EVALUATING FORCED ALIGNMENT FOR UNDER-RESOURCED LANGUAGES: A TEST ON SQULIQ ATAYAL DATA

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

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

ABSTRACT

Trainable forced alignment offers feasible solutions to document under-resourced languages. This study aims to assess the performances of a Montreal Forced Aligner (MFA) trained model using a small scale of phonetically transcribed field data in Squliq Atayal, an endangered Austronesian language spoken in Taiwan. Evaluations were implemented by comparing MFA outputs with manual annotations based on (1) the accuracy measurements on the interval boundaries of each segment, and (2) the acoustic measurements, by fitting the formant trajectories through the most common vowels [a, i, u] with generalized additive mixture models (GAMMs). The results showed that the general agreement reached 73.23% of accuracy when with a tolerance of 30 ms misalignment, and no statistical significance was found between the formant trajectories except for F1 trajectories of [a] and F2 trajectories of [u], revealing that MFA outcomes were fairly consistent with manual annotations when little but comprehensively labeled data were provided.

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

1. INTRODUCTION

Recently, forced alignment (FA) systems have been providing feasible solutions to enhance the workflow of phonetic analyses. A prototypical pipeline of FA includes (1) integrating the audio files consisting of speech signals and the TextGrid files containing orthographic transcriptions at the utterance level to form a corpus. (2) Conducting the alignment with the help of the acoustic model, which calculates how likely a phone is given the acoustic features; and the pronunciation dictionary, which provides the reference for the grapheme-to-phoneme (G2P) mapping. (3) Generating a time-aligned TextGrid file with both word and phone tiers as the output.

Previous works on FA for under-resourced languages often adapted existing acoustic models made for well-documented languages [1, 2, 3, 4]. Nevertheless, the feasibility of such manipulation may be reduced due to the mismatch of sound inventories or orthographic systems between the two distinct languages. In order to solve the cross-linguistic issues, with the assistance of automatic speech recognition (ASR) toolkits such as Kaldi [5] or HTK [6], researchers can also consider training a new, language-specific acoustic model based on a relatively small corpus [2, 7, 8, 9]. 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.

The overall performance of the FA can be assessed once the alignment is completed. In order to evaluate the FA outcomes, corresponding manually-aligned data are required and treated as the reference. There are several evaluation methods. One of the most common and essential ways is to measure the agreement (in percentage) at the onset and offset boundaries [1, 3, 4, 7, 8], which refers to the proportion of the FA and manual boundaries from the same segment that agree within a given threshold of tolerance (in ms). Intuitively, the larger the threshold, the better the agreement.

Alternatively, other studies proposed that such temporal-based accuracy measurement of the alignment can also be performed by calculating the overlap rate (in percentage) and the robustness (in ms) [9, 10, 11, 12, 13]. The former refers to the proportion of overlap between the intervals established by FA and human annotators. Similar to the analysis of agreement, the greater the overlap, the higher the accuracy; while the latter focuses on the actual displacement between the midpoints of the FA and manual intervals. Specifically, the mean displacement of error tokens (i.e., the segments whose midpoints lie beyond a pre-determined threshold; e.g., 20 ms) was calculated; the less the displacement, the better the FA performances.

Furthermore, FA performances can also be evaluated by measuring acoustic features of the output [2, 14]. For instance, the location of pitch (F0) peak through the words, the vowel space constructed by the first and second formants (F1 and F2), and the consonant VOT. An FA model would be considered robust when statistical significance is absent among these measurements.

For the objectives of this study, 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 the aforementioned measurements, including the agreement (AG), the overlap rate (OR), the midpoint displacement (MD, inspired by the robustness), and the formant measurements.

2. METHODOLOGY

The Montreal Forced Aligner (MFA) [15] was employed in this study. MFA is a Kaldi-based, open-source, cross-platform system that not only provides pre-trained pronunciation dictionaries and acoustic/G2P models in multiple languages but also offers a sophisticated API for training new dictionaries and models, operated in a user-friendly command-line interface.

2.1. Squliq Atayal dataset and model training

The dataset adopted as the source of the FA acoustic model training corpus was the complete recordings (in .wav format, mono at 16-bit/44.1kHz sampling rate) from a series of fieldwork sessions. All elicited utterances produced by one female Squliq Atayal native speaker (excluding the speech produced in the contact language) were manually labeled at both word and phone levels by two trained phoneticians using Praat [16].

To build the pronunciation dictionary, since every TextGrid file in the current dataset has a phonetically transcribed phone tier, the pronunciation dictionary was generated using Python [17] by concatenating the word and phone tiers in each manually-annotated transcription. Consequently, all words that occurred in the corpus were included and combined as a comprehensive pronunciation dictionary.

The acoustic model training and alignment procedures were operated simultaneously, with 31 pairs of audio/TextGrid files and the pronunciation dictionary as the inputs. Note that each TextGrid file could only include the word tier, as the MFA would recognize multiple tiers in a file as different speakers. The output files included a .zip file as a newly-trained acoustic model and 31 aligned TextGrid files. The overall workflow took approximately 40 minutes on a Linux server.

2.2. Evaluation of the MFA outputs

Once the aligned TextGrid file was retrieved, they were loaded using the `PraatIO` [18] package in Python along with their manually-annotated counterparts. Here, a customized algorithm was established to locate the identical segment in both manual and MFA phone tiers automatically. Nevertheless, an alignment issue was discovered: since the manual annotations of the segments were narrow transcriptions which reflect detailed surface representation of sounds, many allophonic variants of identical words were consequently appended into the pronunciation dictionary. Such variants could severely impact the workability of the algorithm, leading to segment mismatches between the manual and MFA phone tiers. In this case, only one pair of audio/TextGrid files was included in the current study (*n* = 2348; duration = 1065.75 seconds).

Regarding the evaluation procedures, the temporal-based measurements of the alignment accuracy (i.e., AG, OR, and MD) were calculated by comparing MFA outputs with manual annotations using Python. As for the acoustic measurements, the F1 and F2 frequencies among each phone interval from both MFA outputs and manual annotations were obtained using a customized Praat script [19] and were then imported to R [20] for statistical modeling. Particularly, the formant trajectories through the most common vowels [a, i, u] constructed by the 30 data points were fitted by six (2 formants * 3 vowels) generalized additive mixture models (GAMMs) [21]. GAMMs are particularly useful for analyzing non-linear relationships between variables. They also involve fitting smooth functions to the predictor variables (i.e., manual vs. MFA) and accommodate non-normal response variables (i.e., the acoustic data points), making it applicable to our study.

3. RESULTS

Figure 1 presents an example from the MFA output merged with its corresponding manual transcriptions. The first two tiers were annotated by a human annotator, and the lower two tiers were aligned by MFA. By eyeballing, the automatic alignment results between human and MFA were fairly consistent, except for the unexpected empty intervals generated between words. This may be the case that MFA prefers utterance-level to word-level transcriptions, as suggested by its documentation.

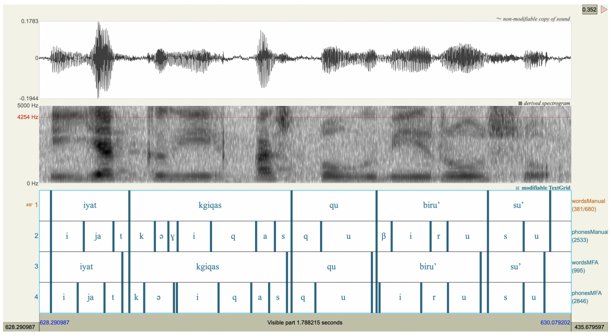

Figure 1: Annotated results by a human annotator (top two tiers) and MFA (bottom two tiers).

3.1. Measurements of alignment accuracy

Table 1 lists the agreements (AGs) at different thresholds, separated by vowel and consonant onset/offset; in addition to the mean AGs at both boundaries, as well as the overall AGs. As revealed by Table 1, better AGs were observed in vowel onsets than offsets. On the other hand, better AGs were observed in consonant offsets than onsets when with a higher tolerance. One possible reason is that the canonical syllable structure of Squliq Atayal is C(G)V(C), suggesting that both consonant clusters and hiatus are prohibited [22, 23]. As such, there should be mostly consonant-vowel sequences that occur throughout the entire phone tier. Moreover, the mean AGs of consonants outperform that of vowels, which is different from the previous studies [1, 2]. For the criteria of implementing the thresholds, according to the FA-related literature, 20 ms is considered the most robust threshold [24]; while the literature on under-resourced languages also reports at 30, or even 50 ms thresholds [1].

Table 2 lists the mean overlap rates (ORs) and midpoint displacements (MDs) of the following five different categories: vowels, consonants, overall performances, and the most common vowels [a, i, u]. Note that despite the algorithm suggested in [9, 13], the calculation of MDs here includes all segments instead of the error tokens pre-defined by a given threshold, in order to provide a relatively comprehensive view of the measurements. Overall, the results are in correspondence to those of AGs, where consonants have slightly higher ORs and narrower MDs than vowels. The results of ORs/MDs are also consistent among the three most common vowels. Specifically, [i] are usually better aligned; while [a] are less aligned.

Table 2: The mean overlap rates (ORs) and midpoint displacements (MDs) of different segment categories.

Category	OR	MD
vowel	53.26%	35.03 ms
consonant	55.20%	29.28 ms
overall	54.29%	31.99 ms
[a]	53.08%	39.63 ms
[i]	71.56%	19.43 ms
[u]	61.88%	28.05 ms

3.2. Measurements of acoustic features

Figure 2 renders the formant trajectories through vowels [a, i, u] fitted by GAMMs. The duration was normalized based on the 30 data points along with the shaded areas representing 95% CIs. The red dashed lines illustrate the formant trajectories of MFA-aligned vowels, and the blue solid lines depict those of manually-annotated vowels. Crucially, the overlap between the fits and confidence intervals suggest no difference between manual and MFA annotations.

Table 1: The agreements (AGs) of vowels and consonants at different thresholds.

Threshold	Vowel			Consonant			Overall
	Onset	Offset	Mean	Onset	Offset	Mean	
10 ms	30.62%	23.94%	27.28%	54.15%	22.64%	38.40%	32.84%
20 ms	56.82%	46.97%	51.90%	65.03%	49.15%	57.09%	54.49%
30 ms	79.67%	66.03%	72.85%	73.81%	73.41%	73.61%	73.23%
40 ms	89.97%	69.74%	79.86%	77.76%	82.59%	80.18%	80.02%
50 ms	91.15%	71.91%	81.53%	79.61%	84.21%	81.91%	81.72%

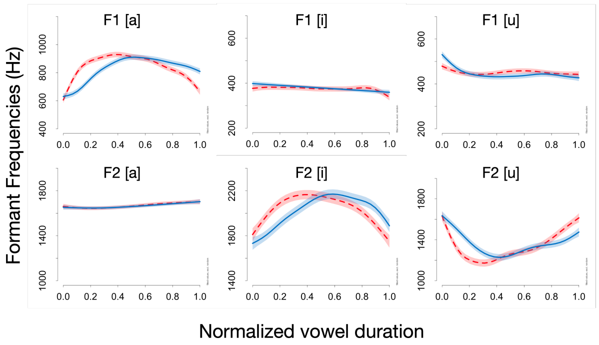

Figure 2: Formant trajectories over the normalized [a, i, u] fitted by GAMMs (MFA = red dashed lines; manual = blue solid lines).

As revealed in Figure 2, no statistical significance was found between MFA outputs and manual annotations except for F1 trajectories of [a] and F2 trajectories of [u]. Note that in the F2 [i] condition, despite that only limited overlap was found between two annotations, no statistical significance was reported. Overall, the acoustic results are not only consistent with the OR/MD results that the alignment of [i] outperforms that of [a] and [u], but also positively support the reliability of the current MFA model.

4. DISCUSSION

The discrepancy between current work and previous studies regarding the assertion that vowels were associated with better alignments than consonants may be accounted for by the effect of segment types. Table 3 further divides the vowel and consonant categories into the following six types: full vowels, weak vowel, complex vowels, plosives, nasals & liquids, and fricatives & affricates. In particular, full vowels include the canonical vowels [a, i, u, e, o]; weak vowels consist of pretonic reduced vowels [ə, ɨ]; while complex vowels are the combinations of vowels [a, i, u] and onglides/offglides [w, j].

Table 3: The mean ORs and MDs of different segment types.

Type	OR	MD
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

According to Table 3, it is obvious that complex vowels have the lowest mean OR and the largest mean MD. An additional linear regression analysis also reveals a statistical significance [$F(5, 2342) = 43.05$, $p < .001$ ***]. As a result, the relatively low performance of vowel alignments is likely to be affected by those segments classified as complex vowels. Previous studies such as [10, 25] also suggested that FA algorithms vary overtly in terms of their accuracy in different types of segments – specifically, semivowels had the greatest displacement among other segment types, which echoes our findings.

On the other hand, the asymmetry between the absence of statistical significance and the less overlapping fitted trajectories occurred in the GAMM results of F2 [i] condition in Figure 2 was likely to be accounted for by the normalization of the vowel duration. As shown in Figure 3, the two critically-misaligned vowels [a] in the word *kinlabang* "very wide" may be resulted from the intrinsic difference in duration. After normalization, formant trajectories might end up being stretched or compressed horizontally. Not to mention the displacement issue – the captured contours could be substantially off. These factors eventually influenced the fitting results of GAMMs.

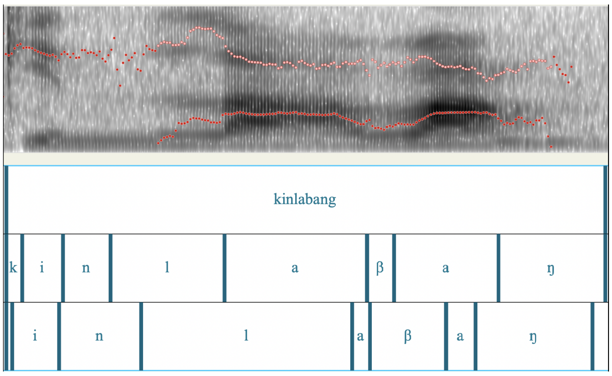

Figure 3: The F1 and F2 trajectories over the word *kinlabang* "very wide". Regarding the two phone tiers, upper tier: manual annotation; lower tier: MFA output.

Overall, the current results reveal that MFA outputs fairly correspond to manual annotations when little but comprehensively labeled data were provided. Furthermore, with the employment of the G2P model training, a function provided by the MFA, the minimum requirement of corpus size for constructing an acoustic model is also worth examining via the data recursion method proposed by [9]. This would call for future studies.

5. REFERENCES

[1] 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.

[2] 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*, 2019, pp. 1–12.

[3] N. J. Young and M. McGarrah, "Forced alignment for nordic languages: Rapidly constructing a high-quality prototype," *Nordic Journal of Linguistics*, pp. 1–27, 2021.

[4] J. Leinonen, N. Partanen, S. Virpioja, and M. Kurimo, "Semiautomatic speech alignment for under-resourced languages," in *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, 2022, pp. 17–21.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011, pp. 1–4.

[6] S. J. Young, "The htk hidden markov model toolkit: Design and philosophy," Ph.D. dissertation, Department of Engineering, University of Cambridge, 1993.

[7] L. M. Johnson, M. Di Paolo, and A. Bell, "Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data," *Language Documentation & Conservation*, vol. 12, pp. 80–123, 2018.

[8] K. Tang and R. Bennett, "Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan)," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 2019, pp. 1719–1723.

[9] 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*, 2018, pp. 145–148.

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

[11] R. Fromont and K. Watson, "Factors influencing automatic segmental alignment of sociophonetic corpora," *Corpora*, vol. 11, no. 3, pp. 401–431, 2016.

[12] R. Coto-Solano and S. F. Solórzano, "Comparison of two forced alignment systems for aligning bribri speech." *CLEI Electron. J.*, vol. 20, no. 1, pp. 2–1, 2017.

[13] D. Barth, J. Grama, S. Gonzalez, and C. Travis, "Using forced alignment for sociophonetic research on a minority language," *University of Pennsylvania Working Papers in Linguistics*, vol. 25, no. 2, article 2, 2020.

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

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

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2021. [Online]. Available: http://www.praat.org/.

[17] Python Core Team, *Python: A dynamic, open source programming language*, Python Software Foundation, 2022. [Online]. Available: https://www.python.org/.

[18] T. Mahrt, "Praatio," 2016. [Online]. Available: https://github.com/timmahrt/praatIO.

[19] W. Styler, "On the acoustical features of vowel nasality in english and french," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/.

[21] M. Wieling, "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*, vol. 70, pp. 86–116, 2018.

[22] P. J. kuei Li, "The phonological rules of atayal dialects," *Bulletin of IHP*, vol. 51, no. 2, pp. 349–405, 1980.

[23] H.-c. J. Huang, "Squliq atayal syllable onset: Simple or complex," *Streams converging into an ocean: Festschrift in honor of Professor Paul Jen-kuei Li on his 70th birthday*, pp. 489–505, 2006.

[24] C. DiCanio, H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. García, "Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.

[25] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in *Proceedings of the 2nd European Conference on Speech Communication and Technology*, 1991, pp. 693–696.