Cross-language Forensic Voice Comparison of Hong Kong Trilingual Speakers Using Filled Pauses and an Automatic Speaker Recognition System

Grace Wenling Cao[1], Vincent Hughes[2], Bruce Xiao Wang[3], Peggy Mok[1]

[1] The Chinese University of Hong Kong

[2] University of York

[3] The Hong Kong Polytechnic University

gracecao@cuhk.edu.hk, vincent.hughes@york.ac.uk, brucex.wang@polyu.edu.hk, peggymok@cuhk.edu.hk

2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3315-1682-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCSLP63861.2024.10800518

Abstract

Forensic voice comparison (FVC) involving bilingual speakers presents a substantial challenge to forensic practitioners. Previous FVC research suggests that discriminatory power is, unsurprisingly, weaker in language mismatch conditions than in the language match conditions. The present paper extends on previous work examining bilingual speakers to consider trilingual speakers. Specifically, we conduct FVC tests of Cantonese-English-Mandarin trilinguals using acoustic-phonetic data extracted from filled pauses and using an automatic speaker recognition (ASR) system. A different language effect was found in the two systems, such that the Mandarin-English condition produced the best performance in the acoustic-phonetic based system, while the Cantonese-English condition produced the worst performance in the ASR system. Individual-speaker analysis suggests that trilingual speakers' proficiency in L2 and L3 seems to affect how well they perform in cross-language FVC.

Index Terms: forensic voice comparison, cross-language comparison, filled pauses, automatic speaker recognition, trilingualism.

1. Introduction

The Codes of Practice of the International Association for Forensic Phonetics and Acoustics suggest that members should exercise particular caution with cross-language comparisons [1], reflecting a challenge of conducting forensic voice comparisons (FVC) with multilingual speakers. Echoing such a challenge, increasingly, studies have started to examine forensic phonetics in bilingual or multilingual contexts [2-8].

Among these studies, there are two lines of research. The first involves testing cross-language FVC using filled pauses. A few studies suggest that bilingual speakers use the same set of filled pauses in their L1 and L2 [3-4]. This is because filled pauses are difficult to manipulate consistently [9] as they are thought to be produced below the level of consciousness. This makes them good candidate features for cross-language speaker identification. However, others found contrasting results [2], [5], suggesting that bilingual speakers use language-specific filled pauses in L1 and L2, and thus filled pauses would not be very useful for cross-language FVC. In particular, [6] examined filled pauses of an Italian-German-English trilingual and found that the trilingual speaker tended to use different sets of filled pauses in the three languages. As [6] only examined one trilingual speaker, it is unclear whether similar patterns would be found using a larger sample.

Another line of research is more forensic in focus, investigating the speaker-discriminatory power of certain speech features using different statistical models following the likelihood ratio (LR) framework [7], [10], [11]. [11] tested cross-language FVC based on English-French bilingual speakers' long-term formant distributions (LTFDs) using the multivariate kernel density (MVKD) [12] and Gaussian mixture model-Universal background model (GMM-UBM) [13] approaches. Results suggest that cross-language comparisons in general generated much weaker strength of evidence than same-language comparisons, particularly when the test data and reference data were mismatched. A similar pattern was also observed in [11]'s ASR results using an x-vector system, except that the system validity was in general very good even in the language mismatch conditions ($C_{llr} < 0.05$). [7] tested whether and how the mismatch between test samples, and between the test samples and calibration data would affect the overall performance of ASR system. Testing on English-French bilingual speech data, they found that the language mismatch conditions produced poorer overall performance than the language match conditions. More severe miscalibration was found where there was a language mismatch between the test and calibration sets. Apart from the system-level analysis, individual speaker-level analysis can also provide diagnostic insights into system performance [10]. For instance, Zooplot analysis can identify problematic speakers using the biometric menagerie [14-15].

2. The current study

2.1 Research Questions

The current study extends previous work on multilingual FVC with a focus on trilingual speakers. Specifically, we explore the following two research questions:

- Are filled pauses good features for cross-language FVC with trilingual speakers?
- Do individual trilingual speakers show different behaviour within the ASR system when comparing different mismatch conditions of the three languages?

The novelty of the current project is threefold. First, the project examines cross-language FVC using filled pauses from a larger set of trilingual speech data compared to [6], filling the gap of research on the phonetic properties of trilingual filled pauses. Second, unlike [7], [11] where English-French bilinguals were tested, the current project tests Cantonese-English-Mandarin trilingual speakers from Hong Kong,

[1] Corresponding author: Grace Wenling Cao

providing evidence from non-Indo-European languages. For example, Cantonese (the Hong Kong trilinguals' L1) and Mandarin (L3) are both Sinitic languages and share many similarities, whereas English (L2) is distinct from Cantonese and Mandarin in its phonological and syntactic features. These differences allow many interesting comparisons across the three languages. Third, while [10] examined individual variability of FVC on different parameters, this project allows us to examine individual speaker-level variability across different mismatch conditions. This leads to one very interesting question: as trilingual speakers speak Cantonese, English and Mandarin, which mismatch condition would generate the stronger discriminative power?

Based on the data from [16], the current project first examines FVC of filled pauses using acoustic-phonetic approach (i.e. Study 1). Cross-language FVC was also conducted using an ASR system with the same set of interview data (i.e. Study 2).

2.2 Participants & Experiment design

Twenty-one female Cantonese-English-Mandarin trilingual speakers participated in three mock police interviews. They were all undergraduate students (age range: 18-27 years old) from a university in Hong Kong and spoke Cantonese as their first language. Participants attended three mock police interviews as part of a larger forensic phonetic project. They had an interview in Cantonese first, followed by Mandarin and English interviews. To help participants adjust their language mode for the coming interviews, they had a 5-minute break after each interview and watched a 3-minute video in Mandarin or English before they started the next task. Interviews were conducted in a sound-proof recording booth. An H4N recorder was placed about 30 cm away from the participant and was used to record the interviews. Each interview was between 6 and 8 minutes long. Participants also completed a questionnaire about their linguistic background. As all the speech data were collected on the same day, non-contemporaneous data is not available for the current project. Regarding their language proficiency, their English proficiency was IELTS (International English Language Testing System) score 6-9, the age of learning English (AOL) was 1-6 years old (mean = 3), and the length of English learning (LOL) was 15-21 years. Regarding their Mandarin, the Mandarin AOL was 1-7 years old (mean = 5), and the Mandarin LOL was 9-21 years. Participants also self-reported their daily use of the three languages: on average, they used 91% Cantonese, 3% Mandarin and 6% English. To gain a comparable proficiency reference for participants' English and Mandarin, a separate set of native speakers of English (N = 15) and Chinese Mandarin (N = 20) rated participants' English and Mandarin on a 10-point scale: the average rating for their English proficiency was 6.42/10 and for their Mandarin was 6.84/10.

3. Study 1

3.1 Data processing and methods

Study 1 involves FVC with filled pauses from three languages. First, all the filled pauses of *-uh* in the three interviews were segmented by two research assistants and two student helpers using Praat [17]. The Cantonese interviews were segmented manually by a research assistant who is a native speaker of Cantonese, and the Mandarin and English interviews were marked by research assistants who are Mandarin-English bilinguals. In total, 1204 tokens of *-uh* including 425 Cantonese tokens, 440 Mandarin tokens and 339 English tokens were coded. On average, each participant has 19 tokens per language. The duration of the vowel was calculated. The formant values of F1, F2 and F3 were extracted on the temporal midpoint of the vowels. The F0 value was extracted at 10% intervals across the vowel, and the average F0 was also calculated. In total, five acoustic-phonetic features including midpoint F1, F2, F3, F0 mean and vowel duration were used as the input for MVKD.

Midpoint F1, F2, F3, F0 mean and vowel duration were used to generate LR-like scores for 21 same-speaker and 210 different-speaker comparisons for each condition using the MVKD approach [12]. Due to the limited number of speakers, cross-validation was implemented. Scores were then calibrated using cross-validated logistic regression [18]. As non-contemporaneous data were not available, within-language comparisons were not tested. 18 language mismatch conditions with single and mixed language as background populations were tested (see Table 1 for details). To fully test the mismatch conditions, the languages used by the nominal offender (KS) and the nominal suspect (QS) were switched and tested. For each mismatch condition, different combinations of features were tested (i.e. F0 + duration + F1 + F2 + F3, duration + F1 + F2 + F3, F0 + F1 + F2 + F3, F0 + duration, F1 + F2 + F3, F1, F2, F3, F1 + F2, F2 + F3, F1 + F3). System performance was evaluated using C_{llr} and EER (equal error rate). Systems with a C_{llr} value less than 1 provide some discriminative power. The lower the C_{llr} value and EER, the better the system performance.

3.2 Results

Figure 1 shows the F1~F2 plot of all the filled pause across Cantonese (CAN), Mandarin (MAN) and English (ENG). First, the CAN-*uh*s are distinguished from the MAN-*uh*s and ENG-*uh*s on both F1 and F2 dimensions, suggesting a fronter and higher vowel for CAN.

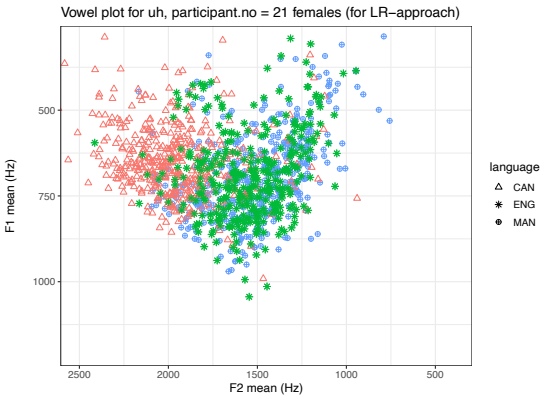

Figure 1. F1~F2 plot of *-uh* in the three languages.

Phonetic transcription of the CAN-*uh* includes variants like [ɛ] and [ə]. Second, MAN-*uh*s and ENG-*uh*s are largely overlapped, and their phonetic transcriptions include variants along [a]~ [ə].

Among the 11 combinations of parameters, F1 and F1 + F3 consistently showed the best performance compared to other conditions. Due to the space limit, only the results of F1 + F3 for the 18 mismatch conditions are reported in Table 1. Firstly, the C_{llr} value of all the conditions was close to 1, suggesting that models using filled pauses as input performed poorly in cross-language FVC. This is also reflected in high EERs, indicating high discrimination error. Second, including both languages in

the background population does not seem to increase the performance compared with using a single language in the background population. Third, the best performance was found in the condition where Mandarin was used as the nominal suspect (QS) and the background population, and English was used as the nominal offender (KS, C_{llr} = 0.81, EER = 28.8%). This might be due to the similar vowels that these trilingual speakers used in their filled pauses of English and Mandarin, resulting in a lower within-speaker variability, as shown in Figure 1.

Table 1. Summary of MVKD models of filled pauses *-uh* (parameters = F1 + F3 of vowel midpoint)

Condition	QS	KS	Background	C_{llr}	EER
Mismatch CAN + MAN	CAN	MAN	MAN	0.97	47.6%
			CAN	0.99	47.6%
			CAN + MAN	0.97	47.6%
	MAN	CAN	MAN	0.94	42.9%
			CAN	0.97	42.6%
			CAN + MAN	0.95	45.7%
Mismatch CAN + ENG	CAN	ENG	ENG	0.94	44.1%
			CAN	0.94	47.6%
			CAN + ENG	0.94	42.9%
	ENG	CAN	ENG	0.89	37.9%
			CAN	0.87	37.6%
			CAN + ENG	0.89	33.8%
Mismatch MAN + ENG	**MAN**	**ENG**	ENG	0.81	33.3%
			MAN	**0.81**	**28.8%**
			MAN + ENG	0.80	33.3%
	ENG	MAN	ENG	0.80	32.9%
			MAN	0.81	32.1%
			MAN + ENG	0.80	33.1%

4. Study 2

4.1 Data processing and methods

Interview data of the same 21 female participants in Study 1 were used in Study 2. Two samples of 30-second speech were extracted from their interviews in three languages for each speaker. In total, 126 samples were used as input for the ASR system. The commercial x-vector ASR system, Phonexia Voice Inspector (v.4.0.0), was used for testing. The system extracted MFCCs from the 126 samples and generated x-vector speaker models for comparison. Six mismatched conditions were tested, with C_{llr} and EER reported in Table 2. In Study 2, 84 same-speaker (SS) and 1680 different-speaker (DS) scores of each condition were generated in the ASR system, and they were exported and calibrated using the same cross-calibration method in Study 1. Calibration mismatch was not tested in the current study.

4.2 Results

The ASR results suggest an overall good performance as the C_{llr} value of all conditions was lower than 1 and EER was close to 0.

4.2.1 Mismatched-language comparisons

When the KS and QS spoke different languages, the ASR in general produced very good performance as the C_{llr} value of all conditions was lower than 1. The lowest EER was found in the mixed Cantonese-Mandarin and Mandarin-English conditions, with a value of 0.03%. Interestingly, the mixed Cantonese-English conditions produced the highest EER (0.15%), which is five times higher than the other two mixed conditions. It seems that the cross-language condition of Cantonese-English was more difficult than the conditions of Cantonese-Mandarin and Mandarin-English for the ASR system.

Table 2. Summary of ASR models (parameters = MFCC of 30 seconds speech)

Condition	QS	KS	C_{llr}	EER
Mismatch (CAN + MAN)	CAN	MAN	0.0119	0.03%
Mismatch (CAN + ENG)	CAN	ENG	0.0205	0.15%
Mismatch (MAN+ ENG)	MAN	ENG	0.0125	0.03%

4.2.2 Individual speaker-level analysis

Following [15], speakers were categorized into one of the four animal groups based on their SS-LLR (log likelihood ratio) and DS-LLR: doves, who produce the top 25% strongest SS and DS LLRs; worms, who produce the bottom 25% lowest SS and DS LLRs; phantoms, who produce the top 25% strongest DS and bottom 25% lowest SS; chameleons, who produce the top 25% strongest SS and the bottom 25% lowest DS LLRs. Doves are regarded as the 'best' speakers for FVC whereas worms are the problematic speakers. Figure 2 summarizes individual speakers' classification across six mismatch conditions.

There are cases where trilingual speakers were consistently in the same category within the zooplot, for instance, HK31 and HK35 were grouped as chameleons in most of the mismatch conditions, suggesting a language-independent effect for these speakers. There are also cases where trilingual speakers were categorized into different groups when the languages of QS and KS were swapped. For instance, HK2 was a worm when comparing Cantonese-QS and Mandarin-KS, but she became a

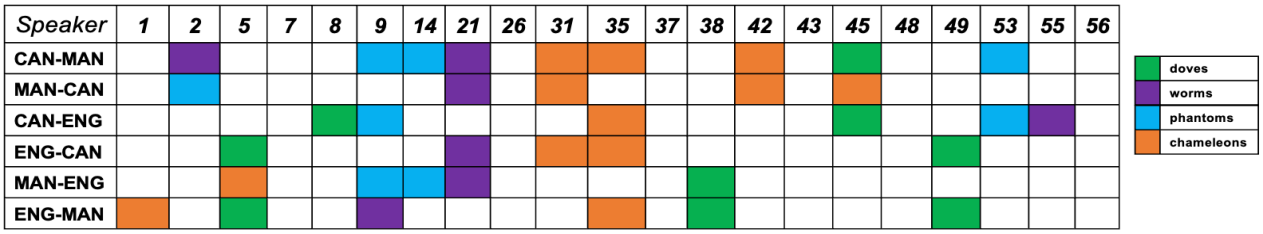

Figure 2. Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (orange) in all systems.

phantom when Cantonese was swapped to the KS. Lastly, there are cases where speakers were grouped into different categories when the mismatch conditions were different. For example, HK45 was a dove when comparing her Cantonese-QS with her English-KS, but she became a chameleon when comparing her Mandarin-QS with her Cantonese-KS.

5. Discussion

The study tested acoustic-phonetic and ASR approaches on cross-language FVC of Cantonese-English-Mandarin trilingual speakers from Hong Kong. Results suggest that the ASR models performed well in FVC when the nominal offender (KS) and the nominal suspect (QS) spoke different languages.

5.1 The performance of the two systems

On the one hand, using around 30s of speech, the ASR system had rich speech material with high variability which allowed it to capture more individual variations. The ASR systems also captured MFCCs which are high-dimensional features, resulting in a low EER in the FVC results. On the other hand, it is not surprising that the acoustic-phonetic approach using filled pauses gained a poor performance in cross-language FVC when it only used around 3 seconds of speech for each speaker per language. The MVKD models used in the acoustic-phonetic approach only had formant values of vowel midpoint of *-uh* as the input features. Adding other dimensions of filled pauses such as F0 mean and vowel duration to the MVKD models did not increase the model performance, suggesting that using filled pauses alone limits the model performance.

Given the different amounts of data used in the ASR system and the acoustic-phonetic approach, a direct comparison of the two systems is not encouraged. However, the presentation of the results of the two systems can shed light on forensic speech studies which explore the fused system of an ASR system and acoustic-phonetic system [19-23]. For instance, [20] found that the ASR system did not outperform the linguistic-phonetic system which was trained on vowel formant values of filled pauses *-um*. A promising improvement was also found when the ASR system and the filled pauses-based acoustic-phonetic system were fused. Although this project did not present a fused system for system comparison, this is one of the directions for future research.

5.2 The language effect

Among the MVKD filled-pause models, although the overall performances were poor, the conditions of mixed Mandarin-English still had the lowest EER (between 28.8% and 33.3%) compared to the other mixed conditions. This between-language effect can be explained by the acoustic analysis of vowels in filled pauses in the three languages. Acoustic results suggested that these trilingual speakers tend to use similar vowels in their Mandarin and English-filled pauses, and their Cantonese-filled pauses used more fronted vowels as shown in Figure 1. Findings in [16] suggest that these Hong Kong trilingual speakers predominantly used [ɛ] (82%) in their CAN-*uh*, but they mostly used [a] in their MAN-*uh* (81%) and ENG-*uh* (85%). As these speakers used similar vowels in their L2 (English) and L3 (Mandarin), the within-speaker variability is likely to be lower in the Mandarin-English condition.

A language effect was also observed in the ASR results – albeit overall performance was extremely good across all sets of tests. The EER in the mismatch Cantonese-English conditions was five times higher (0.15%) than the mismatch conditions of Cantonese-Mandarin (0.03%) and Mandarin-English (0.03%). One possibility is that these trilingual speakers' Cantonese and English are more distinct compared to the pairs of Cantonese-Mandarin, and Mandarin-English. For the former comparison, Cantonese and Mandarin share very similar phonological systems, as they are Chinese varieties. The cross-language variability would likely be lower for the Mandarin-Cantonese condition. For the latter comparison, Mandarin and English are L2 and L3 for these Hong Kong trilingual speakers. One possibility is that these trilingual speakers use similar articulatory settings when speaking their L2 and L3, resulting in a low cross-language variability for Mandarin and English comparisons. For instance, [24] found that Dutch-Turkish bilinguals had similar LTF2 (long-term formant) and LTF3 means for Dutch and Turkish. [10] found a similar LTFD1 for the two languages of English-French bilinguals. [25] found that Canadian Cantonese-English bilinguals have similar spectral properties and lower-dimensional structure in their acoustic voice variation in their English and Cantonese. These studies all demonstrate some language-independent effects. Maybe for these trilingual speakers, the language-independent effects are not between L1 and L2/L3, but between L2 and L3, as shown in the phonetic analysis of filled pauses in Figure 1. Note that although there is a between-language effect for Cantonese-English conditions, the ASR system still performed very well with a C_{llr} value of 0.0205. In other words, although the trilingual speakers presented some cross-language variations, the ASR system can still handle such variations very well in the system.

5.3 Speaker classification

The individual-level analysis shown in Figure 2 suggests that 52% of the participants (11 out of 21) were categorised as the same group across at least two different mismatch conditions, suggesting a consistency of membership for the majority of the trilingual speakers. To explore whether individual patterns were related to the speakers' linguistic backgrounds, further examinations were conducted. HK21 was consistently identified as a worm, suggesting that she was a problematic performer in all conditions. Her English proficiency was rated as 8.8/10, but her Mandarin proficiency was only 5.1/10. She used English orally in 20% of her daily life, while most of the other participants only reported 0 to 10% of oral English use. In a different case, HK38 was classified as a dove when comparing her English with her Mandarin. She had relatively low proficiency in both her English (4.3/10) and Mandarin (4.5/10). Based on these limited observations, we suspect that trilingual speakers' proficiency in their L2 and L3 might affect how well they perform in cross-language FVC. If trilingual speakers' L2 and L3 proficiencies are both relatively low, they might receive similar influence of L1 transfers, leading to a smaller cross-L2/L3 variability which would facilitate FVC as in the case of HK38. If either their L2 or L3 proficiencies are high, trilingual speakers might have developed a distinct articulatory setting for their L2 or L3, therefore cross-L2/L3 variability would be larger, like in the case of HK21. As there isn't a clear pattern in the current study, more studies will be needed to test this hypothesis.

6. Acknowledgements

This project is funded by Hong Kong Research Grants Council Postdoctoral Fellowship #PDFS2122-4H01 to the first author.

7. References

[1] International Association for Forensic Phonetics & Acoustics, "IAFPA CODE OF PRACTICE," 2020.

[2] J. J. H. Lo, "Between Äh(m) and Euh(m): The Distribution and Realization of Filled Pauses in the Speech of German-French Simultaneous Bilinguals," *Lang Speech*, vol. 63, no. 4, pp. 746–768, Dec. 2020, doi: 10.1177/0023830919890068.

[3] S. G. Wong and V. Papp, "Transferability of non-lexical hesitation markers across languages: Evidence from te reo Maori-English bilinguals," in *Proceeding of 26th IAFPA*, 2018, pp. 35–66.

[4] M. M. de Boer and W. F. L. Heeren, " Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English ," *J Acoust Soc Am*, vol. 148, no. 6, pp. 3612–3622, Dec. 2020, doi: 10.1121/10.0002871.

[5] E. de Leeuw, "Hesitation Markers in English, German, and Dutch," *Journal of Germanic Linguistics*, vol. 19, no. 2, pp. 85–114, 2007, doi: 10.1017/S1470542707000049.

[6] L. Spreafico, "Filled pauses in multilingual speech: an acoustic analysis," *Linguistica e Filologia*, vol. 36, pp. 99–116, 2016.

[7] B. Nuttall, P. Harrison, and V. Hughes, "Automatic Speaker Recognition performance with matched and mismatched female bilingual speech data," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2023, pp. 601–605. doi: 10.21437/Interspeech.2023-680.

[8] S. Cho, M. J. Munro, and S. Fraser, "F0, long-term formants and LTAS in Korean-English Bilinguals," Tokyo, Japan: Proceedings of the 31st General Meeting of the Phonetic Society of Japan, Sep. 2017.

[9] K. McDougall and M. Duckworth, "Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of standard southern British english," *International Journal of Speech, Language and the Law*, vol. 25, no. 2, pp. 205–230, 2018, doi: 10.1558/IJSLL.37241.

[10] J. J. H. Lo, "Seeing the trees in the forest: Diagnosing individual performance with acoustic data in likelihood ratio based forensic voice comparison," *Studi AISV*, vol. 8, pp. 77–96, Dec. 2021, doi: 10.17469/O2108AISV000004.

[11] J. J. H. Lo, "Issues of Bilingualism in Likelihood Ratio-based Forensic Voice Comparison," University of York., 2021.

[12] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Appl. Statist*, vol. 53, pp. 109–122, 2004.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1, pp. 19–41, 2000.

[14] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998, p. 608.

[15] T. Dunstone and N. Yager, *Biometric system and data analysis.* . Springer US, 2009.

[16] W. G. Cao and P. Mok, "The Acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech," in *The proceeding of the 20th International Congress of the Phonetic Sciences*, Prague, Aug. 2023, pp. 3814–2818.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. Accessed: Mar. 01, 2022. [Online]. Available: https://www.fon.hum.uva.nl/praat/

[18] N. Brümmer *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," in *Proceedings of IEEE Transactions on Audio, Speech, and Language*, 2007, pp. 2072–2084. [Online]. Available: http://www.nist.gov/speech/tests/

[19] E. Enzinger and G. S. Morrison, "The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems," in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 2012, pp. 137–140.

[20] V. Hughes, P. Foulkes, and S. Wood, "Formant dynamics and durations of -um improve the performance of automatic speaker recognition systems," in *Proceedings of the 16th Australasian Conference on Speech Science and Technology*, 2016.

[21] C. Zhang, G. S. Morrison, and T. Thiruvaran, "FORENSIC VOICE COMPARISON USING CHINESE /iau," in *ICPhS XVII Regular Session Hong Kong*, 2011, pp. 17–21.

[22] C. Zhang, G. S. Morrison, F. Ochoa, and E. Enzinger, "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," *J Acoust Soc Am*, vol. 133, no. 1, pp. EL54–EL60, Jan. 2013, doi: 10.1121/1.4773223.

[23] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison-Female voices," *Speech Commun*, vol. 55, no. 6, pp. 796–813, Jul. 2013, doi: 10.1016/j.specom.2013.01.011.

[24] W. Heeren, D. Van Der Vloed, and J. Vermeulen, "Exploring long-term formants in bilingual speakers," in *Proceedings from the 2014 Conference of the International Association for Forensic Phonetics and Acoustics*, Zurich, Switzerland, 2014.

[25] K. A. Johnson and M. Babel, "The structure of acoustic voice variation in bilingual speech," *J Acoust Soc Am*, vol. 153, no. 6, p. 3221, Jun. 2023, doi: 10.1121/10.0019659.