How well can MFA detect sound change in large-scale sociophonetic research?

Manson Chun Man Wong, Yitian Hong, Peggy Pik Ki Mok
The Chinese University of Hong Kong, Hong Kong SAR

***Introduction**:* In sociophonetics, categorical coding is often used to process speech data to identify the sound variant produced, a method that is appealing for large-scale studies [e.g., 1]. Yet categorical coding is labour-intensive and often involves disagreements, increasing processing time. Existing automatic speech recognition (ASR) tools generally provide reliable word-level transcriptions, but their primary objective is not fine-grained phone-level alignment. Force aligners (FAs), albeit not developed for detecting sound variants, may offer an alternative to fully manual coding. If reliable time alignment and variant detection can be achieved simultaneously using FAs, the time cost of processing large-scale speech data can be substantially reduced. Previous evaluations of FA have largely focused on temporal precision [e.g., 2], with little attention on their ability to detect phonetic variants in languages undergoing sound changes. This study thus asks: How well can the commonly used Montreal Forced Aligner (MFA) detect phonetic variants? Specifically, it examines the performance of a Hong Kong Cantonese (HKC) MFA model [3]. The ongoing sound change features in HKC provide an ideal testing ground for assessing the potential of MFA for future large-scale sociophonetic research.

***Method**:* Speech data were drawn from an ongoing project on HKC sound change across three generations (children aged 9-12, young adults, and elderly), with 16 L1-speakers per group, balanced for gender. Since MFA can only detect variants of monosyllabic words for which pronunciations were specified in the dictionary, words with only one variant listed were excluded. The final dataset comprises of 1684 single-word utterances covering six sound mergers listed in the model (see Figure 1). Outcomes of a coding procedure were compared between the fully manual coding (baseline) and the MFA-assisted (MFA) coding conditions to evaluate MFA performance. In the baseline condition, the initial coding was conducted independently by two human coders, while in the MFA condition, it was conducted by one human coder and MFA. MFA was incorporated at this stage in order to maximize its contribution. Subsequently, for both conditions, all discrepancies were resolved by an additional human coder, ensuring that human remained the final decision-maker. The coded variant for each token was determined by majority rule. MFA performance was assessed by calculating the percentage of tokens with matching labels between the fully manual and MFA conditions.

***Results and Discussion**:* Overall, MFA-assisted coding yielded results comparable to those obtained through fully manual coding. Across all age-by-gender subgroups, at least three of the six sound mergers exhibited exactly identical variant labels (see Figure 1). However, MFA performance was not uniform across social groups or sound categories. When results were stratified by speaker age and gender, MFA detection was more reliable for children's and female speech data respectively. While five sound mergers reached similarity of 80% or above in each subgroup, substantial variation was observed for the [ʃ-]–[s-] merger, with similarity ranging from 29% to 100%. Taken together, MFA approximates human coding to a large extent, showing its potential as a complementary tool in large-scale sociophonetic research. The social group- and sound-specific performance, possibly stemming from biases in the training data (e.g., a higher proportion of female speech reported as the training data in this MFA) [e.g., 4 in ASR], highlights the need for model evaluation that explicitly targets variant-detection accuracy. Ongoing work extends this analysis to the full dataset (n = 120) and incorporates acoustic analyses.

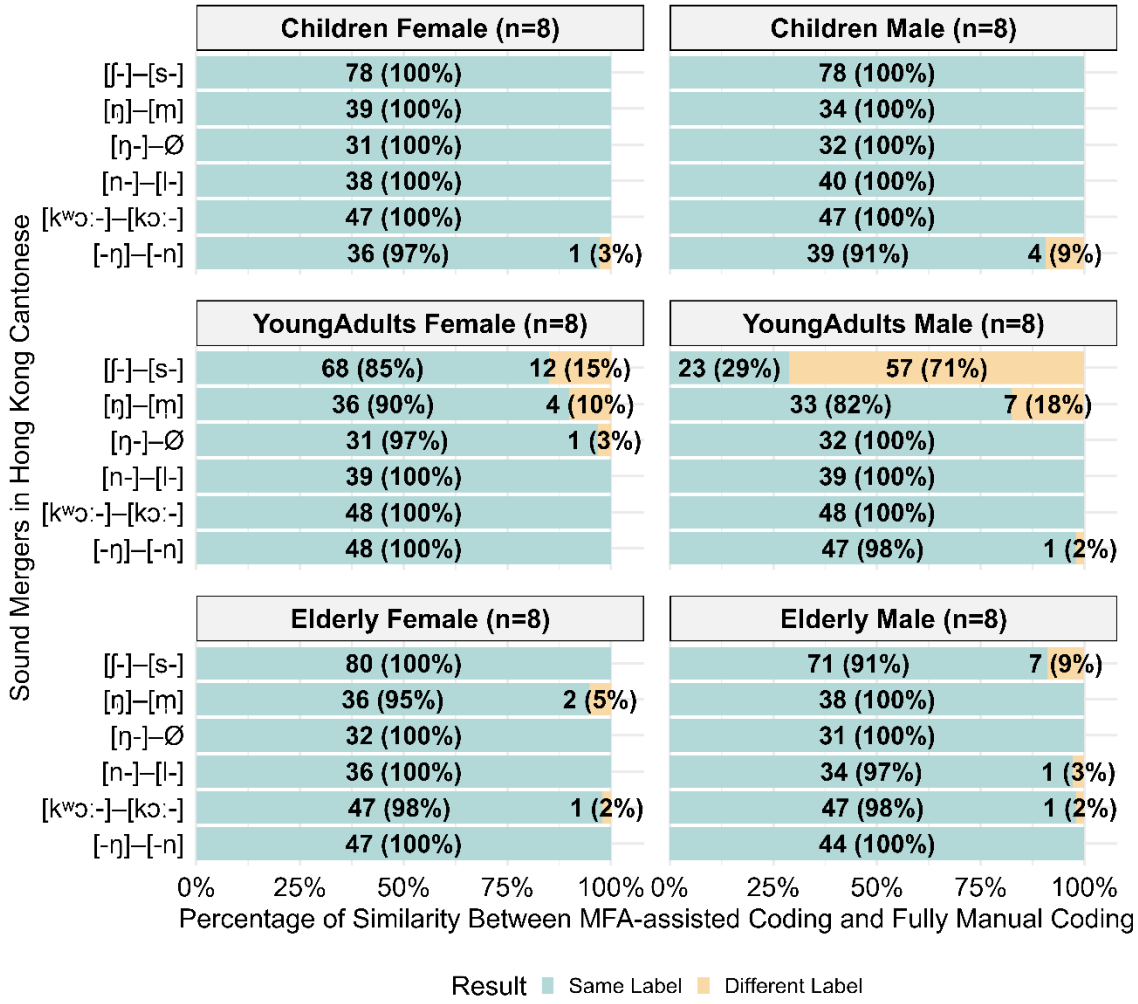

Figure 1. Comparison of similarity between MFA-assisted coding to fully manual coding

References

[1] J. N. Stanford, *New England English: Large-Scale Acoustic Sociophonetics and Dialectology*. Oxford University Press, 2019. doi: 10.1093/oso/9780190625658.001.0001.

[2] S. Gonzalez Ochoa, J. Grama, and C. Travis, "Comparing the performance of forced aligners used in sociophonetic research," 2020, [Online]. Available: http://hdl.handle.net/1885/261248

[3] C. Xu, *HKCantonese_models*. (2024). [Montreal Forced Aligner]. Available: https://github.com/chenchenzi/HKCantonese_models

[4] M. Garnerin, S. Rossato, and L. Besacier, "Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance," Aug. 23, 2019, *arXiv*: arXiv:1908.08717. doi: 10.48550/arXiv.1908.08717.