S2ST

RISHIT SHRIVASTAVA

May 2023

1 Introduction

This paper discusses the evolution of Speech-to-speech translation (S2ST) technology, which aims to translate speech from one language into another. Traditionally, S2ST involved concatenating three systems: automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). However, recent advancements in end-to-end speech-to-text translation (S2T) and text-to-speech translation (T2ST) have simplified the S2ST pipeline into two stages, improving efficiency and reducing error propagation issues. Additionally, researchers have built one-stage S2ST systems that jointly optimize intermediate text generation and target speech generation steps or remove the dependency on text completely.
The article also highlights the challenges of S2ST for unwritten languages, which remain unexplored due to the lack of training data. As an example of an unwritten language, Taiwanese Hokkien is studied for S2ST between English and Hokkien. The article uses a discrete unit-based S2ST approach to build En-Hokkien systems. The article extends HuBERT-based discrete unit extraction to support En→Hokkien translation and examines the feasibility of unit-to-waveform generation for tonal languages. The article leverages the unit-based speech normalization technique to remove non-linguistic variations in speech from multiple speakers and studies two S2ST model training strategies, speech-to-unit translation (S2UT) with a single decoder or a two-pass decoding process that leverages Mandarin as a written language similar to Hokkien to provide extra text supervision.

2 Models and Challenges

2.1 HuBERT

: HuBERT is a variant of the popular BERT (Bidirectional Encoder Representations from Transformers) model that was specifically designed for speech recognition tasks. HuBERT was trained on a large corpus of audio data, which makes it effective at processing spoken language. One challenge with using Hu-

BERT is that it requires a significant amount of computational resources to train and use effectively.

2.2 Wav2Vec

2.0: Wav2Vec 2.0 is a model developed by Facebook AI that is designed to learn representations directly from raw audio data, without requiring any transcription or linguistic information. This makes it useful for tasks like speech recognition and speaker identification. One challenge with using Wav2Vec 2.0 is that it can be computationally expensive to train and use effectively.

2.3 MelNet

: MelNet is a model developed by NVIDIA that is designed to generate high-quality audio samples. It uses a hierarchical structure to model the relationships between different levels of audio features, and has been shown to be effective at generating realistic speech samples. One challenge with using MelNet is that it can require significant amounts of computational resources to train and use effectively.

To address some of the challenges associated with these models, researchers could consider using techniques like transfer learning or model compression. Transfer learning involves using a pre-trained model and fine-tuning it on a specific task, which can reduce the amount of computational resources needed for training. Model compression involves reducing the size of the model by removing unnecessary parameters, which can also reduce the amount of computational resources needed for training and inference.

Another approach to addressing these challenges could be to use hardware accelerators like GPUs or TPUs, which can speed up the training and inference process significantly. Finally, researchers could also consider using distributed computing techniques, which involve spreading the computational workload across multiple machines, to reduce the amount of time and resources required for training and inference.

3 Architectures

The first architecture is a single-pass decoding process that directly translates source speech to the target. The second architecture relies on target text (Mandarin text in the case of Hokkien speech) to provide extra supervision and performs two-pass decoding. Both architectures predict discrete units as the target, and the speech encoder and text or unit decoders are pre-trained with unlabeled speech or text data.

To train the S2ST models, the authors created parallel S2ST training data from human annotations. They also leveraged speech data mining and created weakly supervised data through pseudo-labeling. Speech data mining involves

finding parallel speech data by searching for pairs of similar audio clips in different languages. Weakly supervised data is created by training an initial ASR model and then using it to transcribe and generate pseudo-labels for the speech data.

3.1 How do we generate Pseudo-labeled data?

To create weakly supervised data using pseudo-labeling, an initial ASR model is trained on a large amount of unlabeled speech data. Then, the model is used to transcribe speech data, generating pseudo-labels for the speech utterances. These pseudo-labels are then used to train an S2ST model in a supervised manner.

The process is as follows:

Train an initial ASR model on a large amount of unlabeled speech data. Use the trained ASR model to transcribe speech data, generating pseudo-labels for the speech utterances. Train an S2ST model using the pseudo-labeled speech data in a supervised manner. Fine-tune the S2ST model on a smaller set of labeled data to improve its accuracy. The pseudo-labels generated by the ASR model are not perfect and may contain errors, but they can provide a large amount of weak supervision to train an S2ST model, which can then be fine-tuned on a smaller set of labeled data to improve its accuracy.

3.2 Speech to Unit

"Speech-to-Unit" (S2UT) refers to the process of converting the speech signal into discrete units or subword units. The idea behind this approach is to break down the continuous speech signal into smaller units that can be more easily translated into text or speech, as opposed to directly translating the continuous speech signal.

In the S2UT approach, the speech signal is first encoded into a fixed-dimensional vector representation, which is then clustered using k-means to obtain discrete units. These discrete units can be thought of as building blocks of the language, and they can be combined to form words and sentences. By using S2UT, the model can capture the important acoustic and phonetic information from the speech signal, while also simplifying the translation task by breaking down the continuous speech signal into smaller units.

3.3 Single pass decoding S2UT

In the single-pass decoding S2UT approach, a single unit decoder is used to perform speech-to-unit translation. The unit decoder is trained using standard cross-entropy loss and initialized with a pre-trained decoder obtained from mBART training. The mBART training involves a sequence-to-sequence autoencoder trained with a denoising objective across monolingual text in multiple languages using discrete units extracted from unlabeled speech with consecu-

tive duplicate units removed. During decoding, a single-pass beam search is performed with the unit decoder.

3.4 Two-pass decoding S2UT: UnitY

The UnitY model, proposed by Inaguma et al. (2022), also performs speech-to-unit translation, but it differs from the S2UT model in that it includes a target text decoder and a target text encoder before the unit decoder. This allows the model to incorporate target text prediction as an auxiliary loss during training. The model is trained jointly, and when training the En→Hokkien model, Mandarin is used as the target text due to its proximity to Hokkien and abundance in text data. The model also applies R-Drop regularization during training and initializes the target text decoder with a text mBART model pre-trained on the combination of English and Mandarin monolingual text data.