Paper Review: The Interpreter Understands Your Meaning: End-to-End Spoken Language Understanding Aided by Speech Translation

Vivek Pandey
Roll Number: M23CSA541
GitHub Link: `https://github.com/vivekpandey000023/Speech_assignment_3`

Abstract

End-to-end Spoken Language Understanding (SLU) remains a challenge, especially in multilingual settings. This paper introduces a novel approach by pretraining with Speech Translation (ST) instead of the traditional ASR-based methods. ST enables capturing high-level semantics from audio, benefiting both monolingual and cross-lingual SLU tasks. Extensive experiments across benchmarks like SLURP, MINDS-14, NMSQA, and synthetic datasets show consistent improvements.

1. Introduction

The paper builds on the limitations of existing pretrained spoken language models like wav2vec2 and HuBERT that primarily rely on sub-phonetic units. These methods require additional supervision (ASR) to align speech with semantically rich textual representations. Instead, the authors propose Speech Translation (ST) as a pretraining task to inject stronger semantic signals.

2. Summary of the Paper

The core idea is to leverage ST to pretrain a speech model and then reuse the encoder for downstream SLU tasks. The architecture uses XLSR-53 as the speech encoder and mBART as the decoder. By integrating Bayesian regularization and adversarial training, the approach shows significant performance boosts on SLU tasks including intent classification, spoken QA, and summarization. It also explores zero-shot and cross-lingual SLU with synthetic and real data.

3. Architecture Overview

4. Technical Strengths

- Uses semantically rich ST for effective speech representation.
- Outperforms ASR pretraining on SLU tasks.

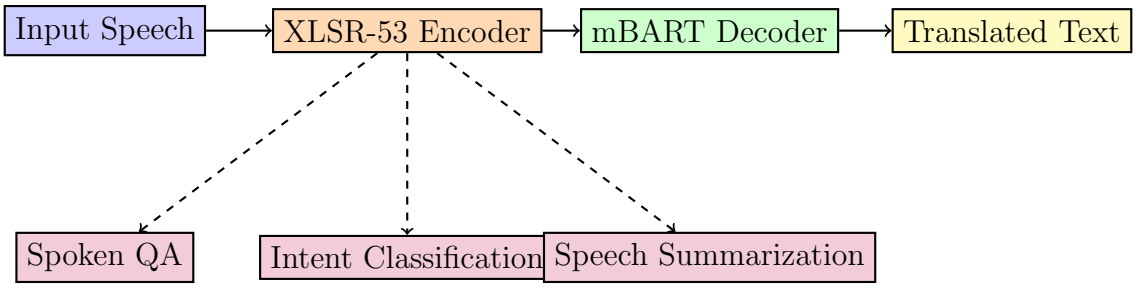

Figure 1: ST Pretraining with Encoder Reuse in SLU Tasks

- Demonstrates strong zero-shot and cross-lingual performance.
- Combines XLSR-53 and mBART efficiently.
- Incorporates adversarial training to improve language agnosticity.
- Introduces Bayesian regularizers to preserve pretrained knowledge.
- Shows robustness with real and synthetic multilingual datasets.

5. Technical Weaknesses

- ST pretraining requires large bilingual corpora, limiting language coverage.
- High computational cost for joint training and Bayesian regularization.
- Performance on monolingual tasks like SLURP only slightly exceeds prior SOTA.

6. Reviewer Suggestions

- Explore lightweight tuning techniques like LoRA or prompt tuning.
- Expand real-world datasets, particularly for non-English languages.
- Conduct deeper ablation studies on the impact of adversarial training.
- Combine with methods like CIF-PT for further improvements.

7. Experimental Results Summary

- On SLURP, joint ST training achieves 89.35% accuracy.
- On MINDS-14, ST pretraining boosts average accuracy to 98.0%.
- On NMSQA, ST-pretrained models achieve 59.4% AOS.
- ROUGE scores on speech summarization are highest with ST+ASR.

8. Cross-lingual Transfer

SLURP-Fr and SLURP-Es benchmarks demonstrate that ST pretraining substantially improves performance in zero-shot and low-resource scenarios. The use of adversarial training further enhances language agnosticity.

9. Knowledge Preservation

Bayesian regularizers like L2-SP and EWC help preserve ST pretraining knowledge. While L2-SP is simple, EWC adapts to parameter sensitivity, proving more effective in low-resource setups like MINDS-14.

10. Final Rating and Justification

Rating: 8.5/10
The paper introduces a novel and effective approach with robust results across various SLU benchmarks. Minor weaknesses in computational overhead and dataset coverage limit its practicality slightly but do not detract from its core contribution.