Adapting ASR Models to Technical Scenarios: A Case Study in the Brazilian Automotive Repair Domain

Daniel Ribeiro da Silva[1,2], **Maria Eduarda Silva Borba**[1,2], **Gustavo dos Reis Oliveira**[2], **Pedro Reis Pimenta**[1,2], **Állan Christoffer Pereira Silva**[2], **Guilherme Correia Dutra**[2], **Sávio Salvarino Teles de Oliveira**[1,2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74001-970 – Goiânia, GO – Brazil

[2]Center of Excellence in Artificial Intelligence (CEIA/UFG)
Goiânia, GO – Brazil

{daniel.ribeiro, maria.borba, pedroreis2}@discente.ufg.br

{gustavodreis.dev, guilhermecorreiadutra}@gmail.com

allansilva@egresso.ufg.br, savioteles@ufg.br

Abstract. *This work proposes a pipeline for adapting automatic speech recognition models to the domain of automotive repair shops in Brazilian Portuguese. The process includes real data collection, manual transcription, and dataset curation, with adjustments to the Conformer, and Wav2Vec* 2.0 *models, while Whisper model served as a comparative baseline. The approach aims to improve speech recognition accuracy in noisy environments with specific technical vocabulary. The Conformer model achieves the best results, with a word error rate of* 12.97 *percent and a character error rate of* 5.46 *percent, surpassing the Whisper Large-v3 in both transcription accuracy and inference speed.*

1. Introduction

Automatic Speech Recognition (ASR) is a field within artificial intelligence aiming to transcribe human speech into text accurately and efficiently. This process involves converting audio signals into sequences of words using computational models trained on large datasets. Advancements in deep learning, particularly end-to-end models, significantly improve the accuracy and robustness of ASR systems [de Azevedo et al. 2022]. Models such as Transformers and Conformers excel in handling speech variability and environmental noise, expanding ASR applications across various contexts, from virtual assistants to automatic captioning systems [Ahlawat et al. 2025].

Despite progress in ASR for languages such as English, developing effective systems for Brazilian Portuguese still faces challenges. Limited access to publicly available annotated data restricts the development of comprehensive models capable of generalizing well. The diversity of accents, intonation patterns, and regional expressions further complicates model training. Studies show that even with reduced data volumes, competitive word error rates (WER) emerge using techniques such as Wav2vec 2.0 [Baevski et al. 2020], which rely on self-supervised learning and robust feature extraction [Gris et al. 2022].

The environment of mechanical workshops presents specific challenges for ASR systems. The constant presence of noise from machines, tools, and simultaneous conversations creates a complex acoustic scenario. In addition, the use of technical jargon and regional linguistic variations requires ASR systems to be adapted to understand this specialized vocabulary. These characteristics make automatic speech transcription in this domain particularly challenging. The lack of domain-specific data and the need to understand technical terms and regional linguistic variations highlight the importance of developing model adaptation approaches for this specific context.

While studies exist on adapting ASR models to different domains [Maulik et al. 2025, Gonçalves et al. 2024], there is a significant gap in the literature regarding the application of these techniques to the context of mechanical workshops in Brazilian Portuguese. The central hypothesis of this study is that, when working with audio data from challenging contexts, careful preparation and annotation of these data are fundamental to adjusting smaller ASR models that can perform well in this specific scenario. It is believed that a well-structured preprocessing and annotation workflow can compensate for data limitations and enhance the effectiveness of adapted models.

This work aims to study a workflow for adapting ASR models to specific contexts, using the domain of mechanical workshops as a case study. It seeks to develop and evaluate a process that includes data collection, preparation, and annotation, as well as the adjustment of existing models to improve speech recognition in this environment. The originality of this study lies in the application of ASR adaptation techniques to an underexplored domain, considering the specificities of Brazilian Portuguese. The relevance of the topic is highlighted by the growing need for automatic transcription systems in industrial environments, where accurate documentation of processes and communications is essential for operational efficiency and safety.

2. Related Works

The creation of high-performance Automatic Speech Recognition (ASR) systems is intrinsically linked to the availability of large and representative datasets. In this context, the advent of CORAA ASR (Corpus of Annotated Audios) [Candido Junior et al. 2022] represents a significant advancement. This corpus, comprising 290 hours of Brazilian Portuguese (BP) speech with manually validated audio-transcription pairs, is of utmost importance for training ASR models geared towards Portuguese, as it follows rigorous annotation standards to ensure robustness across the entire dataset.

The performance of ASR systems is highly dependent on the quality and quantity of training data. Research highlights that recognizing speech in Portuguese, particularly Brazilian Portuguese (BP), faces challenges due to a comparative scarcity of data when contrasted with more resourced languages such as English. Studies investigating the impact of data on ASR systems for BP emphasize the importance of understanding data quality to optimize language model utilization, seeking a balance between performance and computational cost [Alvarenga et al. 2023]. Similarly, for European Portuguese, deep learning-based ASR systems have shown that while models created from scratch often underperform due to data scarcity, transfer learning approaches significantly improve performance, underscoring the crucial role of data quantity and quality [Medeiros et al. 2023].

These findings collectively underscore that comprehensive datasets are not merely an asset but a fundamental requirement for achieving state-of-the-art ASR capabilities across different Portuguese variants. In this scenario of data and model optimization, ASR architectures have evolved considerably.

The Whisper model, developed by OpenAI, revolutionized the field of ASR by being trained on a vast amount of supervised audio data, covering various languages and tasks. Its robust architecture and the scale of its training allowed Whisper to demonstrate impressive capabilities in multilingual speech recognition and in noisy environments, in addition to being proficient in tasks such as language detection and speech-to-text translation [Radford et al. 2023]. Whisper's main contribution lies in its generalization capability and superior performance in low-resource scenarios, making it a valuable tool for languages that historically had less data available for training ASR models. Its unified approach to recognition and translation establishes it as a milestone in ASR research.

The Conformer architecture emerged as a significant advancement in ASR models, combining the advantages of convolutional neural networks (CNNs) and transformers. This combination allows Conformer to capture both local and global dependencies in audio sequences, resulting in superior performance in speech recognition tasks. CNNs are effective in extracting local features and modeling the fine structure of speech, while transformers excel at understanding long-range dependencies and contexts [Gulati et al. 2020]. Conformer's ability to efficiently process long audio durations and its robustness to speech variations have made it a popular choice for state-of-the-art ASR systems.

Furthermore, Wav2vec 2.0 is a key advancement in ASR, particularly for speech representation learning. Using a self-supervised approach, it learns audio features from large amounts of unlabeled speech through a masked speech unit prediction task, capturing phonetic and contextual information without extensive manual transcriptions. This capability is critical for low-resource languages, as the model can then be fine-tuned with limited labeled data to achieve competitive performance [Baevski et al. 2020]. Its main contribution lies in data efficiency and reducing reliance on large transcribed corpora, facilitating ASR development for diverse languages, including Portuguese.

The applicability of Wav2vec 2.0 for Brazilian Portuguese has been explored in recent works, demonstrating its effectiveness in both low-resource scenarios and with more extensive datasets. Research has validated the ability of the Wav2vec 2.0 XLSR-53 model to build functional ASR systems for BP even with limited amounts of labeled data, such as 1 hour of transcribed speech [Gris et al. 2021]. Subsequently, the use of larger and more diverse datasets, totaling approximately 470 hours of BP speech, allowed for significantly lower word error rates (WER), establishing a new state-of-the-art for open end-to-end (E2E) ASR models in BP [Gris et al. 2022]. These studies exemplify how the Wav2vec 2.0 architecture, combined with the curation and expansion of language-specific data, is fundamental for boosting the performance of automatic speech recognition in Portuguese.

In the field of Automatic Speech Recognition (ASR) for specific domains, recent studies demonstrate significant advancements in improving accuracy, which are applicable to diverse specialized scenarios. For instance, [Maulik et al. 2025] explore enhancements in ASR performance within the medical domain through finetuning and post-

processing using Large Language Models (LLMs), successfully reducing errors in specialized terminology. Similarly, [Gonçalves et al. 2024] investigate the effectiveness of pre-trained ASR models for medical history transcription in Brazilian Portuguese, highlighting the importance of integrating language models to enhance transcription accuracy and semantic fidelity. These works are crucial as they show that domain adaptation and postprocessing strategies, while applied to medicine, offer a robust methodological basis. Such approaches are highly relevant for developing ASR systems in other technical and specific contexts, such as transcribing conversations in auto repair shops, where technical jargon and specific terminologies are equally prevalent and transcription accuracy is fundamental for documentation and operational efficiency.

In summary, the methods use metrics such as WER (Word Error Rate) and CER (Character Error Rate) to assess the quality of speech and text recognition systems. WER measures the proportion of words that must be inserted, deleted, or replaced for the generated transcription to match the reference, reflecting the impact of errors on message comprehension [Chen et al. 1998]. CER performs a similar evaluation at the character level, mainly applied in tasks such as OCR, by calculating the proportion of inserted, deleted, or modified characters relative to the total in the reference transcription [Neudecker et al. 2021]. Both metrics enable objective and standardized performance comparison, supporting error analysis and the selection of more effective solutions.

3. Methodology

Based on the studies presented in the previous section, the processes for developing a pipeline to adapt automatic speech recognition (ASR) models to the automotive repair shop domain are defined. This process encompasses real data collection, preprocessing for audio standardization, manual transcription, and the subsequent creation of a curated dataset. The initial evaluation employs the Whisper model as a baseline, followed by fine-tuning of the Conformer and Wav2Vec 2.0 models with specific adjustments for the domain. The adopted methodology aims to optimize ASR model performance, ensuring their effectiveness and accuracy within a specialized context where terminology and acoustic speech characteristics significantly differ from general domains. A visual representation of this pipeline is detailed in Figure 1.

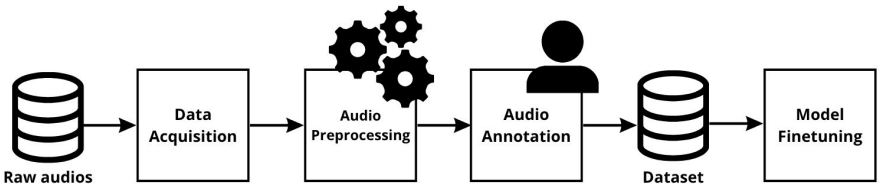

Figure 1. Pipeline to adapt automatic speech recognition (ASR) models.

3.1. Data Acquisition

The audio data used in this study originates from a company that specializes in providing technical support services to professionals in the automotive repair industry. This company facilitates communication between mechanics, auto shop owners, and technical experts with extensive knowledge in vehicle systems. The dataset includes voice messages

exchanged between these professionals during discussions aimed at solving mechanical problems and carrying out diagnostic procedures. These interactions reflect real-life scenarios in which issues are described, hypotheses are proposed, and guidance is provided based on specialized expertise.

In addition to conversational exchanges, the dataset incorporates audio segments extracted from educational video lectures that address various subjects related to automotive repair. These lectures focus on topics such as engine diagnostics, electrical systems, maintenance procedures, and workshop safety. Their inclusion brings structured speech content aligned with the automotive domain and introduces variation in speaking styles, levels of formality, pronunciation patterns, and the use of technical expressions. This combination of instructional and conversational audio contributes to a more comprehensive and challenging dataset for speech processing tasks.

The speech data exhibits a wide range of linguistic and acoustic characteristics, combining spontaneous and semi-structured communication. Common features include hesitations, filled pauses, false starts, and disfluencies, which naturally occur in human dialogue. The recordings also feature terminology and jargon specific to automotive workshops, reflecting the technical nature of the discussions. Audio quality varies considerably, with many samples containing background noise, overlapping speech, reverberation, and other distortions typically encountered in workshop environments. These conditions increase the complexity of the dataset and provide a realistic testbed for evaluating speech recognition systems.

This type of audio content introduces several challenges for automatic speech recognition models. The combination of domain-specific language, informal yet technical dialogue, and acoustic interference impacts the performance of generic ASR systems. Improvements in recognition accuracy require domain adaptation strategies and robust preprocessing pipelines capable of addressing these issues. Table 1 illustrates the characteristics of the dataset by presenting sample audio excerpts along with corresponding transcriptions. These examples highlight the diversity of the speech data and the importance of tailoring ASR models to specialized environments.

Table 1. Examples of audio samples and their corresponding transcriptions.

Audio ID	Source	Transcription
00532	Voice message (support)	*O relé lá, na hora que a gente tá dando partida, aquele relé principal lá dentro da injeção, ele fica tec, tec, tec, tec, tec, batendo.*
06574	Voice message (support)	*Eh... Válvula termostática nova, e o radiador que ele tinha colocado era novo, né?*
01248	Video lecture	*Então eu consigo utilizar também aqui um transdutor de vácuo na admissão e entender aí qual que é a válvula que está com problema, se é admissão ou escapamento.*

3.2. Audio Preprocessing

After data acquisition, the preprocessing procedure applies well-established practices in speech processing to standardize the dataset and prepare it for automatic speech recog-

nition tasks. The adopted methodology aligns with strategies commonly used in the development of Portuguese speech corpora, including resources such as the CORAA ASR dataset [Candido Junior et al. 2022]. Standardization begins by resampling all audio files to 16 kHz, which ensures compatibility with widely used pretrained models and provides a consistent audio quality level across the dataset.

Speech segmentation is performed using a voice activity detection algorithm designed to isolate regions containing speech. The Silero VAD model [1] is used for this purpose due to its efficiency and accuracy in detecting speech boundaries, especially in acoustically challenging environments such as industrial or workshop settings. This segmentation produces manageable audio units, facilitating annotation and further processing.

To support the manual transcription process, preliminary transcriptions are generated using the Whisper model. These transcriptions act as references to improve the efficiency of the annotation workflow and provide guidance in interpreting utterances that involve domain-specific terminology or spontaneous language. This approach significantly reduces the cognitive load on annotators and helps ensure more accurate and consistent labels throughout the dataset. The dataset also undergoes speaker separation to ensure that training and testing subsets contain different speakers. This separation enhances the validity of evaluation results by guaranteeing that model performance reflects its ability to generalize to previously unseen voices, which is essential for realistic deployment scenarios.

3.3. Audio Annotation

The audio annotation process uses the Audino tool [Grover et al. 2020], a web based open source platform designed for the transcription and annotation of audio data. The dataset is divided among 11 annotators, who work independently on separate portions of the corpus. Each annotator accesses the original audio files along with voice activity detection (VAD) timestamps generated by the Silero VAD model and preliminary transcriptions generated by the Whisper model. The platform interface enables the adjustment of VAD timestamps whenever necessary to prevent the segmentation of speech in the middle of words or sentences. This setup supports efficient and precise annotation, allowing annotators to maintain the integrity of spoken utterances while working asynchronously.

All annotators participate in a training phase before starting the annotation task to ensure consistency with the annotation guidelines. The primary goal is to transcribe speech exactly as spoken, maintaining the natural flow and spontaneous characteristics of real conversations. This includes the faithful representation of hesitations, filled pauses, and other speech disfluencies, which are common in everyday dialogue. Annotators receive support from domain specialists for clarification, particularly when handling unclear or ambiguous terms, which reinforces the accuracy and quality of the annotations throughout the process.

Filled pauses such as “eh”, “ah”, and “hm” are standardized in all transcriptions to preserve consistency. Informal expressions and contractions frequently used in spontaneous Brazilian Portuguese, such as “tá” instead of “está” and “pra” instead

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

of “para”, are transcribed exactly as spoken. This approach ensures the preservation of authentic speech and reflects the linguistic traits specific to the domain covered by the dataset. The transcription process follows the conventions defined by the CORAA project [Candido Junior et al. 2022], which prioritize the representation of informal speech. Numerical expressions are fully written out in words rather than symbols to promote uniformity across the dataset and simplify future post processing and manual review stages.

3.4. Final Dataset

The final dataset contains $11{,}014$ segmented utterances derived from $2{,}165$ distinct original audio recordings. Each recording undergoes segmentation based on speech boundaries and speaker changes, producing coherent and natural speech units. To ensure effective evaluation and avoid conversational context overlap, segmentation is organized by audio channels. All utterances from the same channel, such as those from a support agent or customer, along with their corresponding original audio file, are assigned exclusively to either the training or test set. This approach preserves contextual integrity and significantly reduces the possibility of content leakage between subsets of the dataset.

The resulting data split allocates 6.18 hours of audio to the test set and 14.56 hours to the training set. The duration of each subset is presented in Figure 2(A). This setup enforces strict independence between speakers and domains, aligning with real-world deployment scenarios where models must generalize to entirely new content and speakers. Ensuring that each subset reflects realistic and challenging conditions contributes to more reliable model evaluation.

Beyond maintaining independence between speakers and preventing duration imbalance across sets, segment length distributions are also examined. Both subsets display similar characteristics, with the majority of utterances falling within the 5 to 10 second range. This uniformity contributes to stable model performance, as it avoids bias caused by variation in segment lengths. Figure 2(B) illustrates the segment duration histogram for both training and test sets, confirming the consistency of this characteristic throughout the dataset. Finally, due to confidentiality constraints requested by the data provider, the dataset will not be publicly released.

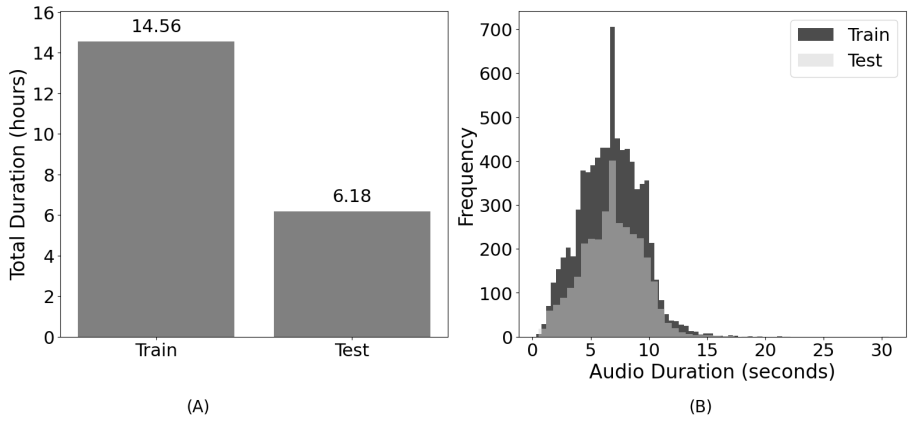

Figure 2. Dataset specifications: (A) Total duration of training and test sets and (B) Distribution of audio segments durations.

3.5. Model Finetuning

To evaluate the performance of more efficient ASR models under constrained computational conditions, two architectures are fine-tuned on a domain-specific dataset composed of manually annotated audio-transcription pairs. Whisper is used as a benchmark model and evaluated in its different size variants to provide reference performance for Portuguese automatic speech recognition (ASR). Two smaller and more efficient architectures are fine-tuned for comparison: Conformer [Gulati et al. 2020] and Wav2Vec 2.0 [Babu et al. 2021].

Pre-trained models in Portuguese were used for fine-tuning in this work. The Conformer model checkpoint [2] was fine-tuned on the dataset using the Connectionist Temporal Classification (CTC) loss function, running for 100 epochs with a batch size of 32 and a learning rate of 0.1. Similarly, the Wav2Vec 2.0 checkpoint [3] was fine-tuned on the same data for 20 epochs with a batch size of 16 and a learning rate of $3e-4$, based on configurations successfully applied in related work [Babu et al. 2021] to improve adaptation to spontaneous speech and variable acoustic conditions. Both fine-tuning processes were conducted on an NVIDIA V100 GPU with 32 GB of VRAM, selecting the best checkpoints according to validation loss. Finally, due to restrictions, these fine-tuned checkpoints will not be publicly released.

4. Results

All test experiments are conducted on the same hardware configuration to ensure consistency across models: an NVIDIA RTX 4090 GPU with 24 GB of VRAM. Table 2 presents the evaluation results of various speech recognition models on the test set. The comparison includes five variants of the Whisper architecture (Tiny, Base, Small, Medium, and Large-v3), along with two alternative models, Wav2Vec 2.0 and Conformer, evaluated both in their original pre-trained versions and after fine-tuning on a domain-specific dataset. Performance is reported in terms of Word Error Rate (WER) and Character Error Rate (CER).

Table 2. Performance comparison of different speech recognition models.

Model	WER	CER
Whisper Tiny	62.54	38.41
Whisper Base	42.99	24.47
Whisper Small	25.68	13.88
Whisper Medium	18.56	10.38
Whisper Large-v3	13.26	6.93
Conformer (pre-trained)	16.89	7.35
Conformer (fine-tuned)	**12.97**	**5.46**
Wav2vec 2.0 (pre-trained)	41.70	16.30
Wav2vec 2.0 (fine-tuned)	22.69	8.72

The results indicate that fine-tuning improves the performance of both Wav2Vec 2.0 and Conformer models. The reduction in WER and CER shows that domain adap-

[2] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/riva/models/speechtotext_pt_br_conformer

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

tation enables the capture of more relevant phonetic and linguistic patterns, resulting in more accurate transcriptions. The fine-tuned Conformer achieves the lowest error rates among the evaluated models, surpassing Whisper Large-v3. This suggests that targeted adaptation with appropriate data can be more effective than increasing model size or complexity. The model's performance may be related to inductive biases for sequence modeling, which support generalization over structured audio inputs. Qualitative analysis shows that its most frequent errors are orthographic, often involving incorrect vowel placement, doubled consonants, or substitutions between phonetically similar sounds. These issues may be linked to the CTC-based decoding process, which aligns phonemes to characters without explicit spelling correction, and to the presence of domain-specific or low-frequency terms that increase phonetic confusions, as illustrated in Table 3.

Table 3. Examples of errors in the transcriptions provided by Conformer model.

Reference transcription	Conformer Transcription
O próprio trambulador, né, a alavanca de marcha, você verificou se bate os pinos?	*o o próprio* ***trangulador*** *né a alavanca de marcha você verificou se bate os pinos*
os indicadores de relés e fusiveis, tudo esquematizado e padronizado	*os indicadores de relés e fusíveis tudo* ***esquemaatizado*** *e* ***pardronizado***

In addition to transcription accuracy, inference efficiency plays a critical role in deployment scenarios. Figure 3 illustrates the inference times for each model when processing the test set. Both Wav2Vec 2.0 and Conformer models deliver significantly faster inference compared to the Whisper architecture. In some cases, inference time improves by up to a factor of 34. This considerable speedup may result from the lower architectural complexity and reduced parameter count of the Conformer and Wav2Vec 2.0 models, particularly when compared to larger Whisper variants. These characteristics make them more suitable for environments with limited processing power or real-time processing requirements.

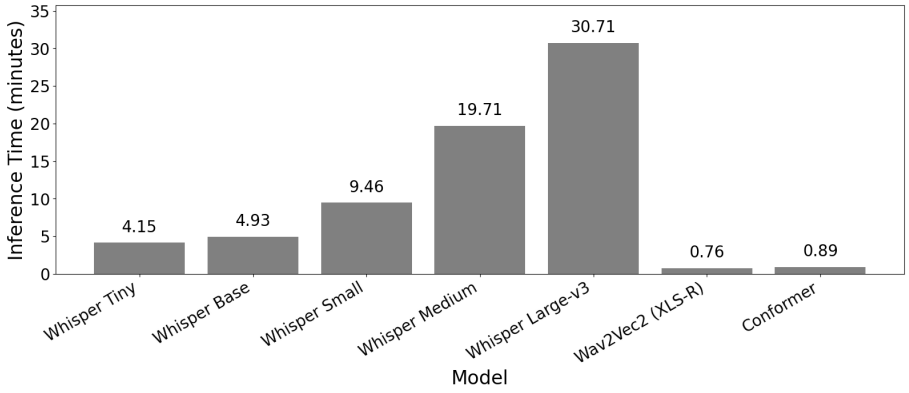

Figure 3. Inference Time by Model.

These results highlight the relevance of domain adaptation through fine tuning, especially in scenarios involving challenging audio environments such as mechanical workshops. The Conformer model demonstrates superior performance in these conditions due to its ability to capture long-range dependencies and model sequential data effectively, making it more resilient to acoustic variations and contextual ambiguities. Fine tuning

enables the model to adjust to the specific phonetic and linguistic patterns present in workshop conversations, which leads to better recognition accuracy. Inference efficiency also remains a crucial factor for real time deployment, and the Conformer combines low latency with high accuracy, making it well suited for practical applications. These findings suggest that carefully adapted and efficient architectures can meet or exceed the performance of large general purpose models. The evidence also indicates that similar improvements can be achieved in other domains where audio data shares characteristics such as background noise, informal speech, and domain specific vocabulary, reinforcing the potential of targeted adaptation strategies.

5. Conclusions and Future Works

This work presents a complete pipeline for developing automatic speech recognition (ASR) systems focused on the domain of automotive repair shops. The process encompasses data collection, preprocessing, audio annotation, and the fine-tuning of compact neural models. Each stage is structured to ensure the resulting system performs efficiently and accurately within the chosen application context. By applying domain-specific adaptations, the models become capable of understanding specialized vocabulary and handling acoustic variability typical of the target environment.

Experiments demonstrate that fine-tuned models, including Conformer and Wav2Vec 2.0, achieve competitive performance compared to larger systems such as Whisper. These compact models operate with lower computational demand and deliver faster inference, while maintaining or surpassing transcription quality in terms of word error rate (WER) and character error rate (CER). Such outcomes demonstrate that smaller ASR models, when tailored to a defined domain, provide a practical solution that balances effectiveness with deployment feasibility in constrained scenarios.

For future work, improving speech recognition in specialized contexts requires strategies that enhance both accuracy and usability. Expanding the training dataset with synthetic audio that reflects the linguistic and acoustic characteristics of the target domain improves robustness, especially in scenarios with limited annotated data [Karl et al. 2024]. Another approach is the integration of external language models during decoding or post-processing, which can correct structural inconsistencies, improve grammatical fluency, and mitigate errors caused by misspellings in model transcriptions [Alvarenga et al. 2023, Maulik et al. 2025]. Combining synthetic data generation with advanced language modeling supports more effective domain-specific adaptation, fostering the development of reliable and context-aware speech recognition systems while addressing current limitations such as reduced robustness when handling terms absent from the training data.

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 the company Oficina Conectada for supplying the audio data and for their support in helping us understand the challenges and context of the problem.

References

[Ahlawat et al. 2025] Ahlawat, H., Aggarwal, N., and Gupta, D. (2025). Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

[Alvarenga et al. 2023] Alvarenga, J. P. R., Merschmann, L. H. d. C., and Luz, E. J. d. S. (2023). A data-centric approach for portuguese speech recognition: Language model and its implications. *IEEE Latin America Transactions*, 21(4):546–556.

[Babu et al. 2021] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Von Platen, P., Saraf, Y., Pino, J., et al. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

[Baevski et al. 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

[Candido Junior et al. 2022] Candido Junior, A., Casanova, E., Soares, A., de Oliveira, F. S., Oliveira, L., Junior, R. C. F., da Silva, D. P. P., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., and Aluísio, S. M. (2022). Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Lang. Resour. Eval.*, 57(3):1139–1171.

[Chen et al. 1998] Chen, S. F., Beeferman, D., and Rosenfeld, R. (1998). Evaluation metrics for language models. Carnegie Mellon University.

[de Azevedo et al. 2022] de Azevedo, D. M., Rodrigues, G. S., and Ladeira, M. (2022). A probabilistically-oriented analysis of the performance of asr systems for brazilian radios and tvs. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part II*, page 169–180, Berlin, Heidelberg. Springer-Verlag.

[Gonçalves et al. 2024] Gonçalves, Y., Alves, J., Sá, B., Silva, L., Macedo, J., and da Silva, T. C. (2024). Speech recognition models in assisting medical history. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 485–497, Porto Alegre, RS, Brasil. SBC.

[Gris et al. 2022] Gris, L. R. S., Casanova, E., de Oliveira, F. S., da Silva Soares, A., and Candido Junior, A. (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.

[Gris et al. 2021] Gris, L. R. S., Casanova, E., Oliveira, F. S. d., Soares, A. d. S., and Candido Junior, A. (2021). Desenvolvimento de um modelo de reconhecimento de voz para o português brasileiro com poucos dados utilizando o wav2vec 2.0. In *Congresso da Sociedade Brasileira de Computação - CSBC*. SBC.

[Grover et al. 2020] Grover, M., Bamdev, P., Singla, Y., Hama, M., and Shah, R. (2020). Audino: A modern annotation tool for audio and speech.

[Gulati et al. 2020] Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

[Karl et al. 2024] Karl, A., Fernandes, G., Pires, L., Serpa, Y., and Caminha, C. (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.

[Maulik et al. 2025] Maulik, U. B., Mitra, P., and Sarkar, S. (2025). Enhancing domain-specific asr performance using finetuning and zero-shot prompting: A study in the medical domain. In *Proceedings of the 2024 Sixth Doctoral Symposium on Intelligence Enabled Research (DoSIER 2024)*, pages 1–6, Jalpaiguri, India.

[Medeiros et al. 2023] Medeiros, E., Corado, L., Rato, L., Quaresma, P., and Salgueiro, P. (2023). Domain adaptation speech-to-text for low-resource european portuguese using deep learning. *Future Internet*, 15(5).

[Neudecker et al. 2021] Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., and Pletschacher, S. (2021). A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

[Radford et al. 2023] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.