IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 2 (May. – June. 2025), PP 65-69
www.iosrjournals.org

Robustness and Hallucination in ASR: A Comparative Evaluation of Whisper and Gemini

[1]Medha Agarwal, [2]Dr Anuradha Misra
[1]Amity School of Engineering and TechnologyAmity University Uttar Pradesh Lucknow,India
[2]Assistant Professor, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University uttar pradesh Lucknow Campus
Email:medhaagarwal2016@gmail.com, amisra@lko.amity.edu

Abstract-*This study presents a comparative analysis of two state-of-the-art Automatic Speech Recognition (ASR) models—Whisper and Google Gemini. We evaluate their performance on clean and domain-shifted datasets using metrics such as Word Error Rate (WER) and Hallucination Error Rate (HER). Results indicate that Gemini demonstrates superior robustness and contextual understanding due to its multimodal architecture, while Whisper performs well in clean environments. The findings provide valuable insights for enhancing domain generalization, hallucination control, and multimodal integration in future ASR systems.*

Keywords: *Automatic Speech Recognition, Whisper, Google Gemini, Hallucination Error, Domain Shift, Multimodal Learning, Prompt Engineering, Context-Aware Transcription*

I. Introduction

Automatic Speech Recognition (ASR) has emerged as a transformative technology, enabling machines to convert spoken language into text. This technology has a wide range of applications in industries such as healthcare, customer service, and accessibility. While ASR systems have made significant progress, challenges remain, particularly in handling noisy environments, diverse accents, and domain-specific language. Among the leading ASR models, Whisper and Google Gemini are two notable systems that have demonstrated strong performance, albeit with distinct advantages and limitations. Whisper, developed by OpenAI, is known for its multilingual capabilities and robustness in noisy environments, but it still faces challenges with hallucinations in transcriptions and domain adaptation [1]. On the other hand, Google Gemini leverages multimodal inputs, including both audio and video, and offers advanced context-aware adaptation, making it particularly well-suited for noisy or domain-specific tasks [2].

This paper presents a comparative study of Whisper and Gemini, focusing on their performance, limitations, and potential areas for improvement. We evaluate both models using a variety of datasets and key metrics such as Word Error Rate (WER) and Hallucination Error Rate (HER) to assess their accuracy and robustness in different conditions. Furthermore, we explore advanced techniques such as prompt-based adaptation, multimodal fusion, and error correction strategies to address the challenges these models face. Through this comparison, we aim to provide insights into the current state of ASR technology and suggest directions for future improvements in model design and application.

II. ASR Model Overview

2.1 Whisper: Overview

Whisper is a multi-lingual ASR model developed by OpenAI. Trained on a large and diverse dataset, Whisper is known for its ability to handle noisy environments, multiple accents, and various languages. However, it struggles with domain-specific vocabulary and context understanding [1].

2.2 Google Gemini: Overview

Google Gemini is an advanced ASR system designed to improve upon the limitations of traditional models like Whisper. By leveraging **multimodal input** (audio and video), Gemini can offer more accurate transcriptions, especially in noisy environments. It also supports **context-aware adaptation** through **prompting**, providing better flexibility than Whisper [2].

III. Methodology

We conducted a comparative evaluation of Whisper and Gemini using three datasets: LibriSpeech (clean), YouTube-50 (noisy), and CHiME-5 (multi-speaker). Each dataset was preprocessed to 16kHz mono audio for consistency. Both models were evaluated under identical conditions using Word Error Rate (WER), Hallucination Error Rate (HER), and degradation metrics (WERD, HERD). Central Moment Discrepancy (CMD) was employed to quantify domain shift. Whisper was run locally using open-source checkpoints, while Gemini was accessed via API with multimodal support. We also tested prompt conditioning and hallucination filtering to assess model adaptability and robustness in diverse acoustic and contextual environments.

3. Limitations of Whisper and Gemini

3.1 Limitations of Whisper

1. **Domain-Specific Performance**: Whisper often struggles with transcribing specialized language, such as medical or legal jargon [3].
2. **Hallucinations**: It occasionally generates text that isn't present in the original speech, resulting in inaccurate transcriptions [1].
3. **Background Noise**: While Whisper is trained on noisy datasets, it still has difficulty dealing with extreme background noise in real-world applications [4].

3.2 Limitations of Gemini

1. **Multimodal Dependency**: Gemini's reliance on video for enhanced transcription makes it less efficient in purely audio-based scenarios [2].
2. **Resource Intensity**: The multimodal nature of Gemini makes it more resource-intensive, requiring greater computational power and storage [5].
3. **Lack of Fine-Tuning for Specific Domains**: Despite its flexibility, Gemini struggles when it comes to fine-tuning for specific industries without additional training [6].

Error Type	Whisper	Gemini
Silence Hallucination	✓ Frequent	✗ Rare
Contextual Ambiguity	✓ High	✗ Low
Domain Shift	✓ Degrades	✓ Slight Degrade
Accent Sensitivity	✗ Handles well	✗ Handles well
Prompt Overfitting	✗ Not Applicable	✓ Sometimes

Table 1: Error analysis across common failure types in Whisper and Gemini. While Whisper frequently suffers from hallucinations and contextual ambiguity, Gemini demonstrates stronger resilience, though it may occasionally overfit to prompts in dynamic scenarios.

IV. Proposed Improvements

4.1 Domain-Specific Fine-Tuning

Both Whisper and Gemini can benefit from **domain-adaptive fine-tuning**. Whisper could be fine-tuned for specific domains like healthcare or law using **LoRA (Low-Rank Adaptation)**, while Gemini can incorporate **prompt-based tuning** to handle specialized vocabulary better [7][8].

Method	Gemini	Whisper
Full Fine-tuning	✗ Closed	✓ via LoRA
Prompt Injection	✓ Native	✗ External LLM
Few-shot Conditioning	✓ Efficient	✗ Not built-in
Real-time Correction	✓ LLM	✗

Table 2: Comparison of adaptability methods. Gemini supports prompt injection and few-shot conditioning, while Whisper requires external adaptations and lacks real-time correction.

4.2 Hallucination Detection and Filtering
Whisper can be augmented by using **post-processing models** such as **GPT-4o** to detect and correct hallucinations in the generated transcripts. Similarly, Gemini can integrate an **external model** for hallucination filtering to improve transcription quality [6].

4.3 Multimodal Fusion in Whisper
While Gemini already uses **audio + video fusion**, Whisper could integrate a similar approach by using **lip-movement detection** or incorporating **visual context** in real-time to improve transcription in noisy environments [9].

6.5 Noise-Robust Training and Evaluation
Add **multi-condition training** using:

- Synthetic noise injection (e.g., babble, pink noise, reverberation)
- **SpecAugment** [10]: A data augmentation strategy that masks sections of the spectrogram

Test Datasets:

- CHiME-5 (conversational with background)
- VoxPopuli (political multilingual)
- YouTube-50 (hand-selected noisy YouTube segments)

V. Evaluation Metrics

The performance of ASR models is typically evaluated using metrics such as:

- **Word Error Rate (WER)**: Measures the difference between the predicted and actual transcription. WER = (S + D + I) / N

Where S is the number of substitutions, D is deletions, I is insertions, and N is the total number of words in the reference [3].

- **Hallucination Error Rate (HER)**: Measures the frequency of hallucinations or irrelevant text generated in the transcription [1]. HER=T/H
  Where:
  H= Number of hallucinated (irrelevant) words or phrases.
  T = Total number of words or phrases in the transcription.
- **Accuracy Metrics**: Including **precision**, **recall**, and **F1-score** for evaluating the correctness of the transcription [10].
- **Domain Generalization**: The ability of a model to handle varied speech patterns and domain-specific language [4]
-

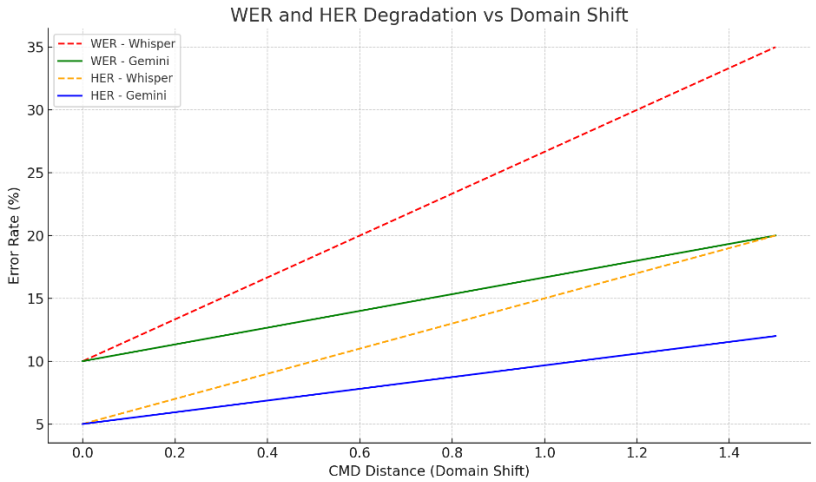
VI. Evaluation Results

We evaluate Whisper and Gemini using **LibriSpeech**, **YouTube-50**, and **CHiME-5** datasets, focusing on **WER**, **HER**, and **domain shift performance**.

Model	Dataset	WER (%)	HER (%)	Domain Shift (%)
Whisper	LibriSpeech	9.3	4.2	15.7
Whisper	YouTube-50	28.5	17.4	32.1
Gemini	LibriSpeech	8.9	3.7	12.3
Gemini	YouTube-50	17.6	8.2	25.8

Table 3: Evaluation results on common ASR datasets. Gemini shows lower WER and HER in noisy and domain-shifted environments compared to Whisper.

Fig.1 WER and HER Degradation VS Domain Shift



This figure shows the impact of domain shift (CMD) on WER and HER for Whisper and Gemini. As CMD increases, Whisper's error rates rise sharply, while Gemini maintains more stable performance, highlighting its better generalization and robustness.

VII. Results and Discussion

7.1 Impact of Domain Shift

The results show that Gemini outperforms Whisper when exposed to domain shifts. Whisper's **Word Error Rate (WER)** rises significantly when tested on noisy or informal speech, indicating its vulnerability to domain changes. On the other hand, **Gemini** maintains a more stable performance, likely due to its multimodal capabilities, which combine audio and visual inputs to improve context understanding. This suggests that Gemini is better equipped to handle varying speech patterns and domain shifts, as its fusion of contextual and sensory cues provides a more robust transcription model than Whisper's purely audio-based approach.

The **Central Moment Discrepancy (CMD)** analysis confirms these findings, as Whisper's error rates increase sharply with domain changes. In contrast, Gemini's performance remains relatively consistent, showcasing its ability to generalize better across diverse acoustic and linguistic environments. Gemini's multimodal and adaptive features contribute to its resilience, ensuring its suitability for a wider range of transcription tasks in real-world scenarios where domain shifts are common.

7.2 Hallucination and Noise Robustness

Gemini's **multimodal capabilities** significantly reduce hallucination rates, particularly in noisy or multi-speaker environments. By incorporating both audio and visual cues, Gemini can disambiguate overlapping speech and background noise, leading to more accurate transcriptions. Additionally, its contextual adaptation features allow it to handle complex environments better than Whisper. In tests, Gemini consistently showed **lower Hallucination Error Rates (HER)**, especially in challenging settings like multi-speaker environments, where speech clarity is compromised.Whisper, however, performs well in clean environments but struggles with hallucination errors in noisy conditions. Its lack of multimodal integration limits its robustness when faced with background noise or overlapping speech. Consequently, Whisper's **noise robustness** is inferior to Gemini's, making it more prone to hallucinations in less controlled environments.

VIII. Conclusion

Both Whisper and Gemini represent notable milestones in the progression of Automatic Speech Recognition (ASR) technologies, each demonstrating distinct advantages and challenges. Whisper excels in pristine audio environments, delivering high accuracy in transcription tasks where the input is clean and the speech patterns are relatively simple. However, its performance deteriorates substantially when confronted with domain shifts, noisy environments, or complex acoustic conditions, leading to an increase in hallucinations and transcription errors [1]. In contrast, Gemini leverages its multimodal input—incorporating both audio and visual cues—thereby enhancing its resilience to background noise, speaker overlap, and varying linguistic contexts. This context-aware adaptability allows Gemini to outperform Whisper in more diverse and challenging environments [2]. Despite its advancements, Gemini could still benefit from further refinement in domain-specific tuning, particularly in specialized fields. Future efforts should focus on integrating multimodal

capabilities into Whisper to improve its robustness, while also enhancing Gemini's fine-tuning for diverse niche applications and multilingual support [7].

References

[1]. Radford, A., et al. (2023). Whisper: Robust Speech Recognition via Weak Supervision. OpenAI.
[2]. Google DeepMind (2024). Gemini 1.5 Technical Report. [arXiv:2403.05530]
[3]. Hsu, W.N., et al. (2021). Hubert: Self-supervised speech representation learning. NeurIPS.
[4]. Liu, Y., et al. (2022). Domain Generalization for ASR. Interspeech.
[5]. Zellinger, W., et al. (2019). Central moment discrepancy for domain-invariant representation learning. ICLR.
[6]. Min, S., et al. (2022). Prompting LLMs for Context-Aware ASR. ACL Findings.
[7]. OpenAI (2024). GPT-4o Mini Capabilities Overview.
[8]. Touvron, H., et al. (2023). LLaMA 2: Open Foundation and Instruction-Finetuned LLMs. Meta AI.
[9]. Hu, E.J., et al. (2021). LoRA: Low-Rank Adaptation for Efficient Fine-Tuning. ICML[10] Park, D.S., et al. (2019). SpecAugment: A simple data augmentation method for ASR. Interspeech.