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

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

[1]Graduate Institute of Linguistics, National Taiwan University [2]School of Medicine, National Cheng Kung University

[3]Neurobiology and Cognitive Science Center, National Taiwan University

Introduction

What is forced alignment (FA)?

In forced alignment, speech and its corresponding orthographic transcription are automatically aligned at the word and phone levels, given a way to map graphemes to phonemes (typically a pronunciation dictionary) and an acoustic model of how phones are realized [1].

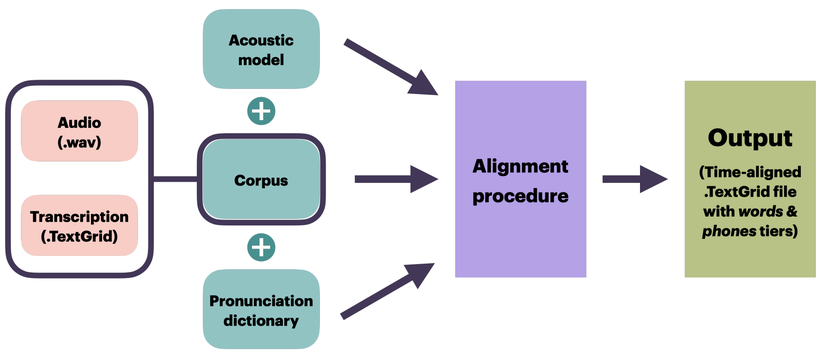


Figure 1. A schematic pipeline of FA

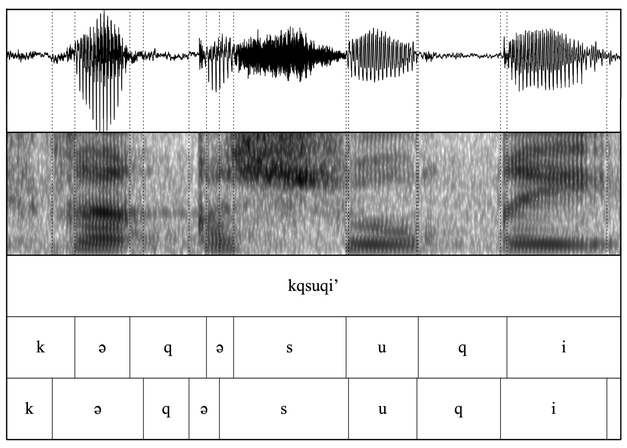


Figure 2. An excerpt from the FA output TextGrid file, including the manually-annotated word tier (upper), the manually-annotated phone tier (middle), and the FA-aligned phone tier (lower).

FA for under-resourced languages:

- Adapting an existing acoustic model made for well-documented languages [2]:
  The feasibility of such manipulation may be reduced due to the mismatch of sound inventories or orthographic systems between the two distinct languages.
- Training a new, language-specific acoustic model using ASR toolkits [3]:
  Since all phonological and phonetic clues of the target language are embedded in a customized model, it is likely to outperform the pre-trained model in terms of the alignment results.

Evaluation of FA (corresponding manually-annotated transcription are required):

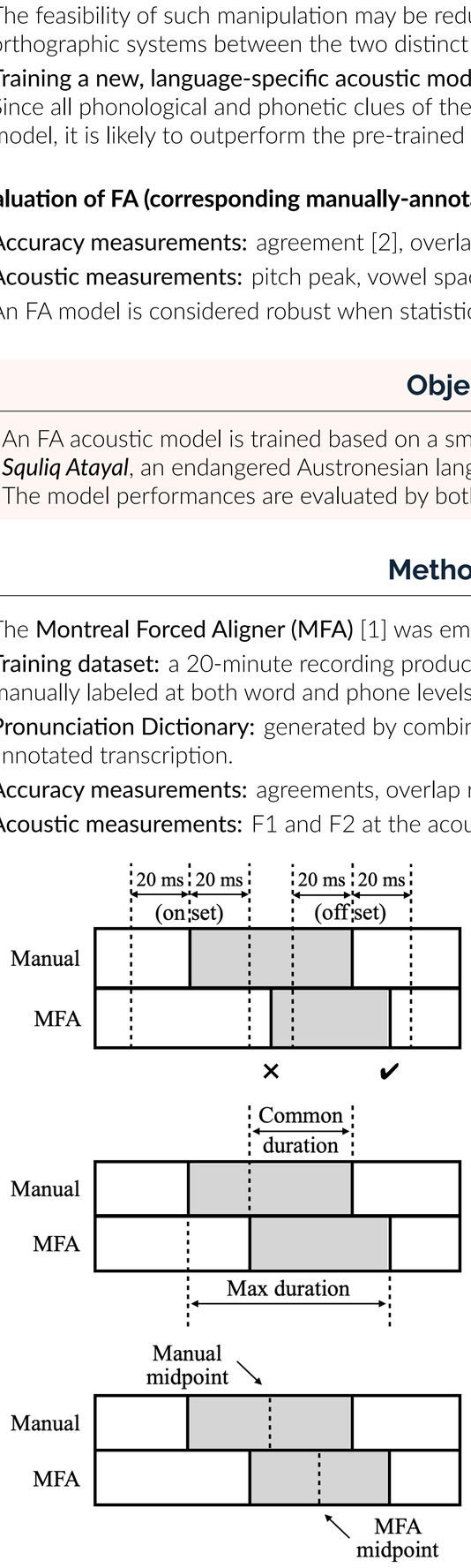
- **Accuracy measurements:** agreement [2], overlap rate, and robustness [4].
- **Acoustic measurements:** pitch peak, vowel space, and consonant VOT [5].
- An FA model is considered robust when statistical significance is absent among the measurements.

Objectives

- An FA acoustic model is trained based on a small scale of phonetically transcribed field data in ***Squliq Atayal***, an endangered Austronesian language spoken in Taiwan.
- The model performances are evaluated by both **accuracy** and **acoustic** measurements.

Methodology

- The **Montreal Forced Aligner (MFA)** [1] was employed in this study.
- **Training dataset:** a 20-minute recording produced by one female Squliq Atayal native speaker, manually labeled at both word and phone levels using Praat.
- **Pronunciation Dictionary:** generated by combining the word and phone tiers in each manually annotated transcription.
- **Accuracy measurements:** agreements, overlap rates, and midpoint displacements (*cf.* robustness).
- **Acoustic measurements:** F1 and F2 at the acoustic midpoints of most common vowels [a, i, u].



$$\text{Agreement} = \frac{\text{Agreed boundaries}}{\text{All boundaries}} \times 100\,\%$$

$$\text{Overlap rate} = \frac{\text{Common duration}}{\text{Max duration}} \times 100\,\%$$

$$\text{Midpoint displacement} = |\text{MFA midpoint} - \text{Manual midpoint}|$$

Figure 3. Representations and calculations of the three accuracy measurements.

Results

Accuracy measurements:

- Agreement: consonants > vowels → different from the previous studies [2, 5].

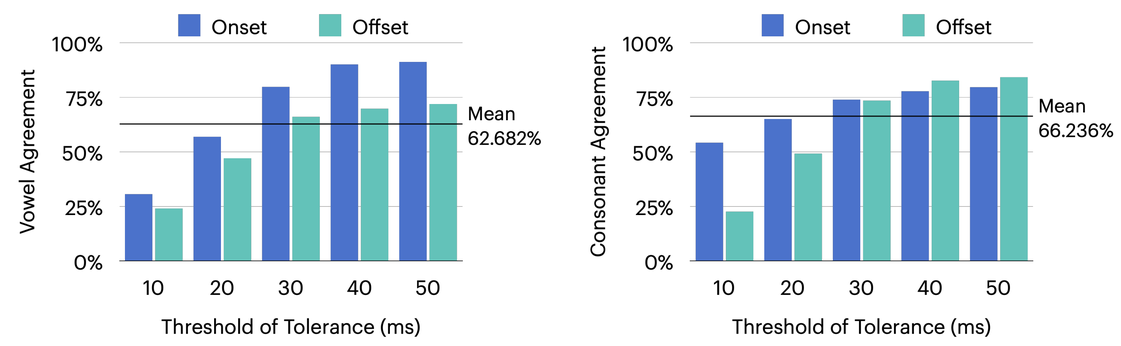


Figure 4. The agreement of vowels (left panel) and consonants (right panel) at different threshold of tolerance.

Table 1. The overall agreement at different threshold of tolerance.

Threshold	10 ms	20 ms	30 ms	40 ms	50 ms
Agreement	32.84%	54.49%	73.23%	80.02%	81.72%

- **Overlap rate:** consonants > vowels; **Midpoint displacement:** vowels > consonants

Table 2. The mean overlap rates and midpoint displacements of different segment categories.

Category	Overlap rate	Midpoint displacement
vowel	53.26%	35.03 ms
consonant	55.20%	29.28 ms
overall	54.29%	31.99 ms

Acoustic measurements:

- Statistical significance was only found in the F1 [u] condition (paired *t*-test, $p < .05$ *).

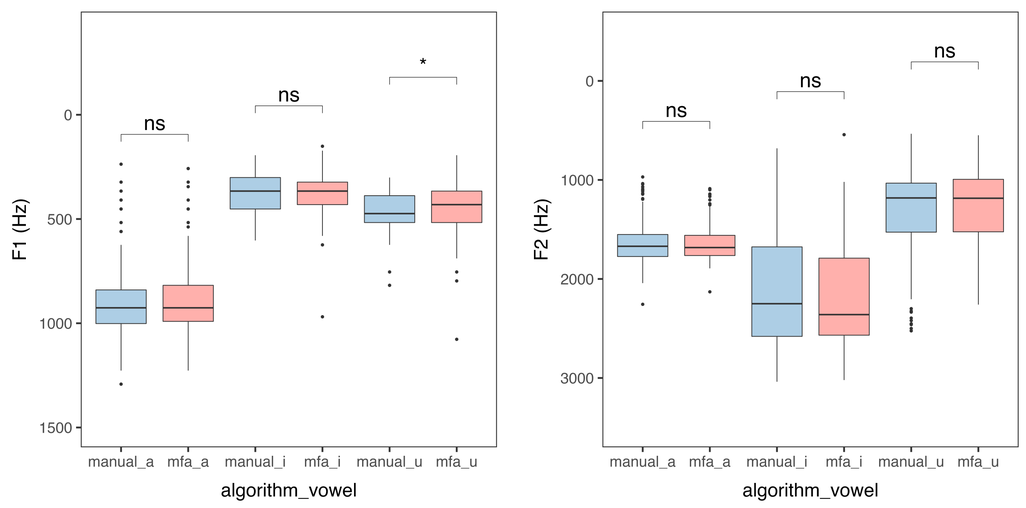


Figure 5. F1 and F2 at the acoustic midpoints of [a, i, u] (Manual annotation = blue; MFA output = red).

Discussion

Previous studies: vowels were associated with better alignments than consonants?

- The effect of segment types: **complex vowels** have the lowest mean overlap rate and the largest mean midpoint displacement [$F(5, 2342) = 43.05$, $p < .001$ ***] → supported by [6].

Table 3. The mean overlap rates and midpoint displacements of different segment types.

Type	Overlap rate	Midpoint displacement
full vowels	60.21%	30.21 ms
weak vowels	39.07%	26.52 ms
complex vowels	38.54%	71.27 ms
plosives	60.86%	18.62 ms
nasals & liquids	46.55%	42.85 ms
fricatives & affricates	54.26%	33.04 ms

Conclusions

- The results of accuracy and acoustic measurements reveal that MFA outputs fairly correspond to manual annotations **when little but comprehensively labeled data were provided.**

References

[1] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, pp. 498–502, 2017.

[2] C. Jones, W. Li, A. Almeida, and A. German, "Evaluating cross-linguistic forced alignment of conversational data in north australian kriol, an under-resourced language," *Language Documentation and Conservation*, pp. 281–299, 2019.

[3] R. Billington, H. Stoakes, and N. Thieberger, "The Pacific Expansion: Optimizing Phonetic Transcription of Archival Corpora," in *Interspeech*, pp. 4029–4033, 2021.

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

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

[6] S. Paulo and L. C. Oliveira, "Automatic phonetic alignment and its confidence measures," in *International Conference on Natural Language Processing*, pp. 36–44, 2004.

Links

Our lab

This poster