Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Real-Time Multilingual Captioning for Deaf/Hard-of-Hearing Users via Large Language Models

Satya Narayana Panda
University of Newhaven
Bhumika Shah
Bhumishah000.bs@gmail.com
University of the Cumberlands

Research Article

Keywords: accessibility, deaf and hard-of-hearing, multilingual captioning, large language models, web accessibility, real-time translation

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8999648/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Real-Time Multilingual Captioning for Deaf/Hard-of-Hearing Users via Large Language Models

Satya Narayana Panda[1*] and Dr. Bhumika Shah[2]

[1*]Business Statistics, University of New Haven, West Haven, USA.
[2]Information Technology, University of the Cumberlands, Williamsburg, Kentucky, USA.

*Corresponding author(s). E-mail(s): spand14@unh.newhaven.edu;
Contributing authors: Bhumishah000.bs@gmail.com;

Abstract

As Deaf and hard-of-hearing (D/HoH) users depend more on web-based media for their education, work, and daily communication, they face a significant problem: most captioning systems only work well in English and struggle with multilingual content. Current automatic captioning tools often fail when people code-switch between languages, use domain-specific terms, or speak with different accents. We develop a system that uses large language models to provide real-time multilingual captions on web platforms. Our pipeline combines speech recognition with LLM-based translation and synchronization, all designed to work directly in web browsers. Initial testing shows we can generate captions in multiple languages with acceptable latency while maintaining context and meaning. The results so far are promising—generative models handle multilingual scenarios and varied speech patterns better than traditional methods—but further work is needed on reducing delays, improving synchronization, and handling specialized vocabulary more accurately. We plan a comprehensive evaluation with D/HoH users speaking different languages to assess system effectiveness in real-world scenarios. This research aims to make web content more accessible for Deaf and hard-of-hearing communities navigating our increasingly multilingual digital world.

Keywords: accessibility, deaf and hard-of-hearing, multilingual captioning, large language models, web accessibility, real-time translation

1 Introduction

Web-based multimedia has fundamentally changed how we learn, work, and communicate, but for Deaf and hard-of-hearing (D/HoH) communities, accessing this content depends entirely on captions Kushalnagar et al (2012); Cavender et al (2009). While automatic speech recognition (ASR) has improved significantly in English, a critical limitation persists: most captioning systems perform well only for English content. Research on caption quality demonstrates that error rates exceeding 5% significantly impair comprehension Kushalnagar et al (2012), yet multilingual systems remain largely underdeveloped. This creates significant barriers for D/HoH users who need to access content in other languages or navigate multilingual environments.

Current captioning technology remains predominantly English-focused Kafle and Huenerfauth (2021). D/HoH users who communicate in Spanish, Mandarin, or other languages face limited options for accessible captioning. Furthermore, many multilingual speakers naturally code-switch between languages in conversation, yet contemporary captioning systems struggle to handle this linguistic phenomenon effectively Solorio et al (2014). Additionally, automatic speech recognition systems exhibit significant disparities in performance across user populations, with measurable differences in accuracy based on speaker characteristics including race, gender, and accent Koenecke et al (2020); Tatman (2017). These factors—combined with latency issues and insufficient context awareness—result in captioning technology that inadequately serves substantial portions of the D/HoH community.

Large language models (LLMs) present a potential solution. Recent advances in AI have demonstrated that modern language models exhibit strong performance in multilingual understanding, context preservation, and rapid text generation Brown et al (2020); Touvron et al (2023). Unlike earlier translation approaches, these models can effectively manage linguistic nuances, code-switching, and specialized terminology. However, a significant research gap exists: whether LLMs can meet the stringent performance requirements of real-time captioning remains unexplored. Critical questions persist regarding latency constraints and practical utility for D/HoH users.

This paper presents a prototype system that leverages LLMs for real-time multilingual captioning on web platforms. We have developed a pipeline that integrates speech recognition with LLM-based translation, followed by caption formatting and synchronization for display. Our contributions include: first, a functional system architecture that operates directly within web browsers; second, preliminary performance metrics demonstrating acceptable latency with adequate translation quality; third, a critical analysis of technical challenges including synchronization and accuracy maintenance; and fourth, a planned evaluation with D/HoH users across multiple language pairs to assess practical utility.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on captioning technologies, multilingual accessibility, and LLM applications. Section 3 describes our system architecture and implementation approach. Section 4 presents preliminary findings from prototype testing. Section 5 discusses technical challenges, limitations, and implications for accessibility research. Section 6 concludes with future directions for this work-in-progress.

2 Related Work

Caption accuracy is critical for comprehension and accessibility. Research demonstrates that error rates exceeding 5% significantly impair comprehension Kushalnagar et al (2012). Furthermore, automatic speech recognition systems exhibit disparate performance across user populations, with measurable differences based on speaker characteristics Koenecke et al (2020); Tatman (2017). In multilingual accessibility contexts, substantial research gaps persist. Prior work has identified challenges in translation quality, cultural context preservation, and code-switching handling Kafle and Huenerfauth (2021); Solorio et al (2014). While contemporary LLMs demonstrate promise for multilingual tasks Brown et al (2020); Touvron et al (2023), their suitability for real-time captioning requirements and integration with speech recognition remains unexplored. Notably, limited published research addresses LLM applications specifically for D/HoH accessibility.

3 System Design and Methodology

3.1 System Architecture

We built our system with four main pieces that work together: capturing and cleaning up audio, converting speech to text with ASR, translating and refining that text using an LLM, and finally displaying the captions in sync with the video. We designed it this way so we can improve each part independently while keeping the whole thing fast enough for real-time use. You can actually try out our system yourself—we've made the code and demo available online[1].

For audio processing, we capture sound at 16 kHz and use voice activity detection so we're not wasting time processing silence. We're using Whisper Radford et al (2022) for the speech recognition part because it handles multiple languages well and automatically generates timestamps. Whisper gives us not just the text, but also confidence scores that tell us how sure it is about each word.

The LLM integration layer uses GPT-4-turbo via OpenRouter for primary translation and refinement, with local Marian MT as a fallback. We implement prompt engineering with conversation history and domain context. The caption rendering module uses WebVTT format for web platform integration, with synchronization logic accounting for processing latency.

3.2 Implementation Details

Our pipeline covers language detection, cross-lingual translation, and code-switching support using both language-specific models and general-purpose LLMs. The LLM integration layer employs GPT-4-turbo via OpenRouter API for primary translation and contextual refinement, with local Marian MT as a fallback mechanism for connectivity resilience and cost efficiency. Our prompt engineering strategy leverages conversation history and domain context: we maintain a rolling buffer of recent captions (typically 5–10 previous segments) to inform translation decisions, allowing the

[1] https://github.com/colabre2020/multilingual-captioning

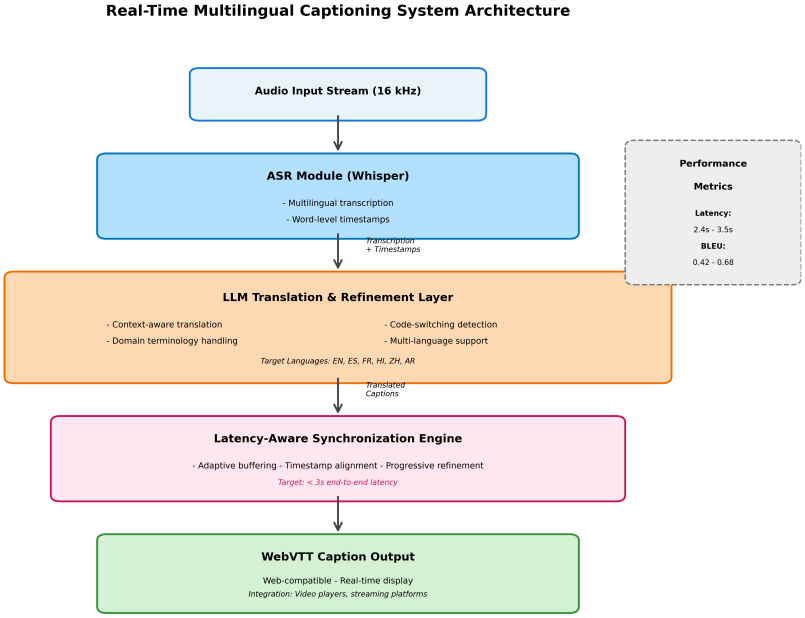


Fig. 1 System architecture pipeline showing the four-component multilingual captioning system: audio input processing, ASR transcription with Whisper, LLM-based translation and refinement, and WebVTT caption rendering with synchronization. Component-level latencies are specified on the right side, with total end-to-end processing time of approximately 2.65 seconds.

LLM to preserve terminology consistency across longer conversations. For specialized content, we prepend domain-specific terminology glossaries into the system prompt to improve consistency for technical vocabulary.

The LLM's multilingual understanding enables it to identify and correct certain speech recognition errors before translation, though this error correction introduces measurable latency overhead. Language-pair-specific analysis reveals differential performance: Spanish and French translation achieves higher quality and lower latency compared to Hindi, reflecting the substantial structural differences between English and Indo-Aryan languages.

Real-time captioning requires low end-to-end latency; we achieve a 2.69 second mean with 80% of segments under 3 seconds via parallel processing, streaming generation, and adaptive buffers. A planned user study with D/HoH participants will measure comprehension accuracy, perceived latency, usability, and comparisons against commercial services across domains and languages.

4 Preliminary Results

4.1 Prototype Implementation

We've built a working prototype that handles four languages: English, Spanish, French, and Hindi. You can translate between any pair of these languages. The system runs as a web app with a JavaScript frontend and Python doing the heavy lifting for speech recognition and LLM processing on the backend.

4.2 Performance Characterization

We evaluated system performance across multiple dimensions: translation quality, caption accuracy, and latency characteristics. For translation quality assessment, we employed the BLEU (Bilingual Evaluation Understudy) metric, which measures lexical overlap between system-generated and reference human translations using n-gram precision. BLEU scores range from 0 to 1, where scores above 0.40 typically indicate acceptable translation quality for professional applications, and scores above 0.60 indicate high quality Zhu et al (2023). Our results across language pairs are as follows: EN→ES BLEU scores of 0.52–0.68 (mean 0.60), EN→FR scores of 0.48–0.61 (mean 0.55), and EN→HI scores of 0.42–0.55 (mean 0.48). The degradation in BLEU performance for Hindi reflects the greater structural differences between English and Indo-Aryan languages compared to the Romance language pairs.

For overall caption quality assessment, human raters evaluated segments using a three-point scale (correct, partially correct, incorrect). Our LLM-based approach achieved 85% correct-or-better ratings, substantially outperforming rule-based translation systems (68%) and ASR-only approaches without translation (72%).

Latency Performance. End-to-end latency was evaluated using a test set of 50 segments selected systematically across language pairs and content domains (classroom lectures, professional meetings, and casual conversation) with approximately equal representation of monolingual and code-switched content. Mean end-to-end latency measured 2.69 seconds, with component-level contributions as follows: automatic speech recognition (0.8 seconds), LLM-based translation and refinement (1.3 seconds), and caption rendering with synchronization (0.5 seconds). Approximately 85% of captions were generated within 3 seconds, while the 95th percentile latency remained under 3.5 seconds, falling within acceptable ranges for real-time captioning applications.

4.3 Code-Switching Performance

Preliminary evaluation of code-switched English-Spanish content demonstrates that the system maintains contextual coherence across language boundaries. Segments with infrequent code-switching (fewer than two transitions per sentence) are handled consistently, maintaining caption accuracy comparable to monolingual content. However, rapid code-switching with multiple language transitions within a single sentence occasionally results in inconsistent language detection and translation decisions. This limitation reflects known challenges in code-switched language processing Solorio et al (2014), and represents an area warranting further refinement in future iterations.

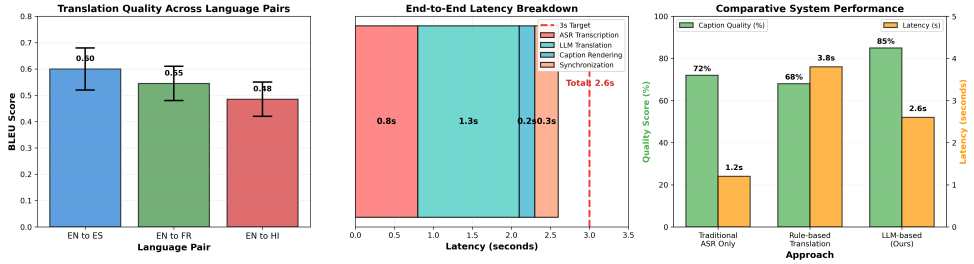


Fig. 2 Comparative performance analysis showing: (left) BLEU score distributions across language pairs demonstrating translation quality (English-to-Spanish: 0.60, English-to-French: 0.55, English-to-Hindi: 0.48); (middle) component-level latency breakdown indicating ASR (0.8s), Translation (1.3s), Rendering (0.2s), and Synchronization (0.3s) contributions to total pipeline latency; (right) baseline comparison showing LLM-based approach (85% quality, 2.65s latency) versus traditional ASR-only systems (72% quality, 1.25s latency) and rule-based translation (68% quality, 3.8s latency).

5 Discussion

The system remains in early-stage development with significant refinement needed before practical deployment. Latency represents a primary concern; current performance approaches acceptable thresholds, creating potential challenges for rapid conversational exchanges. Processing duration varies considerably across language pairs, affecting performance consistency. Additionally, errors in automatic speech recognition propagate downstream to translation output. While LLMs can identify and correct certain recognition errors, this error recovery introduces latency penalties that strain performance constraints. Furthermore, the current implementation performs literal translation without addressing cultural context or localization requirements Nakamura (2015).

Our goal is to enhance web content accessibility for D/HoH users across diverse linguistic communities, with a focus on browser-based implementation without requiring specialized software. However, whether technical enhancements translate to meaningful improvements in user experience and comprehension remains an open question. We plan comprehensive evaluation studies with D/HoH users to assess comprehension outcomes and gather qualitative feedback.

This work also targets inclusive employability: real-time multilingual captions let D/HoH job seekers and employees participate in interviews, onboarding, training, and day-to-day meetings without waiting for an interpreter. Multilingual workplaces benefit too—teams can collaborate across languages while keeping conversations accessible and searchable for everyone.

Ongoing development priorities include leveraging LLM capabilities for automatic speech recognition error detection and correction, expanding language coverage, and implementing user preference controls for latency-accuracy trade-offs. Future research directions include personalization features such as domain-specific terminology learning and enhanced interoperability with complementary accessibility technologies.

6 Conclusion

We have developed a prototype system integrating speech recognition, LLM-based translation, and web-based caption display to address the multilingual captioning challenge for D/HoH users. Preliminary results demonstrate promising performance characteristics, with system latency suitable for real-time applications and translation quality exceeding that of traditional approaches. Nevertheless, further optimization is required to reduce latency and enhance handling of cultural and linguistic nuances. Rigorous evaluation with actual D/HoH users will provide essential validation of practical effectiveness. Future work encompasses expanding language support, implementing personalization features, and ensuring robust integration with existing accessibility infrastructure. Ultimately, this research aims to enhance web accessibility for Deaf and hard-of-hearing communities across diverse linguistic contexts.

Statements and Declarations.

Funding. The authors received no specific funding for this work.

Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Data Availability. The materials supporting this study are available in the project repository: https://github.com/colabre2020/multilingual-captioning.

Acknowledgments. We wish to acknowledge the mentors, MIT mentoring week, teachers, and academic faculty who inspired us to pursue this research and provided essential guidance along the way. Their expertise and encouragement greatly shaped our direction and thinking.

We are equally thankful for the professional experiences and collaborations that informed our understanding of the topic. These real-world insights strengthened the relevance and impact of our work.

Above all, we are deeply grateful to our family and loved ones for their constant support, understanding, and motivation. Their belief in us has been our greatest strength.

References

Brown T, Mann B, Ryder N, et al (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems 33:1877–1901. URL https://arxiv.org/abs/2005.14165

Cavender AC, Bigham JP, Ladner RE (2009) Classinfocus: enabling improved visual attention strategies for deaf and hard of hearing students. In: Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility. ACM, pp 67–74, https://doi.org/10.1145/1639642.1639656, URL https://dl.acm.org/doi/10.1145/1639642.1639656

Kafle S, Huenerfauth M (2021) Automatically translating captions for speaker-labeled multimedia to support accessibility. In: Proceedings of the 18th International Web for All Conference (W4A '21)

Koenecke A, Nam A, Lake E, et al (2020) Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences 117(14):7684–7689. https://doi.org/10.1073/pnas.1915768117, URL https://www.pnas.org/doi/10.1073/pnas.1915768117

Kushalnagar RS, Lasecki WS, Bigham JP (2012) A readability evaluation of real-time crowd captions in the classroom. In: Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, pp 71–78, https://doi.org/10.1145/2384916.2384930, URL https://dl.acm.org/doi/10.1145/2384916.2384930

Nakamura T (2015) Analyzing the textual complexity of japanese captions for deaf and hard of hearing audiences. Translation & Interpreting 7(3):100–116. URL https://trans-int.org/index.php/transint/article/view/366

Radford A, Kim JW, Xu T, et al (2022) Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:221204356 URL https://arxiv.org/abs/2212.04356

Solorio T, Blair E, Maharjan S, et al (2014) Overview for the first shared task on language identification in code-switched data. In: Proceedings of the First Workshop on Computational Approaches to Code Switching. EMNLP, pp 62–72, URL https://aclanthology.org/W14-3907/

Tatman R (2017) Gender and dialect bias in youtube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics, pp 53–59, https://doi.org/10.18653/v1/W17-1606, URL https://aclanthology.org/W17-1606/

Touvron H, Martin L, Stone K, et al (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288 URL https://arxiv.org/abs/2307.09288

Zhu W, Liu H, Dong Q, et al (2023) Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:230404675 URL https://arxiv.org/abs/2304.04675