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

Sociophonetic Studies - Very Diverse

- → Different purposes, features, and languages

Research Purpose	Feature(s)	Language/Situation	Studies
Regional Variation	Insertion of [x]	A contact variety in Hohhot	Wang, 2019
	Interdentals	Arabic variety in Hijaz	Al-Essa, 2022
Stylistic Variation	Release of /t/	American English (AmE) political speech	Podesva et al., 2015
	Release of /t/, /-ŋ/ in '-ing'	AmE political speech	D'Onofrio & Stecker, 2020
	3 sound change features	Hong Kong Cantonese (HKC) in speech registers	Bourgerie, 1990
Synchronic Variation	9 sound change features	Cross-generational comparison of HKC in the mid-2000s	To et al., 2015
	3 sound change features	Cross-generational comparison of HKC in the mid-2010s	Cheng et al., 2022
	Rhoticity	New England English	Stanford, 2019

- → Despite the diversity, 2 things are similar…

Sociophonetic Studies - Something Common

- → 1. Often involve **a LOT of datapoints** **from** **multiple speakers**

- Regional/Dialectal Variation, e.g.,
  - **1957** tokens, **67** speakers (Wang, 2019)
  - **5386** tokens, **122** speakers (Al-Essa, 2022)
- Stylistic Variation, e.g.,
  - **5444** tokens, **6** speakers (Podesva et al., 2015)
  - **13465** tokens, **3** speakers (D'Onofrio & Stecker, 2020)
  - **~3500** tokens, **14** speakers (Bourgerie, 1990)
- Synchronic Variation, e.g.,
  - **6750** tokens, **250** speakers (To et al., 2015)
  - **1122** tokens, **51** speakers (Cheng et al., 2022)
  - **32734** tokens, **367** speakers (Stanford, 2019)

Sociophonetic Studies - Something Common

- → 2. More importantly, the data are all **manually** and **categorically coded**!
- Regional/Dialectal Variation, e.g.,
  - "all tokens were … **hand coded**" (Wang, 2019)
  - "The data were **auditorily analysed**" (Al-Essa, 2022)
- Stylistic Variation, e.g.,
  - "Each token … was **coded** for its realization by **auditory** analysis" (Podesva et al., 2015)
  - "All collected speeches were **coded**…" (D'Onofrio & Stecker, 2020)
  - "All three … are … **dichotomous** (i.e., …only two possible realizations)" (Bourgerie, 1990)
- Synchronic Variation (Sound change), e.g.,
  - "**two SLPs re-transcribed** and scored all the speech samples…" (To et al., 2015)
  - "To assess production, **two … speakers** … **coded** the onset…" (Cheng et al., 2022)
  - "…, each (r) token was **coded** by **two human judges**…" (Stanford, 2019)
- → WHY?

Manual Categorical Coding - Pros

- A method **appealing** for large-scale sociophonetic studies
  1. Easy to implement
     - → Can just do with an Excel sheet
  2. Easy to interpret
     - → Usually percentage-based with additional analysis
  3. Faster than segmentation
     - → No need to manually adjust the boundaries, just play the audio
  4. Less technical
     - → whether phonetically-trained, you can hear the difference as long as you are a native speaker of that language
  5. **Listener-based**
     - → reflect how human hear and process these sociolinguistic variants

Manual Categorical Coding - Cons

- **Time-consuming + Labour-intensive**, especially if
    1. there are a lot of datapoints (very common)
    2. a lot of disagreements are involved
    3. native speakers of the language under studied are scarce in the community
- → Any alternatives beyond **manual** coding?

Manual Coding - Alternatives?

- Yes, Random Forest Classification (Villarreal et al., 2020)
  - Sociolinguistic variants
    - 1. Non-prevocalic /r/; 2. Word-medial intervocalic /t/ in English
  - Results: Overall accuracy ranges from 85.5% to 91.8%
    - → Closely resemble how human perceive the variants

- **Yet, 8907 tokens needed to be hand-coded first to build that classifier**
- → Any (less powerful?) existing alternatives to further simplify the process?

Manual Coding - Forced Aligner as an alternative?

- **Forced Aligner (FA)**
  - Not primarily developed for detecting sound variants
  - If FAs can achieve reliable time alignment and **variant detection** at the same time, the time cost of processing large-scale speech data can be substantially reduced
    - Some FAs have a pronunciation dictionary with variants explicitly encoded

- → **Montreal Forced Aligner (MFA)** offers a feasible alternative (McAuliffe et al., 2017)
  - Often used in large-scale data for time-alignment
  - Offer **variants encoding** in the files (.textgrid) at the same time

Manual Coding - Forced Aligner as an alternative?

- Evaluations of FA have largely focused on <u>temporal precision</u>
  - → Little attention on their ability to detect phonetic variants (e.g., Bailey, 2016, Kendall et al., 2021)

- → **Research Question**: Can MFA be a good candidate for categorical coding?
  - More specifically, we are asking…
    i. Can MFA replace the **<u>whole</u>** manual process? (***<u>Sole</u>*** Coder)
    ii. If not, can MFA at least replace **<u>part</u>** of it? (***<u>Assistant</u>*** Coder)

Method - Language & MFA

- **Hong Kong Cantonese (HKC)**
  1. **Multiple** ongoing sound change features (Synchronic Variation)
  2. **Existing** MFA model for Cantonese (Xu, 2024) covers 6 of the HKC mergers
     a) Onset [ŋ-]-[Ø-], e.g., 牛 [**ŋ** ɐu] or [ɐu]
     b) Onset [n-]-[l-], e.g., 你 [**n** ei] or [**l** ei]
     c) Onset [kʷ-]-[k] before rounded vowels, e.g., 國 [**kʷ** ɔː k] or [**k** ɔː k]
     d) Onset [s-]-[ʃ-] before rounded vowels, e.g., 書 [**s** yː] or [**ʃ** yː]
     e) Coda [-ŋ]-[n], e.g., 想 [s œː **ŋ**] or [s œː **n**]
     f) Syllabic [ṅ]-[m], e.g.,五 [**ṅ**] or [**m̩**]
- → **HKC** provides an ideal testing ground for assessing the potential of **MFA** for future large-scale sociophonetic research

Method - Data

- → Extracted from our larger project on phonetic accommodation

Research Purpose	Feature(s)	Language/Situation	Datapoints, Speakers
Synchronic Variation	1. Onset [ŋ-]-[Ø-] 2. Onset [n-]-[l-] 3. Onset [kʷ-]-[k] 4. Onset [s-]-[ʃ-] 5. Coda [-ŋ]-[n] 6. Syllabic [ŋ̍]-[m]	Cross-generational comparison of HKC in the mid-2020s	**1673**, **48** • 16 Children (10.8±0.8) • 16 Young Adults (22.7±3.6) • 16 Elderly (67.8±4.7)

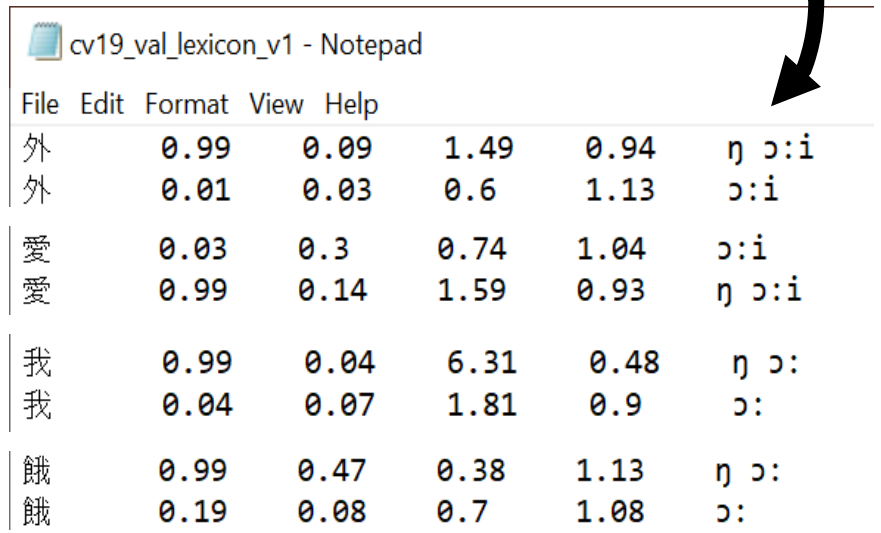
- → As an **exploratory** study
  - **16 participants (8F, 8M)** in each group were selected randomly
  - → A total of **1673** tokens from **48 speakers**
    - Exclude tokens that cannot be coded (e.g., mispronunciation, missing the target phoneme, n = 55)

Feature	Character Count
[ŋ̩]–[m̩]	5
[n-]–[l-]	5
[kʷɔː-]–[kɔː-]	6
[ŋ-]–Ø	4
[ʃ-]–[s-]	10
[-ŋ]–[-n]	6

Method - Coding Extraction in 5 steps

1. Only included characters with both variant pronunciations specified in the dictionary
2. Create .txt file for each character
3. Place the .txt files into each participant's folder
4. Run MFA
5. Extract the variant labels from .textgrid using a Praat script

- **→ ~30 minutes (very fast!)**



Method - MFA's Target

⭐Gold Standard: Full Manual Coding

- Previous work often relied on single-person categorical coding

→ A 2-step procedure was used in this study

- Step 1: **Two coders** evaluated the data independently
- Step 2: **The 3rd coder** made the final decision
- → More robust for upcoming comparisons

1st & 2nd Coder (Main Round)	→	**3rd Coder (Disagreement)**
$Human_A$ + $Human_B$		$Human_C$

Method - MFA's Competitor

📚Literature-based naïve baseline (Reading)

Single-Output variant solely based on literature

- If reading is better than using MFA, MFA probably is not a good idea

	Childre n	Young Adults / Elderly		References
[ŋ̩]–[m̩]	**[m̩]**	**[m̩]**		To et al. (2015); Cheng et al. (2022)
[n̩-]–[l̩-]	**[l̩]**	**[l̩]**		To et al. (2015); Cheng et al. (2022)
[kʷɔː-]–[kɔː-]	**[kʷ]**	**[k]**		To et al. (2015)
[ŋ-]–Ø	**[Ø]**	**[ŋ-]**	**[ŋ-]**	To et al. (2015); Cheng et al. (2022)
[ʃ-]–[s-]	**[s]**	**[s]**		Yu (2016)
[-ŋ]–[-n]	**[ŋ]**	**[ŋ]**		To et al. (2015)

Method - Possible role for MFA

i. MFA as the sole coder

- Context: MFA alone as the **final decision maker** → replace **ALL** the work
- → If …
  - **better** than its peer competitor (‘Literature-based naïve baseline ’) &
  - **comparable** to the target (‘Gold Standard’)
  - → it has the potential to work alone
- → If not, maybe as an assistive role? **→ 2.**

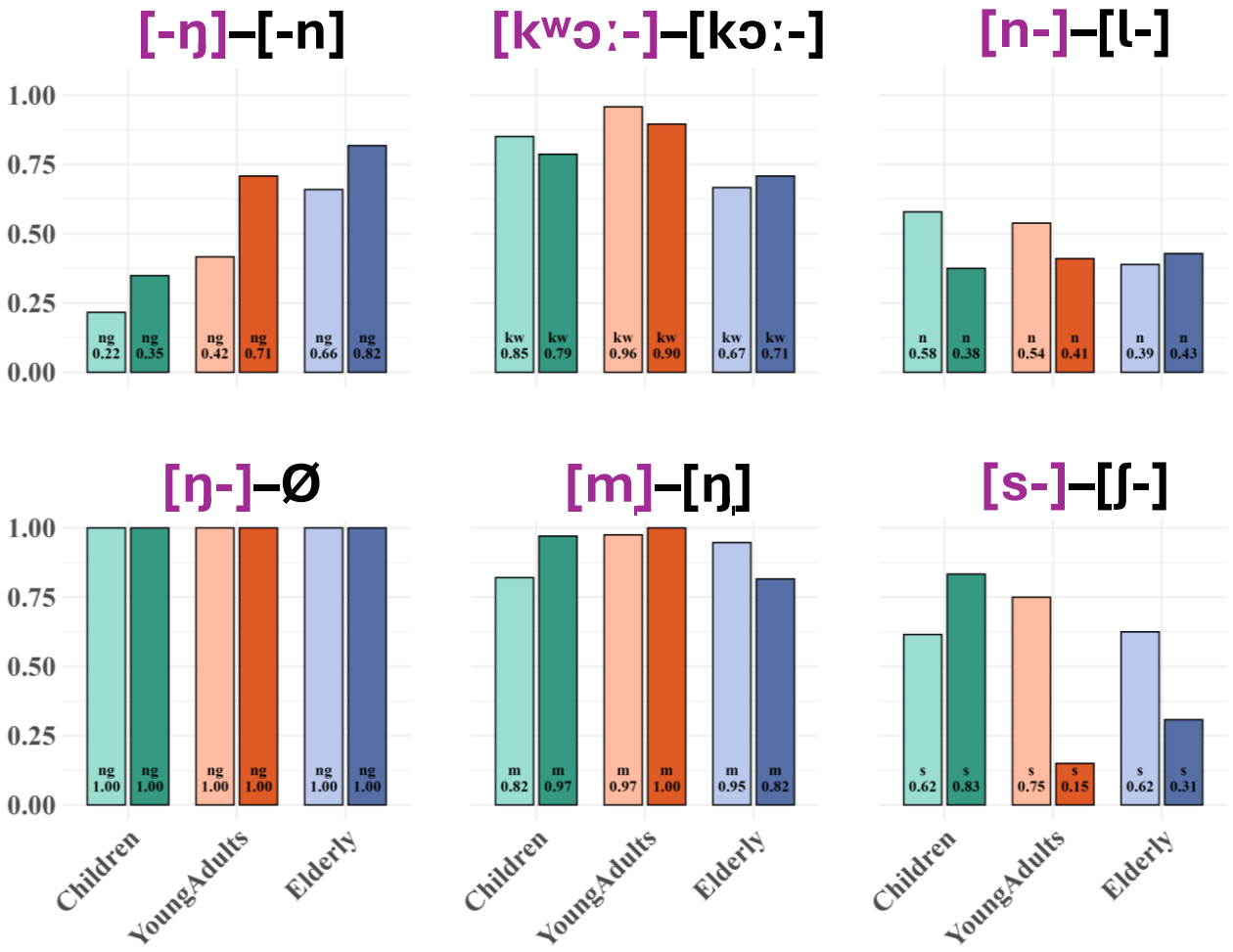
ii. MFA as the assistant coder

- Context: MFA as a **helper** → replace **PART** of the work
- → If …
  - **better** than its peer competitor (‘ Literature-based naïve baseline’) &
  - **comparable** to the target (‘Gold Standard’)
  - → it has the potential work as an assistant

iii. If neither works → Keep it manual!

Results - Is MFA trying?

- Yes, MFA is **attempting** to categorize for most features, **except onset [ŋ-]-[Ø-]**



[ŋ-]–Ø

cv19_val_lexicon_v1 - Notepad

File Edit Format View Help

外	0.99	0.09	1.49	0.94	ŋ ɔ:i
外	0.01	0.03	0.6	1.13	ɔ:i
愛	0.03	0.3	0.74	1.04	ɔ:i
愛	0.99	0.14	1.59	0.93	ŋ ɔ:i
我	0.99	0.04	6.31	0.48	ŋ ɔ:
我	0.04	0.07	1.81	0.9	ɔ:
餓	0.99	0.47	0.38	1.13	ŋ ɔ:
餓	0.19	0.08	0.7	1.08	ɔ:

Note: female (light bar); male (dark bar)

MFA as the sole coder

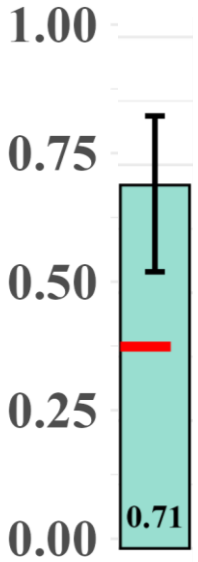
Scenario	1st & 2nd Coder (Main Round)	3rd Coder (Disagreement)	Condition
1	$Human_A$ + $Human_B$	$Human_C$	Fully Manual Coding (human as final decision maker)
2	**MFA**		**Full MFA Coding**

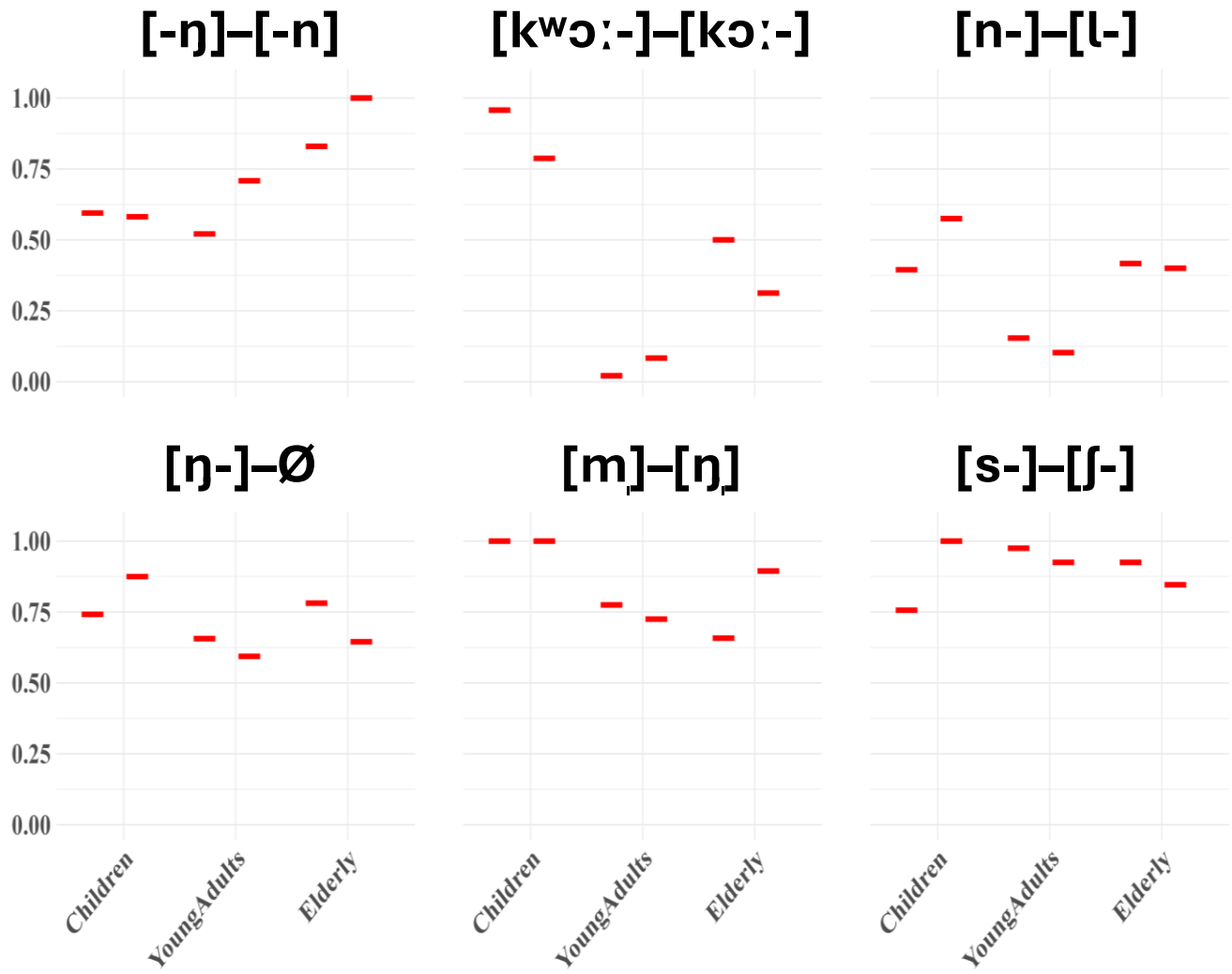
Results - Sole Coder

1. vs. 📚 Literature-based naïve baseline in overlapping with ⭐ Gold Standard

- Mcnemar's Test (McNemar, 1947)
- ---- = 📚's percentage overlap with ⭐
- bar = MFA's percentage overlap with ⭐

- **If MFA's overlap > 📚's overlap**
  - **= MFA is better than reading**
  - E.g.,





Results - Sole Coder

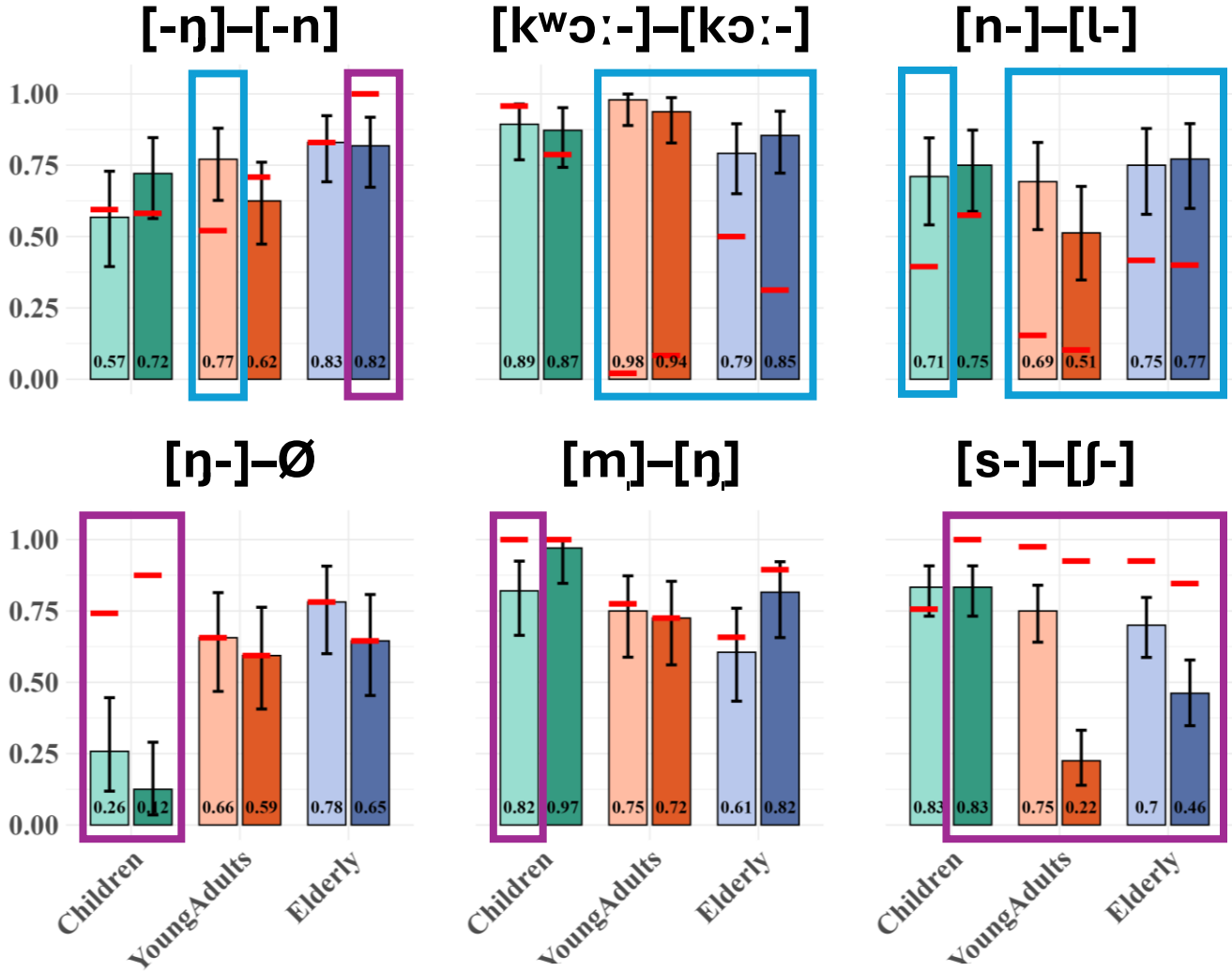
1. vs. 📚Literature-based naïve baseline in overlapping with ⭐Gold Standard

- Mcnemar's Test (McNemar, 1947)
- ---- = 📚's percentage overlap with ⭐
- bar = MFA's percentage overlap with ⭐

MFA performs sig. better than the naïve baseline in 10 out of 36 subgroups

The naïve baseline performs sig. better than MFA in 9 out of 36 subgroups

No diff. between MFA & the naïve baseline in 17 out of 36 subgroups
→ Comparable to the naïve baseline



Note: female (light bar); male (dark bar)

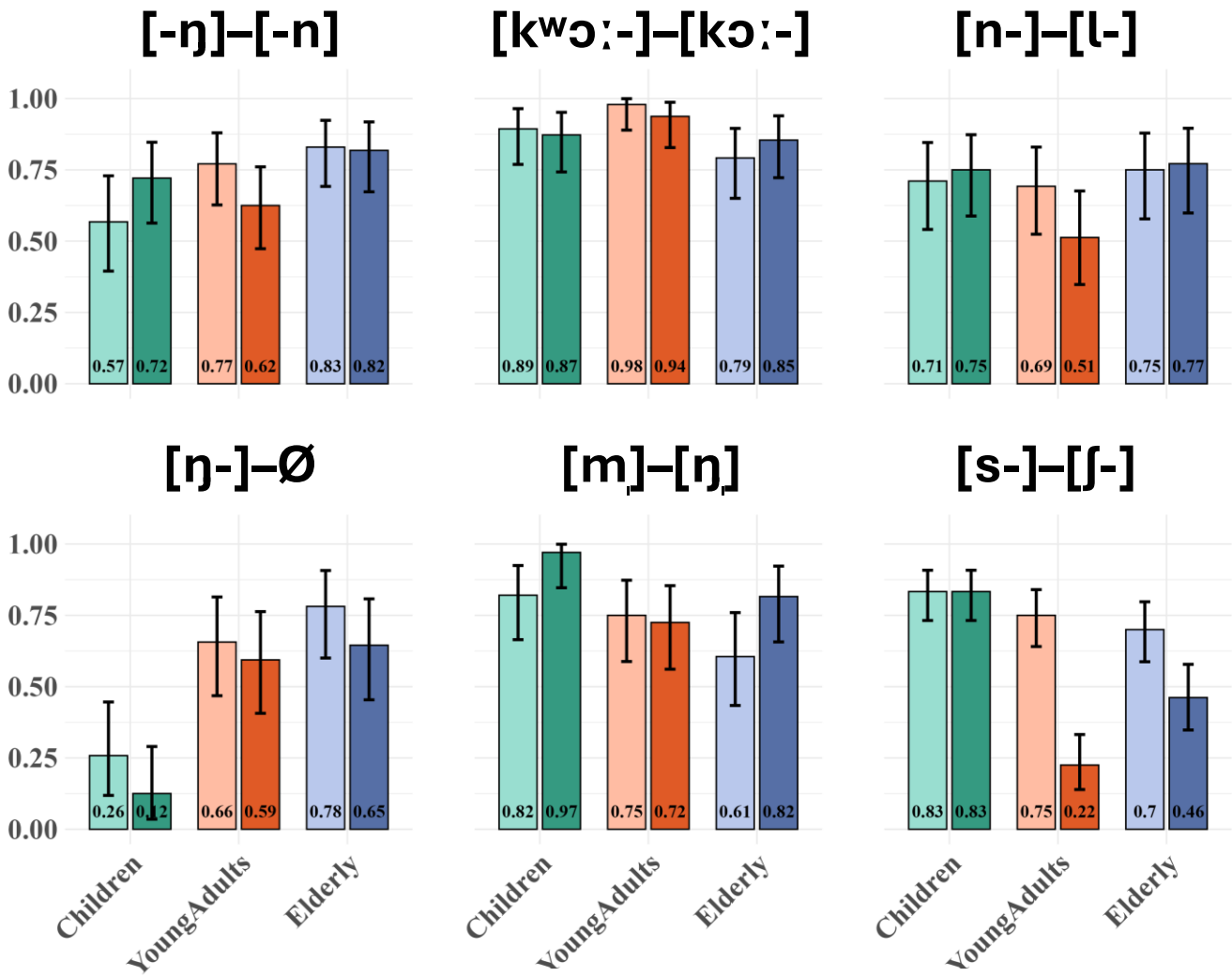
Results - Sole Coder

2. Percentage overlap with ⭐Gold Standard

→ **Substantial** variation by…

Gender	Female: **22% - 98%**
	Male: **12% - 97%**
Age Group	Children: **12% - 97%**
	Young Adults: **22% - 98%**
	Elderly: 46% - 85%
Feature	[-ŋ]-[-n]: 57% - 83%
	[kʷ-]-[k-]: 79% - 98%
	[n-]-[l-]: 51% - 77%
	[ŋ-]-[Ø-]: **12% - 78%**
	[m̩]–[ŋ̩] : 61% - 97%
	[s-]-[ʃ-]: **22% - 83%**

→ **Not comparable to Manual Coding**

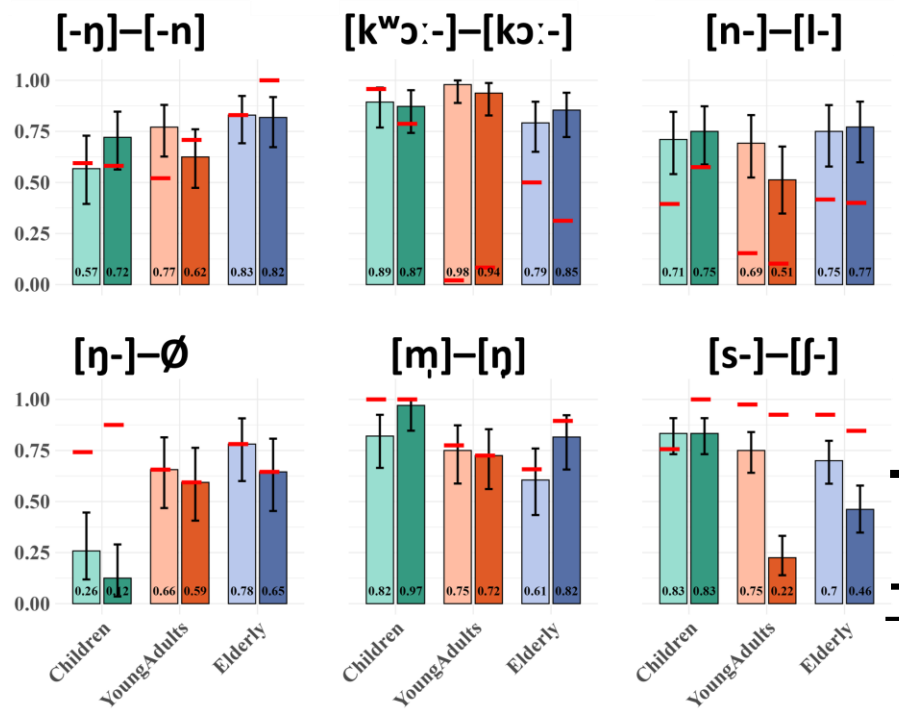


Note: female (light bar); male (dark bar)

Results - Interim Summary

i. MFA as the sole coder

- Fail to detect **1** of the 6 features
- Comparable to 📚 **naïve baseline (----)**
- Substantial variation by social groups & feature
- Not comparable **to** ⭐**Gold Standard**



→ **Not ideal as the sole coder**

→ What about as an assistant? (ii.)

MFA as the assistant coder

→ Given the poor performance of MFA as the sole coder (results in (i)), human should be the final decision maker

Scenario	**1st & 2nd Coder (Main Round)**	**3rd Coder (Disagreement)**	**Condition**
1	$Human_A$ + $Human_B$	$Human_C$	Fully Manual Coding (human as final decision maker)
2	MFA + $Human_A$	$Human_B$	**MFA-Assisted Coding** (human as final decision maker)

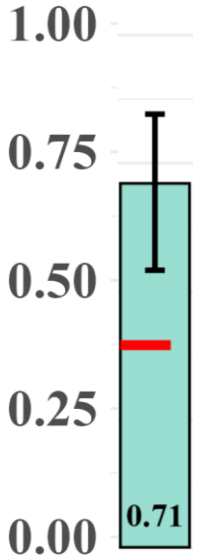
→ Only Scenario 2 will be compared to Scenario 1

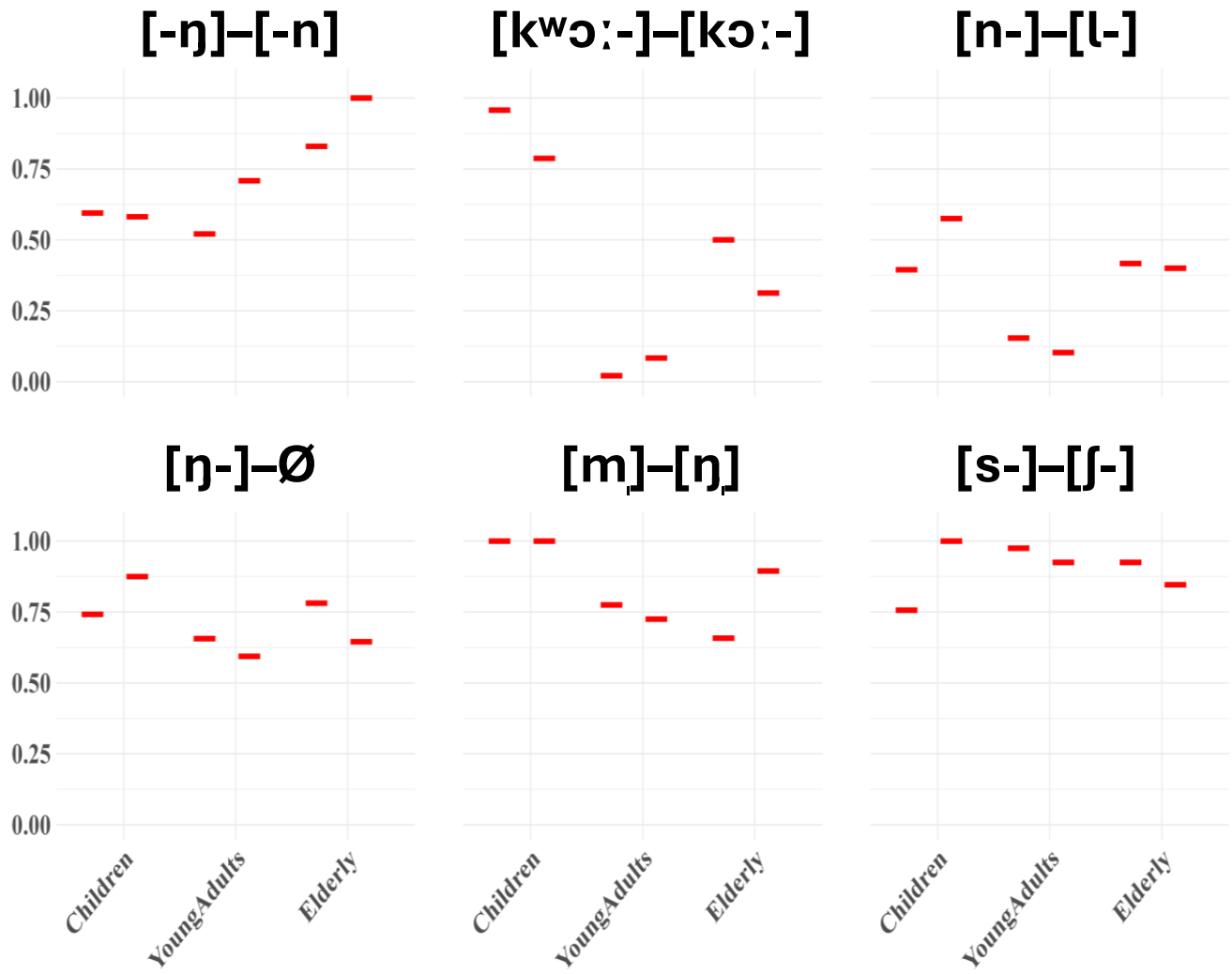
Results - Assistant Coder

1. vs. Literature-based naïve baseline in overlapping with Gold Standard

- Mcnemar's Test (McNemar, 1947)
- ---- = 's overlap with
- bar = MFA's percentage overlap with

- **If MFA's overlap > 's overlap**
  - **= MFA is better than naïve baseline**
  - E.g.,





Results - Assistant Coder

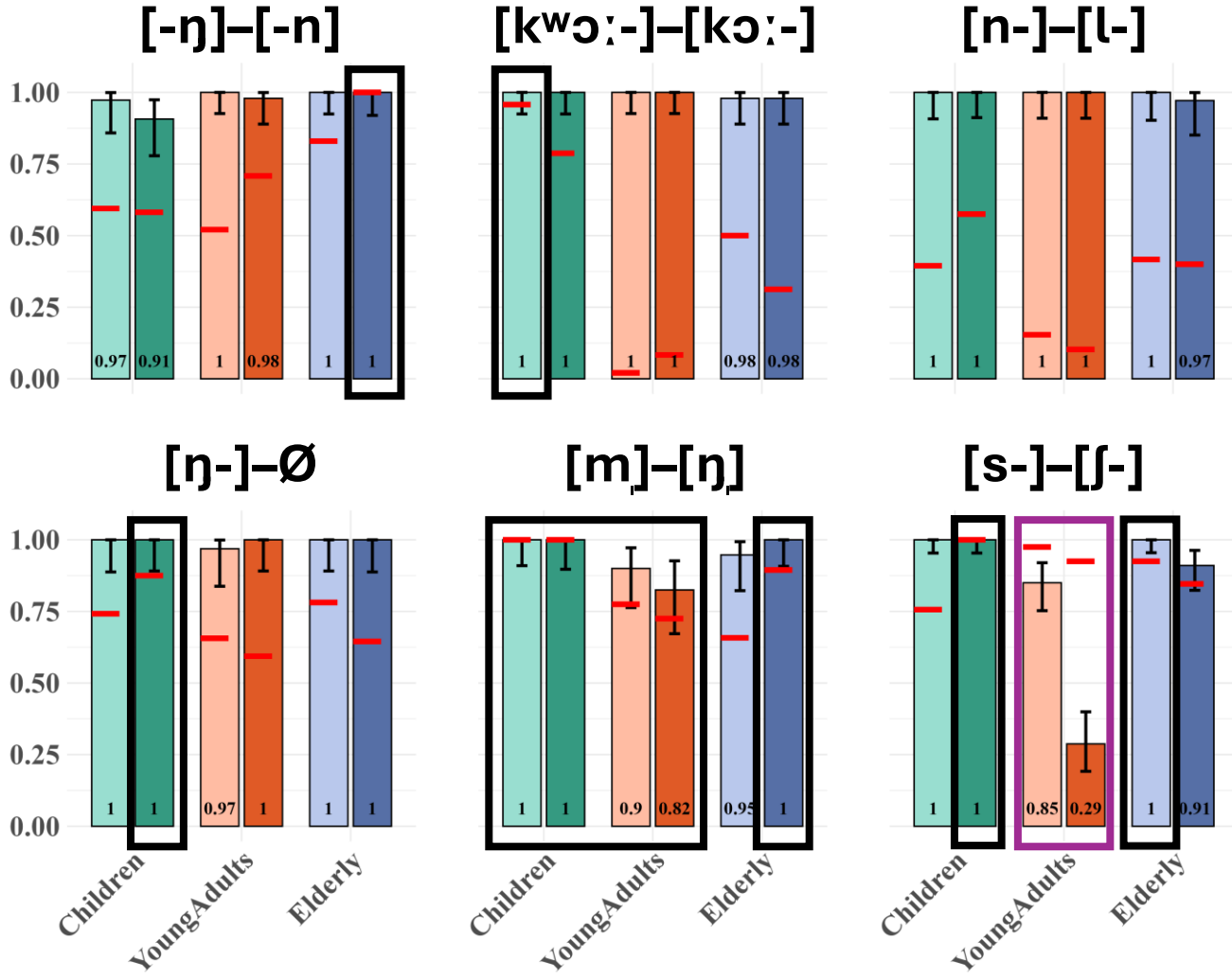
1. vs. 📚 Literature-based naïve baseline in overlapping with ⭐ Gold Standard

- Mcnemar's Test (McNemar, 1947)
- ---- = 📚's overlap with ⭐
- bar = MFA's overlap with ⭐

The naive baseline performs sig. better than MFA in 2 out of 36 subgroups

No diff. between MFA & the naïve baseline in 10 out of 36 subgroups

MFA performs sig. better than the naïve baseline in 24 out of 36 subgroups
→ Much better than the naïve baseline



Note: female (light bar); male (dark bar)

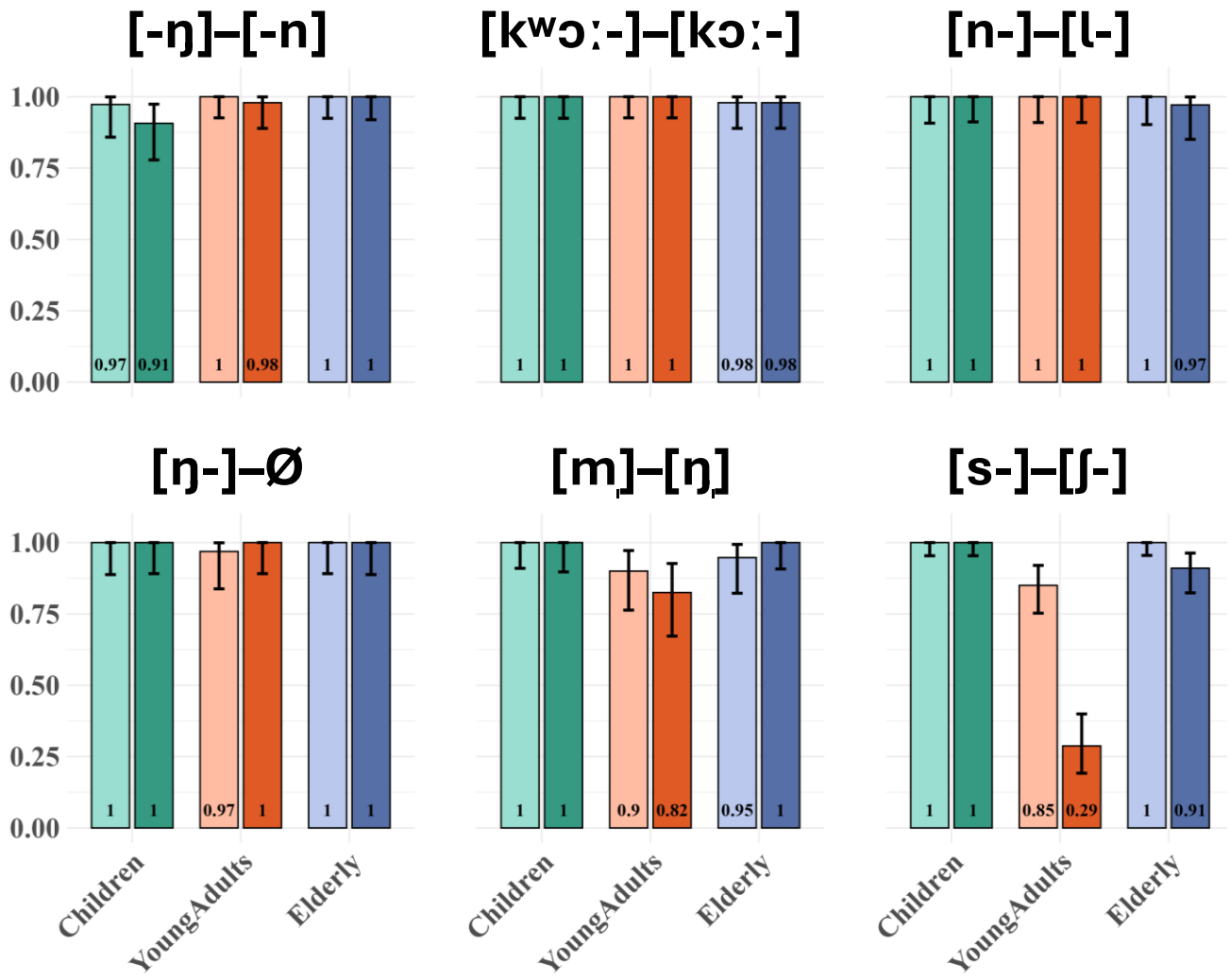
Results - Assistant Coder

2. Percentage Overlap with ⭐Gold Standard

→ **Much less** variation

Gender	Female: **85% - 100%**
	Male: **29% - 100%**
Age Group	Children: **91% - 100%**
	Young Adults: **29% - 100%**
	Elderly: **91% - 100%**
Feature	[-ŋ]-[-n]: **91% - 100%**
	[kʷ-]-[k-]: **98% - 100%**
	[n-]-[l-]: **97% - 100%**
	[ŋ-]-[Ø-]: **97% - 100%**
	[m̩]–[ŋ̩] : **82% - 100%**
	[s-]-[ʃ-]: **29% - 100%**

→ **Comparable to Manual Coding**

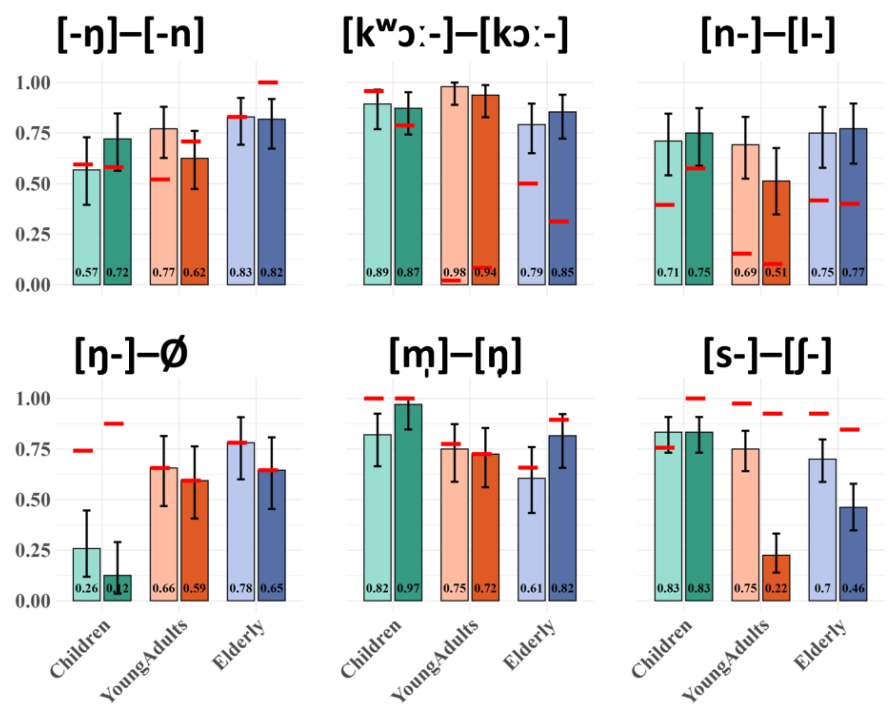


Note: female (light bar); male (dark bar)

Results - Summary

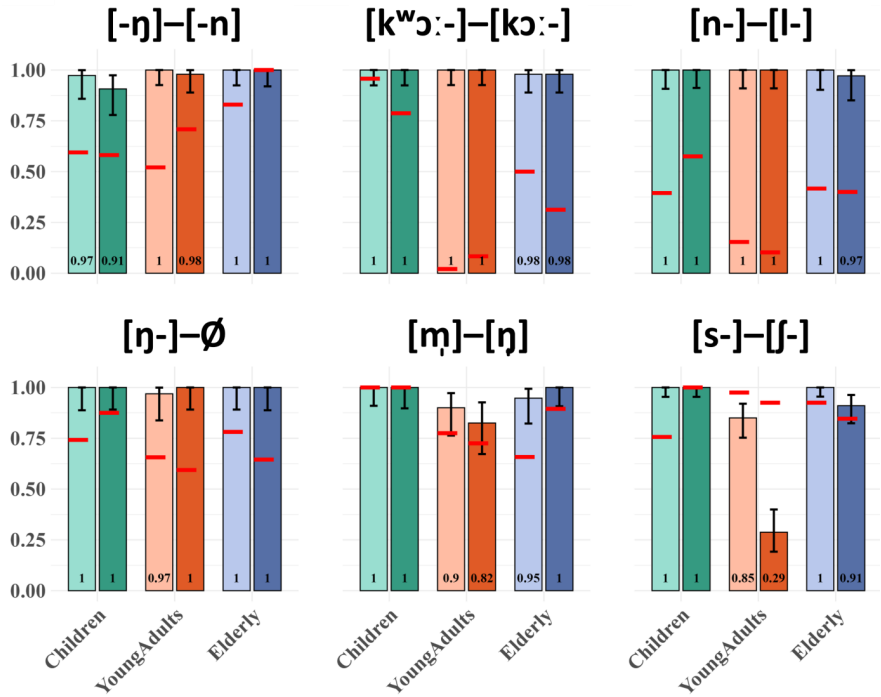
i. MFA as the sole coder

- Fail to detect **1** of the 6 features
- ❌ Comparable to 📚 **naïve baseline (----)**
- ❌ Substantial variation by social groups & feature
- ❌ Not comparable **to** ⭐**Gold Standard**



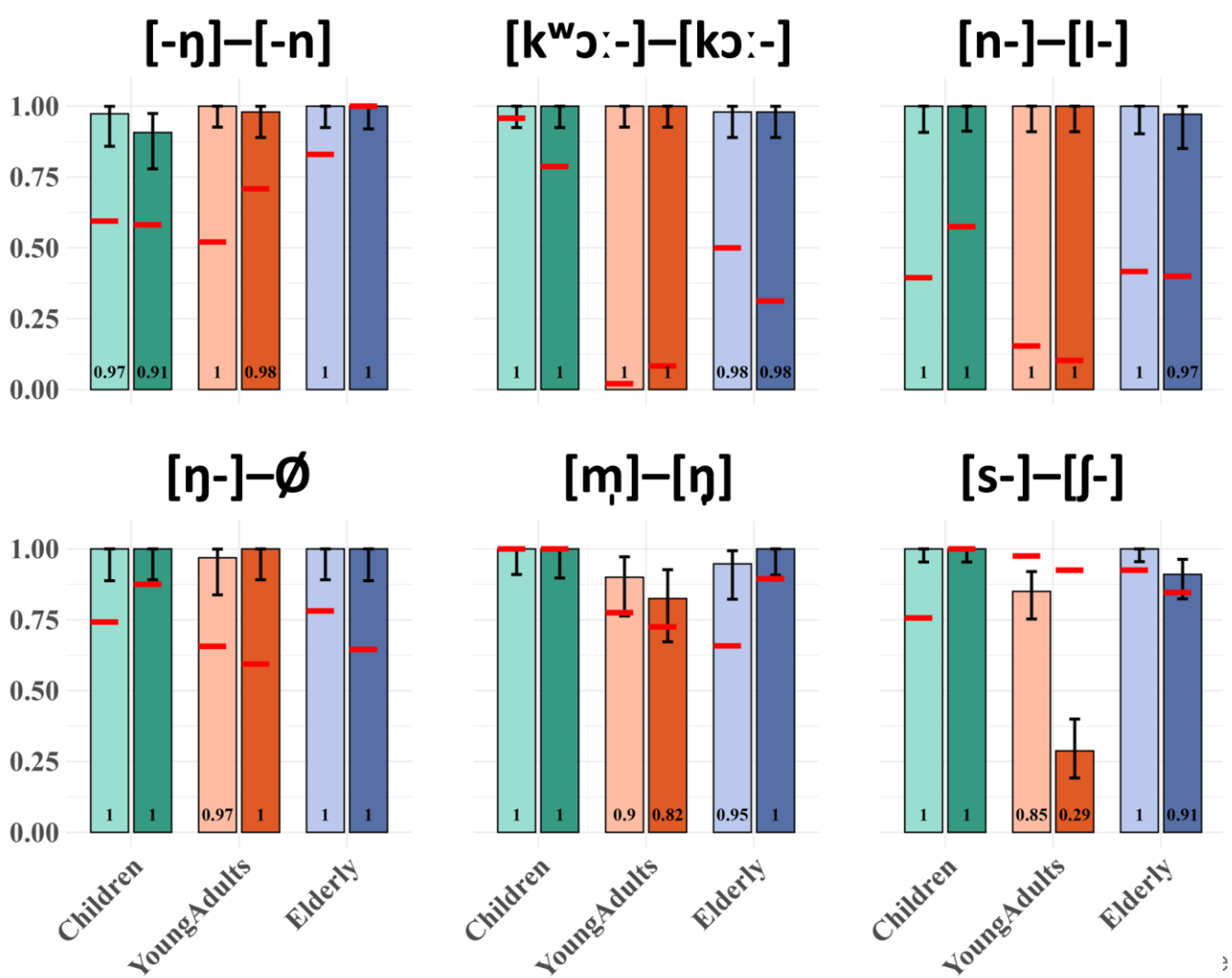
ii. MFA as the assistant coder

- ..... (yet ✓recoverable by human as 3rd coder)
- ✓ Much Better than 📚 **naïve baseline (----)**
- ✓ Much less variation by social groups & feature
- ✓ Comparable **to** ⭐**Gold Standard**



Discussion - MFA as assistant?

1. Despite better performance, ***Lower accuracy*** *for specific subgroups?*

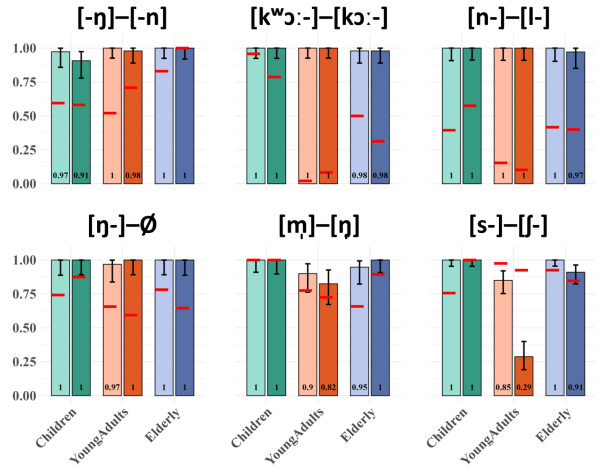


Discussion - MFA as assistant?

*1. Despite better performance, **Lower accuracy** for specific subgroups?*

Match/Mismatch ~ Gender + Age Group + Feature + (1 | Word) + (1 | participant)

- Gender: Female > ***Male***
- Age Group: Children / Elderly > ***Young Adults***
- Feature: [n-]–[l-] / [ŋ-]–Ø / [-ŋ]–[-n] / ([kʷɔː-]–[kɔː-] > [ŋ̩]–[m̩]) > ***[s-]–[ʃ-]***



Potentially due to training data bias? (e.g., Garnerin et al., 2019)

- E.g., fewer male clips → poorer male performance

Training data	**Clips (Female; Male)**	**Age Group (Children; Young Adults; Elderly)**
Hong Kong Chinese Corpus (zh-HK)	20,892; 50,176	1999; 81,658; 600
Cantonese Corpus (yue)	174,436; 47,656	2901; 40,912; 809
	Female > **Male**	***Young Adults*** > Children > Elderly

→ But it cannot explain why performance is poorer in young adults → to be further explored

Conclusion

- **Research Question**: Can MFA be a good candidate for categorical coding?
  - → It depends!

i. MFA as the sole coder (final decision maker)

- → Not a good idea
  - 📚 Literature-based naïve baseline (Reading) maybe more cost-effective

ii. MFA as the assistant coder (main round coder)

- → Much better and more promising, but more work is needed to optimize the model
  - minor flaws (poorer performance towards specific social groups & features)
- → Theoretically more reasonable! (listener-based, human as the final decision maker)

Acknowledgements

Hong Kong Research Grants Council General Research Fund (Project No. 14612023)
Human vs. AI: Speech register and speech accommodation in human-machine interaction

CUHK Research Committee Postdoctoral Fellowship Scheme 2023–24
Hong Kong RGC Postdoctoral Fellowship

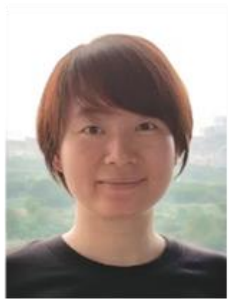
Principal Investigator

Peggy Mok

Collaborators

Grace Cao

Tan Lee



Yusheng Tian

Team Members

Yitian Hong

Xinyi Chen

Yuanlin Yang

Ann To

Karen So

Manson Wong

Miko Ng

Thank you and comments are welcome!

Abstract (& Slides) are available at https://mansonwongcm.github.io/