Pre-training on High-Resource Speech Recognition for Low Research Speech-to-Text Translation

Joost Bottenbley
George Mason University
email@gmu.edu

Jonathan Mbuya
George Mason University
jmbuya@gmu.edu

Jonathan Vasquez
George Mason University
jvasqu6@gmu.edu

1 Introduction

1.1 Task / Research Question Description

We aim to advance the field Speech Translation (ST) for Low Resource Languages (LRL). This task uses audio speech input from a source language and returns a corresponding translation in a target language (i.e., $X \rightarrow Y$). ST has applications in documenting languages, automatic closed captioning, and supporting business relationships across language boundaries (Bansal et al., 2018). Our problem domain assumes that an audio track of a source language is given and must be translated into target text. Previously proposed architectures typically consisted of a pipeline of an automatic speech recognition (ASR) module followed by a machine translation (MT) module. However, current state-of-the-art paradigms typically present end-to-end neural network models as the champion architecture. In these systems there are no intermediate representation, like audio transcripts of the source language, pasted between modules. In this research, we strive to improve upon the existing body of research in this field.

1.2 Related Research

ST requires several hundred hours of parallel trained data consisting of audio files with its corresponding translation texts. While these resources are prevalent and easy to obtain for HRL this kind of data is scarce for LRL (Anastasopoulos and Chiang, 2018). We aim to extend speech-to-text translation models in cases where we do not have enough samples in the corpus to train a model for a given language pair. There has been prior work trying to solve the same problems for different tasks in ST (Anastasopoulos and Chiang, 2018; Bansal et al., 2018; Conneau et al., 2019; Lauscher et al., 2020). The literature we surveyed uses sequence-to-sequence architectures with or without attention. We propose to approach this task using the transformer architecture, which has shown superior performance in various natural language processing tasks.

1.3 Proposed Approach

We present two approach for moving forward with project development. Each with be discussed with our advisor Dr. Antonios Anastasopoulos and selection of which approach to use will be made with his consultation by the end of week of Oct.14.2022.

1. Approach 1: Leverage a HRL pre-trained ASR acoustic encoder to use for initialization in development of the ST model.

2. Approach 2: Leverage large multilingual models, such as XLM-R or mBERT, and fine-tuning on the downstream task of ST where a LRL is target language. The use of custom transformers would be explored.

We plan on doing some experimentation within the framework to see which direction of development might yield the best results under the time allotted. These experiments include the heavy exploration of transformers.

1.4 Likely challenges and mitigation

ST is a complex task. We plan on using popular frameworks in our development process. We plan to be able to get the expected behavior from these frameworks. If we are unsuccessful, we will modify these frameworks to meet our needs. We may need to re-implement different components from scratch. We also plan to use the Hopper cluster from GMU. If the scope is to broad, then we may need to narrow the scope of our project to issues that can be solved in the time frame allotted.

2 Motivation & Limitations of Previous Research

(Bansal et al., 2018) use the parameters of pre-trained models in HRL to initialize encoders/decoders in LRL. Their results show outstanding results even though the data languages are different. Similarly, (Salesky et al., 2021) use pre-trained models in their experiments with similar results. Furthermore, (Anastasopoulos and Chiang, 2018) propose an ensemble architecture where two similar datasets aimed at different goals (one for translation and another for transcription) are merged to obtain better performances for another task. Finally, (Conneau et al., 2019) proposes an unsupervised framework to create cross-lingual speech recognition. However, although the *learning transfer* can provide practical solutions, certain limitations should be considered. In this regard, (Lauscher et al., 2020) proposes four key questions to acknowledge learning transfer and provide answers by performing several experiment setups and the corresponding analysis of the results. Having this, we want to explore different pre-training methods and the impact of pre-training on related and unrelated languages. The primary motivation is that languages with low resources (limited data) do not directly benefit from recent advances in speech transition due to limited access to a corpus resource used for training a model.

3 Experiments

3.1 Datasets

We plan to use the CoVost 2 dataset which is a multilingual corpus in the development process because its publicly available[1]. We also choose this dataset because it has been used in previous tasks by other researchers in giving baseline results (Wang et al., 2020). It also contains high and low-resource languages. Languages included in the data set are: Arabic, Catalan, Welsh, German, Estonian, Persian, Indonesian, Japanese, Swedish, and Chinese.

3.2 Baselines

We will implement the baselines from (Wang et al., 2020), where authors use the CoVoST 2 dataset. We are replicating results depicted in Table 2 (Table 3) for the monolingual (multilingual) translation. In our scenario we are considering 2 models: an end-to-end ST and pre-trained end-to-end ST. Our scenario includes English as either the target language (i.e., $X \rightarrow En$) or the source language (i.e., $En \rightarrow Y$). We pair English with France, Spanish, Catalan, and Persian languages. Note that we select the first two languages because they can be considered high-resource, and we select the other two because they are classified as low-resources.

We will use the fairseq framework (Salesky et al., 2021), built on top of Pytroch. Pytroch is one of the most popular frameworks for deep learning applications. On the other hand, Fairseq is one of the most popular frameworks, particularly for various NLP tasks, including speech translation.

3.3 Timeline

Table 1 illustrates our schedule. The first week aims to retrieve and prepare the large datasets for our experiments as well as decide the final framework. The next two weeks code development for the selected baseline from (Wang et al., 2020). Finally, the last week aims to run the experiment and analyze the results. These will be used to write the report required in Checkpoint 2.

Week	Task
10/03 – 10/14	Download and prepare datasets. Decide if the source, target, or both will be LRL. Decide the approach.
10/15 – 10/21	Reproduce the baseline from (Wang et al., 2020)
10/22 – 10/28	Reproduce the baseline from (Wang et al., 2020)
10/29 – 11/03	Running the model on the prepared corpus and analyze the results. Write the report for Checkpoint 2

Table 1: Week-by-week timeline up to Checkpoint 2.

[1] https://commonvoice.mozilla.org/en/dataset

References

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcrip-

tion in low-resource settings. *arXiv preprint arXiv:1803.08991*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.