The Acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech

Grace Wenling Cao, Peggy Mok

The Chinese University of Hong Kong
gracecao@cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

This paper explored filled pauses *uh*, *um* and *m* in Cantonese-Mandarin-English trilingual speech by 20 Hong Kong trilingual speakers. Duration, frequency and vowels of 1493 filled pauses were examined. Auditory analysis suggested that [ɛ, a, ə] vowels were commonly used for *uh* and *um* in the three languages. A preference for *uh* was found for all three languages, but the distributions of *uh*, *um* and *m* were more similar in the two Chinese languages than in English. For the duration, *uh* was similar across the three languages. For formants of *uh*, the main difference between Cantonese and the other two languages was found on F2, partially supporting the language-specific view in filled pause research. The speakers had similar vowel qualities for *uh* in English and Mandarin. Their parents' L1, primary school background, age and length of learning English and Mandarin significantly predicted their use of filled pauses.

Keywords: forensic phonetics, filled pause, cross-language speaker comparison

1. INTRODUCTION

Previous studies suggested that filled pauses such as *uh* and *um* can be useful for speaker identification in a monolingual context because they appear to be language-specific [1] and consistent for individual speakers [2]. However, whether filled pauses are language-specific in the context of multilingual speech remains unclear. This paper aims to tackle this issue and explores filled pauses in Cantonese-Mandarin-English trilingual speech. In this case, speakers' L2 (e.g. English) is typologically different from L1 Cantonese, whereas L3 (e.g. Mandarin) is typologically similar to their L1.

1.1. Filled pauses in multilingual speech

It has been a debate about whether filled pauses are language-dependent in the context of bilingualism. On the one hand, Wong and Papp [3] argued that filled pauses in English-Maori bilingual speech had a high transferability across languages, suggesting an L1-transfer view. Gosy, Gyarmathy and Beke [4] shared a similar stance, showing that Hungarian-English bilinguals in their study had no perceptual difference in the vowel production of English and Hungarian filled pauses. On the other hand, a few recent studies suggested a language-specific view [5]–[7]. For instance, de Boer and Heeren [5] found consistency in duration and F0 for Dutch-English bilingual speakers, but vowel realization showed some adaptation towards L2. Lo [6] also found that German-French simultaneous bilingual speakers did not use filled pauses with the same distributions and acoustic profiles in the two languages. Spreafico [7] explored filled pauses produced by an Italian-German-English trilingual speaker, and his results supported the language-specific view, where F2 was significantly different across the three languages. Studies which support the language-specific view tend to use bilingual speakers with high proficiency, whereas studies supporting the L1-transfer view had bilingual speakers with various L2 proficiency. None of these studies, however, examined the role of bilingual speakers' linguistic background in their use of filled pauses.

1.2. Present study

The present study explored filled pauses in Cantonese-English-Mandarin trilingual speech. Studies on filled pauses in Chinese are very limited. Among the handful of studies [8]–[10], not all of them conducted an acoustic-phonetic analysis. Common filled pauses in Mandarin include *uh*, *um* and *mm* [8]-[9]. Wu [10] examined L2 Chinese and suggested *uh* with two different representations (呃, 哦) and *um* (嗯). Zhao and Jurafsky [9] suggested that *uh* was more common than *mm* and speakers in the south used more filled pauses than speakers in the north. The lack of phonetic research on filled pauses in Chinese languages and trilingual speech calls for more in-depth investigations. To fill the gap, this study asked two research questions:

- Is vowel quality of filled pauses language-dependent across the three languages?
- Do trilingual speakers' linguistic background influence their use of filled pauses?

Our hypotheses were first, the speakers would use language-specific vowels in three languages; and

second, factors of speakers' linguistic background such as parents' L1, age and length of learning English and Mandarin would influence how they use filled pauses cross-linguistically.

2. METHOD

2.1 Participants

Twenty (8F) Cantonese-English-Mandarin trilingual speakers participated in three mock police interviews. They were all students (age range: 18-27 years old) from a university in Hong Kong and spoke Cantonese as their first language. Seven participants reported that at least one of their parents does not speak Cantonese as their L1, among which six speak Mandarin and one speaks Teochew. Nine participants reported that they attended primary or secondary schools which used Mandarin as the medium of instruction for the Chinese subject. Their English proficiency was IELTS 6-8, the age of learning English was 1-6 years old, and the length of English learning was 15-21 years. Regarding their Mandarin proficiency, none of the participants took the National Putonghua Proficiency Test. Therefore, their proficiencies were estimated by a separate rating task. 100 native Mandarin speakers listened to a 10-second recording of each participant and judged the standardness of the speaker's Mandarin accent. Results suggested that the mean was 5.42 (out of 10), ranging from 2.81 to 8.13. Their age of learning Mandarin was 1-10 years old, and the length of learning Mandarin was 9-27 years. Participants also reported their daily use of the three languages. On average, they used 88% Cantonese, 3% Mandarin and 9% English.

2.2 Experiment

Participants attended three mock police interviews as part of a larger forensic phonetic project. Participants were told that they were involved in an international investment fraud case. A police officer from Hong Kong, Shanghai and London would interrogate them separately in three interviews. The three police officers were role-played by the first author, who is a Cantonese-Mandarin balanced bilingual and speaks English with a near-native British accent. The participants had an interview in Cantonese first, followed by Mandarin and English interviews. To help participants to 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. Participants and the 'police officer' sat face-to-face by a desk. A desktop computer was placed facing the participants to show information about the case. Participants were told that they could look at the screen if they could not remember the case details. An H4N recorder was placed about 30 cm away from the participant and was used to record the interviews. Each interview was about 6 to 8 minutes long. Participants also completed a questionnaire about their linguistic backgrounds.

2.3 Data annotation and analysis

First, filled pauses in the three interviews were coded by two research assistants and two student helpers using Praat [11]. The Cantonese interviews were coded by the research assistant who is a native speaker of Cantonese, whereas the Mandarin and English interviews were marked by the research assistant/student helpers who are Mandarin-English bilinguals. Although code-switching is common for trilingual speakers in Hong Kong, only in the Cantonese interviews intra-sentential code-switching was found, which should not affect the coding for Cantonese interviews. In total, 1493 filled pauses including 413 Cantonese tokens, 471 Mandarin tokens and 609 English tokens were coded. To measure the frequency of filled pauses, participants' interviews were transcribed orthographically and the word/character counts were calculated.

As there is limited research on filled pauses in Chinese languages and trilingual speech, both auditory and acoustic analyses were conducted. The auditory analysis aimed to explore the vowels of filled pauses used in the three languages. First, the first author listened to every single token of the filled pauses in the interviews and annotated the vowels. Another trained forensic phonetician who is also a Cantonese-Mandarin-English trilingual crosschecked all the vowel annotations in three interviews. For any disputed annotations, the first author and the phonetician would discuss to reach an agreement. For the acoustic analysis, a Praat script was used to extract F1, F2 and F3 values at the midpoint of the vowels and calculate the duration of the filled pauses. Data were imported to R along with participants' background information for further analysis.

3. RESULTS

3.1 Counts and frequency

The frequency of *uh*, *um* and *m* over the total word/character counts are shown in Table 1. Note that for Cantonese and Mandarin interviews the number of Chinese characters was calculated, whereas for English the number of words was counted. The comparisons between the Chinese languages and English shall be interpreted cautiously.

	Count	*uh*	*um*	*m*	all fps
Cantonese	15048	2.53%	0.07%	0.14%	2.74%
Mandarin	9804	4.09%	0.42%	0.30%	4.80%
English	17524	2.21%	0.91%	0.35%	3.48%

Table 1. Percentage of filled pauses in the interviews.

The distributions of *uh*, *um* and *m* over all filled pauses are shown in Figure 1. The most frequently used filled pause in all three languages was *uh*. The distributions of filled pauses in Mandarin and Cantonese were similar, whereas the distribution in English suggested a more frequent use of *um* and *m*.

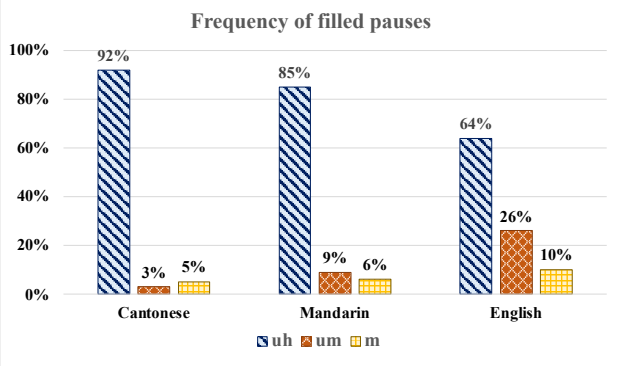

Figure 1. Distribution of *uh*, *um* and *m* in Cantonese, Mandarin and English.

Following Lo [6], the dataset was re-coded for mixed-effects logistic regression models. Tokens of *m* were removed and the use of *uh* and *um* was coded as a binary variable. The glmer() function in lmer package [12] was used to run the models. The basic model included language and sex as fixed effects, and included random intercept by speaker and random slopes by speaker over language. Factors of speakers' linguistic backgrounds such as whether the L1 of participant's parents is Cantonese or not (Parents' L1), age and length of learning Mandarin and English, attending primary schools in which the medium of instruction is Mandarin for the Chinese subject (MAN as MOI), proficiency of English, the experience of learning other languages and the standardness of Mandarin were examined one by one using a forward stepwise approach. Models were then compared with the basic model using the anova function in R. Only factors that significantly improved the model fit would add to the model.

The best model contained the fixed factors shown in Table 2. Results suggest that the distribution of *um* over *uh* was significantly different between Cantonese and English, but not between Cantonese and Mandarin (um/uh ratio for Cantonese: 0.028; for English: 0.409; for Mandarin: 0.102). For participants whose parents' L1 is *not* Cantonese, their *um:uh* ratio (0.29) was higher than those whose parents' L1 is Cantonese (0.11). For participants who had Mandarin as the Medium of Instruction in their primary schools, their *um:uh* ratio (0.28) was higher than those who did not (0.15). Factors which were not included in Table 2 did not improve model fit.

	Est	SE	z	*p*
(Intercept)	-5.3	1.7	-3.1	.002
Language (ENG)	1.9	0.9	2.2	**.03***
Language (MAN)	0.5	1.0	0.5	.60
Parent's L1 is Cantonese (yes)	-2.9	1.3	-2.3	**.02***
MAN as MOI (yes)	2.4	0.8	2.9	**.003***
Sex (male)	-1.3	0.7	-1.7	.09
Standardness of MAN	0.6	0.4	1.4	0.17
Language (ENG) × Parents' L1 is Cantonese (yes)	1.4	1.3	1.1	0.26
Language (MAN) × Parent's L1 is Cantonese (yes)	-1.5	1.7	-0.9	0.38

Table 2. Summary of fixed effects in the mixed-effects logistic regression model for the distribution of *uh* over *um*.

3.2 Acoustic and auditory analysis of vowels in *uh*

Due to the limited space, only acoustic and auditory analysis of vowels in *uh* will be reported here. For F1, F2 and F3, data of males and females were separated into two subsets and linear mixed-effects models were run separately. The basic model contained language as a fixed effect and included random intercept by speaker and random slopes by speaker over language. Factors of participants' linguistic backgrounds were added to the model one by one to test its significance. Results are shown in Table 3.

The most significant cross-linguistic difference was found on F2, where *uh* was articulated with much fronter vowels in Cantonese than in Mandarin and English for both females (about 400 Hz) and males (about 250 Hz). F1 was relatively similar across the three languages, except that female speakers produced vowels with a significantly lower F1 in their Cantonese interviews compared to English ones. For F3, females and males showed different cross-linguistic patterns. Female speakers had no significant change in their F3 across the three languages, whereas male speakers produced vowels with a significantly higher F3 in Mandarin and English than in Cantonese. The durations (in ms) of *uh* were similar across the three languages for female speakers (CAN: 374ms, MAN: 350ms, ENG: 370ms). No significant comparisons were found for male speakers too: English (360ms), Cantonese (337ms) and MAN (311ms).

Auditory analysis suggested that the participants used [ɛ, a, ɐ, ə, ʌ] vowels for most of the *uh* in the interviews. The filled pause *uh*-[ɛ] is commonly found in colloquial Cantonese, the quality of this vowel is similar to the Cantonese /ɛ/ as in 賒 'lend'. The *uh*-[ɛ] is not commonly found in colloquial Mandarin and English. The *uh*-[a] is close

to the Cantonese /ɐ/ as in 濕 'wet'. For simplification, two categories *uh*-[ɛ] and *uh*-[a] were distinguished based on auditory analysis. The *uh*-[a] contained representations vary from [ə] to [ʌ].

F1-females	**Est**	**SE**	**t**	***p***
(Intercept)	546	26	21	**<.001***
Language (MAN)	42	22	2	.109
Language (ENG)	61	21	3	**.019***
MAN as MOI (yes)	-113	51	-2	**.033***
Age of learning MAN	41	12	3	**.002***
F1-males	**Est**	**SE**	**t**	***p***
(Intercept)	318	82	4	**.003***
Language (MAN)	10	23	0.4	.672
Language (ENG)	-33	26	-1.3	.227
Length of learning ENG	15	4.5	3	**.009***
F2-females	**Est**	**SE**	**t**	***p***
(Intercept)	1775	78	23	**<.001***
Language (MAN)	-446	66	-7	**<.001***
Language (ENG)	-408	57	-7	**<.001***
Other languages (yes)	129	47	3	**.012***
F2-males	**Est**	**SE**	**t**	***p***
(Intercept)	1735	68	26	**<.001***
Language (MAN)	-253	53	-5	**<.001***
Language (ENG)	-240	53	-5	**<.001***
MAN as MOI (yes)	-242	102	-2	**.035***
F3-females	**Est**	**SE**	**t**	***p***
(Intercept)	3256	154	21	**<.001***
Language (MAN)	-23	81	-0.3	.787
Language (ENG)	84	64	1.3	.223
Length of learning MAN	-30	8	-4	**.003***
Other languages (yes)	310	40	8	**<.001***
F3-males	**Est**	**SE**	**t**	***p***
(Intercept)	2418	52	46	**<.001***
Language (MAN)	105	43	2.5	**.028***
Language (ENG)	124	31	4	**.005***

Table 3. Summary of fixed effects in the mixed-effects regression models for F1 – F3 of *uh*.

For Cantonese interviews, the frequency of [ɛ] was (82%) much higher than [a] (18%). In contrast, [ɛ] was less common for MAN (19%) and ENG (15%). Figure 2 shows the vowel plots for all these vowels in three languages. The vowel [a] was similar across the three languages, whereas the CAN-[ɛ] vowel was distinguished from the MAN-[ɛ] and ENG-[ɛ] for both male and female participants.

4. DISCUSSION

This study investigated Cantonese-Mandarin-English trilingual speakers' filled pauses in three languages. Different from Lo [6], distributional difference was found between Cantonese and English, but not between Cantonese and Mandarin. This could be because Cantonese and Mandarin are typologically similar, and the speakers might directly transfer the distribution of Cantonese filled pauses to Mandarin. Both Cantonese and Mandarin belong to the Sino-Tibetan language family. Although the two languages differ in phonology and lexicons, they have the same SVO word order and they both are tonal languages. As factors of language proficiency were not significant for the models (Table 2), the difference is unlikely due to speakers' proficiency in English and Mandarin.

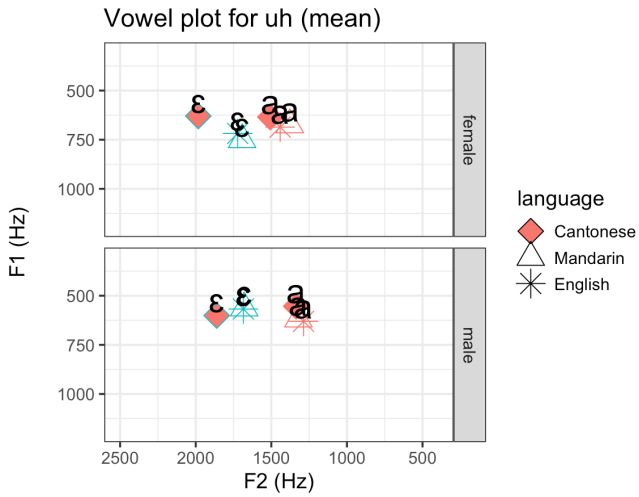

Figure 2. Vowel plot for *uh* by sex and language.

The cross-linguistic difference was mainly found in F2, which is in line with Lo [6] and Spreafico [7]. The significant frontness of Cantonese-*uh* is likely driven by the Cantonese-specific filled pause *uh*-[ɛ]. As shown in Figure 2, Cantonese *uh*-[ɛ] was much fronter than the other two languages, but *uh*-[a] was similar across the three languages. Although F2 was significantly different between Cantonese and the other two languages, no significant difference was found for F1, F2 and F3 between English and Mandarin (except for F1-male), suggesting that the speakers used similar vowels for *uh* in English and Mandarin. In other words, only part of the results supports the language-specific view.

Several factors of linguistic background had a significant impact on the speakers' use of filled pauses, including parents' L1, MAN as MOI, the length and age of learning Mandarin and English, and other languages. This reveals the necessity of taking these factors into consideration for future studies.

5. CONCLUSION

To conclude, this paper partially confirms the language-specific view with trilingual speech data: Cantonese-Mandarin-English trilingual speakers used a fronter CAN-[ɛ] than ENG-[ɛ] and MAN-[ɛ]. Speakers' linguistic backgrounds such as parents' L1 and age and length of learning L2 and L3 are found significant to predict their use of filled pauses in three languages.

6. ACKNOWLEDGEMENTS

This research is supported by Hong Kong RGC Postdoctoral Fellowship. We are grateful to So Ka Ki, Liu Xinyue, Wang Wenhui and Zhou Henan for their assistance with data coding and to Fung Hei for annotation crosschecking.

7. REFERENCES

[1] E. de Leeuw, "Hesitation Markers in English, German, and Dutch," *J. Ger. Linguist.*, vol. 19, no. 2, pp. 85–114, 2007.

[2] H. J. Künzel, "Some general phonetic and forensic aspects of speaking tempo," *Int. J. Speech Lang. Law*, vol. 4, no. 1, pp. 1350–1771, 1997.

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

[4] M. Gósy, D. Gyarmathy, and A. Beke, "Phonetic analysis of filled pauses based on a Hungarian-English learner corpus," *Int. J. Learn. Corpus Res.*, vol. 3, no. 2, pp. 149–174, 2017.

[5] 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, 2020.

[6] 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, 2020.

[7] L. Spreafico, "Filled pauses in multilingual speech: an acoustic analysis," *Linguist. e Filol.*, vol. 36, no. 2016, pp. 99–116, 2016.

[8] X. Wang and X. Jin, "汉语二语学习者口语非流利填充型停顿研究," 東北師大學報, vol. 3, pp. 84–92, 2020.

[9] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses," *Proc. Disfluency Spontaneous Speech, DiSS 2005*, no. September, pp. 179–182, 2005.

[10] C. Wu, "Filled Pauses in L2 Chinese: a comparison of native and non-native speakers," *Proc. 20th North Am. Conf. Chinese Linguist.*, vol. 1, pp. 213–227, 2008.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.