Evaluating forced alignment for under-resourced languages: A test on Squliq Atayal

Chi-Wei Wang[1], Bo-Wei Chen[1], Bo-Xuan Huang[1], Ching-Hung Lai[2] & Chenhao Chiu[1 3]

[1]Graduate Institute of Linguistics, National Taiwan University (Taiwan), [2]School of Medicine, National Cheng Kung University (Taiwan), [3]Neurobiology and Cognitive Science Center, National Taiwan University (Taiwan)

r09142007@ntu.edu.tw, r09142001@ntu.edu.tw, r09142003@ntu.edu.tw, i54051092@gs.ncku.edu.tw, chenhaochiu@ntu.edu.tw

Trainable forced alignment offers feasible solutions to document under-resourced languages. This study aims to assess the performances of a Montreal Forced Aligner (MFA) [1] trained model using a small scale of phonetically transcribed field data in Squliq Atayal, an endangered Austronesian language spoken in Taiwan. Regarding the training dataset, the preliminary corpus consists of a 20-minute recording as an excerpt from a series of fieldwork sessions. All elicited utterances produced by one female Squliq Atayal native speaker were manually labeled at both word and phone levels using Praat [2]. The pronunciation dictionary was generated by combining the word and phone tiers in each manually annotated transcription, ensuring that all words that occurred in the corpus were appended. Once the aligned TextGrid files were retrieved, evaluations were implemented by comparing MFA outputs with manual annotations based on (1) the *accuracy* measurements [3, 4], by measuring the agreements (AG) of interval boundaries at different thresholds, the overlap rates (OvR), and the midpoint displacements (MpD) of each segment; in addition to (2) the *acoustic* measurements [5, 6], by fitting the pitch trajectories through words and the formant trajectories through the most common vowels /a, i, u/ constructed by 30 data points using 7 (F0 + 2 formants * 3 vowels) generalized additive mixture models (GAMMs) [7].

The *accuracy* results suggested that the mean AG of consonants slightly outperformed that of vowels (Table 1) while the mean OvR of consonants was lower, along with the mean MpD being significantly larger when compared to those of vowels (Table 2). Here, the discrepancy might be accounted for by few extreme misalignments. In this case, AG would be more suitable for evaluating overall performances, while OvR and MpD were considered more robust when evaluating the effects of segment types on alignment accuracy. On the other hand, for *acoustic* trajectories (both pitch and formants), no statistical significance was found between the MFA and manual annotations, except for the F2 trajectories of [u], as illustrated in Figures 1 and 2, which positively supported the reliability of the current MFA model. Overall, the current results revealed that MFA outcomes were highly consistent with manual annotations when little but comprehensively labeled data were provided. Moreover, our results also suggested that different evaluation methods may come along with diverged results and should be implemented based on the objectives of the research.

Keywords: forced alignment, phonetic fieldwork, language documentation, Squliq Atayal

Threshold	Vowel	Consonant	Overall
10 ms	36.32%	36.67%	36.49%
20 ms	63.37%	68.66%	66.02%
30 ms	76.26%	83.33%	79.80%
40 ms	79.09%	85.73%	82.41%
50 ms	80.35%	86.40%	83.37%

Table 1. The mean AGs at both boundaries for vowels and consonants, along with the overall AGs, at different thresholds.

Type	OvR	MpD
Vowel	54.67%	33.68 ms
Consonant	41.19%	211.23 ms
Overall	47.37%	129.76 ms
[a]	45.04%	50.62 ms
[i]	59.82%	25.44 ms
[u]	40.73%	39.70 ms

Table 2. The mean OvRs and MpDs of vowels, consonants, overall performances, and the most common vowels [a], [i], [u].

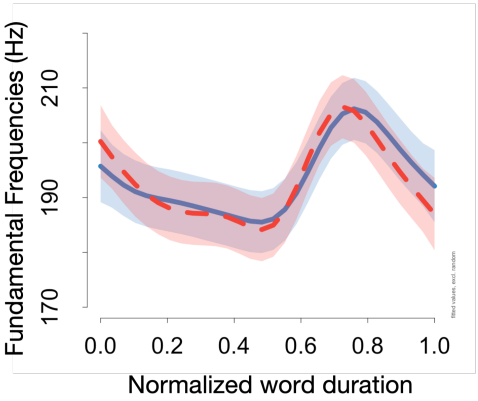

Figure 1. F0 trajectories over normalized word duration fitted by GAMM (MFA = red dashed line; manual = blue solid line).

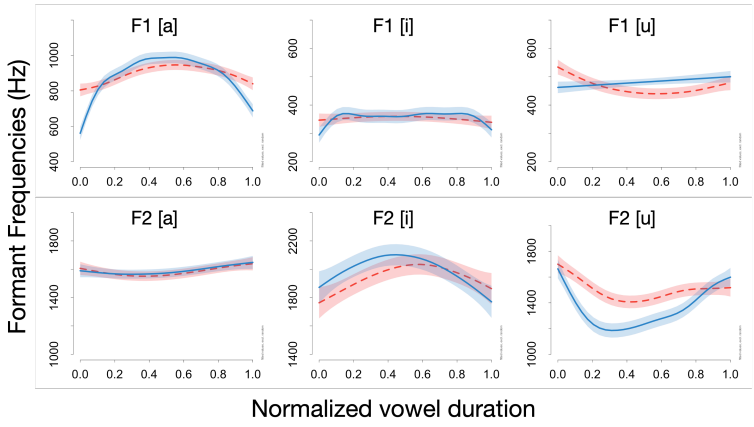

Figure 2. Formant trajectories over the normalized [a, i, u] fitted by GAMMs (MFA = red dashed line; manual = blue solid line).

References

[1] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, 498–502.

[2] Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer.

[3] Jones, C., Li, W., Almeida, A., and German, A. (2019). Evaluating cross-linguistic forced alignment of conversational data in North Australian Kriol, an under-resourced language. *Language Documentation and Conservation*, *13*, 281–299.

[4] Gonzalez, S., Travis, C., Grama, J., Barth, D., and Ananthanarayan, S. (2018). Recursive forced alignment: A test on a minority language. In *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, 145–148.

[5] Babinski, S., Dockum, R., Craft, J. H., Fergus, A., Goldenberg, D., and Bowern, C. (2019). A robin hood approach to forced alignment: English-trained algorithms and their use on Australian languages. In *Proceedings of the Linguistic Society of America*, 1–12.

[6] Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America, 142*(4), 2469–2482.

[7] Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86-116.