Automatic Transcription of Documentation Recordings for Sociolinguistic Analysis: Speech Recognition and Forced-Alignment for Northern Prinmi

Connor Bechler
Department of Linguistics, University of Kentucky | Research Mentor: Dr. Josef Fruehwald

[Project Launch]

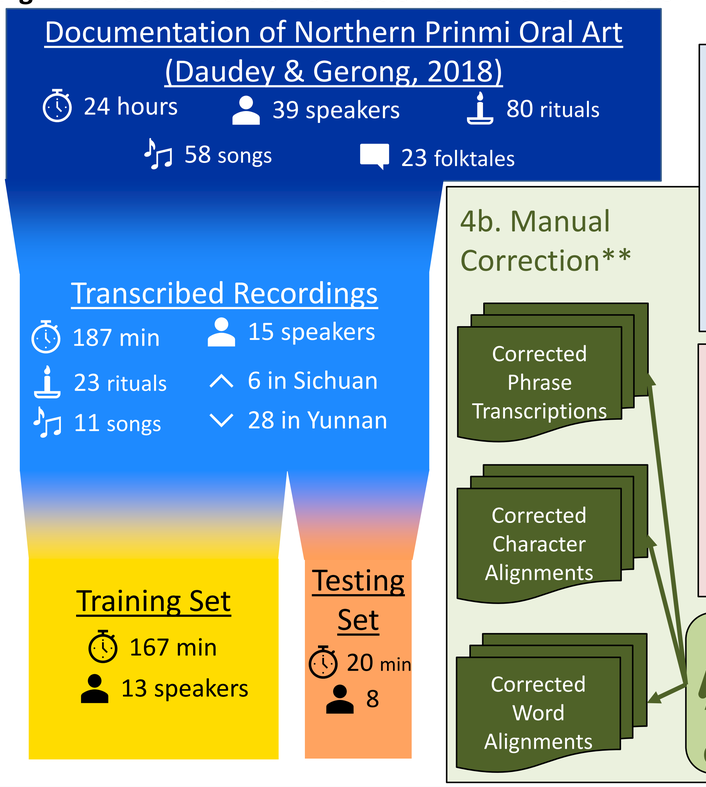
Introduction

This project applies **Automatic Speech Recognition (ASR)** in a novel context, adapting the **XLSR-53 model** (Conneau et al., 2020) to transcribe and align documentation records from **Northern Prinmi**, a Tibeto-Burman Qiangic language spoken in Southwest China (Daudey, 2014)

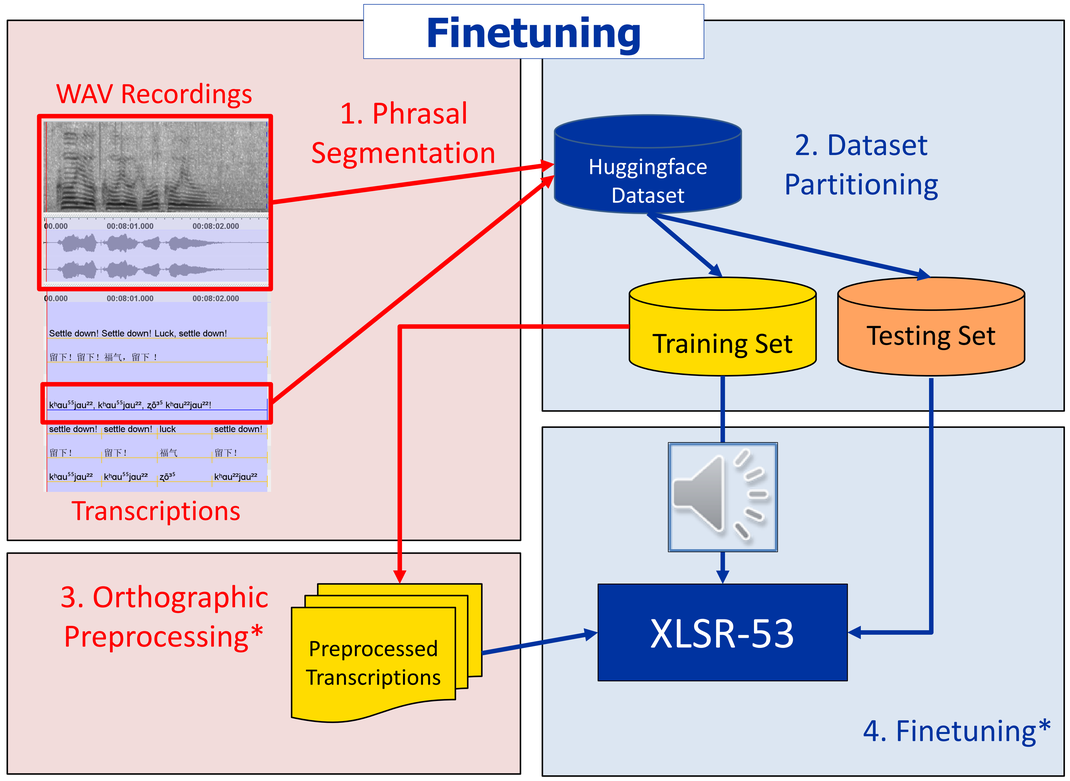
Prinmi

- **Prinmi** is spoken by ~45,000 people in the Sichuan and Yunnan provinces of China (Daudey & Gerong, 2020)
- The region is **extensively multiethnic and multilingual** (Daudey, 2014)
- **Northern Prinmi** is the language's largest dialect group, containing more than half of all speakers (Ding, 2014)
- Prinmi is **unstandardized** and does not have a widely adopted writing system (Daudey & Gerong, 2018)
- Northern Prinmi possesses **42 phonemic consonants**, **14 phonemic vowels**, and **four lexical tones** (Daudey, 2014)

Fig. 1 — Documentation Collection and Dataset Details

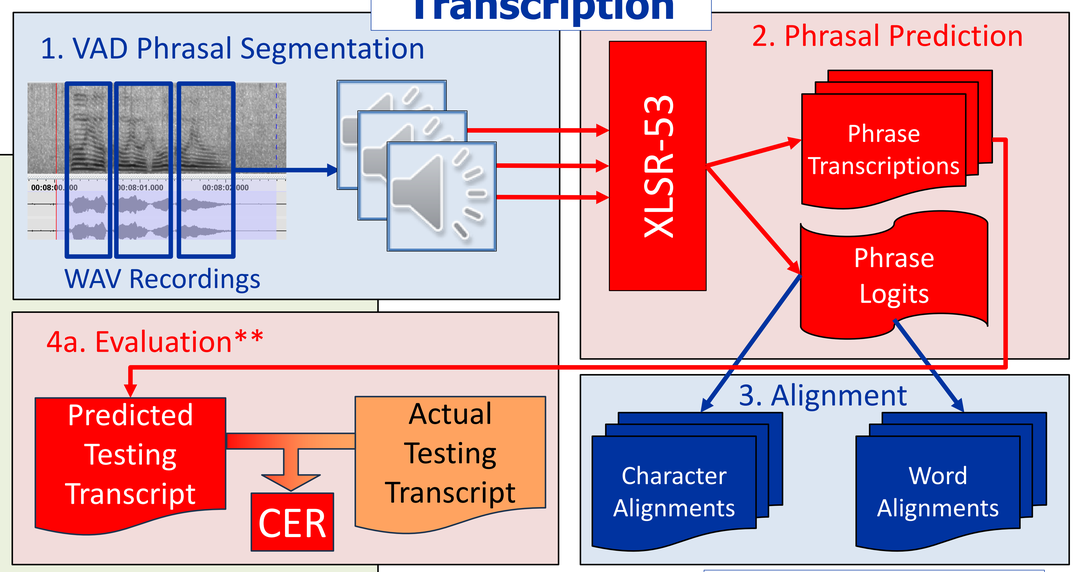


Finetuning



*Steps 3 & 4 were repeated for a range of preprocessing methods

Transcription



**Step 4a occurs during development, step 4b during documentation deployment

Key Findings

1. XLSR-53 model transcription **character error rates are comparable to previous ASR work for sociophonetics** (Coto-Solano et al., 2021)
2. Automatic **transcription quality is highly dependent on audio quality and genre**
3. Phrasal segmentation with **voice activity detection (VAD) results in worse transcriptions** than manual segmentation
4. **Predicted alignments substantially differ** between XLSR-53 and Montreal Forced Aligner (McAuliffe et al., 2017)

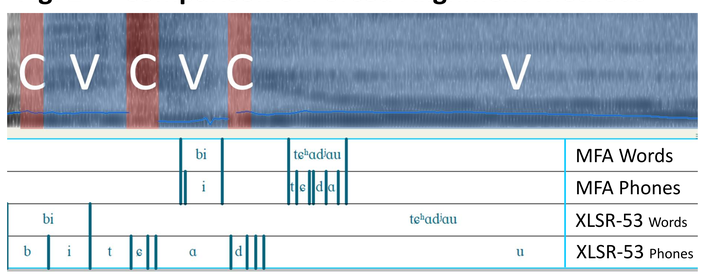
Table 1 — Error Rates by Model and Subsection

Testing Set (TS) Section	% of Training Data (TD)	Model CER		Model WER	
		Predicting Tones	Not Predicting Tones	Predicting Tones	Not Predicting Tones
Full	NA	0.323	0.315	0.876	0.769
Yunnan	84%	0.316	0.311	0.866	0.759
Sichuan	16%	0.351	0.328	0.916	0.807
Rituals	80%	0.318	0.321	0.860	0.767
Songs	20%	0.346	0.284	0.945	0.780

Table 2 — Error Rates for Non-Tone Model With VAD

TS Section	% of TD	CER	WER
Full	NA	0.432	0.867
Yunnan	84%	0.432	0.857
Sichuan	16%	0.429	0.902
Rituals	80%	0.421	0.841
Songs	20%	0.482	0.977

Fig. 2 — Comparison of Forced Alignment Methods



References

Ask me about...

1. Orthographic preprocessing and its effects
2. How well the model handles different classes of sounds
3. Voice activity detection (VAD) and speech diarization
4. Specific tools or choices in my process
5. Applying these methods to a different language/context
6. Incorporating language models with ASR