WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech-to-Text Translation Enhancement using LLM

Omkar Maskar, Omkar Parab, Vaibhav Mane

Mumbai University, ASM Institute of Management & Computer Studies, Higher educational institution in Thane, Maharashtra
omkarmaskar9@gmail.com, omkarparabwork@gmail.com, manevaibhav1701@gmail.com

ABSTRACT

Speech-to-text translation involves converting spoken language in one language into written text. This technology is widely used in various applications such as hands-free communication, dictation, video lecture transcription, and translation. The traditional approach relies heavily on Automatic Speech Recognition (ASR) and Machine Translation (MT) using Large Language Models (LLMs), which facilitate the conversion of spoken words into written text and enable cross-lingual communication. While ASR transcribes spoken language, MT translate the text into the target language. However, these separate processes can lead to error propagation and incur high resource and training costs. Consequently, researchers are exploring end-to-end (E2E) models for speech-to-text translation to address these issues. Despite this progress, the existing E2E ST models have not been thoroughly reviewed. This survey aimed to fill this gap by providing a detailed review of the models, metrics, and datasets used in ST tasks, highlighting challenges and future research directions. This review is intended to aid researchers in various ST-model applications.

Keywords: Speech-to-Text Translation, Automatic Speech Recognition, Machine Translation, Large Language Models

1.Introduction

The goal of speech-to-text (ST) translation is to convert spoken language from one language to written text in another. This task is crucial for applications such as automatic subtitling, dictation, video lecture translation, tourism, and telephone conversations. The ST problem can vary depending on whether ranslation is performed online (simultaneous translation) or offline. Online translation is essential for live video streaming, while offline translation is suitable for applications such as movies, where some delays are acceptable. This challenge is compounded by issues such as noisy inputs, low-resource or code-mixed languages, and multiple speakers.

Several datasets have played crucial roles in advancing both cascade and end-to-end (E2E) speech-to-text (ST) models. The Prabhupadavani dataset offers 54 minutes of code-mixed English-Bengali-Sanskrit speech translated into 25 languages, featuring 130 speakers. TIMIT, with 54 minutes of phonetica Several datasets have played crucial roles in advancing both cascade and end-to-end (E2E) speech-to-text (ST) models. The Prabhupadavani dataset offers 54 minutes of code-mixed English-Bengali-Sanskrit speech translated into 25 languages, featuring 130 speakers. TIMIT, with 54 minutes of phonetically balanced English speech from 630 speakers, is pivotal in ASR and speaker recognition. The Spoken Wikipedia Corpora provides 30 minutes of multilingual speech from Wikipedia articles, offering diverse language samples. NOIZEUS offers 30 minutes of noisy English speech for robust ASR evaluation. ARCTIC features 12 minutes of speech for speech synthesis research.

ESC-50 includes 42 minutes of environmental audio recordings for sound classification tasks. The Google Speech Commands Dataset provides 1.2 minutes of short English utterances for keyword spotting. VoxCeleb1 offers 3 minutes of celebrity speech for speaker verification. VCTK provides 6 minutes of English speech from 109 speakers, supporting ASR and speech synthesis. RAVDESS comprises 3.6 minutes of emotional English speech from 24 speakers, used for emotion recognition and ASR.

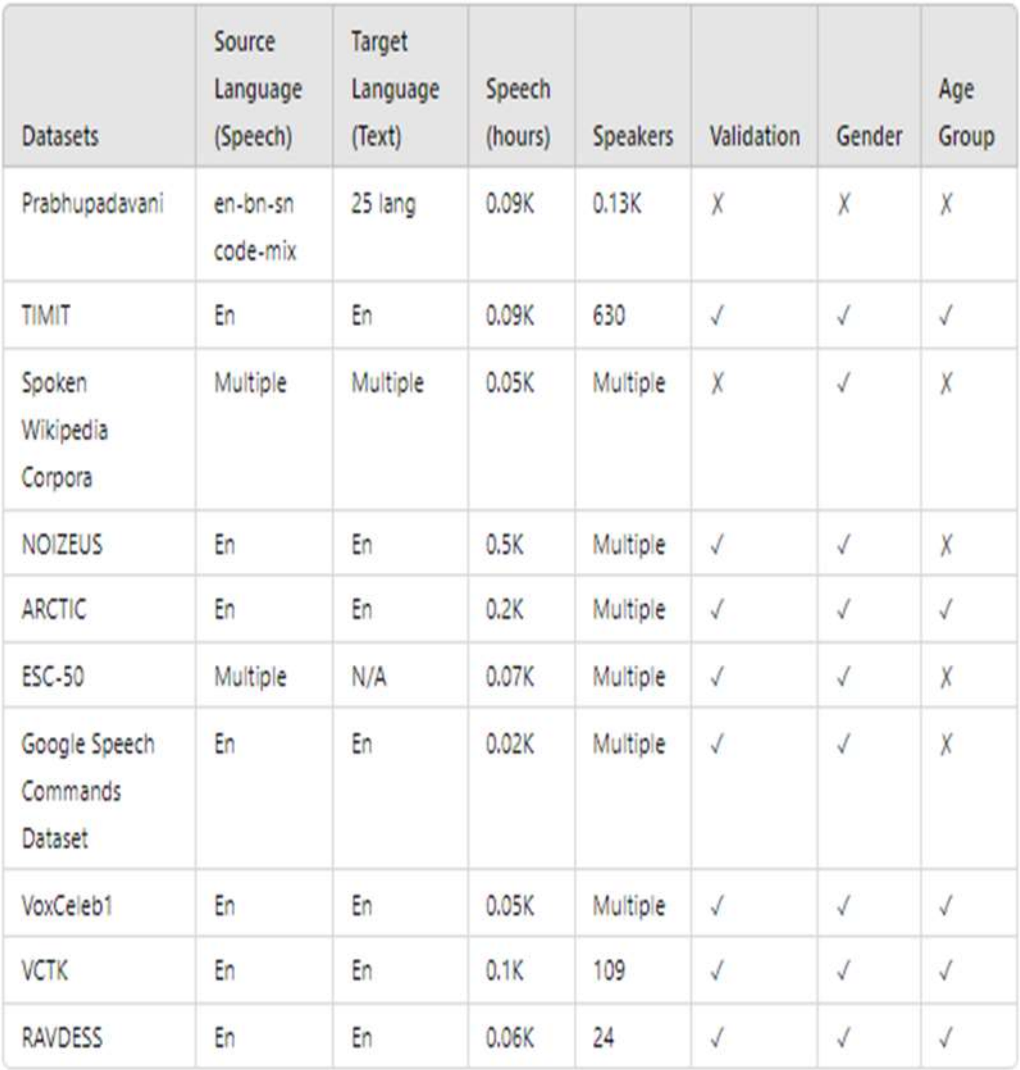
Our review aims to comprehensively analyze ST models and datasets, critically evaluating current research and identifying emerging challenges and directions. This review complements existing literature by focusing on dataset diversity and comparing cascade and E2E model performances. Ultimately, we aim to highlight the strengths and limitations of current methodologies while proposing future research avenues in ST translation

lly balanced English speech from 630 speakers, is pivotal in ASR and speaker recognition. The Spoken Wikipedia Corpora provides 30 minutes of multilingual speech from Wikipedia articles, offering diverse language samples. NOIZEUS offers 30 minutes of noisy English speech for robust ASR evaluation. ARCTIC features 12 minutes of speech for speech synthesis research.

ESC-50 includes 42 minutes of environmental audio recordings for sound classification tasks. The Google Speech Commands Dataset provides 1.2 minutes of short English utterances for keyword spotting. VoxCeleb1 offers 3 minutes of celebrity speech for speaker verification. VCTK provides 6

minutes of English speech from 109 speakers, supporting ASR and speech synthesis. RAVDESS comprises 3.6 minutes of emotional English speech from 24 speakers, used for emotion recognition and ASR.

Our review aims to comprehensively analyze ST models and datasets, critically evaluating current research and identifying emerging challenges and directions. This review complements existing literature by focusing on dataset diversity and comparing cascade and E2E model performances. Ultimately, we aim to highlight the strengths and limitations of current methodologies while proposing future research avenues in ST translation.



Datasets	Source Language (Speech)	Target Language (Text)	Speech (hours)	Speakers	Validation	Gender	Age Group
Prabhupadavani	en-bn-sn code-mix	25 lang	0.09K	0.13K	X	X	X
TIMIT	En	En	0.09K	630	✓	✓	✓
Spoken Wikipedia Corpora	Multiple	Multiple	0.05K	Multiple	X	✓	X
NOIZEUS	En	En	0.5K	Multiple	✓	✓	X
ARCTIC	En	En	0.2K	Multiple	✓	✓	✓
ESC-50	Multiple	N/A	0.07K	Multiple	✓	✓	X
Google Speech Commands Dataset	En	En	0.02K	Multiple	✓	✓	X
VoxCeleb1	En	En	0.05K	Multiple	✓	✓	✓
VCTK	En	En	0.1K	109	✓	✓	✓
RAVDESS	En	En	0.06K	24	✓	✓	✓

2. LLM versions for ST

Large language models (LLMs) have revolutionized research by enabling advanced tasks in natural language processing (NLP) and beyond. Google's Gemini family of LLMs offers a spectrum of capabilities, catering to diverse research needs. This paper explores four prominent models within the Gemini family: Gemini 1.0 Pro, Ultra, and the recently released Gemini 1.5 Pro and Flash. We delve into their strengths, suitable applications in research, and the unique functionalities they bring to the research landscape..

Gemini 1.0 Pro

- Focus: General-purpose performance across a wide range of tasks.
- Strengths:
- Capable of handling various tasks such as complex web searches, code analysis, creative writing, and data analysis.
- Well-suited for research projects requiring strong performance across different NLP applications.
- Offers a balance between power and efficiency, making it accessible for a broader range of research environments compared to Ultra.

Gemini 1.0 Ultra

- Focus: Unparalleled processing power for complex tasks.
- Strengths:
- Most powerful variant in Gemini 1.0, excelling in computationally intensive tasks and multimodal understanding (text, images, audio, video).
- Ideal for research requiring deep learning and analysis of complex data sets.
- Capable of handling large and intricate models, making it valuable for research on the forefront of AI development.

Gemini 1.5 Pro

- Focus: Balance between quality, performance, and cost.
- Strengths:
- Improved performance across various tasks like translation, coding, reasoning, and more compared to Gemini 1.0 Pro.
- Well-suited for a broad range of research applications due to its versatility.
- Long context window: Processes information from a vast amount of text (up to 2 million tokens by waitlist access), enabling comprehensive analysis of sequential data.

Gemini 1.5 Flash

- Focus: Speed and efficiency for high-frequency tasks.
- Strengths:
- Designed for rapid response times, making it ideal for real-time applications and chatbots.
- Lighter-weight model compared to Pro, leading to lower computational cost.
- Maintains a long context window of 1 million tokens, enabling efficient processing of large amounts of text data.

3. Future Directions for Research

This section highlights challenges that need the attention of researchers work- ing on ST problems.

3.1. Cascade vs End-to-End Models: The Prabhupadavani dataset, comprising 54 minutes of code-mixed speech across 25 languages, could be used to compare cascade and end-to-end (E2E) models for low-resource language scenarios. Given its multilingual nature and code-mixed content, evaluating both model types on this dataset can provide insights into handling diverse linguistic inputs efficiently.

3.2. Domain-Invariant Models: Datasets like TIMIT (54 minutes of English speech) and ARCTIC (12 minutes of English speech for speech synthesis) can be utilized to explore domain adaptation challenges. These datasets offer controlled environments for training models in specific domains (e.g., phonetic balance or speech synthesis) and testing their adaptability to other domains.

3.3. Discrepancy between Automatic and Human Evaluation: NOIZEUS, which includes 30 minutes of noisy English speech, is suitable for assessing how well ST models handle ambient noise and background interference. Evaluating model outputs against human perception of speech quality and semantic coherence can highlight discrepancies and guide metric development.

3.4. Handling Ambient Noise: NOIZEUS and ESC-50 (42 minutes of environmental audio recordings) are ideal for studying the impact of ambient noise on ST model performance. These datasets simulate real-world scenarios where distinguishing between speech and nonverbal sounds is crucial.

3.5. Handling Multiple Speakers: VoxCeleb1 (3 minutes of celebrity speech) and RAVDESS (3.6 minutes of emotional speech from multiple speakers) can be used to explore speaker variability challenges. These datasets feature multiple speakers with varying accents, emotions, and speaking styles, posing challenges for speaker adaptation and speech separation techniques.

3.6. Handling Speaker Diarization: Currently, the datasets provided do not explicitly include speaker diarization annotations. However, enhancing them with speaker boundary marks, especially in multilingual settings like Prabhupadavani, could facilitate testing the robustness of ST models in identifying and separating speakers.

3.7. Multilingual and Simultaneous ST: Prabhupadavani, with its code-mixed content across 25 languages, is suitable for evaluating multilingual ST models. Additionally, exploring simultaneous multilingual ST scenarios, such as those discussed in the context of conferences or diverse audience settings, can leverage datasets with varied language representations.

3.8. Low-resource ST Datasets and Models: While many datasets listed focus on high-resource languages, addressing low-resource language challenges requires datasets like Prabhupadavani to develop transfer learning models. Initiatives focusing on transferring knowledge from high-resource pairs (e.g., TIMIT) to low-resource settings (e.g., Mboshi-French) can advance ST research in underserved linguistic communities.

4. Conclusion

This survey paper delves into the most recent advancements in STT translations. Our discussion includes the models, evaluation metrics, and datasets used to train the ST models. We review various LLM versions for ST models and highlight previous research in this field. ST models are categorized based on the type of data they handle and the models employed. In addition, we discuss potential future directions for improving speech-to-text translations. Our findings suggest that the gap between the cascade and E2E system performance in both online and offline settings is narrow. However, for some language pairs, the gap remains wide; therefore, additional work is warranted. Our goal in the present ST survey was to offer valuable insight into this topic and drive advancements in ST research.

5. Acknowledgement

This paper provides an in-depth analysis of the latest advancements in speech-to-text (STT) translations. Specifically, we examine the models, evaluation metrics, and datasets utilized in training the ST models. We also review various large language models (LLMs) employed in ST models and discuss previous research conducted in this field. Based on the type of data they handle and the models they utilize, ST models are categorized. Furthermore, we explore potential future directions for enhancing speech-to-text translations. Our findings indicate that the performance gap between cascade and end-to-end (E2E) systems in both online and offline settings has narrowed, but there are still some language pairs for which the gap remains wide. Our primary objective in conducting this ST survey is to offer valuable insights into this topic and contribute to advancements in STT research

6.References

Abbott, L.F., 1999. Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Research Bulletin 50, 303–304

Alastruey, B., Ferrando, J., Gállego, G.I., Costa-jussà, M.R., 2022. On the locality of attention in direct speech translation. arXiv preprint

Kahn, J., Lee, A., Hannun, A., Kiros, J., 2020. "Self-Training for End-to-End Speech Recognition." This paper discusses the benefits of pre-training in speech recognition, particularly for low-resource scenarios

Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Banerjee, S., Lavie, A., 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., Goldwater, S., 2018. Pre- training on high-resource speech recognition improves low-resource speech- t