Speech-to-Text Recognition: Bridging the Gap between Spoken and Written Communication

Nishant Kumar
Sr. No.- 21495

Mid-Term MTech Project Report

Abstract

Speech Data to Text Data Translation (ST) is the translation of spoken language into written text and has gained significant attention recently. End-to-end ST is a novel methodology that has the potential to be faster, employ smaller models, and produce less errors than the conventional method. However, developing this sort of model without intermediate transcriptions is challenging.

Most prior efforts have focused on enhancing translation quality through multi-task learning to train the End-to-end ST model in conjunction with ASR (Automatic Speech Recognition). Still, the degree to which these jobs may improve upon one another is limited. Another approach gradually combines two different models, slowing down training and the system's ability to comprehend and react.

This project is a new way to improve ASRs' and STs' work together. Introduce an interactive attention mechanism that lets both tasks happen at the same time in one model. This means the system relies on something other than what it predicted before but also on what it predicts for the other task. I tested our approach on TED speech translation data, and the results show that my model performs better than solid baseline models in both speech translation quality and speech recognition accuracy[1].

1 INTRODUCTION

Communication between speakers of various languages is facilitated by speech translation, which is the process of converting spoken words or sentences in one language into written text in another language. There are two models for doing this type of translation: one for machine translation (MT) and another for automatic speech recognition (ASR), which recognises spoken words. However, this method can be sluggish and prone to mistakes [2].

On the other hand, there's a newer method called end-to-end ST that combines these tasks into a single model. This could reduce delays, avoid repeating unnecessary steps, and prevent mistakes from spreading. Recent studies on end-to-end ST have been progressing quickly and are showing promising results[3].

While End-to-end Speech to text (ST) models offer advantages, their performance could be more robust, and they are easier to develop by employing transcripts as intermediate stages. Prior research attempted to enhance them through pretraining or multitasking[3]. They may utilise a speech-understanding pre-trained component or combine speech recognition and translation training. They can only learn so much from each other in the portions they exchange between duties. Some research proposes a two-step paradigm to address this, in which the first stage establishes a hidden state and recognises speech. Then, utilising that concealed state, the second portion translates. This facilitates better translation, but training and comprehension are slowed since the second half must wait until the speech is recognised. It should be noted that this model can only employ translation data from the recognition phase, not the other way around[4].

The procedures of developing text translation (ST) and speech-to-text (ASR) may really help each other. How to do it is as follows: (1) creating translated text from voice gets more straightforward when you have more information from transcribed words, not just only the speech itself, and (2) these converted/translated words may also assist in the process of recognising the spoken words better. Imagine that the ASR and ST tasks co-

operate while you speak a whole English sentence.

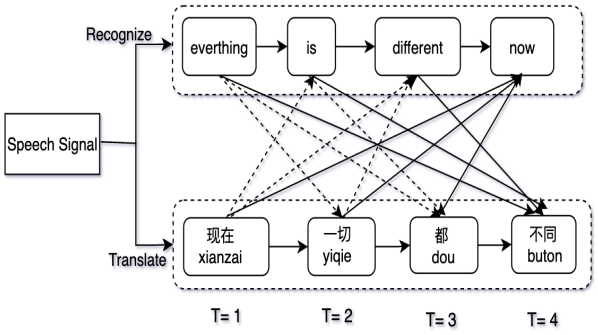

Figure 1: An example of voice recognition in English and its translation into Chinese is shown, with the outcomes of both tasks being compatible[1].

In the example, the phrase "everything" from step T = 1 in the transcription job gives more context for translating the Chinese term "yiqie" (meaning everything) in the ST task at step T = 2. In the same way, for the ASR job, "now" at step T = 4 may be recognised with the help of the translated phrase "xianzai" at step T = 1. Hence, transcription and translation quality can be raised if the processes of the two activities cooperate.

This is a novel approach to teaching a computer to interpret and comprehend spoken language. In addition to helping each other learn more effectively, my model can recognise and interpret speech simultaneously. Tasks in conventional models are independent, but in our method, they exchange information. i am employ a novel technique known as interactive attention, in which the translation component aids with word recognition prediction and vice versa. This allows both activities to learn from one another continuously[4]. Using fresh data on English voice translations into German, French, Chinese, and Japanese, I wish to test our theory. Conducted several tests to ensure that our method is effective. We want to improve our technique further in the future by doing additional research.
The main things accomplished were as follows:

- Developed a learning model that does both voice recognition and translation concurrently, improving upon each job.
- Technique can do simultaneous transcription and translation in a same model, unlike normal models that perform these tasks separately.
- Model outperformed alternative approaches, such as the standard procedure, a pre-trained model, a regular multi-task learning model, and a two-step model, according to our testing on four language pairings.

2 Related Work

Speech Translation: Voice-to-text translation uses different models for voice recognition (ASR) and neural machine translation[5]. However, integrating these phases into a single End-to-end model has been investigated in recent research. Though the concept of End-to-end speech translation was initially proposed in 1999, an utterly end-to-end model without source transcriptions was not created until 2016. Nevertheless, developing an End-to-end speech translation model is complex, and the results could also be more robust. Some used a technique known as multi-task learning to train both voice translation and speech recognition (ASR)[6]. Some people pretrained the speech translation model using target phrases to improve the language model or using more ASR data to improve the acoustic model. Some even went so far as to educate the voice translation model on how to perform a process known as knowledge distillation using a text translation model[6].

One theory is that interpreting speech could be more straightforward if the model is aware of the speech's textual transcription. Thus, two-stage models were recommended by several researchers. These models have a first portion that identifies the transcriptions and a second part that uses the first part's information to translate. It is slower because the second part must wait for the entire transcription to be recognised. Furthermore, voice recognition is not aided by it; transcriptions are the sole tool to enhance translation. However, as Figure 1 illustrates, voice recognition and translation outcomes can support one another. Therefore, it makes sense to figure out how to improve both activities by using what each other has to teach people.

Synchronous Inference: Refers to the simultaneous processing or production of information in a computer model. In this case, it frequently entails bidirectional or parallel inference, in which many model orientations or aspects cooperate simultaneously. By enabling the model to take into account input from several sources or directions concurrently, the objective is to improve efficiency and performance[6]. A build model that can concurrently process and produce data from left to right and vice versa. In this manner, the model may gain from the data on both

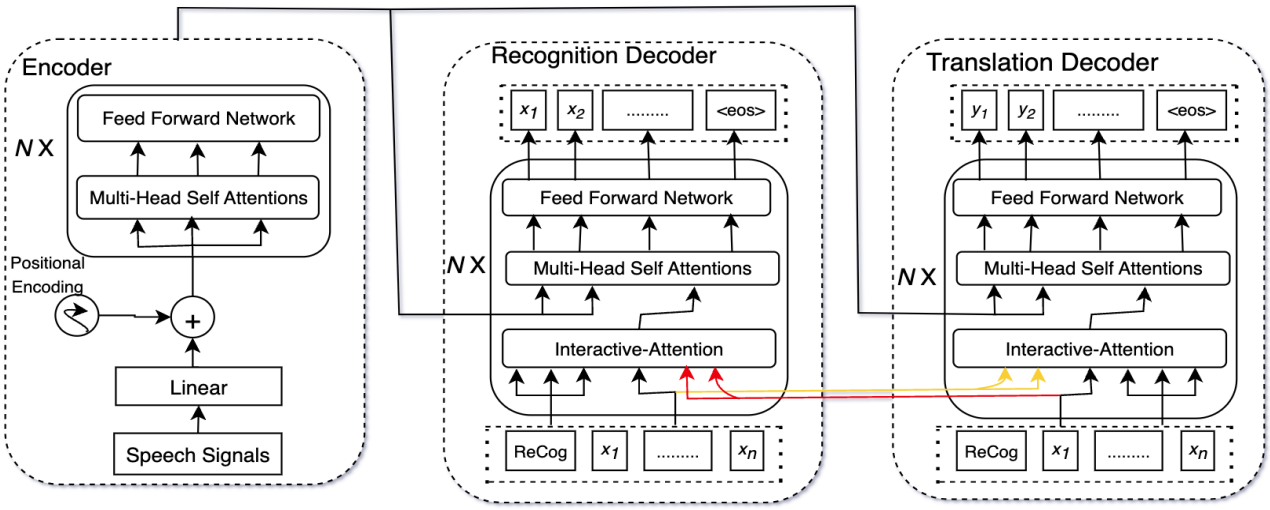

Figure 2: This is the architecture model of my technique, which is transformer based. Speech characteristics are fed into the left portion, which is known as the encoder. The voice translation model and the speech recognition model share this encoder. The translation decoder is on the right, while the 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[1].

sides and raise the output's quality. When the method was first applied to translation jobs, it significantly improved such tasks. Afterwards, with successful outcomes, this strategy was also used for other jobs, including summarising[7]. While earlier research concentrated on jobs that provide outputs in several directions, our work is more directly associated with those that simultaneously translate text in numerous languages. In this instance, we are focusing on two tasks: voice translation and recognition, and our goal is to ensure that they function well together in a single model.

3 Background

I selected the Transformer model due to its current state-of-the-art performance in both automatic voice recognition and machine translation. Not just the Transformer but any similar model with an encoder and a decoder may be used with our new technique[8].

An encoder and a decoder are part of the conventional framework of the Transformer model. An input sequence, such as the words in a phrase, is fed into the encoder, which converts it into a series of continuous representations. Next, the decoder creates an output sequence using these representations, one word at a time.

In the Transformer, both the Encoder and Decoder have multiple layers with N numbers of FF-Network and Mutli-Head Self Attentions similarly in the encoder part of the transformer. Each layer in the encoder has two parts: one that pays attention to different parts of the input sequence (self-attention) and another that processes this information (feed-forward). The decoder layers have three parts:

- Masked self-attention (to prevent looking at future positions during training).
- Attention to the encoder's output.
- Another feed-forward step.

Each part has a connection to the input called residual connection and is normalized.

The three attention sub-layers computation procedure may be expressed using the following formula[8]:

$$Attention(X, Y, Z) = Softmax(\frac{XY^T}{\sqrt{d_k}})Z \quad (1)$$

To put it another way, query, key, and value are represented by the symbols Q, K, and V in the Transformer model. The dimension of the key is represented as d_k. The output for a layer is then created by performing a feed-forward operation. Ultimately, the final output is obtained, and predictions are made using a softmax function[8].

4 Approch

In this section, I introduce a novel method for allowing voice recognition and speech translation models to learn from one another during training and output generation. Figure 2 shows the framework that we recommend. However, before entering this new strategy, let's first discuss how we use the Transformer model for voice recognition, machine translation, and speech translation jobs.

The Transformer model may handle speech recognition(SR), text machine translation(MT), and speech translation(ST) tasks. On the other hand, the input and output sequences (I and O) for every job vary.

4.1 ASR Task:

For the ASR task, we start with a sequence of speech features (denoted as I or S), where each feature represents a small unit of the speech signal. These features are obtained by converting the raw speech signal using a method called log-Mel filterbanks. Additionally, we perform mean and variance normalization. To manage the input size, we use a technique called frame stacking and downsampling, which reduces the length of the input sequence. The resulting sequence has a specific dimension.

The output sequence (denoted as O or X) is the corresponding transcription, representing the source sentence.

4.2 MT Task:

In machine translation (MT), the input sequence (Text Inputs(I)) is the source language transcription, represented as X with individual elements x_1 to x_N. The output sequence (Text Output(O)) is the translated text of the target language, represented as Y with elements y_1 to y_M, where M is the total length of the sentence that is translated by the model.

4.3 ST Task:

In End-to-end Speech to text (ST) tasks, the input sequence (I) are the speech signals, represented as S with elements s_1 to s_T, just like in Automatic Speech Recognition (ASR) tasks. The output sequence (text(O)) is the translated text of the target language, represented as Y with elements y_1 to y_M.

4.4 Interactive Learning

Different tasks are trained separately but share certain parameters in the usual multi-task learning. The output from one task can help predict the output of another task, so it makes sense to improve both tasks by letting them exchange information. Additionally, traditional multi-task learning can only do one task at a time during inference, but there are situations where we need both transcription and translation simultaneously. To address these issues, we suggest an interactive learning model where two tasks not only learn from each other but also make predictions at the same time.

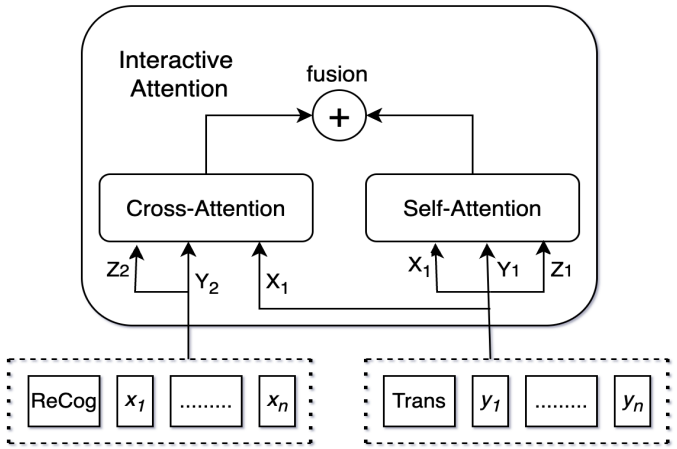

Figure 3: As seen in the illustration, the interactive attention sub-layer comprises two parts: the self-attention sub-layer and the cross-attention sub-layer. These elements work together to get data from the other task.

Figure 2 depicts the model's main structural elements. First, a series of acoustic characteristics are created by processing the voice input. Next, a linear layer is used to alter this sequence, changing its dimensions to match the hidden size known as dmodel. Subsequently, the encoder transforms this sequence into a more sophisticated audio representation. For distinct tasks, two decoders are used: one for speech translation and the other for voice recognition.

We change a part of the Transformer decoder to make the two decoders learn from each other. We replace the self-attention sub-layer with our new interactive attention sub-layer. In Figure 3, you can see that the interactive attention sub-layer includes a self-attention part and a cross-attention part[8].

The self-attention part uses the hidden representation from task 1 to learn a higher representation. The cross-attention part uses the confidential representation from task 1 as the query and the hidden representation from task 2 as the key and value to bring in

information from the other task. Both these parts are in the same layer. Mathematically, we can represent this as:

$$S_{\text{self}} = \text{Attention}(X_1, Y_1, Z_1) \quad (2)$$

$$S_{\text{cross}} = \text{Attention}(X_1, Y_2, Z_2) \quad (3)$$

Then, we combine the output of the self-attention part and the cross-attention part using a fusion function to get the final representation:

$$S_{\text{final}} = \text{Fusion}(S_{\text{self}}, S_{\text{cross}}) \quad (4)$$

We use a linear interpolation as the fusion function, which can be calculated as:

$$S_{\text{final}} = S_{\text{self}} + \lambda \times S_{\text{cross}} \quad (5)$$

Here, λ (lambda) is a parameter that controls how much information from the other task is considered. Both decoders then use this combined representation, which has information from both tasks[9].

5 Dataset

A collection of talks given at TED (Technology, Entertainment, Design) conferences is called the TED Talk dataset. These presentations touch on many subjects, such as science, technology, education, and more. The dataset is often utilised for various machine learning and natural language processing (NLP) applications, including sentiment analysis, speech recognition, and language translation. It offers a broad and varied supply of spoken language data for development and research needs.

The TED Talk dataset is a bunch of speeches from TED conferences, where people discuss topics like science, technology, and education. People use this dataset for computer tasks like understanding feelings, recognizing speech, and translating languages. It gives a wide range of spoken language data for research and technological improvement.

For the language pairings English-German (En-De), English-French, English-Chinese, and English-Japanese, we gathered 235,000 triplets. Speech recordings, handwritten transcriptions, and translations are included for every triplet. Following the IWSLT split, we separated the data into development (Dev) and test sets, with tst2014 as the development set and tst2015 as the test set. Training is done using the remaining data. Visit http://www.nlpr.ia.ac.cn/cip/dataset.html to view this dataset.

All Pre-trained and my own models in our study are built upon the Transformer architecture, explicitly using the configuration called "transformer base" introduced by Vaswani et al. (2017)[8]. This configuration involves, by default, six layers of encoders and six layers of decoders, each having hidden layers of size 512 dimensions. My model uses the Adam optimizer (Kingma and Ba, 2015) on NVIDIA RTx 3080 GPUs during training[10].

6 Baseline models

Used the following baseline models to evaluate our technique against:

1. **Pipeline System:** Initially, we train the machine translation (MT) and automatic speech recognition (ASR) models separately. Next, we feed the MT model with the results from the ASR model.

2. **Pretrained Speech Translation (ST) Model:** Using ASR data, we train the encoder of the end-to-end ST model to start from scratch. Afterwards, we use voice translation data to refine the model.

3. **Multi-task Learning Model:** By sharing encoder parameters, pre-train ASR and ST models simultaneously.

4. **Dual-phase Framework:** This approach is divided into two stages: transcriptions are produced in the first stage and translations in the second. We reused the Transformer-based fundamental model, adhering to this methodology. Training on ASR data initiates the first step.

7 Results

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
myModel	13.91	19.60

Table 1: The outcomes demonstrate how well English-Japanese (En-Ja) voice recognition and translation

perform on TED datasets. "E2E" stands for the End-to-end speech translation model that has already been trained, and "Interactive" is our freshly suggested interactive learning strategy.

In the provided table, the WER values are expressed as percentages (↓ indicates lower is better), while BLEU scores are presented as percentages (↑ means higher is better). These metrics help assess the accuracy and quality of machine translation models.

Table 1 shows the critical speech recognition and translation outcomes on English to Japanese TED datasets. The BLEU metric scores in the top row indicate translation results using the text MT model with pristine manual transcriptions as inputs, representing an upper limit for the speech translation task. We've set λ to 0.3 for the interactive learning model and k to 3.

8 Conclusions and Future Work

The suggested paradigm makes simultaneous and interactive speech recognition and translation possible here; I'm using the transformer base model architecture that works in 2 stages but in a single model, first speech recognition and second converting or translation of the given speech words text of different or target language. It improves overall performance by enabling the recognition and translation processes to profit from each other's generated outputs. Experiments conducted on a range of language pairings confirm the efficacy of our method.

In the future, I aim to develop a translation model from scratch using acoustic features of speech with a self-attention model architecture, directly converting speech into text without using speech transcripts.

The main focus of this project, which I want to use in a different project, the project is speech-to-Indian sign language translation. That project is divided into two stages. The first stage is converting the speech of other languages into English language text. The second stage of the project is using that converted language(English) or target language into Indian signs to communicate with unique disabled persons who have a hard time listening. Still, these people are good at understanding the signs.

References

[1] Liu, Y., Zhang, J., Xiong, H., Zhou, L., He, Z., Wu, H., Wang, H., & Zong, C. (2019). Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding. ArXiv. /abs/1912.07240.

[2] 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.

[3] 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.

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

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

[6] Zong, C.; Huang, T.; Xu, B.; and Xu, B. 1999. The technical analysis on automatic spoken language translation systems (in chinese). In Journal of Chinese Information Processing.

[7] Wang, Y.; Zhang, J.; Zhou, L.; Liu, Y.; and Zong, C. 2019. Synchronously generating two languages with inter- active decoding. In Proceedings of EMNLP 3341–3346.

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

[9] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of CVPR 770–778.

[10] Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In Proceedings of ICLR.