# Combining Real and Synthetic Speech for ASR Adaptation in Brazilian Portuguese

**Daniel R. da Silva**[1,2], **Maria Eduarda S. Borba**[1,2], **Állan C. P. Silva**[2], **Maria Carolina S. Barreto**[3], **Arthur F. de Morais**[3], **Paulo V. dos Santos**[2], **Guilherme C. Dutra**[1,2], **Sávio S. T. de Oliveira**[1,2], **Anderson da S. Soares**[1,2]

[1]Institute of Informatics, Federal University of Goiás (UFG), Goiás, Brazil
[2]Center of Excellence in Artificial Intelligence (CEIA/UFG), Goiás, Brazil
[3]Aeronautics Institute of Technology (ITA), São José dos Campos, Brazil

daniel.ribeiro@discente.ufg.br

## Abstract

Automatic Speech Recognition (ASR) systems require large amounts of annotated speech, which are difficult to obtain in specialized domains. This paper introduces GARAGEM: General Automotive Real and Artificial speech corpus for Garage Environments and Maintenance in brazilian portuguese, a domain specific ASR dataset for Brazilian Portuguese focused on automotive repair, combining real speech collected from online sources with synthetic speech generated from curated technical terminology. A reproducible methodology is proposed, encompassing real data acquisition, domain guided synthetic data generation, dataset consolidation, and ASR model fine-tuning. Experiments conducted with the Whisper, Wav2vec 2.0, and Conformer models show that synthetic data provides improvements when used to complement real recordings. Quantitative and qualitative analyses show reductions in Word Error Rate (WER) and Character Error Rate (CER) and improved recognition of domain specific terms absent from the real training set. The results indicate that domain guided synthetic speech is an effective data augmentation strategy for ASR adaptation in specialized and low resource scenarios.

## 1 Introduction

Automatic Speech Recognition (ASR) is a central task in natural language processing, enabling the conversion of acoustic signals into textual representations (Kheddar et al., 2024). The development of robust ASR systems depends on large volumes of annotated speech that cover speaker variability, acoustic conditions, and linguistic diversity (Yadav and Sitaram, 2022; Stefanel Gris et al., 2022). This requirement poses a major challenge in specialized domains, where labeled data is scarce, costly to produce, and often insufficient to represent domain specific terminology and acoustic characteristics (Bartelds et al., 2023). As a result, ASR performance degrades in scenarios involving technical vocabulary, non standard recording environments, and domain specific speech patterns.

Several strategies aim to reduce the reliance on large annotated corpora, including transfer learning (Tran et al., 2025), data augmentation (Rossenbach et al., 2019), and synthetic speech generation (Karl et al., 2024). Among these, synthetic data produced by Text-to-Speech (TTS) systems offers a scalable alternative for expanding training material without manual annotation costs (Do et al., 2024). Prior studies report performance gains when synthetic and real data are combined (Zheng et al., 2021). However, a portion of this literature relies on simulated conditions, with limited investigation of synthetic data grounded in real domain specific content and acoustic contexts.

This work addresses this gap by focusing on a realistic application scenario in which synthetic data generation is explicitly guided by domain specific linguistic and acoustic properties. Rather than relying on generic corpora, the study integrates real domain data to inform the creation of synthetic speech, enabling a more faithful approximation of real world usage conditions and a more reliable assessment of its contribution to ASR adaptation.

The central hypothesis is that domain specific synthetic speech, generated from curated terminology and contextualized sentences, can effectively support the adaptation of ASR models to specialized domains. By aligning lexical content, speaking style, and acoustic variability with real domain data, synthetic speech is expected to reduce domain mismatch and improve recognition performance when used alongside limited real data.

Brazilian Portuguese is selected as the target language due to the limited availability of domain specific ASR resources and the substantial linguistic variation observed across regions and professional contexts (Candido Junior et al., 2022). The automotive repair domain is chosen because it combines spontaneous speech, specialized technical terminol-



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 838–846
April 13-16, 2026 

ogy, and challenging acoustic conditions, including background noise and non professional recordings. These characteristics define a demanding scenario for evaluating domain adaptation strategies in automatic speech recognition.

The objective of this study is to construct an ASR dataset for Brazilian Portuguese focused on the automotive repair domain, combining real speech collected from online sources with synthetically generated audio. The work analyzes whether the proposed synthetic data supports the fine-tuning of ASR models for this domain. Experiments are conducted using the Whisper (Radford et al., 2023), Wav2vec 2.0 (Baevski et al., 2020), and Conformer (Gulati et al., 2020) models, which encompass different architectural paradigms in modern speech recognition.

The originality of this work lies in the joint release of a real and synthetic domain specific ASR dataset for Brazilian Portuguese named GARAGEM: General Automotive Real and Artificial speech corpus for Garage Environments and Maintenance in brazilian portuguese, and in the evaluation of domain guided synthetic speech for ASR adaptation. The study contributes evidence on how synthetic data grounded in real domain characteristics can complement real speech and improve ASR robustness in specialized and low resource scenarios.

## 2 Related Work

The use of synthetic speech generated by TTS systems for training and fine-tuning ASR models attracts growing interest due to its ability to mitigate the scarcity of annotated data in low resource languages (Bartelds et al., 2023) and specialized domains (Rossenbach et al., 2019; Karl et al., 2024; Tran et al., 2025). Prior work explores combinations of synthetic and real speech (Do et al., 2024), strategies for mixing heterogeneous speakers and sources (Werchniak et al., 2021; Xue et al., 2022), and the influence of acoustic, prosodic, and linguistic characteristics of generated audio on recognition performance (Rossenbach et al., 2023). These studies show that synthetic data can yield significant gains when real data is limited, motivating systematic analysis of the conditions under which synthetic speech is most effective.

Domain adaptation using synthetic data proves beneficial in scenarios involving accent variation (Do et al., 2024), languages with complex phonetic inventories (Xue et al., 2022), and specialized vocabularies (Karl et al., 2024). Do et al. (Do et al., 2024) show that synthetic accented speech generated with VITS improves recognition of accented English when used to fine-tune Wav2vec 2.0, without degrading performance on standard benchmarks.

Xue et al. (2022) report that high fidelity synthetic Mandarin speech generated with AdaSpeech enhances results on AISHELL by modeling speaker specific traits. Karl et al. (2024) demonstrate that fine-tuning WhisperX with synthetic radiology speech leads to substantial gains even in the absence of real domain specific audio. These results indicate that synthetic data supports domain transfer while reducing annotation requirements, although real speech still defines the upper performance bound.

Evidence from Rossenbach et al. (2019) further highlights the value of synthetic data under restricted supervision. Their experiments with attention based ASR models trained on synthetic LibriSpeech show that combining generated audio with data augmentation and external language models improves recognition, especially when labeled data is scarce. The gains are more pronounced in small data regimes, suggesting that synthetic speech increases linguistic and acoustic diversity. Similar conclusions arise in studies on minority languages, where synthetic data compensates for limited real speech resources.

Bartelds et al. (2023) investigate low resource languages such as Groningsian, West Frisian, Besemah, and Nasal using self learning frameworks in which a teacher model generates transcriptions to train a student model. Their results show that XLSR based ASR models achieve up to 25.5% relative WER reduction through self training with synthetic data. This work demonstrates that synthetic speech can approximate the benefits of much larger real datasets and serve as an effective complement in severely constrained scenarios.

Werchniak et al. (2021) study keyword spotting and show that speaker diversity in synthetic data impacts generalization more strongly than perceptual audio quality. Performance peaks when synthetic speech represents approximately seventy five percent of the training mixture, indicating that diversity rather than realism alone drives robustness. These findings suggest that optimal real to synthetic ratios depend on task requirements and the representational richness of synthetic voices.

Rossenbach et al. (2023) analyze the role of acoustic and prosodic properties of synthetic speech, focusing on phoneme duration modeling. Their results show that low variance and underestimated durations produced by non autoregressive TTS systems degrade ASR performance, while aligning duration distributions with real speech improves recognition. Tran et al. (2025) complement these findings by proposing domain adaptation strategies for Whisper using synthetic speech generated from large language model outputs. They report improvements of up to seventeen percent without real training audio and show that adapting only the decoder with lightweight matrices increases efficiency, indicating that targeted adaptation strategies are preferable to full model updates.

Huang et al. (2024) examine cases where synthetic data degrades ASR performance due to distributional mismatch with real speech. Their study on Paraformer based systems proposes techniques such as prompt design and transcription manipulation to mitigate negative effects. The results show that ASR models learn systematic differences between synthetic and real audio, and that careful input design stabilizes training. These observations clarify how acoustic mismatch influences learning dynamics and provide guidance for controlled use of synthetic speech.

The use of TTS and voice conversion for ASR data augmentation in extremely low resource settings is investigated by Casanova et al. (2024). Their work introduces zero shot and multilingual TTS models capable of generating multi speaker synthetic data from a single real speaker. Experiments show that ASR models trained with synthetic speech achieve competitive performance, demonstrating that TTS based augmentation can effectively replace large multi speaker corpora in constrained scenarios.

This line of research is referenced by Casanova et al. (2023), who combine cross lingual multi speaker TTS and cross lingual voice conversion for ASR data augmentation. Their experiments in Portuguese and Russian show large WER reductions in extreme settings with only one real speaker, confirming that synthetic diversity derived from other languages significantly improves generalization. This work establishes synthetic speech as a viable strategy for ASR development in severely under resourced languages.

## 3 Methodology

Figure 1 summarizes the proposed methodology, which is structured into four stages, each corresponding to a subsection of this section. The workflow is designed to support reproducibility, both for repeating the same experiments and for adapting the methodology to other application domains beyond automotive repair. Each stage addresses a distinct component of dataset construction and model adaptation, forming a modular pipeline that can be replicated with minimal domain specific adjustments.

The methodology combines real data acquisition from online sources with synthetic data generation guided by domain knowledge. Real speech is collected from publicly available platforms, such as YouTube, enabling the capture of domain relevant audio produced in authentic working environments with natural acoustic conditions. Synthetic data generation is based on a domain specific dictionary and the use of language and speech generation models to produce contextualized sentences and corresponding audio. This design supports transfer to other domains by redefining the term inventory, identifying domain relevant real audio from online sources, and reusing the same data generation and processing stages.

### 3.1 Real Data Acquisition and Preparation

This section refers to Stage 1 of the proposed methodology shown in Figure 1, which focuses on the construction of a real speech dataset. Real data collection starts with the identification of YouTube channels (YouTube video search) focused on automotive mechanics, as indicated by domain specialists. Video selection prioritizes representative repair shop content recorded in real work environments, where acoustic conditions such as reverberation, tool noise, and non professional recording setups are present. Manual inspection ensures inclusion of material with high density of automotive technical vocabulary and detailed explanations of mechanical procedures, increasing the relevance of the dataset.

Audio preprocessing relies on the Silero VAD [1] (Voice Activity Detection) model for automatic speech segmentation. The VAD identifies time intervals containing voice activity, removing extended silence and facilitating manual annotation. This process reduces the volume of audio to be

[1] https://github.com/snakers4/silero-vad

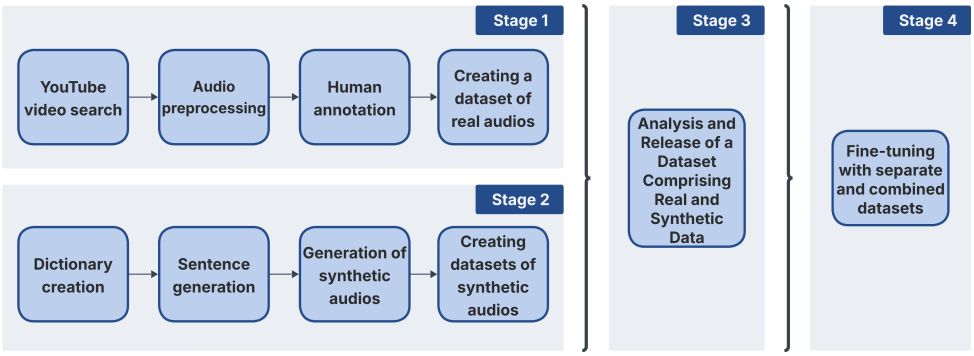


Figure 1: Overview of the proposed methodology composed of four stages: (i) real data acquisition and preparation, (ii) synthetic data generation, (iii) dataset consolidation and release, and (iv) ASR model fine-tuning and evaluation. Each stage corresponds to a subsection of the Methodology and is designed to support reproducibility across different application domains.

transcribed while preserving natural prosodic characteristics of speech, including pauses and hesitations common in workshop communication. The resulting segments have variable duration and follow natural speech units detected by the VAD.

The division of data into training and test sets follows a speaker separation strategy widely adopted in ASR research to ensure reliable evaluation. Speakers assigned to the training set do not appear in the test set, preventing the model from exploiting idiosyncratic voice traits such as timbre, speech rate, or prosodic patterns. This setting encourages the learning of generalizable phonetic and linguistic representations (Yu and Deng, 2014).

Human annotation is carried out by 11 trained annotators following unified transcription guidelines, with careful treatment of technical terminology, automotive acronyms, and numerical expressions. Automatic pre transcriptions generated by the Whisper model serve as initial drafts, reducing annotation effort per segment. Annotators may adjust segment boundaries to avoid truncated words and to preserve semantic coherence. Transcription follows a verbatim approach, retaining hesitations, repetitions, and discourse markers typical of spontaneous speech. The guidelines align with principles adopted in initiatives such as CORAA (Candido Junior et al., 2022), and NURC-SP (Rodrigues et al., 2024) ensuring annotation consistency and quality.

### 3.2 Synthetic Data Generation

This section refers to Stage 2 of the proposed methodology shown in Figure 1, which concerns the construction of a synthetic speech dataset. Synthetic data creation starts with the compilation of a domain specific dictionary for automotive repair shops (Dictionary creation), organized into four categories: car models, automotive parts, product brands, and generic domain terms. Term selection covers official manufacturer nomenclature, commonly used mechanic denominations, and regional variants. This structured dictionary supports the generation of contextualized sentences that reflect authentic workshop communication, including diagnostics, maintenance procedures, and customer interactions.

Contextualized sentence generation relies on large language models, namely Gemini 2.5 (Comanici et al., 2025) and GPT 4 (Achiam et al., 2023), configured to emulate spontaneous and informal speech produced by experienced Brazilian Portuguese speaking mechanics. The generation strategy encourages natural and diverse utterances for each dictionary term, incorporating features of orality such as vocalized pauses, slang, contractions, and spoken syntactic structures. Sentences alternate between explicit mentions and simplified references, capturing variability characteristic of professional oral communication. For each dictionary term, the models produce between five and ten distinct sentences, while generic terms may also appear incidentally in sentences associated with other entries.

An example of the prompt used for sentence generation is presented below, preserving the Portuguese domain term.

*You are an experienced Brazilian Portuguese speaking mechanic explaining things in a spontaneous and informal way. Create five natural and varied sentences as if they were everyday workshop speech, in which the term appears naturally in con-*

*text. The sentences should sound spoken, including pauses such as hm, ah, eh, slang, and contractions like tá and pra.*

*Term: Veículos a Combustão*

*Definition: Veículos a combustão are vehicles called this way because they use internal combustion engines. They receive this name because they rely on fuel burning reactions as a source of chemical energy, which is converted into mechanical energy to move the vehicle.*

Audio synthesis (Generation of synthetic audios) employs two TTS systems, Gemini 2.5 TTS (Comanici et al., 2025) and Google Cloud TTS based on Chirp 3 [2]. Both systems generate speech using the same set of speakers, ensuring consistency in voice identity across synthetic data. The main difference lies in synthesis quality, as Gemini 2.5 aims to produce more natural and less robotic intonation compared to Google Cloud TTS. This design enables analysis of whether ASR performance benefits from synthetic speech with more realistic prosodic patterns. All synthetic audio is generated at a 16 kHz sampling rate, preserving acoustic detail relevant for speech recognition.

### 3.3 Dataset Consolidation and Release

This section refers to Stage 3 of the proposed methodology shown in Figure 1, which addresses the analysis and public release of the complete real and synthetic datasets. The real YouTube corpus contains 4,412 audio segments totaling 9 hours and 45 minutes of transcribed speech. The data split assigns 3,121 segments (6 hours and 51 minutes) to training and 1,291 segments (2 hours and 54 minutes) to testing, corresponding to an approximate 70% and 30% division. This partition provides sufficient material for model training while preserving a representative evaluation set.

Synthetic data generated with Gemini TTS amount to 12 hours and 49 minutes of audio, while Google Cloud TTS yields 11 hours and 15 minutes. Each TTS system produces 6,367 audio files with 18 distinct speakers, resulting in two complete synthetic subsets. Figure 2 illustrates the temporal distribution of the real and synthetic subsets, providing a clear visualization of the relative amount of audio across each portion of the dataset. Figure 3 shows the distribution of segment durations, indicating that most audio segments fall between 4 and 12 seconds, a range that supports complete sentence coverage and representative speech samples.

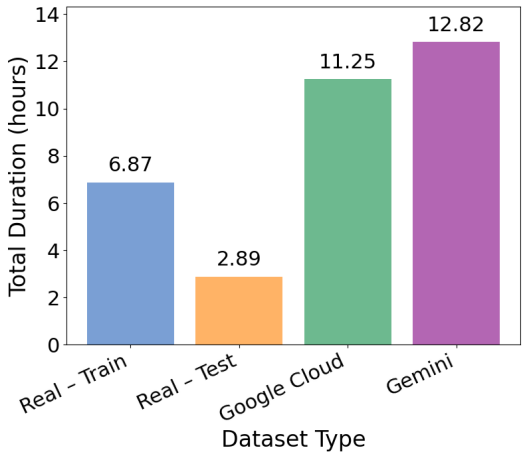


Figure 2: Temporal distribution of the real and synthetic subsets of the GARAGEM.

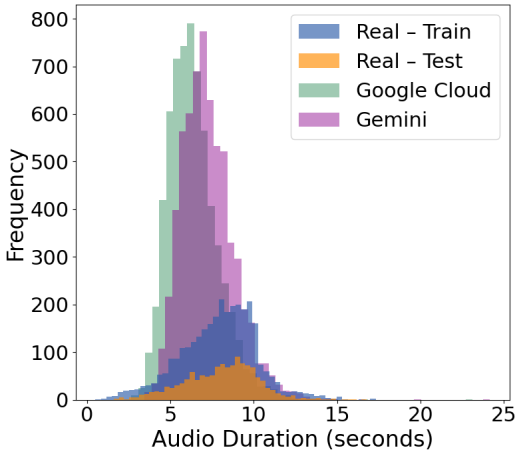


Figure 3: Distribution of segment durations across real and synthetic subsets of the GARAGEM.

The technical terms dictionary includes 721 entries distributed across four categories: 136 car model names, 142 generic automotive domain terms, 113 product or part brands, and 330 specific automotive part names. This composition ensures broad coverage of domain specific vocabulary, encompassing both frequent terminology and less common specialized expressions. Orthographic variants and alternative denominations for identical components are included to reflect linguistic variation in professional mechanic communication.

Dataset release follows protocols that ensure copyright compliance and reproducibility. The corpus GARAGEM: General Automotive Real and Ar-

[2] https://docs.cloud.google.com/text-to-speech/docs/chirp3-hd

tificial speech corpus for Garage Environments and Maintenance in brazilian portuguese, is publicly available on Zenodo under a Creative Commons Attribution 4.0 International license with DOI: https://doi.org/10.5281/zenodo.18202081. The YouTube based subset is distributed through metadata containing the original video links, segment timestamps, and corresponding transcriptions, enabling legal data reconstruction. The two synthetic subsets are released as 16 kHz audio files with associated transcriptions, together with the technical terms dictionary. All materials may also be requested directly from the authors.

### 3.4 Model Fine-Tuning Experiments

This section refers to Stage 4 of the proposed methodology shown in Figure 1, which focuses on fine-tuning the Whisper, Wav2vec 2.0, and Conformer models using the datasets described in the previous sections.

The Whisper medium model checkpoint [3] contains 769 million parameters and is pre-trained on $680,000$ hours of multilingual speech. The architecture follows a Transformer-based encoder-decoder design, where the encoder processes mel-spectrogram representations and the decoder generates text transcriptions autoregressively (Radford et al., 2023). Fine-tuning runs for 8 epochs with a learning rate of $2e-5$, batch size of 8, and the AdamW optimizer. Training minimizes cross-entropy loss over generated token sequences, with gradient accumulation applied to improve stability.

The Wav2vec 2.0 model checkpoint [4] contains 300 million parameters, and was trained on approximately 490 hours of Brazilian Portuguese audio. The architecture employs a convolutional feature encoder followed by a Transformer-based contextualized representation learner, using contrastive learning during pre-training to learn representations from raw audio (Baevski et al., 2020). Fine-tuning runs for 10 epochs with a learning rate of $3e-4$, batch size of 32, and CTC loss as the objective function.

The Conformer model checkpoint [5] contains approximately 120 million parameters, and was trained on 3200 hours of Brazilian Portuguese audio. The Conformer model combines convolution and self-attention mechanisms to capture both local and global dependencies in speech (Gulati et al., 2020). Fine-tuning uses CTC loss with a learning rate of 0.1, batch size of 16, and runs for 30 epochs.

Training runs on an NVIDIA V100 GPU with 32 GB of VRAM. Audio inputs are standardized to a 16 kHz sampling rate to ensure compatibility with the pre-trained models. The best checkpoints are selected based on validation performance monitored using the WER metric. Performance evaluation relies on Word Error Rate (WER) and Character Error Rate (CER) metrics computed on a held-out test set that remains isolated throughout training. Each data configuration, including different synthetic proportions and combinations with real data, is evaluated independently.

## 4 Results and Discussion

Table 1 reports WER and CER obtained from fine-tuning Whisper, Wav2vec 2.0, and Conformer with the datasets introduced in the methodology section. The use of three models with different architectures ensures robustness in the experimental evaluation. Fine-tuning with synthetic data only leads to poorer performance on the real test set when compared to both the non fine-tuned models and the models adapted using only real YouTube data. This result indicates that synthetic speech, when used in isolation, does not provide sufficient acoustic and contextual variability to support generalization to real workshop recordings.

When synthetic data is combined with real YouTube data, both WER and CER are reduced across all models. The improvement suggests that synthetic speech is effective as a complement to real data, increasing exposure to domain specific lexical items while preserving acoustic patterns learned from authentic recordings. This behavior supports the use of synthetic data as an augmentation strategy rather than as a standalone training source.

The two synthetic datasets yield similar gains when combined with real data. Although the YouTube plus Google Cloud configuration achieves the lowest error rates, the difference relative to Gemini TTS remains small. This proximity indicates that performance improvements are primarily driven by domain targeted lexical coverage and sentence diversity, rather than by differences in synthesis quality between the two TTS systems.

Table 2 provides qualitative evidence for these

[3] https://huggingface.co/openai/whisper-medium

[4] https://huggingface.co/lgris/wav2vec2-large-xlsr-open-brazilian-portuguese-v2

[5] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/speechtotext_pt_br_conformer

| Fine-tuning Dataset | Whisper | | Wav2Vec 2.0 | | Conformer | |
|---|---|---|---|---|---|---|
| | **WER** | **CER** | **WER** | **CER** | **WER** | **CER** |
| No Fine-tuning | 16.10 | 8.07 | 37.64 | 14.00 | 16.19 | 6.96 |
| YouTube | 13.42 | 5.86 | 22.72 | 8.24 | 14.86 | 5.82 |
| Synthetic (Google Cloud) | 25.53 | 11.83 | 36.84 | 13.78 | 27.19 | 12.99 |
| Synthetic (Gemini) | 27.08 | 12.82 | 38.15 | 14.79 | 23.87 | 11.08 |
| YouTube + Synthetic (Google Cloud) | <u>12.40</u> | <u>5.38</u> | **21.38** | **7.67** | **12.72** | **5.17** |
| YouTube + Synthetic (Gemini) | **12.36** | **5.32** | <u>21.51</u> | <u>7.94</u> | <u>13.05</u> | <u>5.81</u> |

Table 1: Whisper, Wav2vec 2.0, and Conformer fine-tuning results evaluated on the real test set using WER and CER metrics. The best results for each model are highlighted in bold, and the second best results are underlined.

observations. The models fine-tuned with YouTube and synthetic data produce more accurate transcriptions of domain specific terms such as *monobloco*, *para-barra*, *rede can*, and *corolla*. Several of these terms are misrecognized or distorted when training relies only on real data. In addition, some corrected terms are absent from the real training set, indicating that synthetic data contributes new lexical items that improve recognition in the target domain.

Overall, the results show that domain guided synthetic data strengthens ASR adaptation when integrated with real speech, improving recognition of specialized vocabulary without replacing authentic recordings.

## 5 Conclusion and Future Work

This work presents the GARAGEM, a domain specific ASR corpus for Brazilian Portuguese that combines real speech collected from online sources with synthetic speech generated from curated automotive terminology. The proposed methodology demonstrates that domain guided synthetic data, when used as a complement to real recordings, improves ASR performance by reducing recognition errors and increasing coverage of specialized vocabulary. The experimental results show gains in WER and CER when synthetic data augments real data, while synthetic speech alone proves insufficient for reliable generalization to real test conditions.

The qualitative analysis further supports these findings by showing improvements in the recognition of technical terms that are absent or underrepresented in the real training set. This behavior indicates that synthetic data contributes primarily at the lexical level, reinforcing domain specific knowledge without replacing the acoustic diversity provided by authentic recordings. The similar performance observed between different TTS systems suggests that lexical relevance and sentence diversity play a more significant role than synthesis quality differences in this setting.

Limitations of this study include the non exploration of data augmentation strategies applied to synthetic speech and the use of fixed proportions between real and synthetic data. Future work includes investigating the impact of additional augmentation techniques such as noise injection and reverberation, as well as systematic analysis of optimal synthetic to real data ratios. Further experiments may also extend the methodology to other specialized domains to validate its generality and reproducibility.

## Acknowledgments

We acknowledge the support provided by the Center of Excellence in Artificial Intelligence (CEIA/UFG). This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1. We also thank Oficina Conectada for their collaboration in the data curation process, helping to ensure the audio samples were representative of the domain, as well as for their valuable insights into the challenges and context of the problem.

| Audio ID | Reference Transcription | Transcription (YouTube Fine-tuned) | Transcription (YouTube + Synthetic Fine-tuned) |
|---|---|---|---|
| 368 | *o veículo também apresenta vibração do conjunto monobloco com carga* | *veículo também apresenta vibração do conjunto* ***o monobloc*** *com carga* | *o veículo também apresenta vibração do conjunto* ***monobloco*** *com carga* |
| 889 | *colocar como a gente tirou o parabarra aqui tem que deixar tudo originalzinho tá ó todos parafusos tudo colocado certinho tá* | *colocar como a gente tirou* ***para a barra*** *aqui tem que deixar tudo originalzinho tá ó todos os parafusos tudo colocado certinho tá* | *colocar como a gente tirou o* ***parabarra*** *aqui tem que deixar tudo originalzinho tá ó todos parafusos tudo colocado certinho tá* |
| 774 | *medimos a rede can ambas chegando ali alimentação dois e meio a um ponto oito volts então a mesma tá ok* | *medimos a* ***rede cam*** *ambas chegando ali a alimentação dois e meio a um ponto oito volts então a mesma tá ok* | *medimos a* ***rede can*** *ambas chegando ali alimentação dois e meio a um ponto oito volts então a mesma tá ok* |
| 573 | *que vedou legal encaixa bem direitinho tudo certinho com certeza substitui muito bem a original beleza* | *que vedou legal* ***o caixa*** *bem direitinho tudo certinho com certeza* ***substititu*** *muito bem a original beleza* | *que vedou legal* ***encaixa*** *bem direitinho tudo certinho com certeza* ***substituir*** *muito bem a original beleza* |
| 21 | *quando trocar substituir a correia auxiliar de acessórios do toyota corolla conhecida também como correia de acionamento* | *quando trocar substituir a correia a auxiliar de acessórios do toyota* ***corola*** *conhecido também como correia de acionamento* | *quando trocar substituir a correia auxiliar de acessórios do toyota* ***corolla*** *conhecido também como correia de acionamento* |

Table 2: Qualitative transcription examples comparing fine-tuning with YouTube data only and with YouTube plus synthetic data, using normalized sentences without punctuation as employed for metric computation. Terms whose recognition improves with the inclusion of synthetic data are highlighted. The Audio ID corresponds to the segment index in the released metadata.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

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. *Language Resources and Evaluation*, 57(3):1139–1171.

Edresson Casanova, Sandra Aluísio, and Moacir Antonelli Ponti. 2024. TTS applied to the generation of datasets for automatic speech recognition. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 633–638, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. In *INTERSPEECH 2023*, pages 1244–1248. ISCA.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Cong-Thanh Do, Shuhei Imai, Rama Doddipatla, and Thomas Hain. 2024. Improving accented speech recognition using data augmentation based on unsupervised text-to-speech synthesis. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 136–140. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang,

Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, pages 5036–5040. ISCA.

Jian Huang, Yancheng Bai, Yang Cai, and Wei Bian. 2024. A study on the adverse impact of synthetic speech on speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10266–10270.

Anderson Karl, Guilherme Fernandes, Leonardo Pires, Yvens Serpa, and Carlos Caminha. 2024. Synthetic ai data pipeline for domain-specific speech-to-text solutions. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 37–47, Porto Alegre, RS, Brasil. SBC.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Inf. Fusion*, 109(C).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ana Carolina Rodrigues, Alessandra A. Macedo, Arnaldo Candido Jr, Flaviane R. F. Svartman, Giovana M. Craveiro, Marli Quadros Leite, Sandra M. Aluísio, Vinícius G. Santos, and Vinícius M. Garcia. 2024. Portal NURC-SP: Design, development, and speech processing corpora resources to support the public dissemination of Portuguese spoken language. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 187–195, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Nick Rossenbach, Benedikt Hilmes, and Ralf Schlüter. 2023. On the relevance of phoneme duration variability of synthesized training data for automatic speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Nick Rossenbach, Albert Zeyer, Ralf Schluter, and Hermann Ney. 2019. Generating synthetic audio data for attention-based speech recognition systems. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073.

Lucas Rafael Stefanel Gris, Edresson Casanova, Frederico Santos de Oliveira, Anderson da Silva Soares, and Arnaldo Candido Junior. 2022. Brazilian portuguese speech recognition using wav2vec 2.0. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 333–343, Berlin, Heidelberg. Springer-Verlag.

Minh Tran, Yutong Pang, Debjyoti Paul, Laxmi Pandey, Kevin Jiang, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, and Xin Lei. 2025. A domain adaptation framework for speech recognition systems with only synthetic data. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Andrew Werchniak, Roberto Barra Chicote, Yuriy Mishchenko, Jasha Droppo, Jeff Condal, Peng Liu, and Anish Shah. 2021. Exploring the application of synthetic audio in training keyword spotters. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7993–7996.

Shaofei Xue, Jian Tang, and Yazhu Liu. 2022. Improving speech recognition with augmented synthesized data and conditional model training. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 443–447.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

D. Yu and L. Deng. 2014. *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. Springer London.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678.