Construction of the First CS-ASR Model on Spanish-Guarani Code-switching (Jopara) using Hybrid Wav2Vec2-Transformer with Language Alignment Loss and Semi-supervised AI-aided Language Identification

Keywords: Code-Switching ASR, Translanguaging, Low-Resource Languages, Jopara, Spanish, Guarani, Language Alignment Loss

TL;DR: This project introduces the first CS-ASR system for Spanish-Guarani (Jopara), using Language Alignment Loss and a hybrid Wav2Vec2-Transformer, and significantly reduces data annotation requirements by relying only on intonation-unit-level transcription.

Abstract

Translanguaging Automatic Speech Recognition (CS-ASR) remains a technological challenge, especially for low-resource language pairs. *Jopara*, the common code-switching mode in Paraguayan daily life, is still underdeveloped in both sociolinguistic and computational research. This project aims to develop the first CS-ASR system for Spanish-Guarani speech (Jopara) using a hybrid Wav2Vec2-Transformer architecture and a novel Language Alignment Loss (LAL) framework.

Our approach introduces two key innovations: (1) adapting LAL to a Romance-Indigenous pair by focusing on acoustic differences, and (2) implementing a semi-automated, time-efficient pipeline for token-level language tagging using OpenAI API preprocessing and manual correction. Crucially, this approach alleviates the data scarcity problem that often limits low-resource code-switching research, since it requires only intonation-unit-level language annotation rather than costly and labor-intensive frame-level labeling. This reduces annotation and force alignment effort while still enabling effective model supervision.

So far, we have successfully completed tokenization and language labeling of code-switched audio, created gold-standard language tags, and benchmarked multilingual ASR baselines such as `facebook/mms-1b-all`, which yielded high error rates (WER 86.74%, CER 49.37%), highlighting the difficulty of the task.

Ongoing work focuses on several fronts. The full implementation of LAL in model training ensures that cross-attention assigns accurate frame-level language labels from token-level annotations. Training optimizes a

combined loss function:

$$\mathcal{L}_{\text{total}} = \alpha \, \mathcal{L}_{\text{CTC}} + (1 - \alpha) \, \mathcal{L}_{\text{Attn}} + \beta \, \mathcal{L}_{\text{LAL}}$$

where α and β control the contributions of CTC, attention, and LAL objectives. We are also tuning batch size, label smoothing, and learning rate schedule for optimal performance.

In addition, we are experimenting with alternative encoder architectures, such as the Conformer model, to better capture the acoustic variability in code-switched speech. Planned post-processing includes evaluating the impact of language-aware large language models (LLMs) for n-best hypothesis rescoring and error correction on ASR outputs.