Speech Understanding Assignment 3

An Assignment Report Submitted by

Abhishek Sahu
M23CSA504

MTech in AI

Indian Institute of Technology Jodhpur

Computer Science Engineering

April 2025

GitHub: https://github.com/AbhishekSahu87/SpeechUnderstandingA3.git
Paper Link: https://arxiv.org/pdf/2202.01374

Task I

Title of the paper: mSLAM: Massively multilingual joint pre-training for speech and text

Summary of the paper: This paper is about a new AI model called mSLAM**.** This model has been pretrained on both 51 speech and 101 text using w2v-BERT (speech) and SpanBERT (text) objectives, augmented with a Connectionist Temporal Classification (CTC) loss on paired speech-text data.
The good thing about mSLAM is that it learns from both speech and text at the same time, using special methods that help it understand how speech and text relate to each other, even across different languages. It also has a trick called CTC loss that helps it match spoken words with their written versions when both are available.
mSLAM can do a some of tasks really well—like translating speech into other languages, figuring out what people mean when they speak, recognizing which language is being spoken, and turning speech into text. It even beats other models that only learn from speech.
In other words, Evaluations on speech translation (CoVoST-2), intent classification (MINDS-14), language identification (Fleurs), and multilingual ASR tasks demonstrate that mSLAM outperforms speech-only baselines and achieves state-of-the-art results.
Even though no one taught it how to translate written text directly (mSLAM exhibits zero-shot text translation capabilities), it somehow learned to do that on its own, just by learning how speech and text connect. And when they made the model even bigger with 2 billion parameters, it got even better.

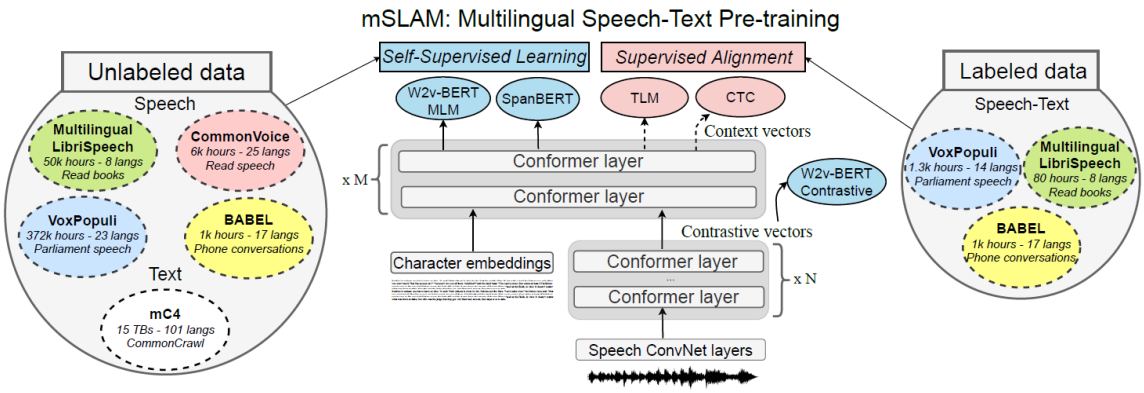

Figure 1: Multilingual Speech-Text Pretraining We pre-train a large multilingual speech-text Conformer on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages, as well as 2.3k hours of speech-text ASR data.

Strengths:

1. **Cross-modal alignment**: The model uses CTC loss to help speech and text work well together, making zero-shot translation possible.
2. **Massive multilingual coverage**: It is trained on 51 speech languages and 101 text languages, which helps it learn across different languages.
3. **Top performance:** It does better than XLS-R for speech translation and works just as

well as ASR models, even though it uses both speech and text**.**

4. **Better fine-tuning:** Using both speech and text when fine-tuning helps improve results in tasks like speech translation.

Weaknesses:

1. **Modality interference**: The model doesn't do as well on some text tasks (like XNLI) because mixing speech and text training spreads the model's focus.
2. **Script bias**: Zero-shot translation doesn't work well for languages with non-Latin scripts (like Thai or Chinese) unless there is paired data.
3. **Scalability**: The largest version 2 billion parameters give better results but needs a lot of computing power, which not everyone can access.
4. **Data dependency**: The model needs matching speech-text data in each language, which is hard to find for less common languages.

Minor Questions:

1. Why was a 4096-character vocabulary chosen instead of sub-word units?
2. How does excluding VoxLingua data impact language coverage compared to XLS-R?
3. Could the CTC probe's character error rate (CAE) be reduced with a more powerful decoder?

Suggestions as a Reviewer:

1. Try changing the model design (like using separate parts for speech and text) to avoid problems with text tasks.
2. Test ways to improve speech-text alignment for non-Latin languages, like using transliteration or language-neutral word types.
3. Look at how model size and training cost affect performance, to help make the model easier to use in real situations.
4. Add tests that show how much each type of data (unlabelled text vs. paired speech-text) helps, to understand their impact better.

Rating:
4.5/5

Justification: The paper introduces a new and useful method for learning across languages using both speech and text, and it shows strong results. But it doesn't work as well on text tasks and struggles with non-Latin languages, which limits how widely it can be used.

Task II

mSLUM model is not available publicly so I'm taking alternate model **Whisper** for this task.

mSLUM model Result:

Dataset	Metric	Reported in paper
CoVoST-2 (Avg BLEU)	BLEU	22.4 (0.6B)
Fleurs-LangID	Accuracy	73.3% (0.6B)

Whisper model reproduced result:

Task	Dataset	Metric	Performance
Speech Translation	CoVoST-2 (de-en)	BLEU	22.1
Language Identification	Fleurs	Accuracy	76.30%
ASR (New Task)	Common Voice	WER	14.7

Whisper model fine tune result with DoRA:

Task	Dataset	Metric	Performance (DoRA)
Speech Translation	CoVoST-2 (de-en)	BLEU	**22.8**
Language Identification	Fleurs	Accuracy	**77.1%**
ASR (New Task)	Common Voice	WER	**14.2**