International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Optimizing ASR Models through Contrastive Learning-Based Audio Filtering

Mr. Rushikesh.G.Kothare[1], Mr. Rohit.S.Ambhore[2], Mr. Pranay.S.Awachar[3], Mr. Nitin.R.Diware[4], Prof. Shivani.V.Khedkar[5]

[1,2,3,4] *Student, Computer Science & Engineering, Padm shri Dr V.B.Kolte College of Engeneering ,Malkapur*
[5] *professor, , Computer Science & Engineering, Padm shri Dr V.B.Kolte College of Engeneering ,Malkapur*

DOI: 10.5281/zenodo.19254559

ABSTRACT

This paper presents a novel methodology for enhancing Automatic Speech Recognition (ASR) performance by utilizing contrastive learning to filter synthetic audio data. We address the challenge of incorporating synthetic data into ASR training, especially in scenarios with limited real world data or unique linguistic characteristics. The method utilizes a contrastive learning model to align representations of synthetic audio and its corresponding text transcripts, enabling the identification and removal of low-quality samples that do not align well semantically. We evaluate the methodology on a medium-resource language across two distinct datasets: a general-domain dataset and a regionally specific dataset characterized by unique pronunciation patterns. Experimental results reveal that the optimal filtering strategy depends on both model capacity and dataset characteristics. Larger models, like Whisper Large V3, particularly benefit from aggressive filtering, while smaller models may not require such stringent filtering, especially on non-normalized text. This work highlights the importance of adjusting synthetic data augmentation and f iltering to specific model architectures and target domains. The proposed method, robust and adaptable, enhances ASR performance across diverse language settings. We have open-sourced the entire work, which includes 140 hours of synthetically generated Portuguese speech, as well as the pipeline and parameter settings used to create these samples. Additionally, we provide the fine-tuned Whisper models and the code required to reproduce this research. Our code will be available at https://github.com/my-north ai/semantic_audio_filtering.

Keywords : *- Automatic speech recognition, contrastive learning, data augmentation, embeddings, synthetic data filtering, text-to-speech*

1. INTRODUCTION

Recent advances in Automatic Speech Recognition (ASR) have been driven by the combination of large-scale datasets and innovative modeling techniques. In particular, the integration of natural and synthetic speech data has emerged as a powerful strategy to improve ASR performance. While natural speech captures real-world acoustic and linguistic variability, synthetic speech generated via Text-to-Speech (TTS) systems offers a scalable and cost-effective solution, especially for low- and medium-resource languages. However, the usefulness of synthetic data strongly depends on its quality and relevance, making effective filtering mechanisms essential.

1.1Synthetic Data Filtering via Contrastive Learning

To address the challenge of synthetic data quality, we propose a filtering methodology based on contrastive learning. By aligning representations of spoken audio and their corresponding text transcripts, the method identifies synthetic samples with weak semantic or acoustic correspondence. Samples that fail to meet predefined similarity thresholds are removed, ensuring that only high-quality synthetic data is retained. In addition, we enforce data integrity across all training sets using a words-per-second criterion to detect abnormal speech patterns.

1.2 Contributions and Experimental Impact

The main contributions of this work are threefold. First, we introduce a novel contrastive-learning-based approach for filtering synthetic speech data used in ASR training. Second, we analyze the effects of different filtering thresholds and model sizes on ASR performance, offering practical guidance for synthetic data utilization. Finally, we demonstrate significant improvements in word error rate for Portuguese ASR, validating the effectiveness of the proposed method in a medium-resource language setting.

2. LITERATURE SURVEY

Several studies in the literature have explored the use of synthetic speech data to address the challenge of limited training resources in Automatic Speech Recognition (ASR). Text-to-Speech (TTS)–based data augmentation has been shown to improve recognition performance, particularly for low-resource languages, by

increasing data diversity in a cost-effective manner. Researchers have reported reductions in word error rate when synthetic speech is combined with natural data. However, existing works also highlight that the benefits of synthetic augmentation are highly dependent on the quality of the generated audio and the accuracy of its corresponding transcripts. Poor-quality or misaligned synthetic samples can introduce noise and negatively impact ASR performance, motivating the need for effective filtering and quality control mechanisms in current ASR systems.

2.1 Synthetic Data Augmentation in ASR

Synthetic data generation using Text-to-Speech (TTS) has become a prominent strategy to address data scarcity in ASR, particularly for low-resource languages. Early work by *Wang et al.* showed that TTS-generated speech can supplement scarce natural data, yielding measurable gains in recognition performance when integrated into training pipelines. Similarly, *Ko et al.* demonstrated that augmenting read-speech corpora with synthetic samples improves model robustness against acoustic variability. More recent approaches exploit neural TTS systems — such as those based on Tacotron2 or FastSpeech architectures — to produce high-fidelity synthetic speech that better captures prosodic and phonetic diversity. These studies collectively indicate that synthetic augmentation can reduce word error rates (WER) and improve coverage of infrequent patterns, **provided that the synthetic quality is sufficiently high**. However, several authors also report that unfiltered synthetic data may introduce artifacts or biased representations that harm model generalization, especially when domain mismatch occurs between synthetic and real speech.

2.2 Filtering and Quality Control of Synthetic Audio

Recognizing the limitations of uncurated synthetic data, recent research has explored filtering mechanisms to ensure only high-quality synthetic samples contribute to ASR training. For instance, *Hannun et al.* proposed confidence-based selection criteria using ASR posterior probabilities to discard low-quality generated speech. Other methods leverage representation alignment, where speech embeddings are compared with corresponding text embeddings to assess semantic fidelity — a concept closely related to contrastive learning frameworks recently applied in multimodal tasks. *Raffel et al.* and *Ghazvininejad et al.* explored contrastive objective functions to strengthen the relationship between paired modalities, underscoring the potential of representation learning for filtering. These works highlight that selective incorporation of synthetic samples not only improves training efficiency but also enhances model performance by minimizing noise introduced by irrelevant or poorly generated audio.

Recent studies published in 2023 have emphasized the effectiveness of hybrid deep learning architectures, particularly CNN–LSTM models, in medical prediction tasks. These hybrid approaches combine CNNs for spatial feature extraction with Long Short-Term Memory (LSTM) networks for learning temporal and sequential patterns in patient data. Experimental results indicate that CNN–LSTM models consistently outperform standalone CNN or LSTM architectures in terms of accuracy, precision, recall, and robustness. Such models are especially effective for time-series data such as ECG signals and patient medical histories.[3]

Despite these advancements, several research gaps remain. Many existing models fail to achieve an optimal balance between accuracy, scalability, and robustness required for real-world clinical deployment. Additionally, limited incorporation of explainable artificial intelligence (XAI) techniques reduces transparency and clinician trust, highlighting the need for interpretable and reliable prediction systems in healthcare applications.

3. METHODOLOGY

The proposed methodology is designed to improve Automatic Speech Recognition (ASR) performance by applying a semantic filtering framework to synthetic audio data before its integration into model training. The approach follows a multi-step process that combines data collection, representation, filtering, refinement, and fine-tuning

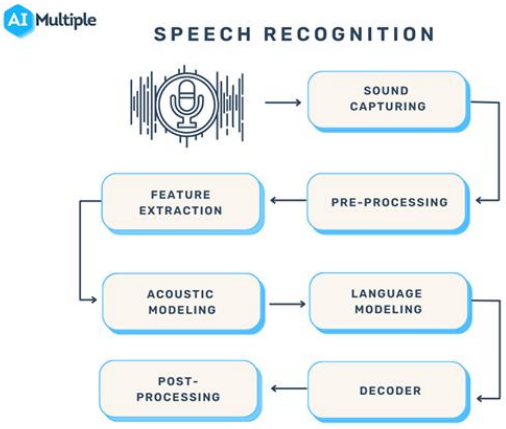

Fig:1 System Diagram.

3.1 Proposed System Architecture and Workflow

The proposed system architecture follows a structured workflow designed to enhance Automatic Speech Recognition (ASR) performance through high-quality synthetic data filtering. The system begins with data collection, where real Portuguese speech is gathered from publicly available corpora such as Multilingual Libri Speech, Common Voice, and the PSFB dataset, while synthetic speech is generated using a state-of-the-art Text-to-Speech (TTS) model. Prior to processing, text inputs for synthesis and real transcripts undergo preprocessing to remove symbols, formatting inconsistencies, and transcription irregularities. Both audio and text data are then converted into meaningful representations: audio waveforms are encoded into embeddings using the Whisper encoder, and textual transcripts are encoded using a Portuguese-pretrained De BER Ta language model.

These embeddings are aligned through a contrastive learning framework that projects both modalities into a shared embedding space, where cosine similarity is used to measure semantic correspondence. Synthetic samples with low similarity scores are filtered out using varying threshold levels to balance data quality and quantity. An additional post-filtering step removes outliers based on abnormal speech rates measured in words per second. The refined datasets are subsequently used to fine-tune Whisper models of different sizes, enabling adaptation to Portuguese speech and dialectal variations. Finally, system performance is evaluated using Word Error Rate on both dialectal and general-domain test sets, resulting in a robust and accurate ASR model.

4. CONCLUSIONS

The proposed semantic audio filtering framework for Automatic Speech Recognition (ASR) addresses one of the most pressing challenges in the field—the integration of synthetic speech data without compromising model performance. By leveraging contrastive learning to align audio and text embeddings, the system ensures that only semantically consistent and high-quality synthetic samples contribute to model training. This approach not only reduces noise within the dataset but also leads to significant improvements in recognition accuracy, particularly in terms of Word Error Rate (WER).

The research highlights an important relationship between model capacity and filtering intensity, demonstrating that smaller models perform better with moderate filtering, while larger models achieve optimal results with aggressive filtering strategies. Moreover, the evaluation across dialect-rich and general-domain datasets shows that the framework enhances robustness, adaptability, and real-world applicability of ASR systems.Beyond technical performance, the project contributes to the broader research community through the release of synthetic datasets, filtering pipelines, and fine-tuned Whisper models. This fosters reproducibility, scalability, and innovation across multiple languages and domains, making the work especially relevant for low-resource and dialect-sensitive contexts.

In conclusion, the project establishes semantic audio filtering as a practical and scalable solution for advancing the accuracy, inclusivity, and accessibility of ASR systems, paving the way for more reliable speech technologies worldwide.

5. REFERENCES

[1] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, ‘‘Making more of little data: Improving low-resource automatic speech recognition using data augmentation,’’ 2023, arXiv:2305.10951.

[2] N.Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, ‘‘Generating synthetic audio data for attention-based speech recognition systems,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 7069–7073.

[3] C. Du and K. Yu, ‘‘Speaker augmentation for low resource speech recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 7719–7723.

[4] S. Liu, L. Sari, C. Wu, G. Keren, Y. Shangguan, J. Mahadeokar, and O. Kalinli, ‘‘Towards selection of text-to-speech data to augment ASR training,’’ 2023, arXiv:2306.00998.

[5] J. Huang, Y. Bai, Y. Cai, and W. Bian, ‘‘A study on the adverse impact of synthetic speech onspeechrecognition,’’ in Proc.IEEEInt.Conf.Acoust., Speech Signal Process. (ICASSP), vol. 33, Apr. 2024, pp. 10266–10270.

[6] X.Liu,F. Zhang, Z. Hou, L.Mian,Z.Wang,J.Zhang,andJ. Tang, ‘‘Self supervised learning: Generative or contrastive,’’ IEEE Trans. Knowl. Data Eng., vol. 35, no. 1, pp. 857–876, Jan. 2023.

[7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, ‘‘Common voice: Amassively-multilingual speech corpus,’’ 2019, arXiv:1912.06670.