

Benchmarking ASR Models on Brazilian Portuguese Accented Speech: a Case Study with Piauí Accent

Felipe Farias, Wilmer Lobato, William Alberto Cruz Castañeda e Marcellus Amadeus
{felipe.farias, wilmer, william, marcellus}@soberania.ai
SoberanIA

Abstract—Voice interaction with artificial intelligence systems has improved to the point that it is increasingly used as the primary interaction between people and those systems. However, this technology doesn't have equal performance across all language variants, such as regional or sociolinguistic dialects. The training of an automatic speech recognition model is heavily data-dependent, which places particular pressure on the selection of speech variants present in the training data. The effect of any variant in speech is currently measured by Word Error Rate (*WER*), the same metric that aggregates errors from an entire dataset. *WER* is a simple and practical metric, at the expense of hiding model behavior that may be useful for analyzing speech variants. In this article, we propose a benchmark for evaluating accented speech, detailing data collection and the use of average *WER*, the 90th percentile of *WER*, and a qualitative analysis of a sample of the transcriptions to locate transcription errors. We demonstrate this approach by creating partitions of speech datasets based on the regional accent of Piauí and evaluating 11 state-of-the-art models on this regional accent. The results show that all models face unique difficulties in transcription due to elements of the regional accent, such as regional expressions and changes in the pronunciation of certain words. These elements are captured in the qualitative analysis. Analyzing the *WER* and comparing it to a dataset of read speech helps in the choice of models and in understanding their limitations, while *WER*_{p90} presents itself as an interesting metric to understand the low end of their performance.

Keywords—automatic speech recognition, regional accent, benchmarking.

I. INTRODUCTION

Automatic Speech Recognition (ASR) systems have reached such a level of maturity that several applications now use voice as interface with humans. This includes voice-first applications, such as home assistants like Alexa, Siri, and Google Assistant [1]. Even applications that traditionally use written text, such as chatbots, now include voice as input [2]. This inclusion expands access to these tools and their potential benefits, accomodating hearing impaired [3] or illiterate users [4].

However, even among speakers of the same language, the performance of ASR models is not uniform. Being a data-intensive application of machine learning, the performance of ASR models reflects the biases of the dataset used to train it [5]. The work of [6] highlights differences in performance according to speaker gender and also with sociolinguistic dialects in English. In this study, the researchers built testing datasets using Chicano, Spanglish and AAVE sociolinguistic accents, evaluated Whisper [7], a state-of-the-art (SOTA) ASR model, and observed performance degradation in accented speech when compared to Standard American English. This

also happens with regional accents, as can be seen in the works from [8], [9], which investigated the performance of ASR models in speech containing regional and sociolinguistic accents. The same pattern can be observed in different languages, such as Spanish [10] and French [11]. This last study hypothesizes that this difference in performance can be due to imbalances in representation of certain accents in the training data, which is a similar point to the one made in studies regarding less represented languages in multilingual datasets [12].

A. Goals

The goal of this work is to quantify the impact of accented speech on the performance of ASR models currently in use in Brazilian Portuguese. For this, we first develop speech datasets that represent spontaneous accented speech in Brazilian Portuguese. These datasets are annotated with the accent of the speaker in order to enable the evaluation of ASR models across several partitions. We also develop a benchmark for ASR models and services in accented speech, utilizing several metrics to investigate the performance of ASR models, in addition to the average Word Error Rate (*WER*).

B. Contributions

The main contributions of this work are: (i) the development of a dataset of spontaneous accented speech collecting 44 hours of speech from audio recordings, local news broadcasts and spontaneous speech datasets. The dataset has two partitions: *Benchmark BR*, that includes speech from 12 Brazilian states, and *Benchmark PI*, that includes only the state of Piauí; (ii) the introduction of a speech transcription benchmark, using three different metrics to evaluate ASR models in raw accuracy (*WER*), accuracy on the worst examples of each dataset (*WER*_{p90}) and the location of types of errors (qualitative analysis); and (iii) the evaluation of 11 SOTA models for Piauí accent, achieving interesting insights on the factors that influence model performance.

II. RELATED WORK

This work is conceptually related to previous research in ASR for accented speech. The work of [8] investigated the effect and performance of different accents of American English, finding differences in recognition among some states. Similarly, [11] present a study for French models were evaluated in the accent spoken in the Canadian province of Quebec, severely underrepresented in the multilingual datasets used to train multilingual models.

The work of [13] uses several metrics to complement *WER* in the evaluation of speech recognition systems, stemming

from studies such as [14], which surface different statistical metrics to calculate WER , such as the median and the 90th percentile, to better understand the distribution of transcription errors, and [11], which uses qualitative analysis of transcriptions to better capture details that are difficult to obtain quantitatively, such as the proportion of hallucinations. The speaker's state of origin is used as a proxy for the regional accent spoken in each utterance, as seen in [8] and more recently in Portuguese in the accent classification work of [15]¹. This approach has lower resolution than that presented in [16], where accents are divided not only by state, but also between capital and countryside, but it shares with it the advantage of not depending on broad classifications that can change over time [17] or the inexactitude of self-reported accents, as in [18].

A. Datasets III. MATERIALS AND METHODS

The first step for this evaluation is the creation of a dataset that can be used to represent accented speech in the Piauí regional accent. For this end, we collect and process data from transcribed audios taken from interviews and reporting programs in Piauí and from selected speech datasets in Brazilian Portuguese. The resultant dataset is divided into two different partitions, *Benchmark BR* and *Benchmark PI*, and its characteristics are presented in Table I.

- 1) **Benchmark PI** dataset represents the regional accent of Piauí, collected from audio recordings and transcripts of local news programs on YouTube. This guarantees examples of spoken language with a specific accent. We separate the audio files in spoken sections, detecting speech activity with Silero VAD [19]. After that, each audio file is transcribed using `whisper-large-v3`. The transcribed set is programatically cleaned to remove silences and correct mistranscriptions. The resulting dataset contains 168k audio/transcription pairs. Thus, we split the resulting dataset into training, test and validation sets using the 80/10/10 ratio. *Benchmark PI* is the validation set.
- 2) **Benchmark BR** dataset includes different accents from Brazilian Portuguese. To improve fairness, it has the same number of audio files of *Benchmark PI*. Table I shows that the duration of the audio files is considerably different. For this reason, we collect data from

¹In Tables, the accents are represented by the abbreviation of each corresponding regional state: Alagoas - AL, Bahia - BA, Ceará - CE, Minas Gerais - MG, Pará - PA, Paraíba - PB, Pernambuco - PE, Piauí - PI, Rio de Janeiro - RJ, Rio Grande do Sul - RS, Rondônia - RO, São Paulo - SP

spontaneous speech datasets, selecting initially CORAA [16], MUPE [20], and some of the data gathered for the Piauí accent in the previous partition. We also process the metadata from the datasets to map the speakers' reported origin to their estimated accent, represented by their state. The data collected contains, on aggregate, speech from 12 different Brazilian states. The number of audios from each State is proportional to the number of natives of each state. The distribution of native speakers across states is gathered from the most recent census performed by the Brazilian Institute of Geography and Statistics [21]. Combining this with the number of audio files, we sampled a number of audio files from each regional accent to form this partition.

B. Experimental Procedures

The evaluation was performed on a machine running the Linux operating system with one Nvidia GPU H100 with 80GB RAM. Each inference was performed in batches of 32 audio samples, except `voxtral` sample, which demanded the batch to be reduced to 16. The transcriptions were normalized by converting each sentence to lowercase and stripping them of punctuation to avoid measurement errors. After the transcription of the whole dataset, WER and WER_{p90} were calculated for each model. After that, the outputs were sampled to perform qualitative analysis. This sample was manually annotated by people with knowledge of the accent, marking accent-related terms and correcting proper nouns.

C. Metrics

WER: The word error rate is a measurement of the difference between the transcription of a speech and the original intended speech. This is specially useful in languages in which word is the basic unit of meaning, instead of characters [13]. WER is calculated as:

$$WER = \frac{S + I + D}{N} \quad (1)$$

where S is the number of substitutions in the transcription, I is the number of insertions, D is the number of deletions and N is the total number of analyzed words. Note that the analyzed words can be one sentence or several. WER of a dataset is calculated aggregating all words, inclusions, deletions and substitutions that occurred in that dataset,

WER 90th Percentile (WER_{p90}): In order to acquire more granularity of the information provided in the calculation of WER , a second statistical measure was taken: the 90th percentile. Together with the average, this metric can inform

Partition	Size (h)	#Audios	Original Data	Accent - #Audios
<i>Benchmark BR</i>	32	16682	MUPE - 9259, CORAA - 7240	AL - 366, BA - 1659, CE - 1031
			Youtube Channel #1 - 316	MG - 2409, PA - 952, PB - 466
			Youtube Channel #2 - 67	PE - 1062, PI - 383, RJ - 1883
<i>Benchmark PI</i>	12	16889	Youtube Channel #1 - 10704	RS - 1276, RO - 185, SP - 5210
			Youtube Channel #2 - 6185	PI - 16889

TABLE I
CHARACTERISTICS OF THE CREATED DATASETS.

Model [reference]	Origin	#Parameters	License
whisper-base [7]	openai	74M	apache-2.0
whisper-medium [7]	openai	769M	apache-2.0
whisper-large-v2 [7]	openai	1.55B	apache-2.0
whisper-large-v3 [7]	openai	1.55B	apache-2.0
qwen-ASR-1.3B [22]	Alibaba	1.3B	apache-2.0
parakeet-TDT-0.6B-v3 [23]	Nvidia	600M	cc-by 4.0
canary-1B-v2 [23]	Nvidia	1B	cc-by 4.0
cohere-transcribe-03-2026 [24]	Cohere	2B	apache-2.0
voxtral-Small-24B-2507 [25]	Mistral	24B	apache-2.0
mms-1b-all	Meta	1B	cc-by-nc 4.0
granite-4.0-1b-speech [26]	IBM	1B	apache-2.0

TABLE II
CHARACTERISTICS OF THE EVALUATED MODELS.

Model	Common Voice	Benchmark PI	Benchmark BR
whisper-base	33,4%	120,4%	99,8%
whisper-medium	10,1%	35,9%	31,5%
whisper-large-v2	8,0%	35,1%	26,7%
whisper-large-v3	6,4%	25,5%	20,6%
qwen-ASR-1.3B	6,5%	20,0%**	20,2%*
parakeet-TDT-0.6B-v3	7,4%	21,6%	19,8%*
canary-1B-v2	12,1%	28,7%	27,1%
cohere-transcribe-03-2026	4,2%**	23,1%	20,5%
voxtral-Small-24B-2507	5,5%	21,8%	21,0%
mms-1b-all	13,0%	156,0%	48,3%
granite-4.0-1b-speech	5,9%	80,3%	69,3%

TABLE III

WER OF ASR MODELS IN COMMON VOICE DATASET AND IN BR AND PI PARTITIONS OF THE PROPOSED ASR BENCHMARK. ** DENOTES $p < 0.005$ WHEN COMPARING WITH ALL MODELS. * DENOTES $p < 0.005$ WHEN COMPARING WITH ALL BUT ONE MODEL.

Model	Common Voice	Benchmark PI	Benchmark BR
whisper-base	100,0%	100,0%	133,3%
whisper-medium	100,0%	83,3%	100,0%
whisper-large-v2	100,0%	85,7%	100,0%
whisper-large-v3	100,0%	71,4%	100,0%
qwen-ASR-1.3B	50,0%	66,7%	100,0%
parakeet-TDT-0.6B-v3	100,0%	75,0%	100,0%
canary-1B-v2	100,0%	83,3%	100,0%
cohere-transcribe-03-2026	33,3%	75,0%	100,0%
voxtral-Small-24B-2507	50,0%	71,4%	100,0%
mms-1b-all	100,0%	100,0%	100,0%
granite-4.0-1b-speech	100,0%	150,0%	133,3%

TABLE IV

WERp90 OF ASR MODELS IN COMMON VOICE DATASET AND IN BR AND PI PARTITIONS OF THE PROPOSED ASR BENCHMARK.

the shape of the distribution of *WER* [14]. To achieve the 90th percentile, we calculate *WER* by sentence and order them. It is important to notice that a value above 1 is possible, considering hallucinations of repetition and audios whose correct answer has only one word.

Qualitative analysis: We investigate the location of the transcription errors in a sample of 100 audio files from the *Benchmark PI* dataset to better understand the impact of accented language [11]. For this, we analyze the transcriptions of each model. For each transcription, we annotate errors in common terms, in proper nouns, in dialectal terms and hallucinations. Proper nouns encompass names of people, places, companies; dialectal terms include words and expressions that are proper from PI accent; hallucinations refer to inclusions of words not presented in the audio file, significant change

of sentence or exclusion of entire pieces of the intended transcription; common terms are words or expressions that can be found in several accents.

D. Models

The requirements to select a model to be evaluated were: SOTA performance in Brazilian Portuguese, recently updated and easy to access. The Open ASR Leaderboard [27] was used as starting point to investigate models. The selected list of models appears in Table II. All of the candidates are ASR models, except for voxtral-Small-24B-2507 [25] and granite-4.0-1b-speech [26]. These models can perform other tasks that involve audio, such as language identification and translation, and are considered *speech-language* (SL) models.

IV. RESULTS

The results of each model in the Benchmark are presented in three situations: global results, presented in Table III for *WER* and in Table IV for *WERp90*; localized results, discriminated by state, in Table V; and qualitative results in Table VI.

Considering global results, the model with the best accuracy in *Benchmark PI* is qwen-ASR-1.3B, with 20.0% *WER* and *WERp90* of 66.7%. This model is the second best in *Benchmark BR*, where parakeet-TDT-0.6B-v3 is the best model with 19.8% *WER*. In this dataset, none of the models got less than 100.0% *WERp90*. The best model in Common Voice was cohere-transcribe-03-2026, be it in *WER* (4.2%) or *WERp90* (33.3%). A statistical significance test was performed on *WER* results comparing the best model with all of the remaining models using bootstrapped pairwise comparison. Results of cohere-transcribe-03-2026 in Common Voice and qwen-ASR-1.3B in Benchmark PI achieve p-value below 0.005 when compared with remaining models. In dataset Benchmark BR, parakeet-TDT-0.6B-v3 could not achieve significant separation from qwen-ASR-1.3B, however, these two models' results achieve p-value below 0.005 when compared with remaining models.

Considering localized results, *WER* split by state from the partition *Benchmark BR* is represented in Table V, with states presented in alphabetical order. The best performance in each state is highlighted in bold. Five different models achieve

Modelo	TOTAL	AL	BA	CE	MG	PA	PB	PE	PI	RJ	RS	RO	SP
whisper-base	99,8%	68,0%	66,3%	63,5%	107,5%	118,9%	149,6%	184,8%	124,9%	62,3%	73,6%	79,2%	137,8%
whisper-medium	31,5%	27,7%	19,0%	22,8%	32,6%	51,9%	40,3%	77,0%	29,5%	19,0%	27,0%	41,6%	36,5%
whisper-large-v2	26,7%	24,3%	17,9%	19,2%	27,4%	35,6%	26,7%	44,8%	36,8%	20,8%	20,8%	25,1%	34,0%
whisper-large-v3	20,6%	19,2%	14,6%	14,2%	23,4%	26,5%	19,9%	32,3%	24,2%	14,0%	18,1%	20,9%	26,0%
qwen-ASR-1.3B	20,2%	24,4%	14,1%	15,5%	22,9%	24,6%	23,3%	29,3%	20,4%	15,5%	28,9%	25,4%	22,9%
parakeet-TDT-0.6B-v3	19,8%	19,3%	15,1%	15,0%	23,7%	23,3%	21,8%	31,3%	20,2%	15,2%	18,4%	19,7%	21,8%
canary-1B-v2	27,1%	29,0%	19,1%	19,7%	29,9%	32,1%	31,9%	44,6%	28,2%	19,1%	23,3%	30,0%	33,2%
cohere-transcribe-03-2026	20,5%	20,1%	15,2%	16,2%	22,8%	23,5%	22,3%	30,2%	21,0%	15,0%	18,2%	21,3%	24,7%
voxtral-Small-24B-2507	21,0%	18,9%	14,1%	13,9%	23,8%	25,1%	21,0%	43,1%	30,3%	15,5%	16,4%	21,3%	25,4%
mms-1b-all	48,3%	55,5%	40,5%	42,4%	50,2%	56,4%	55,3%	62,6%	54,6%	39,8%	44,6%	57,8%	52,8%
granite-4.0-1b-speech	69,3%	77,1%	55,6%	59,0%	68,0%	84,0%	82,5%	79,1%	76,6%	54,9%	72,2%	77,9%	79,2%

TABLE V

WER BY MODEL IN EACH ACCENT PRESENT IN *Benchmark BR*. BEST RESULTS IN BOLD.

Model [reference]	Proper Nouns	Common Terms	Hallucination	Dialectal Terms
whisper-base	11%	42%	15%	9%
whisper-medium	12%	28%	10%	9%
whisper-large-v2	9%	25%	20%	10%
whisper-large-v3	6%	20%	19%	6%
qwen-ASR-1.3B	7%	19%	7%	5%
parakeet-TDT-0.6B-v3	6%	19%	20%	6%
canary-1B-v2	11%	34%	7%	7%
cohere-transcribe-03-2026	4%	22%	13%	6%
voxtral-Small-24B-2507	4%	26%	10%	7%
mms-1b-all	15%	80%	9%	8%
granite-4.0-1b-speech	13%	44%	58%	8%

TABLE VI

PERCENT OF SAMPLES WITH EACH TYPE OF ERROR IN THE 100 SAMPLES OF *Benchmark PI*.

best result in at least one of the represented states. Model parakeet-TDT-0.6B-v3 yields best performance in four states, PA, PI, RO and SP. voxtral-Small-24B-2507 and Whisper-large-v3 are best in 3 states each, while qwen-ASR-1.3B is best in 2 and textttcohere-transcribe-03-2026 is best in one.

Qualitative analysis of the transcription errors shows the type of word where errors were located. Table VI shows that all models include errors in dialectal-specific words. These errors include misrecognition of expressions, such as *rapaz, o homem...* that was transcribed as *Paul Omi* using Qwen, *Palomis* using Whisper-large-v3 and simply was not transcribed by Parakeet. All models presented some errors in these types of words, with a minimum of 5% of the samples using Qwen and a maximum of 10% using Whisper-large-v2.

V. DISCUSSION

The experiments in this study indicate that the performance of the top ASR models in accented speech is far inferior to the performance on clear speech datasets, which are commonly used in evaluating and training ASR models [7], [23]. The average *WER* obtained by all models in clean speech is 10.2%, while the average *WER* in both partitions of the Benchmark are 39.7%, indicating almost four times more errors. These results confirm a trend reported on investigations of the performance of ASR models in Quebecois French, that found 7 of 9 models above 10% *WER* in the accented speech, while in clean speech only 1 of 9 achieves this threshold [11]. The same trend is presented in sociolinguistic accents, as seen in [9], that compares Chicano, Spanglish and other accents formed across social and ethnic divisions instead of geographic ones. This is also ratified by the presence of transcription errors in dialectal terms in the qualitative analysis of the models' response, as presented in Table VI.

The analysis of error location shows that all models presented at least 5% of the sentences with some kind of regional expression or term where they failed to transcribe the speech. Even large models such as whisper-large-v2 accused errors in 10% of the sentences. Examples of errors include getting the verb tense wrong, i.e. *chamar* being transcribed as *chamou*, a mistake that 6 of the 11 models made, or mistranscribing a construction such as *estados que estão aqui*, typical of the Piauí accent, that was transcribed as *estados estão aqui*, the standard construction, by 9 of the 11 models. The error location analysis presented in [6] found 2.5% of the samples in accented English with errors on dialectal terms. The examples presented in the paper indicate the same types of errors. While the experimental results are too sparse to infer more about the impact of these failures in *WER* or in intelligibility, they can be seen as representations of the unique foil that accent represents to ASR models. Another hypothesis is that this is a particular case of low representation of particular accents in the datasets used in the training of the majority of foundation ASR models, as asked in [11]. This phenomenon is analogous to the poor performance in low-resource languages, also under-represented in multilingual datasets [12].

Finally, we discuss performance on accented speech across models. Despite being a good indicator of performance on clean speech, model size, presented on Table II, is not a good indicator of performance in accented speech, as the results in Table III clearly show. The best model on clean speech is cohere-transcribe-03-2026, while the best overall model on accented speech is qwen-ASR-1.3B. Excluding speech models, such as granite-4.0-1b-speech and voxtral-Small-24B-2507, the correlation between model size and performance on accented speech improves a little, but it is still not a reliable indicator. However, granite-4.0-1b-speech presents performance comparable to the best models in clean speech and one of the worst performances in accented speech, full of hallucinations (as seen in Table VI), which indicates lack of generalizability. voxtral-Small-24B-2507 doesn't show the same difference in performance, however, this is the biggest among the evaluated models, which can help explain.

VI. CONCLUSION

In this work, we outlined a broad methodology to evaluate ASR models in accented speech, detailing how to acquire

speech datasets and how to evaluate, using metrics such as WER_{p90} and a qualitative measure of the types of words where transcriptions occur. We added an implementation in Piauí accent in Portuguese, finding that the methodology helps us gain insight on the performance of ASR models in this condition.

The practical evaluation brings to light the gap in performance between ASR models in clean, read speech and in spontaneous speech. Comparing the same models on read and spontaneous speech, we found models yield on average almost 4 times more transcription errors in spontaneous speech. Qualitative analysis accuses the presence of errors exclusively due to regional terms or expressions present in 5% to 10% of the datasets, depending on the model.

The practical results of the implementation of this broad methodology in Piauí accent indicate that neither model size nor performance in read speech are decisive indicators of model performance in accented speech. This highlights the importance of the creation of training datasets specific for the targeted accent in the current architecture of solutions for ASR.

One of the main limitations of this work is that we only evaluated extensively the performance of ASR models in one regional dialect. Results in Table III, as well as the differences in performance presented in [11], indicate that there are differences between accents. Further work can investigate the different effects in different Brazilian Portuguese accents.

The second limitation is that the dataset gathered for our benchmark combines several conditions that impair the performance of ASR models and services: accented speech, informal speech, noise, uncontrolled conditions. We cannot isolate the effect of accent in the observed performance. The presence of 'regional accent' related transcription errors is backed by experimental data, but just the upper bound of their magnitude can be known.

ACKNOWLEDGMENTS

This work was developed within the scope of the SoberanIA initiative, a Brazilian effort initiated in the state of Piauí, focused on ethically sourced, auditable data and technological sovereignty.

REFERENCES

- [1] A. Poushneh, "Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors," *Journal of retailing and consumer services*, vol. 58, p. 102283, 2021.
- [2] C. Rzepka, B. Berger, and T. Hess, "Voice assistant vs. chatbot—examining the fit between conversational agents' interaction modalities and information search tasks," *Information Systems Frontiers*, vol. 24, no. 3, pp. 839–856, 2022.
- [3] P. Patel, S. Pampaniya, A. Ghosh, R. Raj, D. Karuppai, and S. Kandasamy, "Enhancing accessibility through machine learning: A review on visual and hearing impairment technologies," *IEEE Access*, vol. 13, pp. 33286–33307, 2025.
- [4] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran, "Speech recognition for illiterate access to information and technology," in *2006 International conference on information and communication technologies and development*, pp. 83–92, IEEE, 2006.
- [5] A. Koencke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Troups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the national academy of sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.
- [6] C. Harris, C. Mgbaurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15166–15184, 2024.
- [7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.
- [8] A. Aksënova, A. Bruguier, A. Ritchart-Scott, and U. Mendlovic, "Algorithmic exploration of American English dialects," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7374–7378, IEEE, 2020.
- [9] A. Aksënova, Z. Chen, C.-C. Chiu, D. Van Esch, P. Golik, W. Han, L. King, B. Ramabhadran, A. Rosenberg, S. Schwartz, *et al.*, "Accented speech recognition: Benchmarking, pre-training, and diverse data," *arXiv preprint arXiv:2205.08014*, 2022.
- [10] A. Guevara-Rukoz, I. Demirsahin, F. He, S.-H. C. Chu, S. Sarin, K. Papatrisawat, A. Gutkin, A. Butryna, and O. Kjartansson, "Crowdsourcing Latin American Spanish for low-resource text-to-speech," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6504–6513, 2020.
- [11] C. Serrand, G. Boulianne, and A. Morsli, "Benchmarking large pre-trained multilingual models on Québec French speech recognition," *arXiv preprint arXiv:2508.21193*, 2025.
- [12] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, 2021.
- [13] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1–2, pp. 19–28, 2002.
- [14] M. Riviere, J. Copet, and G. Synnaeve, "Asr4real: An extended benchmark for speech models," *arXiv preprint arXiv:2110.08583*, 2021.
- [15] A. Matos, G. Araújo, A. C. Junior, and M. Ponti, "Accent classification is challenging but pre-training helps: a case study with novel Brazilian portuguese datasets," in *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pp. 364–373, 2024.
- [16] A. C. Junior, E. Casanova, A. Soares, F. S. de Oliveira, L. Oliveira, R. C. F. Junior, D. P. P. da Silva, F. G. Fayet, B. B. Carlotto, L. R. S. Gris, *et al.*, "Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese," *arXiv preprint arXiv:2110.15731*, 2021.
- [17] N. A. R. Batista, T. F. Tavares, and P. A. Barbosa, "Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets," in *Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (35.: 2018: Campina Grande, PB)*, 2018.
- [18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [19] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," <https://github.com/snakers4/silero-vad>, 2024.
- [20] S. E. Leal, A. C. Junior, R. Marcacini, E. Casanova, O. Gonçalves, A. S. Soares, R. F. Lima, L. R. S. Gris, and S. Aluísio, "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*, pp. 6076–6087, 2025.
- [21] Instituto Brasileiro de Geografia e Estatística, *Censo Demográfico 2022: população e domicílios: primeiros resultados*. Rio de Janeiro: IBGE, 2023. Acesso em: 01-02-2026.
- [22] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang, *et al.*, "Qwen3-ASR technical report," *arXiv preprint arXiv:2601.21337*, 2026.
- [23] M. Sekoyan, N. R. Koluguri, N. Tadevosyan, P. Zelasko, T. Bartley, N. Karpov, J. Balam, and B. Ginsburg, "Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual ASR and AST," *arXiv preprint arXiv:2509.14128*, 2025.
- [24] J. Mack, E. Ranjan, W. Beller-Morales, B. Venkitesh, and P. Richemond, "cohere-transcribe-03-2026 (revision d96e814)," 2026.
- [25] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy, *et al.*, "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025.
- [26] G. Saon, A. Dekel, A. Brooks, T. Nagano, A. Daniels, A. Satt, A. Mittal, B. Kingsbury, D. Haws, E. Morais, *et al.*, "Granite-speech: open-source speech-aware llms with strong english ASR capabilities," *arXiv preprint arXiv:2505.08699*, 2025.
- [27] V. Srivastav, S. Zheng, E. Bezzam, E. L. Bihan, N. Koluguri, P. Zelasko, S. Majumdar, A. Moumen, and S. Gandhi, "Open ASR leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation," 2026.