



# Performance of Montreal Forced Aligner on Cantonese Spontaneous Speech

*Ka Ki So<sup>1</sup>, Chenzi Xu<sup>2</sup>, Grace Wenling Cao<sup>3</sup>, Peggy Mok<sup>1</sup>*

<sup>1</sup>The Chinese University of Hong Kong,

<sup>2</sup>University of Oxford,

<sup>3</sup>University College Dublin,

kakiso@link.cuhk.edu.hk, chenzi.xu@ling-phil.ox.ac.uk, grace.cao@ucd.ie,  
peggymok@cuhk.edu.hk

## Abstract

The study presents a comprehensive evaluation of the Montreal Forced Aligner (MFA) in aligning phone boundaries of Hong Kong Cantonese (HKC) spontaneous speech. We developed two tailored Cantonese MFA models, designed to address distinct Cantonese phonetic features, such as checked syllables. These models were applied to align the same set of recordings from spontaneous interviews, and their performance was compared against human annotations. Our results reveal that the updated Cantonese MFA models achieved decent alignment accuracy on spontaneous speech, with a satisfactory level of agreement with manually adjusted boundaries in vowels. However, Cantonese-specific features and connected speech process remain major challenges for the current models. This observation allows us to propose specific amendments to the models to improve alignment performance, as well as recommendations on manual boundary adjustments.

**Index Terms:** speech recognition, forced alignment, Cantonese, phonetic segmentation

## 1. Introduction

The use of forced aligners such as FAVE/P2FA, Prosodylab-Aligner (PLA), and Montreal Forced Aligner (MFA) to assist with word and phonemic segmentation has been common in current corpus studies and phonetic research. While these forced aligners require some manual input, they effectively reduce the time required for time-aligned transcription compared to fully manual segmentation, which could easily take hundreds of hours for a small dataset.

### 1.1. Montreal Forced Aligner(MFA)

Montreal Forced Aligner (MFA) is an open-source forced aligner established in 2017 as an update to the Prosodylab-Aligner [1]. Compared to other aligners, MFA uses the Kaldi Speech Recognition Toolkit instead of Hidden Markov Model Toolkit (HTK), and a Gaussian Mixture Model - Hidden Markov Model (GMM - HMM) architecture [1]. MFA has been consistently reported to achieve higher accuracy in prior evaluations among different forced aligners [2][3]. Regarding speech-text alignment at the phone level, MFA is also considered to be an optimal tool, even compared to newer neural network based aligners such as WhisperX and the Massively Multilingual Speech (MMS) Model [3].

Various studies evaluating available MFA models have reported satisfactory results for languages for which the models are well trained, such as English [1], even with mismatched acoustic models for non-standard varieties [4]. The ability of handling dialectal diversities makes MFA a good tool for pho-

netic research. MFA's typical accuracy rates for English boundary placement ranges from 77% to 90% [1][3][4][5]. Nevertheless, these studies predominantly focus on MFA's performance with read speech instead of spontaneous speech data. It is unclear if equally high accuracy rates can be achieved with spontaneous speech data having much more variation.

MFA currently offers pre-trained acoustic models for 41 languages, but many languages and varieties still lack available models for forced alignment. A pilot acoustic model for Cantonese was developed in 2023 to address this gap [6]. The evaluation on this model found that initial plosives and checked syllables in Cantonese posed the largest challenge to automatic alignment using MFA [7].

### 1.2. Cantonese Phonology

There are 19 consonants and 22 vowels (11 diphthongs) in the Cantonese inventory, with nasals (/m/, /n/, /ŋ/) and stops (/p/, /t/, /k/) appearing in both the onset and coda positions [8]. Checked syllables, syllables with an unreleased stop coda, might be a challenge to MFA due to lack of audible release.

### 1.3. The present study

With reference to prior evaluation results, we trained new models by incorporating additional training data and refining the phone sets employed to improve the model's performance on spontaneous speech, particularly the low accuracy for checked syllables. In this study, we present and evaluate the latest pre-trained MFA models for Hong Kong Cantonese, compared to the pilot model established in 2023 [6]. We aim to improve the model's performance on features that are difficult for MFA to align in the new models. The present paper addresses the following research questions:

- How well can the updated MFA models handle spontaneous Cantonese speech?
- How do different segment types influence the performance of the updated aligner?
- How can language-specific features, such as Cantonese checked syllables, be better handled?

## 2. Method and materials

### 2.1. The Cantonese acoustic models

We trained the new acoustic models combining the validated sets from Common Voice Hong Kong Cantonese Corpus (yue) 19.0 and Hong Kong Chinese Corpus (zh-HK) 19.0, comprising 308.7 hours of speech from 3,400 speakers. Dictionary containing the lexicon in the corpora are created using the *CharsiuG2P* tool [9], with Chinese characters as entries. Before the align-

ment, dictionary entries of Out-of-vocabulary items (OOVs) were added to avoid misalignment resulting from OOVs.

The 2023 Cantonese MFA model [6], trained on the older version of Common Voice Hong Kong Cantonese corpus, achieved only 67% accuracy on checked syllables, highlighting a specific area to improve. Two new models employing the new acoustic models but with slightly different phone sets were developed to explore the impact of different approaches on checked syllables. The differences between the three models are summarized in Table 1. We treated checked syllable as a single unit in 2024 Model 1 due to the lack of audible release. For evaluation, we used the output of 2024 Model 1 for overall accuracy calculations, and compared the performance of 2024 Model 1 against human annotations, the 2023 Model, and 2024 Model 2, with a particular focus on checked syllables.

Table 1: *Details of the Cantonese models*

Version	Checked Syllable	Examples
2023 Model	separated segments	/j/-/a/-/p/, /c/-/oe/-/t/
2024 Model 1	as a unit	/j/-/ap/, /c/-/oet/
2024 Model 2	separated segments	/j/-/a/-/p/, /c/-/oe/-/t/

## 2.2. The dataset

Recordings analysed in this study were collected from eight native Hong Kong Cantonese speakers (4F4M), aged 18 or older. All participants were students recruited from a University in Hong Kong, with no reported hearing or language impairment.

Each participant engaged in a one-to-one interview inside a soundproof booth on campus, providing an approximately 3 to 5 minutes of speech covering a range of Hong Kong Cantonese consonants, vowels, and rime types. Both the interviewer and interviewee spoke Hong Kong Cantonese throughout the recording. The nature of interview task minimized overlapping speech, making these recordings suitable for forced alignment. Only the interviewee’s speech was extracted for analysis.

## 2.3. Procedure

### 2.3.1. MFA alignment

First, we used Capcut Pro to generate speech-to-text transcriptions for each recording [10]. A native Cantonese-speaking research assistant reviewed and manually corrected transcription errors. Additionally, a space was inserted between characters before alignment to enable syllable-level alignment. We then employed the preprocessed transcriptions as input for MFA.

### 2.3.2. The Gold Standard set

The first author and the two student helpers with acoustic-phonetic training created a manually verified “Gold Standard” set by manually adjusting the phoneme boundaries in Praat TextGrids [11] generated using the 2023 Model [6], which serves as the benchmark for evaluation. We chose the 2023 Model output as the baseline for the Gold Standard set to minimize bias towards either of the two 2024 models.

The auditory-acoustic approach was used for boundary adjustment, which involves listening to the speech produced by the interviewees, and analysing acoustic information from the waveforms and spectrograms to determine boundary placements. This study primarily follows the principles outlined in

[12], with some modifications to account for Cantonese phonetic features.

The following principles were consistently applied during boundary adjustments:

- The closure phase of stops and affricates was excluded as the onset is often unclear; only the releases were included.
- Periodic waveforms were used to determine the onset and offset of vowels.
- The V-unreleased stop boundary of checked syllables was ignored due to challenges in determining the onset and offset of unreleased stops.
- The V-N boundary was ignored when it is unclear because of coda reduction.

### 2.3.3. Removal of certain types of segments

Given that Chinese characters were used in the dictionary, alignment accuracy could be affected by MFA’s misidentification of polyphonic words. Therefore, we first compared the labels of intervals in the Gold Standard set and MFA outputs for each recording prior to boundary adjustments. Two segments of misidentified words were excluded for the evaluation of phone boundaries.

Certain types of speech were labeled and excluded from the analysis during boundary adjustment process due to the challenges they posed even for human annotators. After screening, a total of 5692 segments were selected for subsequent analysis. The segments that were removed included creaky voice, breathy voice, syllable fusion, and overlapped speech.

These speech intervals were excluded for two reasons. Apart from the theoretical limit that it is not easy to put boundary for coarticulation, this was also a strategic research decision based on the challenges they presented. This study chose to focus on relatively clearer speech, laying a robust groundwork for future evaluations. Nevertheless, how MFA handles these types of speech remains an important question, which had not been thoroughly investigated in previous studies. We will leave this question to future research.

## 2.4. Measurements and calculations

Overall model accuracy was assessed by comparing the time difference between automatic-aligned boundaries of 2024 Model 1 and the human-annotated boundaries. The absolute time difference between boundary displacements served as the primary metric, allowing for overall accuracy calculations and assessments of different segment types.

### 2.4.1. Boundary Difference (BD)

To evaluate the performance of the 2024 models in comparison to human annotation, we measured the absolute time difference of the boundaries for each selected segment in milliseconds. Boundary Difference (BD) is the primary metric we used to access accuracy, and is calculated as follows.

$$BD = |t_{Manual} - t_{MFA}| \quad (1)$$

A lower BD value signifies higher accuracy in MFA’s boundary placement. Specifically, a positive BD indicates that MFA has placed the boundary earlier than human annotators, while a negative BD indicates the opposite. Boundary Difference is further divided into onset boundary difference and offset boundary difference for accuracy calculations.

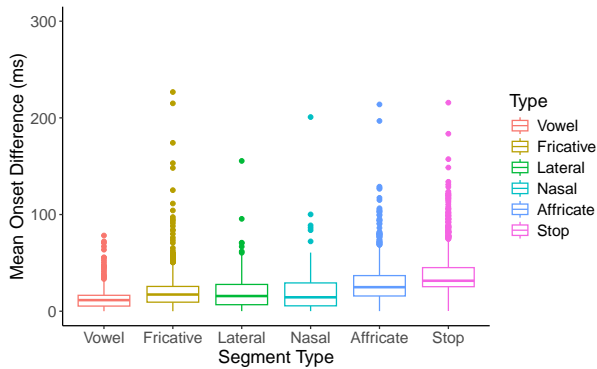


Figure 1: Mean onset difference by segment types

### 2.4.2. Threshold

We evaluated boundary displacement using the golden threshold of 25 ms, as proposed by [1], which was also applied to other MFA evaluation studies [5][13]. Differences greater than 25 ms were considered inaccurate.

## 3. Results

### 3.1. Overall accuracy at phone level

We utilized the output from 2024 Model 1 to assess overall accuracy in comparison to manual annotations. Since the selected segments were not continuous, both onset and offset boundary differences were included in the accuracy calculation.

Within a 25 ms threshold, the 2024 Model 1 achieved a satisfactory accuracy of 73.59% at phone level, which performs slightly better than the 2023 Model (71%) [6]. This closely approximates the phone-level accuracy (77%) reported by [1] and the syllable-level accuracy of Mandarin Chinese (73.49%) [13].

For the eight speakers, the mean onset and end boundary differences were 24 ms (sd = 37) and 20 ms (sd = 39) respectively. Only 3% of the boundaries analyzed were displaced by more than 100 ms, suggesting that the 2024 Model 1 is capable of high accuracy auto-alignment for spontaneous speech. 195 data points with a difference over 100 ms were further examined to investigate the reasons for the high standard deviation. Most of these displacements included obstruent onsets (n = 50) and nasal codas (n = 83). A detailed discussion will be presented in the Discussion section.

### 3.2. Accuracy by segment types

To facilitate direct comparison with the 2023 model, only onset boundary difference was calculated for each segment types. In general, the 2024 Model 1 achieved higher onset accuracy on vowels (93.84%) than consonants (49.12%), with a mean segment onset difference to be 30 ms (sd = 36). Among consonants, the 2024 Model 1 performed the least accurately on stops (23.5%), but better on affricates (50.32%), nasals (67.65%) and laterals (70%). The best performance was found on fricatives (73.64%). Results are summarized in Figure 1. Extreme outliers (Onset Difference >300 ms) were removed from the plot.

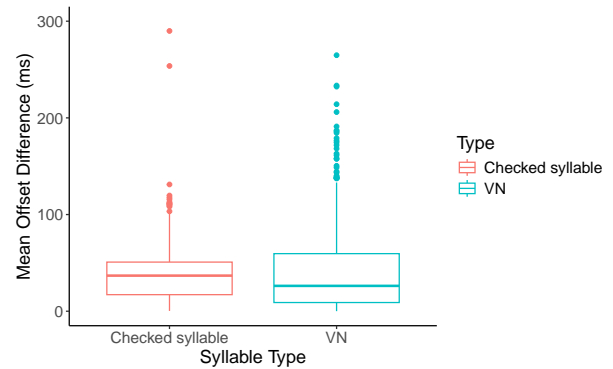


Figure 2: Mean offset difference by syllable types

### 3.3. Accuracy by syllable types

Figure 2 presents 2024 Model 1’s performance of offset accuracy across different syllable types. For VC syllables, i.e. checked and VN syllables, their vowel onsets were comparably high as open syllables (95.51%), with only offset boundary differences (43.15%) showing variation. The mean offset difference of VC syllable is 41 ms (sd = 50). 2024 Model 1 attained an accuracy of 48.9% for VN syllables (e.g. [ɔ̃n] “safe”), which is higher than that for checked syllables (32.15%; e.g. [jɛp] “enter”). Extreme outliers (Offset Difference >300 ms) were removed from the plot.

The performance of 2024 Model 2 on checked syllables was assessed to examine the effectiveness of different segmentation principles. For a fair comparison, we extracted corresponding checked syllables from Model 2’s output and compared them with the human annotations. The 2024 Model 2 attained an accuracy of only 5.54% on checked syllables’ offset boundary accuracy, indicating that treating checked syllables as two separate segments significantly reduce MFA’s performance. The performance disparity between the two models will be addressed in the Discussion section.

## 4. Discussion

Consistent with previous studies, our 2024 Cantonese Model 1 achieved a high auto-alignment accuracy, indicating the ability of MFA models to handle spontaneous speech. However, the presented Cantonese models, designed to account for Cantonese phonetic features, encountered specific difficulties in stops and checked syllables, which is in line with the performance of the initial model [7].

### 4.1. Stop

The notably low accuracy of initial stops may stem from either the phonological context of stop segments or the analytical method employed. Unlike sonorants, the onset of obstruents are inherently less stable and predictable, which is reflected in their low accuracy (22.95%) compared to liquid consonants (68.13%) in the present study. Since all the segments were extracted from continuous speech, most of the initial stops analyzed were intervocalic. The voicing continuation from preceding vowels may affect MFA’s boundary placement, making the model harder to determine stop onsets.

Stops in English, which are phonetically similar to Cantonese stops, however, reported to have higher onset boundary

accuracy in the English models [2][4]. This leads to the consideration of differences in analytical choice between studies. Closures of stops and affricates were mostly included in [4] but excluded in the present study. In our data, 98% of the inaccurate onset boundaries (human-MFA boundary difference greater than 25 ms) had a positive BD, meaning that MFA placed the stop boundaries earlier than human annotators. It is important to note that whether to include the closure of stops and affricates can be an analytical choice rather than a segmentation error. Meanwhile, the plausible effect from different phonological contexts of English and Cantonese stops cannot be ruled out at the point. Further investigations into this issue shall be carried out in the future.

#### 4.2. Checked syllables

Checked syllables continue to pose a major challenge to our Cantonese models. In the 2023 Cantonese models, the acoustic model was trained to treat unreleased final stops as a single segment, independent of the preceding vowels. This approach yielded an offset accuracy of 60%, with 98.7% of the inaccurate boundaries had a negative boundary difference, indicating that their offsets were placed later than manual annotation. The results signified that checked syllables require more human intervention to adjust offset boundaries compared to other segments [7].

The lack of audible releases causes unreleased stops to be often included in its preceding vowel during phonemic segmentation. As we aim to provide a model that can better facilitate segmentation work in Cantonese research, we have made adjustments to better align with practical needs. In 2024 Model 1, we trained the models to treat the vowel-unreleased stop cluster as a single unit, whereas 2024 Model 2 continues to treat the cluster as two single independent segments as the 2023 model did.

Apart from analytical choice, the 2023 model, trained on an older version of the corpus with fewer hours of recording, likely to underperform compared to the 2024 models. When evaluated against human annotations, 2024 Model 1 outperformed Model 2 on checked syllables. To understand the reason for huge difference between the two 2024 models, we extracted the 352 data points which were labelled as inaccurate from the 2024 Model 2 and compared them with the corresponding annotations in the 2024 Model 1. Our results revealed that all inaccurate offset boundaries exhibited a negative BD, indicating that the offset boundaries in the 2024 Model 2 were placed later than human annotation. The discrepancy suggests that by treating the cluster as two segments, the MFA model may have treated the unreleased stops the same way as those released stops, causing a speech-label mismatch that hindered the model’s ability to handle checked syllables. Figure 3 provides an example of how the two 2024 models segment checked syllables differently, illustrating the impact of different segmentation principles on their performance. We also noticed that different segmentation principles may influence their onset boundary accuracy. Our results show that the onset boundary accuracy of vowel followed by an unreleased stop was comparable to the onset boundary accuracy of open and VN syllables when the cluster was treated as one (97.78%) in the 2024 Model 1, but the onset accuracy decreased when it was not (78.05%) in the 2024 Model 2.

#### 4.3. Nasal Codas

The low accuracy on Cantonese nasal coda offsets revealed the challenge for forced aligners to handle connected speech,

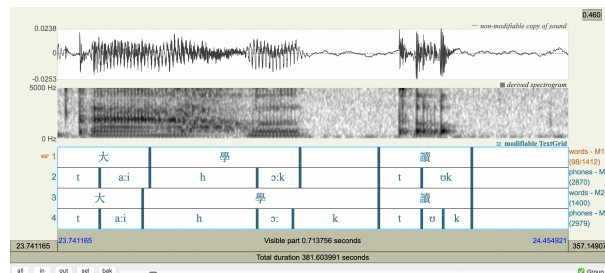


Figure 3: Example of checked syllable annotation by MFA

whereby edges of syllables are blurred by consonant reduction (e.g. incomplete closure or weakening of nasals) or extension (e.g. lengthening of nasal codas). Nasal codas may be realized as nasalized vowels or dropped in rapid spontaneous speech, or assimilate to the place of articulation of the following segment. This phenomenon may create ambiguities of offset boundary for automatic boundary segmentation.

The new models and the 2023 model were trained to treat nasal codas (/m/, /n/, /ŋ/) as separate segments because they typically exhibit clear acoustics cues, such as anti-resonances, low frequencies, and formant transitions. When connected speech process occurs, these typical acoustic cues become inconsistent or absent, leading to mismatch of the acoustic model with coda-reduced syllables. All of the Cantonese models we developed lack training examples of such variants. This may adversely affect MFA’s offset boundary placements by searching for an absent nasal coda.

#### 4.4. Balancing phonetic precision and speech variability

The discussion on speech-label mismatch in checked and VN syllables reveals a challenge in forced alignment—balancing phonetic precision against the articulatory variability of spontaneous speech. Secondary cues, such as vowel nasalization and coarticulatory effects, can also be used to enhance model robustness. Optimizing model accuracy of coarticulation and contextual variability requires training sets with diverse phonetic variants, enabling the model to learn contextual patterns and predict variations. For instance, including triphone model that consider adjacent phonemic contexts, could be a valuable step forward.

### 5. Conclusion

Our latest Cantonese MFA models demonstrate decent automatic alignment at the phone level in Cantonese spontaneous data despite the fact that checked syllable continued to be a major challenge. However, these issues are also challenging for experienced human annotators. It is also suggested incorporating different variants in future models to improve alignment accuracy. The results also indicate the need to prioritize manual adjustment of initial stops and syllable codas.

It is important to note that we pre-screened our speech data, excluded polyphonic words, creaky, breathy, and overlapping speech intervals, to select recordings that are more suitable for forced alignment before evaluation. This means that our accuracy results reflect a relatively optimal situation. Considering the challenges of manually handling spontaneous speech, our Cantonese model still serve as a useful tool for processing a large amount of speech data.

## 6. Acknowledgment

This research was supported by the Postgraduate Studentship (PGS) from the Chinese University of Hong Kong Graduate School, the Leverhulme Trust Early Career Fellowship (ECF-2023-079), and was partially supported by the Hong Kong Research Grants Council General Research Fund (GRF) (Project No. 14612023).

## 7. References

- [1] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldii," in *Interspeech 2017*, 2017, pp. 498–502.
- [2] S. Gonzalez, J. Grama, and C. Travis, "Comparing the performance of forced aligners used in sociophonetic research," *Linguistics Vanguard*, vol. 5, 04 2020.
- [3] R. Rousso, E. Cohen, J. Keshet, and E. Chodroff, "A comparison of modern asr methods for forced alignment," in *Interspeech 2024*, 2024.
- [4] S. Williams, P. Foulkes, and V. Hughes, "Analysis of forced aligner performance on 12 english speech," *Speech Communication*, vol. 158, no. 1, Mar. 2024.
- [5] T. Mahr, V. Berisha, K. Kawabata, J. Liss, and K. Hustad, "Performance of forced-alignment algorithms on children's speech," *Journal of Speech, Language, and Hearing Research*, vol. 64, pp. 2213–2222, 03 2021.
- [6] C. Xu, "Hkcantonese\_models," 2023.
- [7] K. K. K. So, G. W. Cao, C. Xu, and P. Mok, "Analysis of montreal-forced-aligner accuracy on cantonese spontaneous speech," in *The 28th International Conference on Yue Dialects*, 2024.
- [8] R. S. Bauer and P. K. Benedict, *Modern Cantonese Phonology*. De Gruyter Mouton, 1997.
- [9] J. Zhu, C. Zhang, and D. Jurgens, "Byt5 model for massively multilingual grapheme-to-phoneme conversion," 2022. [Online]. Available: <https://arxiv.org/abs/2204.03067>
- [10] "Capcut pro," 2022.
- [11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2025.
- [12] P. Machač and R. Skarnitzl, *Principles of Phonetics Segmentation*. Prague: Epoque Publishing House, 2009.
- [13] H. Wu, J. Yun, X. Li, H. Huang, and C. Liu, "Using a forced aligner for prosody research," *humanities and Social Sciences Communications*, vol. 10, no. 1, p. 429, Jul. 2023.