

香港中文大學
The Chinese University of Hong Kong



UNIVERSITY
of York



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Investigating the evidential value of filled pauses in cross-language forensic voice comparison

Grace Wenling Cao¹, Vincent Hughes², Bruce Xiao Wang,³ Peggy Mok¹

¹Department of Linguistics and Modern Languages, Chinese University of Hong Kong

²Department of Language and Linguistic Science, University of York

³Department of English and Communication, Hong Kong Polytechnic University

INTRODUCTION

Forensic voice comparison (FVC)



VS.



INTRODUCTION

Previous studies tested speaker-discriminatory power of different phonetic features under monolingual context,

- Cantonese /iau, oy, ei/ (Chen & Rose, 2012; Li & Rose, 2012; Pang & Rose, 2012).
- Mandarin /i, y/ (Zhang et al., 2008).
- Japanese syllable-coda nasal /ŋ/, voiceless alveopalatal fricative /ç/ and long back mid-rounded vowel /ɔ:/ (Rose et al., 2004).
- English filled pause (Hughes et al., 2016).

QUESTIONS?

- What if suspect and offender samples are in different languages?

LANGUAGES IN HONG KONG

- Official languages:

English

L2

Chinese



L1

Cantonese

L3

Mandarin



- vernacular in HK
- use in daily life



- The lingua franca in Mainland China
- After the handover (1997): increasing use & status in the education system




PRESENT STUDY



Aim

Are filled pauses good features for cross-language speaker comparison?

Data collection

- **Mock police interviews** in three languages + a **background questionnaire**
 - an international investment fraud case 
 - interrogation with a police officer (role played by the same trilingual researcher)
 - Each interview: 6-8 minutes
 - the order of the interviews was always the same between subjects

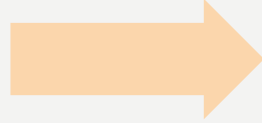


DATA COLLECTION

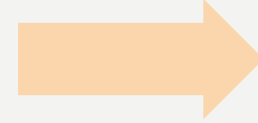
- **Participants:** 21 female Cantonese-English-Mandarin trilingual speakers from Hong Kong 呃 額 嗯 um/un er....uh
- **Age:** 18-27 years old
- **English proficiency:** the average age of learning (AOL) = 3.3 years old, the length of learning (LOL) = 17.5 years; the average score for IELTS = 7 (=CEFR C1)
- **Mandarin proficiency:** AOL = 4.75 years old, and LOL = 13.85 years; the average rating of proficiency = 6.1/10
- **Medium Of Instruction (MOI):** Mandarin as MOI for the Chinese subject in some primary and secondary schools
- **Parent's L1:** identify parents' L1 (e.g. Cantonese, Hakka, Mandarin, Nepali)
- **Other languages:** experience of learning other languages

DATA COLLECTION

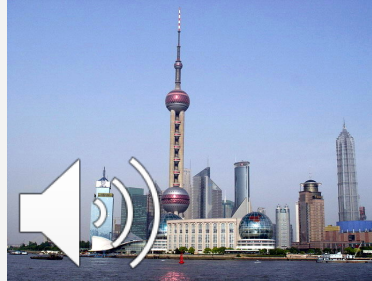
Cantonese



Mandarin



English



Q: 咁你具體係幾號起邊度同黎小高見面嘅咧?
那你具体是几号在哪里跟黎小高见面的?

A: uh...係3月4號。
呃...是3月4号。

(Translation)

Q: When and where did you
meet Li?

A: uh...it's on March 4th.

Q: 有做兼職嗎?

A: uh...兼職是攝影師。

(Translation)

Q: Do you do any part-time
job?

A: uh...my part-time job is
a photographer.

Q: How do you know her?

A: um...She's my working
partner in Hong Kong.

SEGMENTATION

- Filled pauses (FPs) in the three interviews were coded manually in Praat: 2153 tokens = 590 Cantonese + 685 Mandarin + 878 English tokens
- Examples:

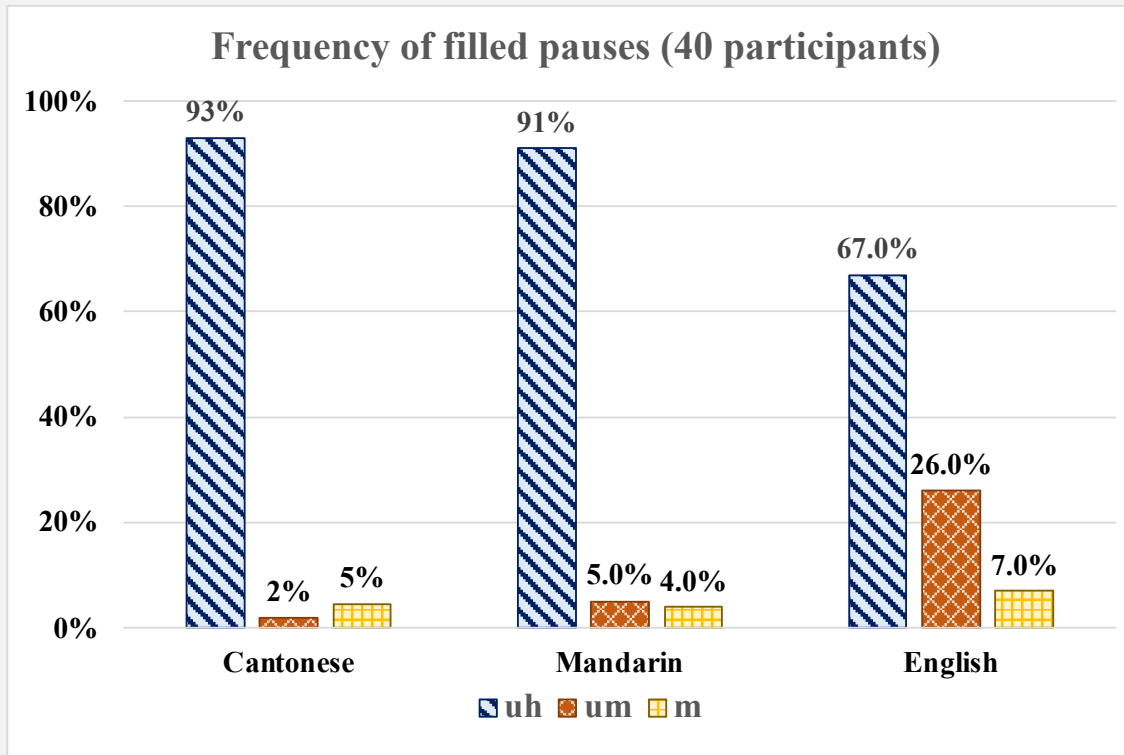


e.g. [ɛ] [e]

DISTRIBUTIONS OF FPS

➤ Distribution of *uh*, *um* and *m*:

$uh\% > um\% > m\%$



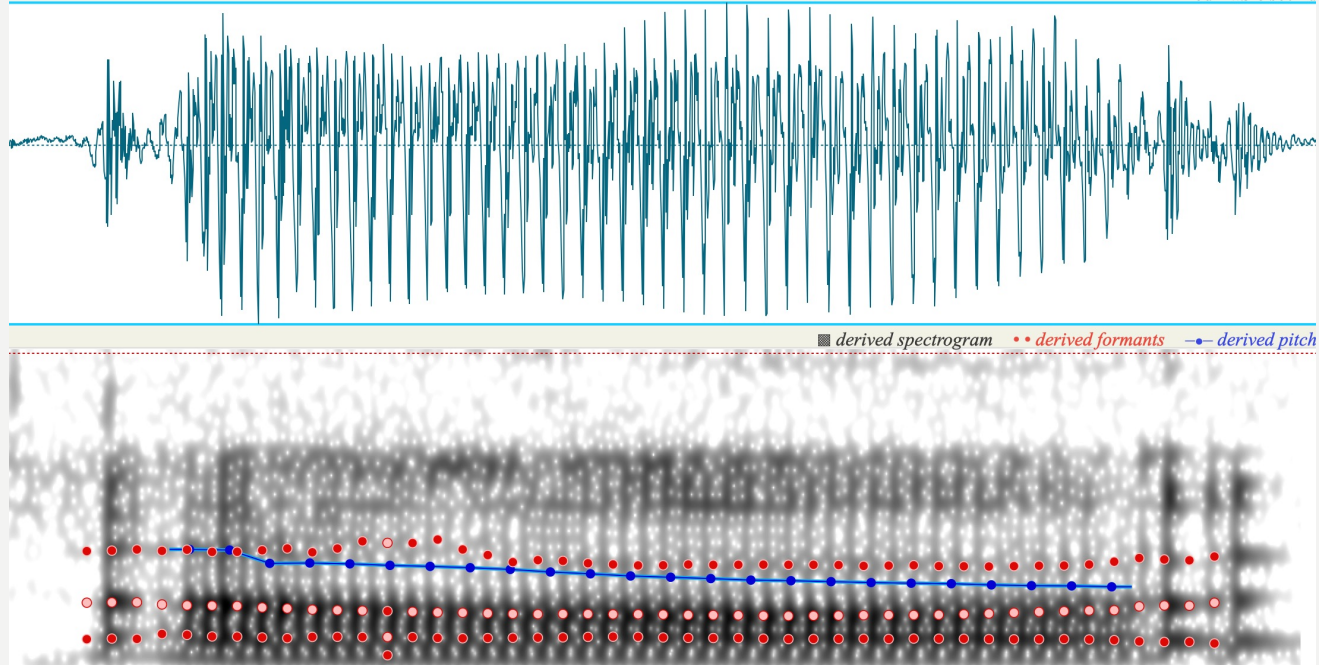
➤ **CAN \approx MAN; CAN&MAN $\not\approx$ ENG**

- The two Chinese languages have similar distributions
- English has more um%

COMPUTATION OF LRS (HUGHES, 2023)

Feature-to-score

- 1024 tokens of *-uh*: 425 Cantonese, 440 Mandarin, 339 English
- 19 tokens per speaker per language
- Acoustic input: **F1-F3 values of the midpoint, F0, vowel duration of *UH***
- The multivariate kernel density (MVKD; Aitken & Lucy, 2004)



/a/

COMPUTATION OF LRS (HUGHES, 2023)

Score-to-LR (calibration)

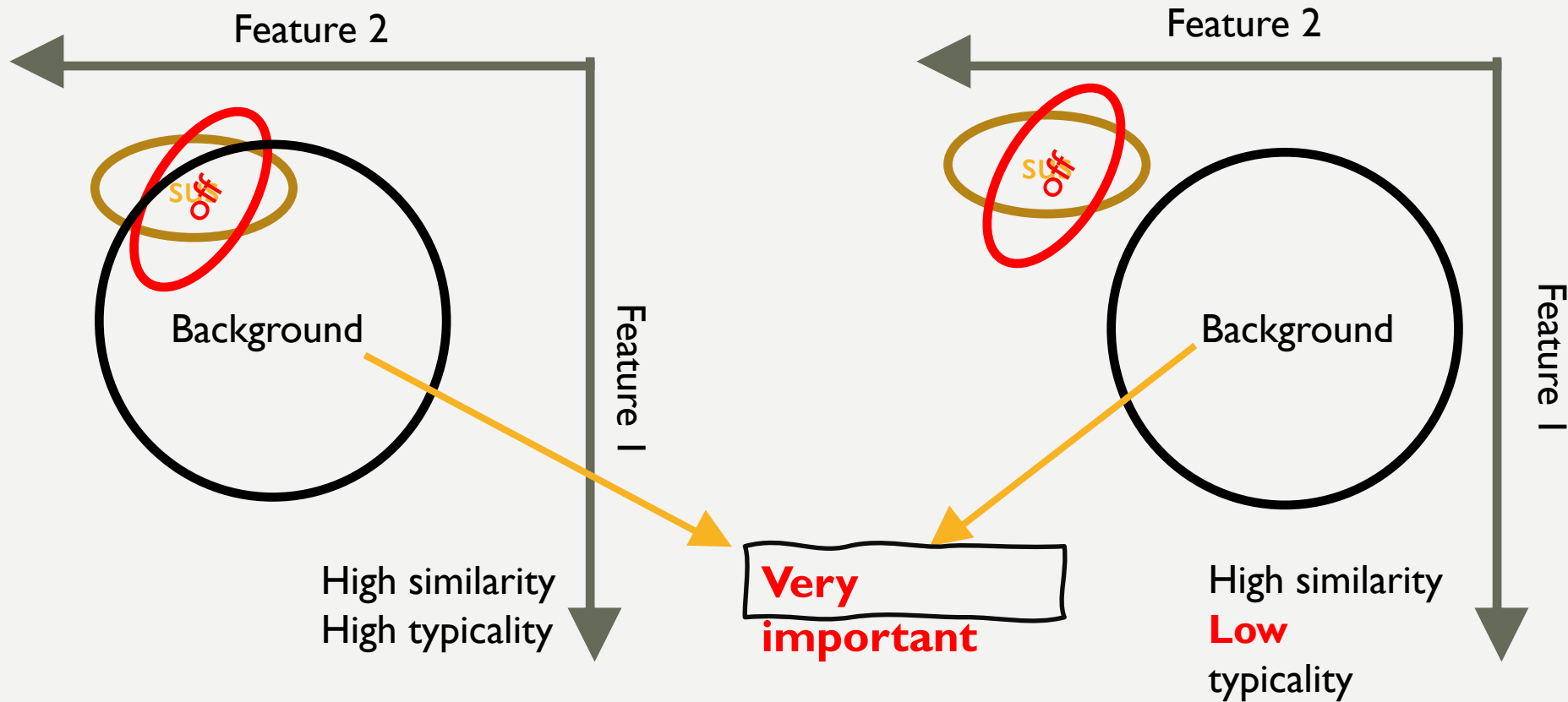
Calibration involves learning the properties of the scores produced by a given system using training data.

- Logistic regression with cross-validation (small number of speakers)

Evaluation metrics

- log-LR cost (C_{llr} ; Brümmer & du Preez, 2006)
- equal error rate (EER)
- The closer to 0 the better

COMPUTATION OF LRS (HUGHES, 2023)



RESULT

Conditions:

Mismatch between KS and DS

* No matched condition is tested due to the lack of data (21 speakers only)

Conditions	KS (suspect)	DS(offender)	Background
CAN + MAN	Cantonese	Mandarin	Mandarin
	Cantonese	Mandarin	Cantonese
CAN + ENG	Cantonese	English	English
	Cantonese	English	Cantonese
MAN + ENG	Mandarin	English	English
	Mandarin	English	Mandarin

An example of a mismatched condition case:

A scam call was made in Mandarin (DS = MAN).

Later the Hong Kong police arrested a suspect and interrogated him/her in Cantonese (KS = CAN).

Background = CAN because in the context of Hong Kong, the vernacular is Cantonese.

RESULT

Tested all the combinations between midpoint F1, F2, F3, F0 and vowel duration,
F1 + F3 tend to have the best performance overall

Conditions	QS (suspect)	KS (offender)	Background	Variables	Cllr	EER
CAN + MAN	Cantonese	Mandarin	Mandarin	F0 + duration + F1-F3	1.011	58.1
				F1 + F3	0.970	47.6
	Cantonese	Mandarin	Cantonese	F0 + duration + F1-F3	1.015	61.9
				F1 + F3	0.985	47.2
CAN + ENG	Cantonese	English	English	F0 + duration + F1-F3	1.008	52.9
				F1 + F3	0.945	44.1
	Cantonese	English	Cantonese	F0 + duration + F1-F3	1.02	38.8
				F1 + F3	0.945	42.9
MAN + ENG	Mandarin	English	English	F0 + duration + F1-F3	0.797	19.3
				F1 + F3	0.811	33.3
	Mandarin	English	Mandarin	F0 + duration + F1-F3	0.819	23.3
				F1 + F3	0.815	28.8

RESULT

- The lower Cllr & EER, the better the system

Conditions	QS (suspect)	KS (offender)	Background	Variables	Cllr	EER
CAN + MAN Cantonese Mandarin	Cantonese	Mandarin	Mandarin	F0 + duration + F1-F3	1.011	58.1
				F1 + F3	0.970	47.6
				F0 + duration + F1-F3	1.015	61.9
				F1 + F3	0.985	47.2
CAN + ENG Cantonese English	Cantonese	English	English	F0 + duration + F1-F3	1.008	52.9
				F1 + F3	0.945	44.1
				F0 + duration + F1-F3	1.02	38.8
				F1 + F3	0.945	42.9
MAN + ENG Mandarin	Mandarin	English	English	F0 + duration + F1-F3	0.797	19.3
				F1 + F3	0.811	33.3
				F0 + duration + F1-F3	0.819	23.3
				F1 + F3	0.815	28.8

- Cllr (the log LR cost function)

Cllr < 1 : the system is giving information, the parameters are valid

Cllr > 1 : the system doesn't work well, parameters are not worth using in FVC

- EER (equal error rate): The EER is calculated by plotting the FAR and the FRR on a graph and determining the point at which they intersect. A lower EER indicates better performance of the speaker identification system.

RESULT

The MAN + ENG condition has the best performance

→ When using trilingual speakers' L2 and L3 for cross-language speaker comparison, the system has the strongest discrimination power.

Conditions	QS (suspect)	KS (offender)	Background	Variables	Cllr	EER
CAN + MAN	Cantonese	Mandarin	Mandarin	F0 + duration + F1-F3	1.011	58.1
				F1+ F3	0.970	47.6
	Cantonese	Mandarin	Cantonese	F0 + duration + F1-F3	1.015	61.9
				F1+ F3	0.985	47.2
CAN + ENG	Cantonese	English	English	F0 + duration + F1-F3	1.008	52.9
				F1+ F3	0.945	44.1
	Cantonese	English	Cantonese	F0 + duration + F1-F3	1.02	38.8
				F1+ F3	0.945	42.9
MAN + ENG	Mandarin	English	English	F0 + duration + F1-F3	0.797	19.3
				F1+ F3	0.811	33.3
	Mandarin	English	Mandarin	F0 + duration + F1-F3	0.819	23.3
				F1+ F3	0.815	28.8

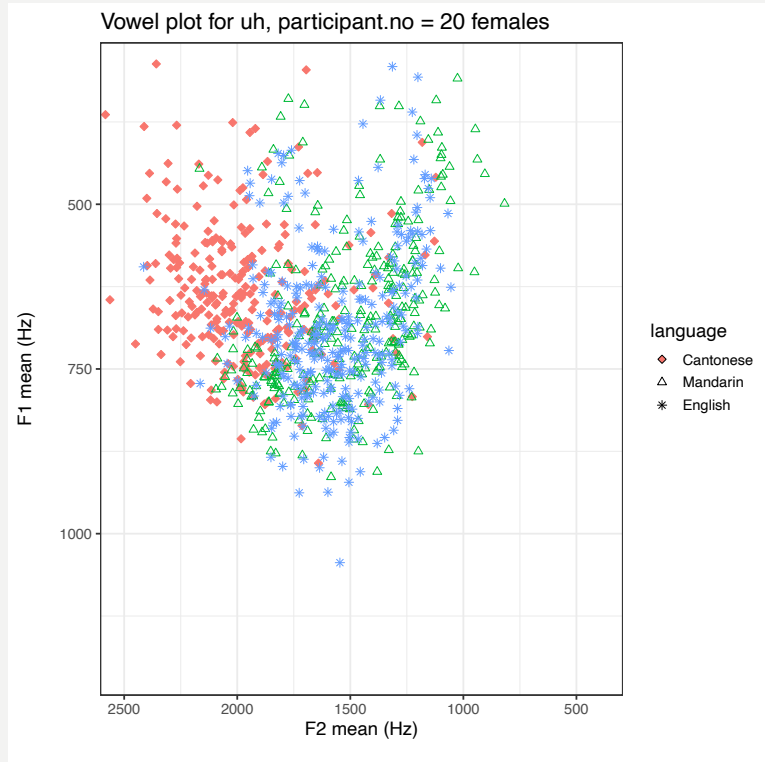
DISCUSSION

Acoustics:

Mandarin and English tend to have similar vowels for fps. (L1-CAN) vs (L3-MAN \approx L2-ENG)

Systems:

The MAN + ENG condition has the best performance.



➤ WHY?

- Vowels of MAN and ENG are similar → meaning smaller within-speaker variability?
- L2 and L3 contain more information about individual variability?

RESULT: INDIVIDUAL VARIABILITY

Condition: MAN + ENG

Same speaker comparisons:
negative values indicate an error

QS	KS	MAN_ENG LR _{ss} (background:MAN)	MAN_ENG LR _{ss} (background: ENG)
1	1	-0.32	-0.38
2	2	0.41	0.39
5	5	0.39	0.42
7	7	0.22	0.22
8	8	0.48	0.51
9	9	-0.60	-0.50
14	14	0.46	0.48
21	21	-0.62	-0.58
26	26	0.68	0.70
31	31	0.36	0.32
35	35	0.31	0.31
37	37	-0.05	-0.07
38	38	0.50	0.53
42	42	0.06	0.04
43	43	0.01	0.02
45	45	0.51	0.50
48	48	0.15	0.09
49	49	0.48	0.48
53	53	0.48	0.57
55	55	0.14	0.14
56	56	0.41	0.39

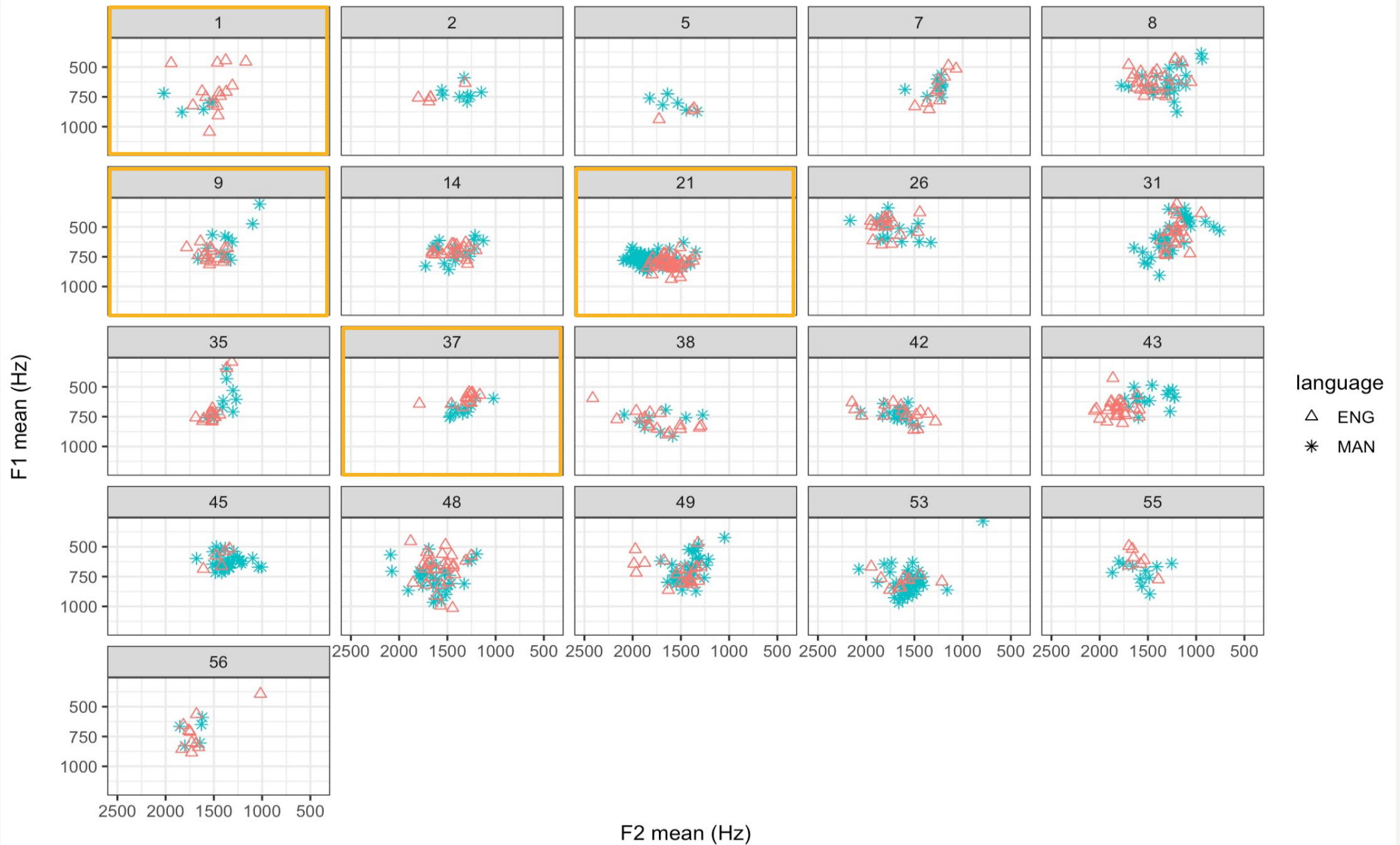
Proficiency between MAN
(7.15/10) and ENG (IELTS 6.5/9)
is large.

Her parent LI is Mandarin →
higher proficiency in Mandarin

Proficiency between MAN
(5.43/10) and ENG (IELTS 9/9) is
large.

- All of these speakers have experience
of learning other languages.

Vowel plot for uh, participant.no = 21 females (for LR-approach), MAN_ENG



TAKE HOME MESSAGE

For Cantonese-Mandarin-English trilingual speakers in Hong Kong:

- the distribution of *uh*, *um* is similar between CAN (L1) and MAN (L3), not with ENG (L2)
- the vowel of *uh* is similar in MAN (L3) and ENG (L2)
- cross-language speaker comparison has generally weak performance with limited speakers (Cllr is 0.8 ~ 1)
- among all comparisons, MAN + ENG mismatched conditions had the best performance

For forensic practitioners:

- given the generally weak performance in the models, filled pauses might not be the best features for cross-language speaker comparisons.
- need to consider suspects/offenders' background of bilingualism. For cases like Cantonese-Mandarin-English trilinguals where Cantonese is the dominant language, English and Mandarin as L2 or L3, comparing L2 with L3 would have a better result.

REFERENCES

- Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1), Article 1. <https://doi.org/10.1046/j.0035-9254.2003.05271.x>
- Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2–3), 230–275. <https://doi.org/10.1016/j.csl.2005.08.001>
- Chen, A., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with the Cantonese Triphthong /iau/. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 197–200.
- Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, 23(1), 99–132. <https://doi.org/10.1558/ijsl.v23i1.29874>
- Li, J., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /oy/ Diphthong. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 201–204.
- Pang, J., & Rose, P. (2012). Likelihood Ratio-Based Forensic Voice Comparison with the Cantonese Diphthong /ei/ F-Pattern. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 205–208.
- Rose, P., Lucy, D., & Osanai, T. (2004). Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects models: A “non-idiot’s Bayes” approach. *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, 492–497.
- Zhang, C., Morrison, G. S., & Rose, P. (2008). Forensic Speaker Recognition in Chinese: A Multivariate Likelihood Ratio Discrimination on /i/ and /y/. *In Proceedings of Interspeech*, 1937–1940.

THANK YOU & QUESTIONS?

This project is funded by Hong Kong Research Grant Council
Postdoctoral Fellowship to Grace Wenling CAO.