Speech to Text Translation: Bridging The Gap Between Spoken and Written Communication

Nishant Kumar, Advisor: Prof. Rathna GN

Mid-Term Poster Presentation

Department of Computer Science and Automation, IISc, Bengaluru

Objective

Develop an end-to-end Speech-to-Text (ST) model with an interactive attention mechanism, aiming for faster, smaller models, and reduced errors. The objective is real-time processing, simultaneously improving both ASR and translation tasks. Results demonstrate superior performance over baseline models..

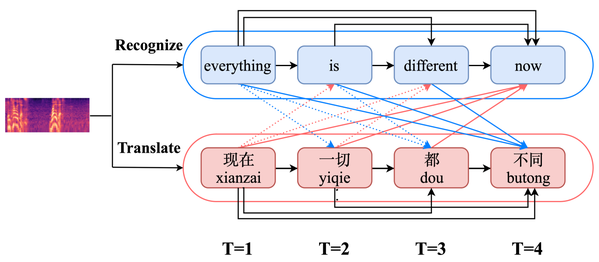

Figure 1: Voice Recognition and Translation

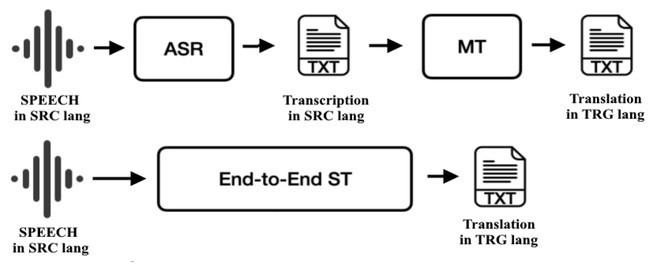

Figure 2: Workflow of Translation system

Motivation

- Communication Bridge: Traditional techniques have difficulties with speed, size, and accuracy, but ST plays a vital role in bringing people who speak various languages together.
- Innovation Drive: The project's motivation stems from the search for creative solutions with the goal of addressing shortcomings in traditional ST techniques.
- Enhanced Efficiency: In order to concurrently increase speed, model size, and mistake rates in both ASR and ST tasks, the emphasis is on utilising an interactive attention mechanism.

Related Work

Voice-to-Text Translation Models:

- Traditionally, voice-to-text translation involves separate models for Automatic Speech Recognition (ASR) and neural machine translation.
- Recent research explores integrating these phases into a single End-to-End model for more seamless processing.

Pretrained ST model:

- The model initiates its encoder by undergoing an initial training using Automatic Speech Recognition (ASR) data.
- The ST model undergoes a fine-tuning process using dedicated speech translation data to enhance its performance specifically for the speech translation task.

Two-Stage Models:

- These models have two parts: one for transcription and another for translation.
- Slower due to waiting for the entire transcription, but transcription aids translation.

Datasets used

1 : TED Talk Dataset

- TED Talk Multilingual Speech Translation Dataset comprises speeches from TED conferences, covering diverse topics.
- Focuses on language pairs such as English-German (En-De), English-French, English-Chinese, and English-Japanese.
- Includes 235,000 triplets: speech recordings, handwritten transcriptions, and translations.
- Used for sentiment analysis, speech recognition, and language translation tasks.
- http://www.nlpr.ia.ac.cn/cip/dataset.html

2 : IWSLT Dataset

- Covering multiple languages, IWSLT datasets support translation tasks for various language pairs.
- Significant inclusion of speech recordings enhances suitability for evaluating spoken language systems.
- Including parallel corpora with source language speech or text and translations into one or more target languages.
- Used as benchmarks in the International Workshop on Spoken Language Translation, offering a standardized platform for system evaluation.
- https://www.iwslt.org

Approch

- Transformer model as the current state-of-the-art in both MT and ASR Task.
- Chooses the Transformer as the core structure for the proposed approach, emphasizing its adaptability to any encoder-decoder architectures.

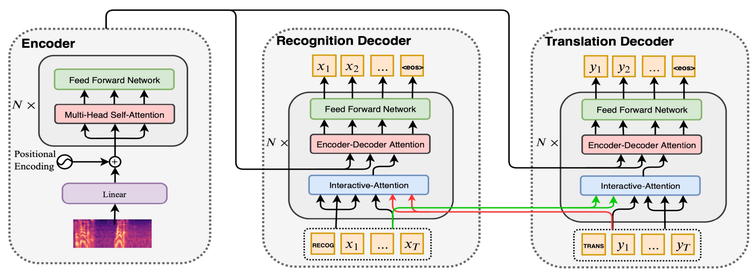

Figure 3: This is the architecture model of technique, which is transformer based. Speech characteristics are fed into the left portion, which is known as the encoder. The voice translation model and the speech recognition model share this encoder. The translation decoder is on the right, while the recognition decoder is in the middle. The two decoders are connected by an interactive attention sub-layer, which enables them to exchange and utilise information

- Two decoder interactively learn from each other simultaneously. by proposing an interactive learning model learn from each other simultaneously.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QKT}{\sqrt{dk}})*V$$

- Where Q, K, and V denotes the Query, Key and Value respectively. Dk is dimension of the key.

Interactive Attention:

- Substituting the self-attention sub-layer in the standard Transformer decoder with an interactive attention sub-layer consists of a self-attention sub-layer and a cross-attention sub-layer.

H_{self} = Attention(Q1,K1,V1), H_{cross} = Attention(Q1,K2,V2)

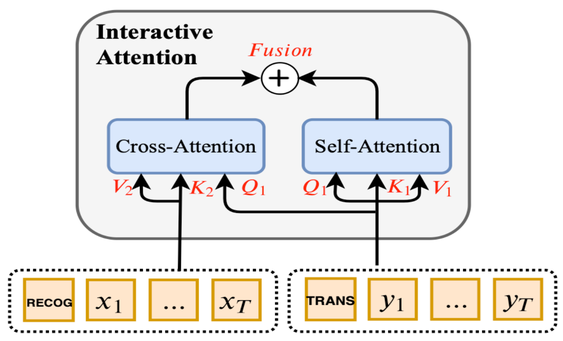

Figure 4: The interactive attention sub-layer consists of a self-attention sub-layer with a cross-attention sub-layer which can capture the information from the other task.

The output of self-attention sub-layer and cross- attention sub-layer can be integrated by a fusion function:

$$H_{final} = \text{Fusion}(H_{self}, H_{cross})$$

- Use a linear Interpolation as fusion function, calculated as:

$$H_{final} = H_{self} + \lambda * H_{cross}$$

- Where λ is a hyper-parameter to control how much information of other task should be taken into consideration.Type equation here.

Contributions

- Innovatively introduces an interactive attention mechanism, allowing ASR and ST tasks to perform synchronously and interactively in a single model.
- Allows predictions to be made in one task based on projected outputs from the other task as well as past outputs, promoting information sharing between the two tasks.

Results

Quantitative Results

- Evaluating a machine translation model for English to Japanese using TED Dataset.
- Highlighting the performance of Interactive model compared to base models.

Models	En-Ja	
	WER(↓)	BLEU(↑)
Pipeline	14.21	20.87
E2E	14.21	16.59
Multi-task	14.01	18.73
Two-stage	14.12	19.32
Interactive	13.91	19.60

Table 1: Results on WER and BELU metrics.

- WER: (low), showing accurate speech recognition.
- BLEU Score: (high), indicating strong translation quality.
- Interactive model outperforms base models, suggesting significant improvements in English to Japanese machine translation.

Conclusions

The suggested paradigm employs a transformer base model, conducting speech recognition and translation in two stages within a single model. This approach enhances overall performance by enabling mutual benefits between the recognition and translation processes, as confirmed by experiments across various language pairings.

Future work

- To develop a speech-to-text model using acoustic features and a self-attention architecture, eliminating the need for speech transcripts also try this model of different data set to translate into English Text.
- This project aligns with a larger initiative for speech-to-Indian sign language translation, involving two stages: converting diverse language speeches to English text and then translating the English text to Indian signs. The goal is to facilitate communication for individuals with hearing difficulties who rely on sign language.

References

[1] Weiss, R. J.; Chorowski, J.; Jaitly, N.; Wu, Y.; and Chen, Z. 2017. Sequence-to-sequence models can directly translate foreign speech. In Proceedings of Interspeech 2625–2629.

[2] Berard, A.; Besacier, L.; Kocabiyikoglu, A. C.; and Pietquin, O. 2018. End-to-end automatic speech translation of audiobooks. In Proceedings of ICASSP 6224–6228.

[3] Sperber, M.; Neubig, G.; Niehues, J.; and Waibel, A. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. Transac- tions of ACL 7:313– 325.

[4] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of ICLR.

[5] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkor- eit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In Proceedings of NIPS 5998–6008.