Zero-Shot Speech-to-text Translation model for low resource language like Dogri
Supervised By: Prof. Mrinmoy Bhattacharjee

Motivation

A Zero-Shot Speech-to-Text Translation (S2TT) model for Dogri is essential because the language spoken by over 2.3 million people lacks proper speech recognition tools. Traditional S2TT models require large labeled datasets, which Dogri does not have, making it difficult to build an efficient system. This model can help Dogri speakers use voice-based technology like Alexa and Google Assistant, bridging the gap in AI accessibility. Additionally, this initiative supports India's Bhashini and AI4Bharat projects, which aim to make AI more inclusive for Indian languages. In real-world applications, a zero-shot S2TT model can transcribe Dogri stories, enable voice search, and assist elderly or non-literate users, ensuring that Dogri remains relevant in the digital age.

Literature Survey

1. Speech-to-Text Translation for Spanish-English

- **Problem**: Traditional speech-to-text translation (S2TT) systems rely on Automatic Speech Recognition (ASR) followed by Machine Translation (MT), but this approach suffers from compounding errors. ASR errors propagate into translation, reducing accuracy.
- **Solution**: The paper uses ASR lattices instead of single-best ASR outputs to improve translation quality.
- **Dataset:** Fisher and Callhome Spanish-English Corpus.
- **Impact**: Demonstrates that leveraging multiple ASR hypotheses (lattices) can improve speech-to-text translation performance, especially for low-resource languages.

2. Speech-to-Text Translation Without ASR

- **Problem**: ASR requires large transcribed datasets, making it impractical for many low-resource languages. Traditional methods fail when there is insufficient labeled speech data.
- **Solution**: Introduces Unsupervised Term Discovery (UTD), a method that detects repeated sound patterns and aligns them with translations, bypassing the need for an ASR step.
- **Dataset:** CALLHOME Spanish-English Corpus.
- **Impact**: Proves that speech-to-text translation is possible without ASR, providing a method to handle languages with little to no transcribed data.

3. Simultaneous Speech-to-Text Translation (SimulSpeech)

- **Problem**: Real-time speech-to-text translation faces a trade-off between accuracy and delay. Traditional systems process ASR first, then MT, increasing processing time.
- **Solution**: Introduces SimulSpeech, an end-to-end model that directly translates speech into text without needing ASR, reducing errors and improving real-time performance.
- **Dataset:** MuST-C (English-Spanish, English-German).
- **Impact**: Shows that a single model can handle speech-to-text translation without explicit transcription, reducing delay and error propagation.

4. Leveraging Large Pre-Trained Multilingual Models for S2TT

- **Problem**: Training ASR and MT separately is computationally expensive and requires large labeled datasets.
- **Solution**: Uses large pre-trained speech models (e.g., Whisper, wav2vec2, mHuBERT-147) combined with a Neural Machine Translation (NMT) model, eliminating the need for language-specific training.
- **Dataset:** CoVoST2, Fleurs datasets.
- **Impact**: Demonstrates that pre-trained multilingual models can perform high-quality speech-to-text translation, even for languages with limited data.

Research Paper	Key Outcome
Speech-to-Text Translation for Spanish-English	ASR lattices improve translation accuracy over single-best ASR outputs, reducing ASR errors in translation.
Speech-to-Text Translation Without ASR	Speech translation can work without ASR using Unsupervised Term Discovery (UTD), useful for low-resource languages.
Simultaneous Speech-to-Text Translation (SimulSpeech)	Real-time translation is possible without ASR, reducing delay and error propagation.
Leveraging Large Pre-Trained Models for S2TT	Pre-trained multilingual speech models enable zero-shot translation without requiring labeled speech data.

Goal

Our goal is to develop a zero-shot speech-to-text translation model for Dogri without relying on large transcribed datasets. Since Dogri lacks sufficient labeled speech data, we will bypass ASR dependency using unsupervised term discovery, which identifies repeating speech patterns and aligns them with text.

We will use pre-trained multilingual speech models like Whisper, wav2vec2, and mHuBERT-147 to extract speech features without extensive training. By incorporating lattice-based translation techniques, we aim to reduce errors and enable direct speech-to-text translation, avoiding traditional pipelines.

This approach ensures a cost-effective and scalable solution, making voice-based AI, automated transcription, and real-time translation accessible for Dogri speakers. It also supports initiatives like Bhashini and AI4Bharat, promoting AI-driven solutions for low-resource languages.

Strategies Towards Goals with Justifications

Dogri is a low-resource language with very little transcribed speech data. Based on insights from the four papers, we can build a Zero-Shot STT model for Dogri using the following approach:

1. Bypass ASR Using UTD & Pre-Trained Models

- Inspired by Paper 2, we avoid explicit ASR training by using Unsupervised Term Discovery (UTD) to detect recurring speech patterns in Dogri.
- Combine this with Paper 4's approach of using pre-trained speech models (Whisper, wav2vec2, mHuBERT-147) to extract meaningful speech embeddings.

2. Direct Speech-to-Translation Using Hybrid Models

- Inspired by SimulSpeech (Paper 3), we train a hybrid end-to-end model that translates Dogri speech into text in another language (e.g., English) without relying on ASR transcriptions.
- Instead of converting speech to text and then translating, we train the model to learn speech-to-text mappings directly.

3. ASR Lattices for Error Reduction

- Inspired by Paper 1, we can use ASR lattices (multiple hypotheses instead of single-best outputs) from pre-trained models to improve Dogri speech transcription before translation.
- This will reduce recognition errors in Dogri speech and lead to more accurate translations.

4. Leverage Large Pre-Trained Multilingual Models

- Inspired by Paper 4, we integrate Whisper, wav2vec2, and mHuBERT-147 with a multilingual NMT model to achieve zero-shot STT for Dogri.

Baseline Model

1. ASR and MT:

We use a two-stage pipeline combining Automatic Speech Recognition (ASR) and Machine Translation (MT). ASR (Whisper Large-v2) which converts speech into text with high accuracy, handling diverse accents and noisy environments. MT (M2M100 by Meta AI) that Translates the transcribed text into target languages like Hindi, French, and Spanish. To demonstrate this approach, Google Colab notebook (BaselineModel1) that implements and evaluates the ASR and MT pipeline, showcasing its effectiveness in multilingual speech processing, could be accessed using the link below.
Model link:

- ASR Model: openai-wispher
- MT Model : m2m100_1.2B

Working Colab Notebook Link : ASR_MT

2. End-to-End with SeamlessM4T v2:

For a more unified solution, we also use Meta's SeamlessM4T v2 model. This single model performs direct speech-to-text translation in one step, simplifying the workflow compared to separate ASR and MT systems. A Google Colab notebook is included to demonstrate our setup, code, and inference results for this end-to-end baseline.
Model link : seamless-m4t-v2
Working colab Notebook link : seamless_colab

References

Jansen, A., & Van Durme, B. (2013). *Speech-to-text translation without ASR.* In Proceedings of the International Workshop on Spoken Language Translation (IWSLT). Retrieved from https://aclanthology.org/2013.iwslt-papers.14/

Bérard, A., Besacier, L., Servan, C., & Linh, V. (2017). *End-to-end automatic speech translation without pre-trained models.* arXiv preprint arXiv:1702.03856. Retrieved from https://arxiv.org/abs/1702.03856

Ma, M., Huang, G., Lal, P., Zhang, J., Cross, J., & Xiong, Y. (2020). *SimulSpeech: End-to-end simultaneous speech-to-text translation.* In Proceedings of the 58th Annual

Meeting of the Association for Computational Linguistics (ACL). Retrieved from https://aclanthology.org/2020.acl-main.350/

Liu, Y., Wang, X., Zhang, Z., Chen, J., & Liu, Q. (2025). *Leveraging large pre-trained multilingual models for speech-to-text translation in industry scenarios.* In Proceedings of the 30th International Conference on Computational Linguistics (COLING). Retrieved from https://aclanthology.org/2025.coling-main.509/

Team

2022ucs0102
(Paper 1, Paper 2 and Baseline model 1)

2022ucs0107
(Paper 3 Paper 4 and Baseline model 2)