Pauses and Pause Fillers in Mandarin Monologue Speech: The Effects of Sex and Proficiency

Jiahong Yuan[1], *Xiaoying Xu*[2], *Wei Lai*[1], *Mark Liberman*[1]

[1] University of Pennsylvania
[2] Beijing Normal University

jiahong@ldc.upenn.edu, xuxiaoying2000@bnu.edu.cn, weilai@sas.upenn.edu, myl@ldc.upenn.edu

Abstract

In this study, we investigate the use of pauses and pause fillers in Mandarin Chinese. Our analysis is based on 267 spoken monologues from a Mandarin proficiency test. We identify two basic pause fillers in Mandarin: *e* and *en*. We find that males use more *e* than females, but there is no difference between them on the frequency of *en*. Therefore, the proportion of nasal-final pause fillers is higher in female than in male speakers, as was found in the studies of Germanic languages. Proficiency, on the other hand, does not affect the frequency of either *e* or *en*. With respect to the use of unfilled pauses, both sex and proficiency have a significant effect. Males and less proficient speakers use more medium and long, but not brief, pauses. Males tend to speak faster than females, they have a shorter *en*, but there is no difference between the two sexes on the duration of *e*. Un-proficient speakers produce shorter pause fillers, both *e* and *en*, than proficient ones. Finally, *en* is longer than *e*, it also precedes and follows a longer pause than *e*.

Index Terms: pauses, pause fillers, disfluency, proficiency, Mandarin Chinese

1. Introduction

Pauses are an important part of human speech [1-2]. According to various studies, the amount of pause time, as a percentage of the total speaking time, is between 5% and 20% in read speech, and between 30% and 46% in spontaneous speech (see review in [3]). Pauses may be filled with hesitation markers including both lexical (e.g., *like* and *you know* in English) and non-lexical items (e.g., *uh* and *um* in English). In this paper, we use the term pauses specifically for unfilled pauses (i.e., silent intervals), and the term pause fillers for non-lexical hesitation markers.

Pauses occur in connected speech for a number of reasons including physical, socio-psychological, communicative, linguistic and cognitive causes [4]. The acoustic silences that are a direct result of articulatory processes (e.g., stop closure) are normally not considered as a pause. A common practice in the literature is to exclude those silent intervals by choosing a minimum cut-off point somewhere between 100 and 300 milliseconds, although there has been a longstanding debate about the threshold [5-7]. Campione and Véronis (2002) analyzed pauses in 5½ hours of read and spontaneous speech in five languages [8]. They found that the distribution of pauses appears as trimodal, suggesting a categorization in brief (< 200 ms), medium (200-1000 ms), and long (>1000 ms) pauses.

Pause frequency and duration have been examined in terms of linguistic and social factors. It has been shown that there is a strong correlation between syntactic and prosodic complexity and pause duration [9-10]. Kendall (2009) showed that region, gender and ethnicity have significant influences on pause duration, and males tend to use longer pauses than females [11]. Clopper and Smiljanic (2011) showed that Southern male speakers use more pauses per intonation phrase (IP) than Southern female speakers and Midland speakers (both male and female) [12].

Like pauses, pause fillers have also been examined in terms of linguistic and social factors in the literature. Modern Germanic languages have two common pause fillers: a neutral vowel in an open syllable and a neutral vowel followed by a final bilabial nasal [13]. In American English they are generally written as *uh* and *um*. Clark and Fox Tree (2002) found that *um* was followed by pauses both more frequently and longer than *uh*, suggesting that the two pause fillers have different functions, i.e., they are used to announce the start of what are expected to be a minor (*uh*) or longer (*um*) delay [14]. Wieling *et al.* (to appear) investigated pause fillers in various Germanic languages and dialects through a quantitative analysis of a range of spoken and written corpora. They found that the use of *um* increased over time relative to the use of *uh*, and that the change is generally led by women and more educated speakers [13]. Tottie (2011) found that men, older people and educated speakers use more fillers than women, younger speakers and less educated speakers, and *um* is used more often by women, young speakers and more educated speakers [15]. Shriberg (2001) also reported that men used more filled pauses than women [16]. Laserna *et al.* (2014), however, found that filled pauses were used at comparable rates across genders and age, whereas discourse markers (*I mean*, *you know*, *like*) were more common among women, younger speakers, and more conscientious speakers [17].

Pauses and pause fillers have also been studied in the context of language learning and language proficiency. It has been consistently shown that pauses and much related speaking rate are the major contributing factors of speech fluency and proficiency [18-21]. Findings in the literature are, however, inconsistent regarding whether there is a correlation between the frequency of occurrence of pause fillers and L2 proficiency [18, 21].

There are relatively few studies on pauses and pause fillers in Mandarin Chinese. Strassel *et al.* (2005) reported their work on annotating spoken corpora in Mandarin Chinese for the purpose of metadata extraction (MDE). They identified four pause fillers in Mandarin: 嗯 (en) 唔 (um) 呃 (eh) 啊 (ah) [22].

Zhao and Jurafsky (2005) found that speakers in the south of China use significantly more pause fillers than those in the north, and the difference mainly lies in the use of *uh* [23]. Wu *et al.* (2012) found a positive correlation between pause duration and pitch reset in Mandarin for women but not men [24].

In this study, we investigate the use of pauses and pause fillers in Mandarin Chinese. We focus on two factors: speaker sex and proficiency. Our analysis is based on 267 spoken monologues from a Mandarin proficiency test, which will be described in Section 2.

2. Data

We used a dataset of *Putonghua Shuiping Ceshi* (PSC) from Beijing Normal University. PSC is the national standard Mandarin proficiency test in China. The test consists of four parts: The first two parts are to read 100 monosyllabic and 50 disyllabic words; the third part is to read an article of 300 characters, randomly selected from a pool of 60 articles; and the last part is to speak freely on a given topic for three minutes. The four parts are graded separately with a numeric score, and the total score (out of 100 points) is converted to a categorical proficiency level. There are six proficiency levels, which are, from high to low: 一级甲等 (Class 1 Level 1), 一级乙等 (Class 1 Level 2), 二级甲等 (Class 2 Level 1), 二级乙等 (Class 2 Level 2), 三级甲等 (Class 3 Level 1), and 三级乙等 (Class 3 Level 2). In order to qualify for teaching K-12, one must pass 二级乙等 (Class 2 Level 2).

Our dataset consists of recordings of college students at Beijing Normal University who took the PSC test in 2011. For this study we used the spoken monologues (the last part of the test) from 267 speakers, 178 female and 89 male, which contain approximately 13 hours of speech. The proficiency levels of the speakers range across four levels, from 一级乙等 to 三级甲等 (hereafter, L1 to L4). Table 1 shows the distribution of the speakers with regards to sex and proficiency.

	一级乙等 (L1)	二级甲等 (L2)	二级乙等 (L3)	三级甲等 (L4)
Male	10	31	30	18
Female	63	80	34	1

Table 1. The distribution of the speakers in the dataset.

3. Transcription and Forced Alignment

The spoken monologues were first transcribed by a professional transcriptionist, and then proofed by the first author for errors and pause fillers (which were ignored in the first pass). The pause fillers were categorized into two types in the transcription: one without nasalization (transcribed as *e*) and one with nasalization (transcribed as *en*).

We then conducted forced alignment to determine the vowel quality and nasal place of articulation in the pause fillers. Based on what we observed in the transcription process, three vowel qualities, /a/ (low), /e/ (central), and /o/ (rounded), were used as pronunciation alternatives for *e*; and two nasals, /m/ and /n/, for *en*. The nasals could either be a syllabic consonant (/m/, /n/) or follow a vowel /e/ (/e m/, /e n/, /en/. Please note that /em/ is not a legitimate rime in Mandarin). The acoustic models used for the alignment were trained on Mandarin Broadcast News Speech (LDC98S73) [25].

The results from forced alignment are shown in Table 2. We can see that the most common pause fillers in Mandarin are /e/ and /en/, a neutral vowel in an open syllable or nasalized with the alveolar nasal /n/. In contrast to Germanic languages, the bilabial nasal /m/ is far less used in Mandarin pause fillers.

	/a/	/e/	/o/
e (1192)	41	1065	86

	/e m/	/e n/	/en/	/m/	/n/
en (1866)	367	134	1126	174	65

Table 2. The phonetic qualities of pause fillers determined by forced alignment.

Forced alignment was also applied to determine and segment pauses in the spoken monologues, through inserting a "tee-model" for possible inter-word silence. A "tee-model" has a direct transition from the entry to the exit node in the HMM; therefore, a silence with a "tee-model" can have "zero" length (i.e., no silent interval). A state-of-the-art Mandarin forced aligner was used for the task [25].

In total, the 267 monologues contain 3,058 pause fillers, of which 1,192 are *e* and 1,866 are *en*; 26, 885 unfilled pauses, and 100,212 words.

4. Analysis and Results

4.1. Frequencies of pauses and pause fillers

The total numbers of pause fillers, pauses, and words for males and females and for different proficiency levels are listed in top part of Table 3.

	Female	Male	L1	L2	L3	L4
# *e*	671	521	384	403	352	53
# *en*	1287	579	568	655	478	165
# pauses	17828	9057	7231	11165	6591	1898
# words	68331	31881	29514	42461	22695	5542
e/(e+en)	***0.343***	***0.474***	0.403	0.381	0.424	0.243
e/words	***0.010***	***0.016***	0.013	0.010	0.016	0.010
en/words	0.019	0.018	0.019	0.015	0.021	0.030
(e+en)/words	0.029	0.035	0.032	0.025	0.037	0.039
pauses/words	***0.261***	***0.284***	***0.245***	***0.263***	***0.290***	***0.343***

Table 3. Frequencies and relative frequencies of pauses and pause fillers.

For each speaker we compute five relative frequencies:

e/(*e*+*en*): the proportion of *e* in pause fillers;

e/words: the number of *e* per word;

en/words: the number of *en* per word;

(e+en)/words: the number of pause fillers per word;

pauses/words: the number of pauses (including all silent intervals) per word.

Mixed-effects logistic regression models [26] were used to assess the effects of sex and proficiency level on the relative frequencies of pauses and pause fillers, in which speaker was treated as a random factor (*glmer(. ~ sex + proficiency + (1|speaker), family=binomial*). The results are shown in the bottom part of Table 3, where the mean values of the five relative frequency measures are listed, with bold and italic numbers representing statistical significance at $p < .05$.

From Table 3 we can see that males use more *e* than females, but there is no difference between them on the frequency of *en*. Therefore, males have a greater percentage of *e* in their pause fillers than females. Proficiency does not appear to affect the frequency of either *e* or *en*. With respect to the use of pauses, both sex and proficiency are a significant factor. Males use more pauses than females, and less proficient speakers also use more pauses.

In Table 4 we group pauses into three categories based on their duration, following Campione and Véronis (2002): brief (< 200 ms), medium (200-1000 ms), and long (> 1000 ms), and conduct mixed-effects logistic regression on the relative frequency for each of the pause categories. As above, bold and italic numbers represent statistical significance $p < .05$. We can see that sex and proficiency have a significant effect on the relative frequency of medium and long pauses but not brief pauses.

	Female	Male	一级乙等	二级甲等	二级乙等	三级甲等
brief	0.062	0.059	0.056	0.061	0.066	0.063
medium	***0.175***	***0.190***	***0.171***	***0.176***	***0.189***	***0.207***
long	***0.033***	***0.050***	***0.026***	***0.035***	***0.047***	***0.076***

Table 4. Relative frequencies of brief, medium and long pauses.

4.2. Durations of pauses and pause fillers

To assess the effects of sex and proficiency level on the durations of pause fillers, pauses, and syllables, we used mixed-effects linear regression models [26], in which speaker was treated as a random factor (*lmer(. ~ sex + proficiency + (1|speaker)*). It has been demonstrated in the literature that the distribution of speech segment duration, including pauses, is skewed and close to lognormal, and therefore logarithmic values should be used for statistical analysis [27]. In the mixed-effects regression analysis, we used log milliseconds for duration. Table 5 lists the mean durations (in seconds) for males and females and for different proficiency levels, with bold and italic numbers representing statistical significance at $p < .05$.

	Female	Male	L1	L2	L3	L4
e	0.237	0.241	0.240	0.239	0.244	***0.203***
en	***0.282***	***0.223***	0.273	0.263	0.284	***0.176***
pauses	***0.511***	***0.615***	***0.487***	***0.531***	***0.582***	***0.734***
syllables	***0.200***	***0.185***	0.195	0.194	0.196	0.195

Table 5. Average durations of pauses, pause fillers, and syllables.

From Table 5 we can see that males tend to speak faster than females (they have shorter average syllable duration after excluding pauses, i.e., faster "articulation rate"), but make longer pauses. Proficiency does not affect articulation rate. However, like male speakers, less proficient speakers make longer pauses. Males have a shorter *en* than females, but there is no difference between them on the duration of *e*. The effect of proficiency on the duration of the two pause fillers is interesting and puzzling. The speakers in L4 (三级甲等, Class 3 Level 1) have a shorter *e* and *en* than the other speakers. These speakers did not qualify for teaching K-12. They are, generally speaking, not proficient. The result suggests that un-proficient speakers produce shorter pause fillers than proficient ones.

QQ plots were drawn in Figure 1 to compare the distribution of pause duration between males and females, and between proficiency levels, respectively. From the figure we can conclude that males use more longer pauses than females, and less proficient speakers use more longer pauses than more proficient ones.

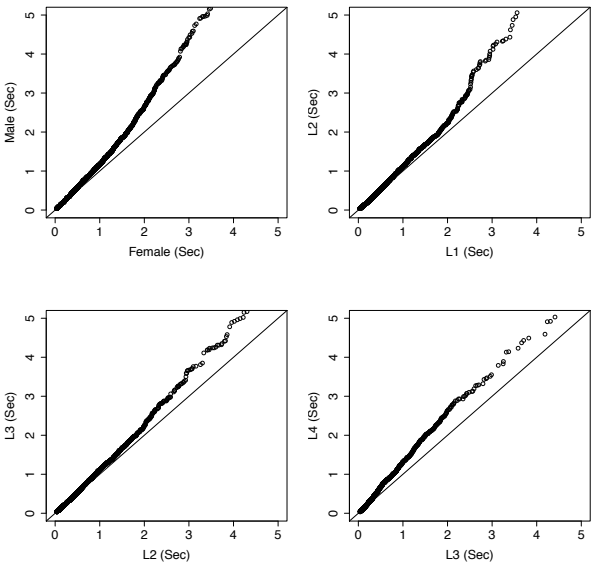

Figure 1. QQ plots of pause duration between males and females and between proficiency levels.

Figure 2 draws the mean durations of the two pause fillers, the mean durations of pauses following a pause filler (the duration is 0 if there is no such pause), and the mean durations of pauses before a pause filler (the duration is 0 if there is no such pause). We can see that *en* is longer than *e*, and the pause is longer both before and after *en* than *e*.

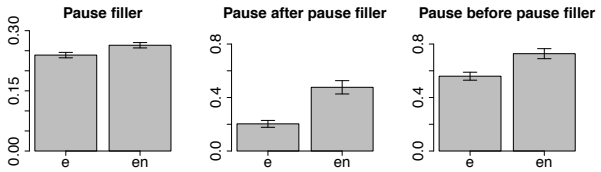

Figure 2. Mean durations of pause fillers and pauses before and after a pause filler.

5. Conclusions

Our study was based on 13 hours of monologue speech from 267 speakers. Through the combination of manual word transcription and phonetic forced alignment, we identified two

basic pause fillers in Mandarin Chinese: *e* and *en*. Different from English and other Germanic languages, the bilabial nasal /m/ is far less used than the alveolar nasal /n/ in Mandarin pause fillers.

The study further demonstrates the distinction between the two basic pause fillers in speech. First, *en* is longer than *e*, it also precedes and follows a longer pause than *e*; Secondly, males have a shorter *en* than females, whereas there is no difference between the two sexes on the duration of *e*; Thirdly, males use more *e* than females, but there is no difference between them on the frequency of *en*.

The proportion of nasal-final pause fillers is higher in female than in male speakers, as was found in the studies of Germanic languages. Proficiency, on the other hand, does not affect the frequency and proportion of *e* and *en*.

Both sex and proficiency have a significant effect on the occurrence of pauses. Males and less proficient speakers use more medium and long pauses, suggesting an interesting interaction between social and behavioral factors in language production.

Un-proficient speakers produce shorter pause fillers, both *e* and *en*, than proficient ones. Further research is needed to explain this result. Finally, males tend to speak faster than females, which is consistent with some previous studies [28-29].

6. References

[1] F. G. Eisler, *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, 1968.

[2] J. Fletcher, "The prosody of speech: Timing and rhythm," *The Handbook of Phonetic Sciences, 2nd Edition*, pp. 521-602, 2010.

[3] J. Trouvain, *Tempo variation in speech production: Implications for speech synthesis*, PhD thesis, University of Saarland, 2003.

[4] A. Esposito, V. Stejskal, Z. Smékal, and N. Bourbakis, "The significance of empty speech pauses: Cognitive and algorithmic issues," *Advances in Brain, Vision, and Artificial Intelligence*, Springer Berlin Heidelberg, pp. 542-554, 2007

[5] A. E. Hieke, K. Sabine, and D. C. O'Connell, "The trouble with ' articulatory' pauses," *Language and Speech*, 26, pp. 203-214, 1983.

[6] N. H. De Jong and H. R. Bosker, "Choosing a threshold for silent pauses to measure second language fluency," *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, pp. 17-20, 2013.

[7] M. Wlodarczak and P. Wagner, "Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps," *Interspeech 2013*, pp. 1434–1437, 2013.

[8] E. Campione and J. Véronis. "A large-scale multilingual study of silent pause duration," *Speech prosody 2002,* pp. 199-202, 2002.

[9] F. Grosjean, L. Grosjean, and H. Lane. "The patterns of silence: Performance structures in sentence production." *Cognitive psychology*, 11, pp. 58-81, 1979.

[10] J. Krivokapić, "Prosodic planning: Effects of phrasal length and complexity on pause duration," *Journal of phonetics*, 35, pp. 162-179, 2007.

[11] T. Kendall, *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis*, PhD thesis, Duke University, 2009.

[12] C. G. Clopper and R. Smiljanic, "Effects of gender and regional dialect on prosodic patterns in American English." *Journal of phonetics*, 39, pp. 237-245, 2011.

[13] M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman, "Variation and change in the use of hesitation markers in Germanic languages," to appear in *Language Dynamics and Change*.

[14] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, 84, pp. 73-111, 2002.

[15] G. Tottie, "Uh and um as sociolinguistic markers in British English," *International Journal of Corpus Linguistics*, 16, pp. 173-197, 2011.

[16] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association* 31, pp. 153-169, 2001.

[17] C. M. Laserna, Y. Seih, and J. W. Pennebaker, "Um... Who Like Says You Know Filler Word Use as a Function of Age, Gender, and Personality," *Journal of Language and Social Psychology*, 33, pp. 328-338, 2014.

[18] P. Lennon, "Investigating fluency in EFL: A quantitative approach," *Language learning*, 40, pp. 387-417, 1990.

[19] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America* 107 pp. 989-999, 2000.

[20] T. M. Derwing, , M. J. Rossiter, M. J. Munro, and R. I. Thomson, "Second language fluency: Judgments on different tasks," *Language learning* 54, pp. 655-679, 2004.

[21] N. Iwashita, A. Brown, T. McNamara, and S. O'Hagan, "Assessed levels of second language speaking proficiency: How distinct?" *Applied linguistics*, pp. 1-26, 2007.

[22] S. Strassel, J. Kolář, Z. Song, L. Barclay, and M. Glenn, "Structural metadata annotation: Moving beyond English," *Interspeech 2005, pp. 1545-1548,* 2005.

[23] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses," *Disfluency in Spontaneous Speech*, 2005.

[24] Q. Wu, B. Wang, and X. Zhang, "Effect of topic structure and sentence length on pause in Mandarin Chinese: Comparing female with male speakers," *Speech Prosody*, 2012.

[25] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone," *Acoustics, Speech and Signal Processing (ICASSP),* pp. 2539-2543, 2014.

[26] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67, pp. 1-48, 2015.

[27] K. M. Rosen, "Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison," *Journal of Phonetics*, 33, pp. 411-426, 2005.

[28] H. Quené, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *The Journal of the Acoustical Society of America,* 123, pp. 1104-1113, 2008.

[29] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," *Interspeech 2006*, pp. 541-544, 2006.