Evaluating wav2vec 2.0 Speech Recognition and Forced Alignment on a Multi-Varietal Language Documentation Collection

Applying sociolinguistic analysis in documentary contexts offers many benefits (Meyerhoff, 2019), but also poses several challenges: many documentation collections lack sufficient metapragmatic information (Di Carlo et al., 2021), include low-fidelity and noisy recordings (Amith et al., 2021; Ćavar et al., 2016), and remain partially unannotated or untranscribed. One solution to the latter two challenges is partial automation of the baseline annotation tasks required in documentation through automatic speech recognition (ASR), forced alignment (FA), and natural language processing (Ćavar et al., 2016; He et al., 2024; Jimerson et al., 2023; Tsoukala et al., 2023). However, as Coto-Solano (2022) notes, adapting these tools for minority and indigenous languages is still "difficult and expensive."

This work is part of an ongoing project exploring semi-automatic annotation of the Northern Prinmi Oral Art Collection, a multi-genre and multi-varietal documentation collection (Daudey & Gerong, 2018). The project aims to assist in the transcription and analysis of an existing documentation collection, stress-test semi-automatic annotation tools in a challenging context, and increase the accessibility of these tools for non-programmers. To achieve these aims, I developed a Python tool, wav2vec2fasr. Wav2vec2fasr includes functions for describing transcribed audio corpora, preprocessing transcripts and audio, and training, applying, and evaluating wav2vec2 models for ASR and FA.

I applied wav2vec2fasr to the Northern Prinmi Oral Art collection, fine-tuning a variety of models with different tokenization schemes and hyperparameters. Analyzing model performance on automatic transcription of previously transcribed documentation recordings, the best model achieved an overall character error rate (CER) of .325, comparable to previous work on automatic transcription for sociophonetic analysis (Coto-Solano et al., 2021), but worse than models from similar projects, which range from .05 to .25 (Coto-Solano et al., 2022; Guillaume et al., 2022; Macaire et al., 2022). CER varied widely by recording, correlating most obviously with average utterance duration, recording location, and recording genre (Fig. 1). Internal regional variation within Northern Prinmi may impact model performance, as there are at least four varieties present in the collection (Fig. 2; drawn from Daudey and Gerong, personal communication, 2024 April 9).

To evaluate the performance of wav2vec2 for FA, I aligned the transcribed recordings from the documentation corpus with both wav2vec2fasr and Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). Following Chodroff et al. (2024), I applied MFA with an English acoustic model. As the corpus did not include word or phone alignments, I calculated inter-aligner agreement between wav2vec2 and MFA as a proxy for aligner performance. Alignments differed substantially, with a median word onset difference of 80 milliseconds and 90% interaligner agreement on word onset boundaries only occurring at 410 milliseconds. Recording genre strongly correlates with inter-aligner agreement (Fig. 3). Examining specific alignments, Wav2vec2 appears less precise in terms of phone boundaries, severely truncating consonant duration, while MFA struggles to align recordings of songs and deletes numerous words. This suggests that Chodroff et al.'s finding that MFA performs more consistently than wav2vec2 alignments on extremely small datasets may be influenced by the speech genre of the dataset.

Supplemental Figures

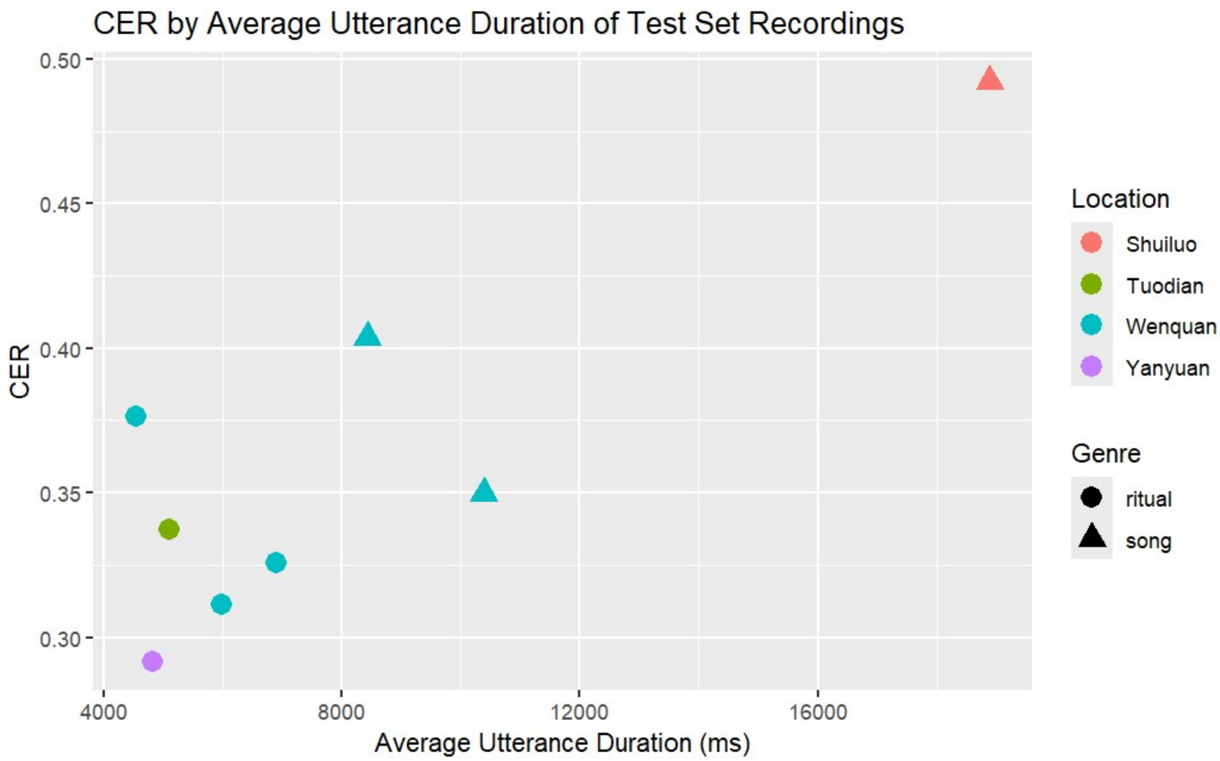

Fig. 1 — Wav2vec2 CER by average utterance duration; each point represents a recording

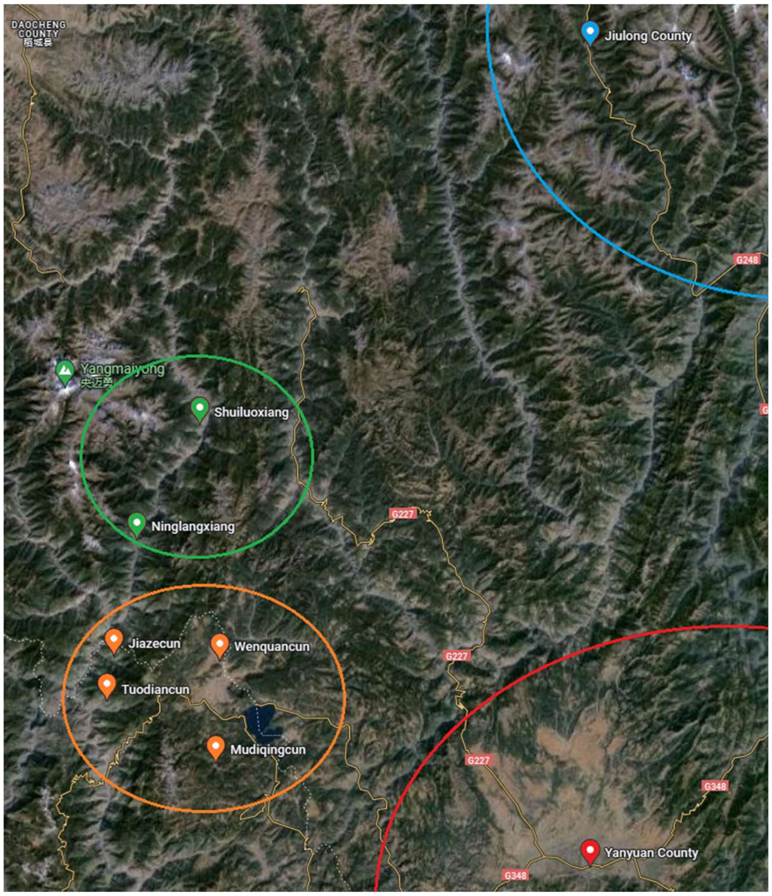

Fig. 2 — Locations and approximate varietal groupings of recordings from the Northern Prinmi Oral Art Collection

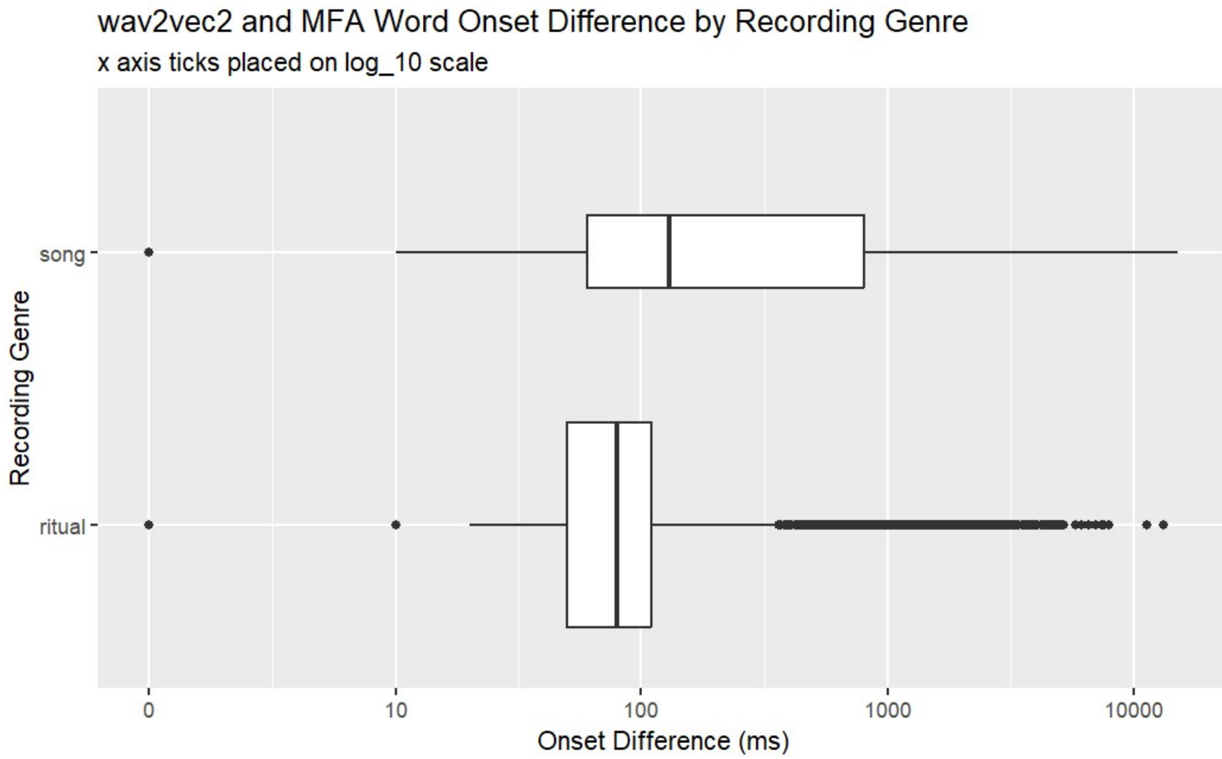

Fig. 3 — Difference between wav2vec2 and MFA word onset boundary alignments by genre

Bibliography

Amith, J. D., Shi, J., & Castillo Garcia, R. (2021). End-to-End Automatic Speech Recognition: Its Impact on the Workflow for Documenting Yoloxóchitl Mixtec. *First Workshop on NLP for Indigenous Languages of the Americas. 11 June 2021. Https://Www.Aclweb.Org/Anthology/2021.Americasnlp-1.8.Pdf*. https://par.nsf.gov/biblio/10281120-end-end-automatic-speech-recognition-its-impact-workflow-documenting-yoloxochitl-mixtec

Ćavar, M., Ćavar, D., & Cruz, H. (2016). Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4004–4011. https://aclanthology.org/L16-1632

Chodroff, E., Ahn, E., & Dolatian, H. (2024). Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*.

Coto-Solano, R. (2022). Computational sociophonetics using automatic speech recognition. *Language and Linguistics Compass*, *16*(9), e12474. https://doi.org/10.1111/lnc3.12474

Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., Koka'ua, L., Tanveer, S., & Feldman, I. (2022, June 20). Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. *Language Resources and Evaluation (LREC) Conference 2022*. https://mro.massey.ac.nz/handle/10179/17252

Coto-Solano, R., Stanford, J. N., & Reddy, S. K. (2021). Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems With DARLA. *Frontiers in Artificial Intelligence*, *4*. https://www.frontiersin.org/articles/10.3389/frai.2021.662097

Daudey, H., & Gerong, P. (2018). *Documentation of Northern Prinmi oral art, with a special focus on ritual speech*. Endangered Languages Archive. Handle: http://hdl.handle.net/2196/00-0000-0000-0010-8820-B

Di Carlo, P., Ojong Diba, R. A., & Good, J. (2021). Towards a coherent methodology for the documentation of small-scale multilingualism: Dealing with speech data. *International Journal of Bilingualism*, *25*(4), 860–877.

Guillaume, S., Wisniewski, G., Macaire, C., Jacques, G., Michaud, A., Galliot, B., Coavoux, M., Rossato, S., Nguyên, M.-C., & Fily, M. (2022). Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 170–178. https://doi.org/10.18653/v1/2022.computel-1.21

He, T., Choi, K., Tjuatja, L., Robinson, N. R., Shi, J., Watanabe, S., Neubig, G., Mortensen, D. R., & Levin, L. (2024). Wav2Gloss: Generating Interlinear Glossed Text from Speech. *arXiv Preprint arXiv:2403.13169*.

Jimerson, R., Liu, Z., & Prud'hommeaux, E. (2023). An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1008–1016. https://doi.org/10.18653/v1/2023.acl-short.87

Macaire, C., Schwab, D., Lecouteux, B., & Schang, E. (2022). *Automatic Speech Recognition and Query By Example for Creole Languages Documentation*. Findings of the Association for Computational Linguistics: ACL 2022. https://hal.science/hal-03625303

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, *2017*, 498–502.

Meyerhoff, M. (2019). Unnatural bedfellows? The sociolinguistic analysis of variation and language documentation. *Journal of the Royal Society of New Zealand*, *49*(2), 229–241. https://doi.org/10.1080/03036758.2019.1619599

Tsoukala, C., Kritsis, K., Douros, I., Katsamanis, A., Kokkas, N., Arampatzakis, V., Sevetlidis, V., Markantonatou, S., & Pavlidis, G. (2023). ASR pipeline for low-resourced languages: A case study on Pomak. *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, 40–45.