Abstract

Speech translation is a vital technology that bridges language barriers, enabling people from different linguistic backgrounds to communicate effectively. However, developing speech translation systems for low-resource languages poses significant challenges. Low-resource languages are those with limited available data for training models, making it difficult to create accurate and reliable translation systems.

The primary objective of this research is to enhance the performance of a direct speech translation model tailored for low-resource languages. A direct speech translation model translates spoken language directly from the source language to the target language without relying on intermediate text representations. The model architecture consists of three main components: an encoder, a dimensionality reduction module, and a transformer decoder layer.

A key contribution of this work is the exploration and implementation of two novel pretraining techniques. These techniques are derived from recent advancements in speech representation learning and are designed to improve the quality of the encoder's understanding of the source language audio. The first technique focuses on creating semantically aligned, multimodal, cross-lingual speech representations that enhance the model's ability to understand and translate spoken language across different languages. The second pretraining technique employs self-distillation and online clustering to learn robust and meaningful speech representations without requiring extensive labeled data. In this study, both pretraining techniques are applied to the encoder using audio data from a low-resource language. Specifically, the audio data of the Tamasheq language, a Niger-Congo language spoken in parts of Mali, Algeria, and Niger, is used. The entire direct speech translation architecture is fine-tuned after pretraining the encoder with the chosen techniques.

The research utilizes data from the IWSLT2024 competition, specifically focusing on the low-resource speech translation task involving the Tamasheq-French language pair. The IWSLT (International Workshop on Spoken Language Translation) competition provides a standardized benchmark for evaluating speech translation systems, allowing for consistent and objective comparisons of different models and techniques.

To evaluate the effectiveness of the two pretraining techniques, BLEU score metric is used. The results of this research are expected to demonstrate that both pretraining techniques significantly improve the performance of the direct speech translation model for the low-resource Tamasheq-French language pair. Moreover, the research highlights the potential of applying these pretraining techniques to other low-resource language pairs, contributing to the larger goal of making speech translation technologies more accessible.