER/CER) and subjective (MOS) metrics can be applied in a fair and reproducible manner across future studies.

Together, these findings confirm that *Certas Palavras* offers a valuable and appealing testbed for developing and comparing TTS systems that more closely reflect real-world speech variability, broadcast-style dynamics, and interactional complexity, while providing a shared reference point for future methodological and modeling advances.

## 8 Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support from the São Paulo Research Foundation (FAPESP, grant #201907665-4) and IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovations, with resources from Law No. 8.248 of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residency in ICT 13, DOU 01245.0102222022-44. Acknowledgments are also due to the Academic Excellence Program (PROEX) of the Coordination for the Improvement of Higher Education Personnel, Brazil (CAPES), grant No. 88887.8412582023-00.

## References


Gustavo Evangelista Araújo, Julio Cesar Galdino, Rodrigo de Freitas Lima, Leonardo Ishida, Gustavo Wadas Lopes, Miguel Jose Alves De Oliveira Junior, Arnaldo Candido Junior, Sandra Maria Aluísio, and Moacir Antonelli Ponti. 2024. EyetrackingMOS: proposta de um método de avaliação online para modelos de síntese de fala. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 87–96, Porto Alegre, RS, Brasil. SBC.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493, Dublin, Ireland.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023*, pages 1983–1987, Dublin, Ireland.

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2022. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Lang Resources & Evaluation*, 57(3):1139–1171.

Carmem Silvia P. Teixeira. 1997. *Acervo Certas Palavras Programa de Livros e Ideias*. Editora Unicamp, Campinas, São Paulo.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Edresson Casanova, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, João Paulo Teixeira, Moacir Antonelli Ponti, and Sandra Aluísio. 2022a. TTS-Portuguese Corpus: a corpus for speech synthesis in brazilian portuguese. *Lang. Resour. Eval.*, 56(3):1043–1055.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022b. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025a. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025b. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

ITU - T. 1996. Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union.

Ryan Langman, Xuesong Yang, Paarth Neekhara, Shehzeen Hussain, Edresson Casanova, Evelina Bakhturina, and Jason Li. 2025. HiFiTTS-2: A Large-Scale High Bandwidth Speech Dataset. In *Interspeech 2025*, pages 4778–4782.

Sidney Evaldo Leal, Arnaldo Candido Junior, Ricardo Marcacini, Edresson Casanova, Odilon Gonçalves, Anderson Silva Soares, Rodrigo Freitas Lima, Lucas Rafael Stefanel Gris, and Sandra Aluísio. 2025. MuPe life stories dataset: Spontaneous speech in Brazilian Portuguese with a case study evaluation on ASR bias against speakers groups and topic modeling. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6076–6087, Abu Dhabi, UAE. Association for Computational Linguistics.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Rodrigo Lima, Sidney E. Leal, Arnaldo Candido Junior, and Sandra M. Aluísio. 2025. A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Intelligent Systems: 34th Brazilian Conference, BRACIS 2024, Belém Do Pará, Brazil, November 17–21, 2024, Proceedings, Part I*, pages 33–47, Berlin, Heidelberg. Springer-Verlag.

Philipos C. Loizou. 2011. Speech quality assessment. In Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, and Haohong Wang, editors, *Multimedia Analysis, Processing and Communications*, pages 623–654. Springer Berlin Heidelberg, Berlin, Heidelberg.

Igor Macedo Quintanilha, Sergio Lima Netto, and Luiz Wagner Pereira Biscainho. 2020. An open-source end-to-end asr system for brazilian portuguese using dnns built from newly assembled corpora. *Journal of Communication and Information Systems*, 35(1):230–242.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. CML-TTS: A multilingual dataset for speech synthesis in low-resource languages. In *Text, Speech, and Dialogue*, pages 188–199, Cham. Springer Nature Switzerland.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Interspeech 2023*, pages 3222–3226, Dublin, Ireland.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761, Shanghai, China. ISCA.

Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2416–2419. IEEE.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual TEDx corpus for speech recognition and translation. In *Interspeech 2021*, pages 3655–3659, Brno, Czechia.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. 2025. SAM audio: Segment anything in audio. *Preprint*, arXiv:2512.18099.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *Preprint*, arXiv:2106.15561.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791, Suzhou, China. Association for Computational Linguistics.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530, Graz, Austria.

# A Appendix

| data_id | text | audio_path | flags | speaker_name | sex | city_birth | state_birth | country |
|---|---|---|---|---|---|---|---|---|
| CP552_001 | Começa agora mais uma edição do programa Certas Palavras, o programa de livros e ideias da Central Brasileira de Notícias. | 0001-CP552_00.000-00.000.wav | | jorge_vasconcellos | M | Rio de Janeiro | RJ | Brasil |
| CP552_010 | É uma história, (risos) uma história rocambolesca. | 0010-CP552_00.000-00.000.wav | [risos] | claudinei_ferreira | M | Mariana | MG | Brasil |
| CP650_003 | Mata... O Alba, veja bem, vo é legal é interessante você falar isso, que é exatamente o outro dado que eu ia colocar, que não está tanto no seu livro, mas que merece reflexão, que é um detalhe da vida no no Brasil, que a mim me assusta muito, exatamente porque eu acho que um dia eu vou cometer uma um uma loucura, ou serei acometido de uma (risos) uma loucura de um próximo. | 0003-CP650_00.000-00.000.wav | [risos] | claudiney_ferreira | M | Rio de Janeiro | RJ | Brasil |
| CP664_003 | Talvez acho que a a aquilo que mais chama atenção quando você se debruça sobre essa conjuntura, sobre os anos cinquenta, em particular sobre os anos o ano cinquenta e quatro (tosse), é a paixão, a a a maneira apaixonada como se vivia, vivenciava a a a política. | 0003-CP664_00.000-00.000.wav | [tosse] | fernando_weltman | M | Mooca | SP | Brasil |

Table 6: Example of Certas Palavras dataset metadata structure.