Speech Communication 101 (2018) 1–10

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Orthographic effects on the perception and production of L2 mandarin tones☆

Peggy Pik Ki Mok[a,*], Albert Lee[a,b], Joanne Jingwen Li[a], Robert Bo Xu[a]

[a] Department of Linguistics and Modern Languages, Leung Kau Kui Building, The Chinese University of Hong Kong, Shatin, Hong Kong
[b] Department of English Language, University of Tokyo, Komaba 3-8-1, Tokyo 153-8902, Japan

ARTICLE INFO

Keywords:
Orthography
Tone
Perception
Production
Acquisition
Cantonese
Mandarin

ABSTRACT

Recent studies on orthographic effects on L2 phonology have typically investigated alphabetic writing systems and segmental contrasts with novice learners. The current study extends such investigation to compare orthographic effects of an opaque logographic system (Chinese characters) and a transparent schematic system (*pinyin*) on a suprasegmental feature (lexical tones) with experienced learners. A perception experiment of Mandarin tones by Cantonese L2 learners shows that *pinyin* was more beneficial for tone perception in monosyllabic words, while tones were better perceived in characters for disyllabic words. A production experiment reveals a similar pattern. Additionally, low performance learners were affected by orthographic differences more than high performance learners. The findings suggest that orthographic effects are not limited to alphabetic systems, and are dependent on task nature and learner proficiency. A transparent system may not always be easier than an opaque system.

1. Effects of orthography on L2 phonology

Studies on the acquisition of L2 phonology abound, but very few have emphasized the effects of orthography because of the primacy of spoken input in acquisition research. Nevertheless, as rightly pointed out by Bassetti (2008), L2 learners are often simultaneously exposed to written and spoken input from the beginning of L2 learning. This stands in stark contrast to child first language acquisition in which input is solely spoken in nature. Children only start to learn the writing system after they have acquired the phonology of their first language. The question of how orthography affects L2 phonological acquisition is a valid, and yet underexplored, one. Researchers have started to investigate the roles of orthographic input on L2 phonological acquisition in the past two decades. Mixed results are reported among these studies covering different languages.

Some studies have demonstrated a positive effect of orthography in helping learners to discriminate L2 phonological contrasts that are otherwise difficult to distinguish. For example, Dutch learners of English often find the /æ/ and /ɛ/ contrast difficult. Using an eye-tracking paradigm, Escudero et al. (2008) showed that Dutch learners could differentiate the confusable English /æ/ and /ɛ/ contrast in non-words if they were exposed to both the auditory and the spelled forms of the words during training, as opposed to those only exposed to auditory forms. Erdener and Burnham (2005) tested the effects of orthographic depth on non-native speech production. They mentioned that orthographic depth can be defined as the degree to which an alphabetic system deviates from simple one-to-one grapheme-to-phoneme correspondences. Writing systems vary along a continuum of orthographic depth, some having very regular and unambiguous grapheme-phoneme correspondence (transparent) while others do not (opaque). Erdener and Burnham (2005) compared the production of non-words in Spanish (transparent) and Irish (opaque) by Australian English and Turkish speakers with and without audiovisual cues. English has an opaque orthography while Turkish has a transparent orthography. They found that the presence of transparent orthography enhanced production accuracy in general, and that orthographic information, when provided, even overrode the general facilitative effect of visual information. Furthermore, orthography was beneficial for transparent Turkish speakers on transparent Spanish but not on opaque Irish, while there was little difference for opaque Australian English speakers on Spanish and Irish. The results of Escudero et al. (2008) and Erdener and Burnham (2005) demonstrate that orthographic input can facilitate both the perception and the production of non-native speech.

Other studies found that orthographic input can induce non-target-like errors which cannot be explained by the spoken input. For instance, beginning learners of Chinese are often taught the official Chinese

☆ We thank Donghui Zuo for her help in the early stages of the project, and Xi Chen for analyzing the production data. We also thank the participants for taking part in the experiments.
* Corresponding author.
E-mail address: peggymok@cuhk.edu.hk (P.P.K. Mok).

https://doi.org/10.1016/j.specom.2018.05.002
Received 21 August 2017; Received in revised form 17 April 2018; Accepted 9 May 2018
Available online 16 May 2018

Romanization system of *pinyin*. For some triphthongal rimes, the main vowels are omitted in *pinyin* spelling. Using a phoneme counting task and a phoneme segmentation task, Bassetti (2006) found that English learners of Chinese would omit the main vowel when it is not represented in *pinyin* (e.g. the same /iou/ sound was counted as two phonemes when spelled as *-iu* but counted as three when spelled as *you*). In addition, two studies on English learners of German also demonstrated that orthography could hinder the learning of L2 phonology. Word-final obstruents in German are devoiced, but they are represented using letters for voiced sounds in spelling, e.g. ⟨bund⟩ for [bʊnt]. Young-Scholten (2002) found that the amount of exposure to orthographic input was inversely related to the rate of final devoicing in the production of English learners of German. Young-Scholten and Langer (2015) reported another interesting case. The German ⟨s⟩ is pronounced as [z] word-initially. Three English teenagers who were exchange students in Germany for one year learning German in an immersion environment, despite having ample correct auditory input from native speakers, pronounced ⟨s⟩ as [s] throughout their exchange period. These results showed that orthographic representation could interfere with the mental representations of L2 phonology even with correct auditory inputs.

In addition, many studies using a word learning paradigm also demonstrated the effects of orthography. Hayes-Harb et al. (2010) manipulated the congruency of orthography on word learning. Three English participant groups in their study received the same auditory input and pictures, but they differed in the written input received during training. One group saw spelling consistent with English conventions (congruent, e.g. ⟨kamad⟩ for [kɑməd]); another group saw wrong spelling inconsistent with English conventions (incongruent, e.g. ⟨kamand⟩ for [kɑməd]), and the auditory-only group saw only ⟨XXXX⟩. Their results showed that the group seeing inconsistent spelling performed the least accurately overall due to the wrong letter spelling. Similarly, the effects of congruent versus incongruent spellings were borne out in Hayes-Harb and Cheng (2016). They asked native English speakers to learn novel Mandarin words of some real object drawings. The speakers were assigned to two types of orthographic input: *pinyin* (Romanized spellings) and *zhuyin* (a semi-syllabary system not using Roman letters). Some Mandarin words have *pinyin* forms congruent with English spelling, e.g. ⟨nai⟩ for [nai], while others have incongruent forms with English spelling, e.g. ⟨xiu⟩ for [ɕiou] (the corresponding English pronunciation for ⟨x⟩ would be [z] or [ks]). They found that the *zhuyin* group outperformed the *pinyin* group on incongruent items due to the poorer performance of the *pinyin* group on such items, while both groups did not differ in their ability to distinguish the relevant sounds auditorily. They argued that despite the familiarity with *pinyin* (Roman letters), native English speakers had to suppress the grapheme-phoneme conversion in their L1 for the incongruent items, which led to poorer performance than those who had to learn an entirely new writing system (*zhuyin*).

A number of studies showed that orthography might have no or only limited influence on L2 phonology. Simon et al. (2010) tested whether English speakers' discrimination of the French /y/ and /u/ contrast would be enhanced by the presence of orthographic representations, but found no difference with and without orthographic input. Showalter and Hayes-Harb (2015) found that English speakers could not benefit from an unfamiliar script when learning a novel and difficult uvular-velar contrast (/q k/) in Arabic. Escudero et al. (2014) working with Spanish listeners of Dutch and Escudero (2015) working with Spanish and English listeners of Dutch both found that orthographic input was beneficial only when orthography is congruent between L1 and L2, or only for easy contrasts.

Many of the previous studies on orthographic effects used a word learning or recognition paradigm testing listeners' perceptual performance on foreign contrasts. Erdener and Burnham (2005) mentioned above illustrated that orthographic effects can be found on learners' production as well. Recently, Hayes-Harb et al. (2017) tested naive English speakers' production of final devoicing in German using a word learning paradigm. Participants who were exposed to the written forms during the learning phase were more likely to produce final voiced obstruents. An explicit instruction about the misleading nature of the orthographic input had no effect on participants' production of final voiced obstruents. This indicates the powerful influence of orthographic input, echoing the findings of Young-Scholten (2002) and Young-Scholten and Langer (2015) also on the production of German final devoicing discussed earlier.

Despite the fact that various results have been observed, one general conclusion that can be drawn from the above studies is that, transparency and congruence are important factors modulating the effects of orthographic input (if any) on L2 phonological acquisition: transparent and congruent orthographic forms can be positive while opaque and incongruent forms can be negative.

Most of the previous studies on orthographic influence were understandably on segmental contrasts, as these contrasts can be clearly captured by different spellings. However, the conclusion based on studies examining segmental contrasts can be extended to suprasegmental contrasts as well.

Two suprasegmental aspects have been examined: lexical stress and lexical tone, but the findings were mixed. Both inexperienced and experienced English learners of Russian did not benefit from the provision of stress marks, or from the use of Latin or Cyrillic script in the acquisition of Russian lexical stress contrasts (Hayes-Harb and Hacking, 2015). In contrast, Showalter and Hayes-Harb (2013) tested naive English speakers' learning of Mandarin tones with and without tone marks. One group was given *pinyin* together with tone marks as diacritics (e.g. < gí >) while the other group was only given *pinyin* with no tone mark (e.g. <gi>). The tone marks are schematic representations of the pitch contours of the four Mandarin tones: level tone [55] (ū), rising tone [35] (ú), dipping tone [214] (ǔ) and falling [51] tone (ù), where the numbers in [] are tone values on a 1–5 scale, with 1 corresponding to the lowest pitch level of a speaker's normal pitch range and vice versa (Chao, 1930). The tone marks are novel symbols to English speakers while *pinyin* resembles English spelling otherwise. The tone-mark group outperformed the non-tone-mark group across tones and across experiments in Showalter and Hayes-Harb's (2013) study. Their findings suggest that orthographic effects in L2 phonology are not limited to segmental contrasts only. Nevertheless, given the contrary findings and very few studies on suprasegmental contrasts, more investigation is needed for a comprehensive understanding of the effects of orthographic inputs.

2. Chinese characters and phonology

Most of the studies reviewed above dealt with alphabetic writing systems. They typically showed that an opaque orthography was a hindrance to L2 phonology. It follows that an opaque logographic writing system like Chinese characters, which does not have regular grapheme-to-phoneme correspondence or indicate lexical tones, will pose difficulties to L2 learners of Mandarin, compared to the alphabetic system of *pinyin*. This may be the case for genuine beginning learners of Mandarin, although Hayes-Harb and Cheng (2016) showed that naive English speakers being trained with the *zhuyin* system (with symbols resembling parts of a Chinese character) outperformed those being trained with *pinyin* for incongruent items. A complete understanding of orthographic effects on L2 phonology requires looking beyond alphabetic writing systems, but the challenges of doing experiments using logographic Chinese characters with beginning learners cannot be underestimated. A possible alternative to this problem is to approach the research question from a different perspective: using learners who are already familiar with Chinese characters and *pinyin*. There are many L2 learners of Mandarin whose first language is a Chinese dialect (e.g. Cantonese) which is quite different from Mandarin. Cantonese and Mandarin are mutually unintelligible, but share the same writing system of Chinese characters. These L2 learners already have a good knowledge of the opaque writing system of Chinese characters and have firmly established the character-to-sound mappings in their L1. In this case, would a more transparent system (*pinyin)* be beneficial to them when they learn Mandarin? How do the effects of a transparent orthography (*pinyin*) compare with those of L1 phonological knowledge (retrieved from Chinese characters) for these learners? To the best of our knowledge, there is no study investigating the effects of these two divergent orthographic systems on learners' phonological performance. Our study can widen the investigation of orthographic effects on L2 phonology by incorporating a non-alphabetic writing system as well.

Many character recognition studies have shown that Chinese native speakers retrieve phonological information instantly from the printed Chinese characters (e.g. Perfetti and Zhang, 1991; Tan et al., 1995; Zhou and Marslen-Wilson, 1999). For instance, Perfetti and Zhang (1991)(1991) showed that phonetic masksw (e.g. 事 /shi/ 'matter') led to facilitation of the identification of target characters (e.g. 視 /shi/ 'see') to a similar degree as semantic masks (e.g. 看 /kan/ 'see') do. Some scholars argued that phonological code activation might even precede the retrieval of semantic information associated with particular characters (e.g. Perfetti and Tan, 1998; Zhang and Weekes, 2009). If phonological codes are activated instantly during character recognition, then for Chinese learners of Mandarin the influence from L1 (Cantonese) phonology would be stronger when they see Chinese characters than when they see *pinyin*. Since Chinese characters are opaque with regards to tone marking compared with *pinyin,* it follows that learners would have better performance in both speech production and perception when they are presented with transparent *pinyin* than when they are presented with opaque Chinese characters. However, so far, there is no study comparing how Chinese characters and *pinyin* would affect the performance of learners of Mandarin who speak another Chinese language as their L1.

3. The present study

In order to examine the above prediction, we decided to work on Mandarin tones for two reasons. First, most previous studies on orthographic effects have dealt with segmental contrasts. Recently, Showalter and Hayes-Harb (2013) showed that orthographic effects could be found in the acquisition of tones as well. It will be useful to extend their findings on suprasegmentals with different methods and groups of speakers to both tone production and perception. Second, Mandarin tones are represented transparently and schematically in *pinyin*, while they are opaque in Chinese characters. This can ensure the maximum difference in transparency between the two orthographic systems.

We focus on the perception and the production of Mandarin tones by Cantonese learners of Mandarin in Hong Kong. Unlike many other dialectal speakers in Mainland China, many Cantonese speakers in Hong Kong receive education mostly in Cantonese (at least during primary school education), and they use mainly Cantonese in their daily life. They learn Mandarin as a subject in school, but with only a couple of lessons a week. Thus, Mandarin is clearly an L2 for them, despite their exposure to it from a very young age. Early introduction of the *pinyin* system accompanies Chinese characters even in elementary Mandarin textbooks. Moreover, Hong Kong students were not taught any Romanization system of Cantonese. They learnt Chinese (as a standard written language) in Cantonese purely by using characters. Thus, there would be no interference from Cantonese Romanization when they see *pinyin*.

Cantonese and Mandarin differ in their tone systems. As mentioned above, there are four tones in Mandarin which differ in pitch contour (Duanmu, 2007; Howie, 1976): T1 (level [55], ū), T2 (rising [35], ú), T3 (dipping [214], ǔ) and T4 (falling [51], ù). The tone system of Cantonese is more complicated with six tones differing in both pitch contour and pitch height (Bauer and Benedict, 1997): T1 (high-level, [55]), T2 (high-rising, [25]), T3 (mid-level, [33]), T4 (low-falling, [21]), T5 (low-rising, [23]) and T6 (low-level, [22]). The two historically related languages have some regular rules of tonal correspondence, as shown in Table 1 (Cheung and Gao, 2000; Tsang-Cheung, 1988). However, learners and also teachers are generally not aware of such tonal correspondence. We can predict that when Cantonese learners of Mandarin are presented with Chinese characters, they can be easily influenced by the tonal mappings in their L1. Chinese characters following the tonal correspondence in Table 1 can be considered 'congruent' for Cantonese learners, while those not following the tonal correspondence can be considered 'incongruent' for them. Meanwhile, the effect of tonal correspondence would not be relevant for *pinyin* as this system is unrelated to their L1.

Table 1
Rules of tonal correspondence between Cantonese and Mandarin based on Cheung and Gao (2000) and Tsang-Cheung (1988).

Mandarin tones	Cantonese tones
T1 [55]	T1 [55] (84%)
T2 [35]	T4 [21] (76%), T6 [22] (12%)
T3 [214]	T2 [25] (60%), T5 [23] (25%)
T4 [51]	T6 [22] (47%), T3 [33] (40%)

A tone perception and a follow-up tone production experiment were conducted to compare Cantonese learners' performance in Chinese characters and in *pinyin* conditions. Previous studies investigated orthographic effects on perception and production separately. We would like to see if similar orthographic effects can be found in both aspects of speech. There were two sets of materials in each experiment: monosyllabic words and disyllabic words. The two sets of materials were originally designed for different research questions. This explains the discrepancy in their designs, e.g. number of choices (see detail below). Nevertheless, we saw the value of reporting these two sets of materials together in this study as this will give us a more comprehensive understanding on our problem at hand. In what follows, we will present the two sets of data separately first, and then present them for a general comparison. No statistical comparison was done to compare the two sets of data. Caution should also be exercised in interpreting their patterns.

4. The perception study

4.1. Participants

The participants were 49 undergraduate students at a university in Hong Kong. All were native speakers of Hong Kong Cantonese. Most of them started to learn Mandarin through formal education in primary school, but with varying amount of Mandarin instruction. There were at least two Mandarin lessons per week, but schools could vary a lot in the amount of extra-curricular Mandarin activities. In addition, some were taught by Mandarin teachers from mainland China, while others were taught by local Mandarin teachers. Since Mandarin is not a compulsory subject in public examinations in Hong Kong, there was no uniform score to objectively measure their Mandarin proficiency which can vary to some extent. They mainly spoke Cantonese in daily life, Cantonese and/or English in classroom settings, and only spoke Mandarin when necessary, for example, in Mandarin classes or talking to Mandarin speakers. *Pinyin* was taught as an assisting tool in the Mandarin classroom. None of the participants reported any history of hearing loss or speech impairment. They all received course credits for participation. All 49 listeners participated in the monosyllabic task, and 38 of them also participated in the disyllabic task.

4.2. Stimuli

Both monosyllabic and disyllabic real words were included in the perception experiment. Pseudo words could not be used because they do not have any tonal correspondence as in Table 1. For monosyllabic words, there were four minimal tone quartets, where each item was a real word in Mandarin. Two quartets were presented in Chinese characters, the other two in *pinyin*. In total there were 16 items in the monosyllabic task. The disyllabic words were minimal pairs that only differed in the tone of one syllable, either the first or the second one. There were six possible different tone-pair contrasts in Mandarin (e.g. T1T2, T3T4, but not T1T1), each presented visually in either Chinese characters or *pinyin*. There were eight items for each condition, yielding a total of 192 items for the disyllabic task (2 syllables × 6 tone-pair contrasts × 2 writing systems × 8 items). All perception materials were produced by a female Mandarin native speaker. It was confirmed that word frequency (mean = 2.69, SD 2.09, based on McEnery and Xiao, 2004) was not correlated with perception accuracy rates in the disyllabic task ($p > .05$, Pearson's correlation) whereas for the monosyllabic task we could not obtain this correlation due to the small number of target words.

4.3. Procedure

The experiment took place in a sound-treated room using the E-Prime 2.0 Professional software. Instructions were given both visually on the screen and orally by the experimenter. The participants listened to a Mandarin word via headphone, and chose the corresponding word on the screen in characters or *pinyin* by pressing buttons on a response box.

After a practice session, the participants finished four blocks of different tasks. The first two blocks, one presented in *pinyin*, the other in characters, required the listeners to choose the monosyllabic word they heard from four choices on the screen. There were altogether 16 trials in the monosyllabic blocks. In each trial, the four choices were segmentally identical and differed in the tones only, e.g. 'yīn yín yǐn yìn' in *pinyin* or '翻 反 犯 凡' in characters. In the following two blocks for disyllabic words, the two words in the minimal pairs (*pinyin* in one block, characters in the other) were shown on the screen for the listeners to choose from, e.g. 'gǔ shī gǔ shí' in *pinyin* or '剪刀 尖刀' in characters. In the disyllabic part, there were altogether 192 trials. All items were randomized in the blocks.

4.4. Results

We tested which orthography led to higher identification accuracy in monosyllabic and disyllabic words. The monosyllabic and disyllabic results will be examined separately. Below the first analysis is for the monosyllabic words, and the second is for the disyllabic words.

The average identification accuracy rate of monosyllabic words by all the listeners was higher in the *pinyin* condition (M = 91.07%, SD = 11%) than in the character condition (M = 76.27%, SD = 22%).

Table 2
Overall error rates (%) by orthography condition (pinyin vs. character) in the monosyllabic perception task.

	Responses								
	Pinyin				Character				
	T1	T2	T3	T4	T1	T2	T3	T4	
Stimulus	T1		0	0	0.26		1.79	2.04	0.26
	T2	0		0.51	0	0.77		3.57	2.81
	T3	0.77	1.53		0.26	2.55	4.34		0.51
	T4	3.06	0.77	1.28		1.02	3.32	0.26	

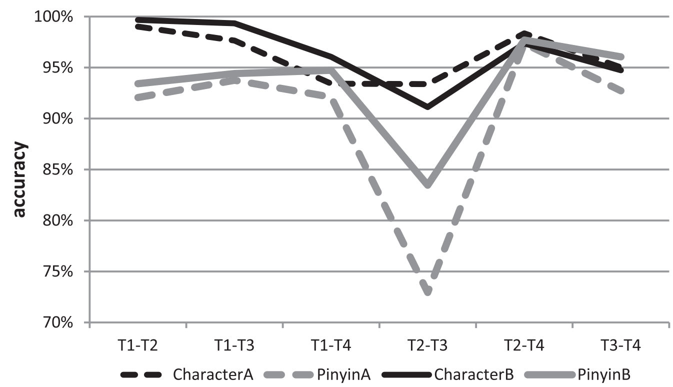

Fig. 1. Average identification accuracy for different tone pairs in four orthography × syllable conditions with disyllabic words.

Table 2 shows the overall error rates by all the listeners under the *pinyin* and character conditions separately. In the *pinyin* condition, listeners mainly misidentified T4 as T1. However, in the character condition, listeners made most confusion in the T2-T3 contrast.

We fitted a logistic mixed-effects model in R using the lme4 package (Bates et al., 2015, ver. 1.1–13). The best-fitting model was selected with a bottom-up approach based on Akaike Information Criteria (AIC) with the alpha level of 0.05. The fixed effect was Orthography (character vs. *pinyin*). Random intercepts for subjects and items were included. This model revealed that the main effect of orthography was significant ($\chi^2(1) = 4.381$, $p = .036$). The addition of other fixed factors (namely Target Tone and Word Frequency) did not lead to an improvement in model fit compared to the model without, as indicated by chi-square log-likelihood tests ($p > .05$).

Fig. 1 shows the average accuracy of different tone contrast conditions in the disyllabic task. In the figure legend, CharacterA and PinyinA refer to the first syllable of the disyllabic words, and CharacterB and PinyinB refer to the second syllable. The accuracy rates of the character items (lines in black) were higher than those of the *pinyin* items (lines in grey) in most cases. The most striking pattern in the figure is the drop in accuracy for the T2–T3 contrast, showing that this was the most difficult tone pair to be distinguished by the listeners. Moreover, the drop is especially obvious when the words were visually presented in *pinyin*.

We fitted another logistic mixed-effects model for verification, following the same procedure as the previous model. In the best fitting model (Table 3), fixed factors included Orthography, Tone Contrast, and their interaction. Random intercepts for subjects and items, as well as by-subject random slopes for Orthography and Tone Pair, were included. However, in this model the main effect of Syllable Position (1 or 2) was non-significant, as adding it to the model did not lead to significantly better model fit ($p > .05$, chi-square log-likelihood test). Table 3 shows that all tone contrasts were significantly better identified than T2-T3 ($p < .05$). Tones were also better identified in the character condition than in the *pinyin* condition ($\beta = 1.069$, SE = 0.321, Z = 3.333).

In order to have a general picture of the effects of orthography on the two sets of materials, Fig. 2 shows the average accuracy and reaction time (RT) data of both the monosyllabic and disyllabic tasks collapsed across syllable positions. As the participants chose among four options in the monosyllabic task, and between two options in the disyllabic task, their RTs in the two tasks were not directly compared. Nevertheless, Fig. 2 can still give us some indication of how they responded to the tasks. It is clear that the task difference was mainly found in the character condition. Both accuracy and reaction times were stable across tasks in the *pinyin* condition. Differentiating monosyllabic tones presented in characters was the most difficult for the listeners, resulting in the lowest accuracy and longest RTs. On the

Table 3
Model summary (disyllabic perception task): Correct response ~ Tone contrast * Orthography + (Tone contrast + Orthography|Subject) + (1|Item). P-values: * < 0.05, ** < 0.01, *** < 0.001.

Parameters	Fixed				Random	
	β	*SE*	*Z*		Subject SD	Item SD
(Intercept)	2.447	0.200	12.231	***	0.820	0.701
Tone contrast (T1T2–T2T3)	1.769	0.323	5.480	***	0.564	
Tone contrast (T1T3–T2T3)	1.509	0.297	5.075	***	0.588	
Tone contrast (T1T4–T2T3)	1.084	0.288	3.767	***	0.878	
Tone contrast (T2T4–T2T3)	1.597	0.305	5.243	***	0.513	
Tone contrast (T3T4–T2T3)	0.868	0.252	3.439	***	0.488	
Orthography	1.069	0.321	3.333	***	0.744	
Tone contrast (T1T2–T2T3): Orthography	1.270	0.570	2.227	*		
Tone contrast (T1T3–T2T3): Orthography	0.398	0.521	0.763			
Tone contrast (T1T4–T2T3): Orthography	−0.193	0.454	−0.424			
Tone contrast (T2T4–T2T3): Orthography	−1.088	0.509	−2.137	*		
Tone contrast (T3T4–T2T3): Orthography	−0.588	0.456	−1.290			

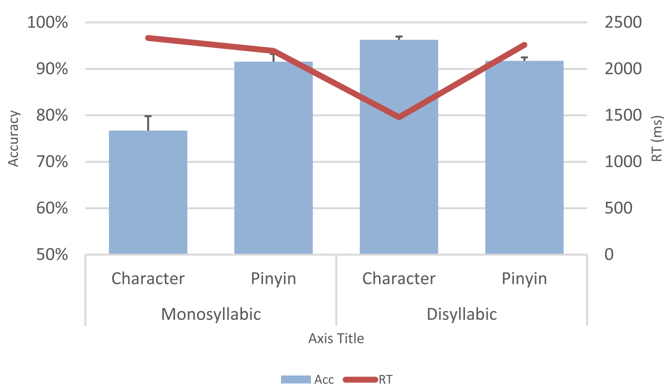

Fig. 2. Averaged identification accuracy and reaction times in different orthography × word length conditions. Error bars show one standard error.

contrary, disyllabic tones in characters were differentiated even better than those in *pinyin*, resulting in higher accuracy and much shorter RTs.

Finally, we tested the effect of tonal correspondence (cf. Table 1) on perception accuracy in disyllabic words presented in characters. Data from the monosyllabic blocks were not analysed as all monosyllabic target words had a regular Cantonese-Mandarin tonal correspondence (i.e. for a given character the tone in Cantonese and that in Mandarin conform to the patterns in Table 1). The mean perception accuracies of words with regular and irregular tonal correspondence were respectively 96.1% and 85.7%. A one-way ANOVA shows that the main effect of Tonal Correspondence was significant ($F(1,94) = 4.855$, $p = .03$, Levene's test $p = .181$). On the other hand, the effect of Tonal Correspondence was non-significant on RTs ($p = .124$).

4.5. Discussion

The results of the monosyllabic and the disyllabic tasks both show significant effects of orthography, but unexpectedly, the patterns were opposite. For monosyllabic tones, the items presented in *pinyin* were perceived significantly more accurately than those in characters. For disyllabic tones, the items presented in characters were significantly better recognized instead. This contrary pattern indicates that the orthography effect is task-dependent.

In the monosyllabic task, the listeners were to choose the target tone from four possible answers. When presented with *pinyin*, they did not need to access the conceptual route (i.e., they did not need to know exactly which words the *pinyin* spellings represent), but only paid attention to the phonetic difference of the tones they were hearing. In addition, the four choices on the screen had the same spelling and differed only in the diacritics that signified lexical tones. As a result, *pinyin*, with a transparent and schematic representation of the tones, seems beneficial during this task. The characters and phonological information are in an opaque correspondence. The task was more difficult because the listeners must identify the characters first before they could decide on their corresponding Mandarin tones. Thus, strong character-L1 tone mappings might have influenced listeners' judgment. It was also possible that the listeners were uncertain about the correct corresponding Mandarin tones for these characters.

The facilitating effect of *pinyin* over characters disappeared in the disyllabic task. Although tones in characters were better recognized than those in *pinyin*, the difference in perception accuracy between the two orthographies in the disyllabic task was much smaller than that in the monosyllabic task. As shown in Fig.2, the accuracy of *pinyin* was very similar across tasks, whereas for characters there was a large difference. Various reasons may have contributed to such discrepancy. Listeners had to choose a correct answer among four possibilities in the monosyllabic task, while they only had to choose between two answers in the disyllabic task. The aforementioned difficulty of the character condition was thus reduced in the disyllabic task. In turn, the relative ease of recognising the different tone diacritics in *pinyin* in the monosyllabic task was reduced in disyllabic words. Nevertheless, the large difference in RTs between the two tasks in the character condition but not in the *pinyin* condition reveals that the number of choices was not the main reason, as the number of choices was the same in the two orthography conditions.

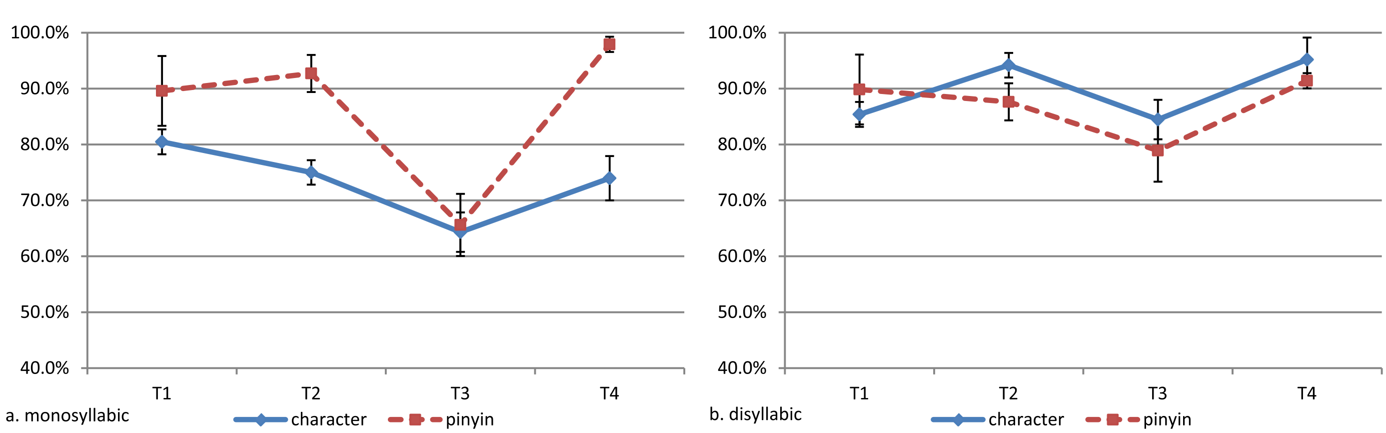

Fig. 3. Average production accuracy (%) for the four tones with two orthographies for (a) monosyllabic and (b) disyllabic words.

The RT data may give us some insights into participants' responses. It should be noted again that the monosyllabic and disyllabic tasks were different in many aspects. The following discussion should be treated with caution. The aforementioned difficulty of perceiving monosyllabic tones in the character condition was reflected in the rather long average RT, in addition to the low accuracy. The equally long RTs in the *pinyin* condition shows that reading *pinyin* was probably a controlled process for the participants, even when the transparency of *pinyin* had helped them in recognizing the tones accurately. Thus, what is worth noting is the facilitation of opaque characters in disyllabic words which resulted in an accuracy rate even higher than that in the transparent *pinyin* condition, and a much shorter average RT.

One possible reason of this facilitation is that when listening to disyllabic words presented in characters, listeners would perceive the words holistically and immediately knew which two meaningful words were being contrasted. The disyllabic words together with their tone combinations were stored as lexical units in listeners' mind. In turn, they could access the meanings and the tone combinations of the words quickly. In this way, the disyllabic task in characters involved both bottom-up and top-down processes, and listeners could get information both semantically and phonologically. Also presented with two choices, *pinyin* makes it harder for the listeners to determine which two words were involved, as there are many homophones in Chinese, and listeners might not realize which words they were from the Romanized spelling immediately. They could only get information phonologically.

Another possibility is that the high correspondence between Mandarin tones and Cantonese tones as shown in Table 1 could facilitate the perception with characters, as listeners could instantly gain access to the Cantonese pronunciation from the written characters. The high tonal correspondence may provide additional source of information of guessing the correct tones in Mandarin. In addition, since there were many more disyllabic than monosyllabic tokens in the experiment, and they involved more complicated segmental combinations, the listeners had to put in more effort in reading *pinyin* spellings for disyllabic words, whereas they were very proficient in recognizing Chinese characters. All these factors could contribute to listeners' better performance in perceiving disyllabic tones presented in characters than in *pinyin*.

5. The production study

Given the intriguing patterns found in tone perception, we extended the study to investigate the orthography effects on tone production as well.

5.1. Participants

As the results in the perception study revealed that the participants performed generally quite well, in order to obtain clear patterns in production, we chose only 16 speakers who participated in the perception study above for the production study. We first calculated the averaged perception accuracy for all participants, then, based on their perception accuracy, we chose the eight speakers with the highest accuracy and the eight with the lowest to form two performance groups. The perception accuracy for the high performance group ranged from 96.2% to 98.1% (mean = 97.4%), and from 81.6% to 92.3% (mean = 88.4%) for the low performance group.

5.2. Materials and recording

Similar to the perception experiment, the production experiment consisted of two conditions: a *pinyin* condition and a Chinese character condition. In order to be comparable to the perception design, both monosyllabic words and disyllabic words were used, and they were all real words. The production materials were different from the perception materials. As the participants were generally very good at producing the four citation tones in *pinyin*, two monosyllabic syllables with all four tones were selected as the targets for the *pinyin* condition, resulting in eight items (2 syllables × 4 tones). For disyllabic words in *pinyin*, we included all possible tone combinations with six items for each, resulting in 96 items (4 tones × 4 tones × 6 items). These 96 items formed 48 minimal pairs that contrasted only in the lexical tone of one syllable, either the first or the second. The *pinyin* tokens were all presented to the subjects in the standard form of *pinyin* (e.g. xīng), where the symbol above the letters represent the tone. For the character condition, 32 monosyllabic words (4 tones × 8 items) and 96 disyllabic words were used. It was confirmed that for the tokens in the disyllabic character condition, word frequency and production accuracy were not significantly correlated with each other ($p > .05$, Pearson's correlation).

All the materials were shown to the speakers on paper. The monosyllabic *pinyin* block was followed by the monosyllabic character block, which was in turn followed by the disyllabic *pinyin* and the disyllabic character blocks. Tokens in each group were randomized. The production experiment was conducted in a quiet room at a university in Hong Kong. The recordings were taken with a solid state recorder with a sampling rate of 44,100 Hz. Three repetitions were recorded.

5.3. Data analysis

Three native Mandarin speakers with training in phonetics transcribed the tones produced in all the recordings. Transcriptions were agreed upon by at least two transcribers as the same tone for use; otherwise, the token was excluded from the data analysis. In the *pinyin* condition, only one token was excluded without agreement; another three tokens were considered by all three transcribers as not being any of the four Mandarin tones. In the character condition, two tokens were not agreed on by any two of the three transcribers and were excluded for further analysis.

5.4. Results

Table 4 shows the average accuracy of tone production by high and low performance speakers in both *pinyin* and character conditions. We can see that for monosyllabic words, the accuracy of the *pinyin* condition was higher than the character condition for both groups of speakers, similar to the perception results. For disyllabic words, the advantage of *pinyin* over character was considerably reduced for high performance speakers, while the pattern was reversed for low performance speakers.

A logistic mixed-effects model was fitted with random intercepts for subjects and items. In the best-fitting model (Table 5), fixed factors included Performance ($\chi^2(1) = 16.064$, $p < .001$), as well as the interactions between Orthography and Performance ($\chi^2(1) = 15.171$, $p < .001$) and between Performance and Target Tone ($\chi^2(3) = 25.173$, $p < .001$). The addition of other fixed factors (namely the main effect of Target Tone and Word Frequency) did not lead to an improvement in model fit compared to the model without, as indicated by chi-square log-likelihood tests ($p > .05$). Table 5 shows that high performance learners were significantly more accurate in monosyllabic production than their low performance peers ($\beta = 2.033$, SE = 0.416, Z = 4.882). The significant interaction between Performance and Tone for T2–T1

Table 4
Average production accuracy (%) in different conditions.

Performance group	monosyllabic				disyllabic			
	pinyin		character		*pinyin*		character	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
High	96.6	10.6	79.6	13.2	96.2	4.2	94.8	5.0
Low	76.0	31.2	67.3	18.8	77.7	17.7	84.8	12.8

Table 5
Model summary (monosyllabic production task): Correct response ∼ Performance + Performance:Orthography + Performance:Tone + (1 | Speaker) + (1 | Item). P-values: * < 0.05, ** < 0.01, *** < 0.001.

Parameters	Fixed				Random	
	β	*SE*	*Z*		Subject SD	Item SD
(Intercept)	1.845	0.272	6.791	***	0.324	1.536
Performance	2.033	0.416	4.882	***		
Performance: Orthography	−1.274	0.458	−2.782	**		
Performance: Tone (T2–T1)	−1.367	0.418	−3.268	**		
Performance: Tone (T3–T1)	0.358	0.431	0.830			
Performance: Tone (T4–T1)	−0.810	0.431	−1.879	.		

Table 6
Model summary (disyllabic production task): Correct response ∼ Performance + Tone + Syllable + Tone:Orthography + (Tone + Syllable|Subject) + (Performance |Item). P-values: * < 0.05, ** < 0.01, *** < 0.001.

Parameters	Fixed				Random	
	β	*SE*	*Z*		Subject SD	Item SD
(Intercept)	3.213	0.254	12.651	***	0.811	1.899
Performance	2.576	0.360	7.154	***		2.242
Tone (T2–T1)	0.704	0.246	2.859	**	0.897	
Tone (T3–T1)	−0.750	0.214	−3.500	***	0.756	
Tone (T4–T1)	1.663	0.302	5.515	***	1.128	
Syllable	−0.822	0.186	−4.430	***	0.776	
Tone: Orthography	−0.129	0.212	−0.608			
Tone (T2–T1): Orthography	1.148	0.234	4.905	***		
Tone (T3–T1): Orthography	0.590	0.207	2.850	**		
Tone (T4–T1): Orthography	0.614	0.242	2.541	*		

($\beta = -1.367$, SE = 0.418, Z = −3.268) is attributed to the higher accuracy of the Low performance group (75.5%) compared to the High performance group (74.5%) in T2; in all other tones the opposite was true.

In the best model for disyllabic production (Table 6), fixed effects included Performance ($\chi^2(1) = 13.78$, $p < .001$), Target Tone ($\chi^2(3) = 400.50$, $p < .001$), Syllable ($\chi^2(1) = 81.58$, $p < .001$) and the interaction between Target Tone and Orthography ($\chi^2(4) = 34.76$, $p < .001$). Random effects included intercepts for subjects and items, as well as by-subject slopes for Target Tone and Syllable and by-item slopes for Performance. Models with additional fixed or random effects did not converge.

Table 6 shows that high proficiency learners were significantly more accurate in disyllabic production than their low proficiency peers ($\beta = 2.576$, SE = 0.360, Z = 7.154). The production accuracy of all lexical tones was significantly different from that of T1. In general, the second syllable of a target word was significantly more accurately produced than the first syllable ($\beta = -0.822$, SE = 0.186, Z = 4.430).

The significant interactions between Tone and Orthography is attributable to the fact that whereas T1 was more accurately produced in the *pinyin* condition (89.7%) than in character (84.1%), the opposite was true for T2 (character = 91.4%, *pinyin* = 87.4%), T3 (character = 76.6%, *pinyin* = 70.1%), and T4 (character = 95.7%, *pinyin* = 91.4%).

Fig. 4 shows the error patterns in different conditions by the two groups of speakers. The number pairs on the horizontal axis represent target tone and error patterns. For example, the pair '12' means that the target T1 was mis-produced as T2; '23' means that the target T2 was mis-produced as T3, and so on. We can see that the errors mainly involved T2 versus T3, and T1 versus T4. For monosyllabic words (see Fig. 4a), the high performance speakers made very few errors in *pinyin*, while their error patterns for characters were quite similar to those of low performance speakers, just to a lesser degree. Low performance speakers also made fewer errors in *pinyin* than in character, except for '32'. The general error patterns in disyllabic words of low performance speakers (see Fig. 4b) were quite comparable to those in monosyllabic words. It is worth noting that T3 presented in *pinyin* was particularly difficult for low performance speakers in both monosyllabic and disyllabic words. High performance speakers made very few disyllabic mistakes in either *pinyin* or character.

Finally we also tested the effect of Tonal Correspondence on production accuracy in monosyllabic and disyllabic words (see Table 7). A two-way ANOVA shows that for monosyllabic words there were significant main effects of Tonal Correspondence ($F(1, 60) = 71.749$, $p < .001$) and proficiency $F(1, 60) = 8.040$, $p = .006$, although their interaction was non-significant ($p = .354$). Post-hoc Bonferroni tests confirm that production accuracy was significantly higher when Tonal Correspondence was regular ($p < .001$), and higher in the high performance group compared with the low performance group ($p = .006$). For disyllabic words, there were significant main effects of Tonal Correspondence $F(1,380) = 28.26$, $p < .001$, and Performance $F(1,380) = 29.70$, $p < .001$, on production accuracy, as well as a significant interaction between these two effects $F(1,380) = 18.16$), $p < .001$. Post-hoc Bonferroni tests confirm that production accuracy was significantly higher when Tonal Correspondence was regular ($p < .001$), and higher in the high performance group compared with the low performance group ($p < .001$). For the low performance group, irregular Tonal Correspondence led to much bigger drop in production accuracy (87.8% → 48.3%) compared with otherwise (91.2% → 86.5%).

5.5. Discussion

Similar to the perception results, the production data also demonstrate non-uniform effects of orthography. For monosyllabic words, items presented in *pinyin* were produced significantly more accurately than those in characters, except for T3 which was equally difficult in both orthographies. For disyllabic words, items presented in characters were instead produced slightly better than those in *pinyin*, except for T1. Again, the accuracy differences in the *pinyin* and the character conditions were larger for monosyllabic words than disyllabic words.

The monosyllabic and the disyllabic tasks were in effect very different tasks. The monosyllabic task asked the speakers to produce one of four tonemes (without semantic information, purely phonological); the disyllabic task asked the speakers to produce a lexical item either as translation of their Cantonese lexicon (character), or recognize one from their Mandarin lexicon (*pinyin*). That the findings in perception and in production echoed each other show that the orthography effects are genuine, and that the possible reasons contributing to the reversing patterns in tone perception discussed above are also at work in tone production. This justifies pulling the results of the monosyllabic and disyllabic materials together for a more comprehensive understanding. Relying on either set of material would only paint a partial picture of the orthography effects.

One unexpected finding is that the production of monosyllabic tones presented in *pinyin* was not as easy as we envisaged for the low performance speakers. All learners of Mandarin were taught how to recite the four Mandarin citation tones on different syllables early on, and our speakers had been learning Mandarin since primary school. That was why we only included two syllables for monosyllabic words in *pinyin* as reference. Nevertheless, the data reveal that even a transparent system may still present difficulties to the learners in some circumstances. In addition, the weaker performance of low performance speakers in *pinyin* for disyllabic words may be due to the fact that they were not very fluent in *pinyin* spellings, so they had to assemble and work out the pronunciation 'on the fly', as it were. The presence of tone marks could

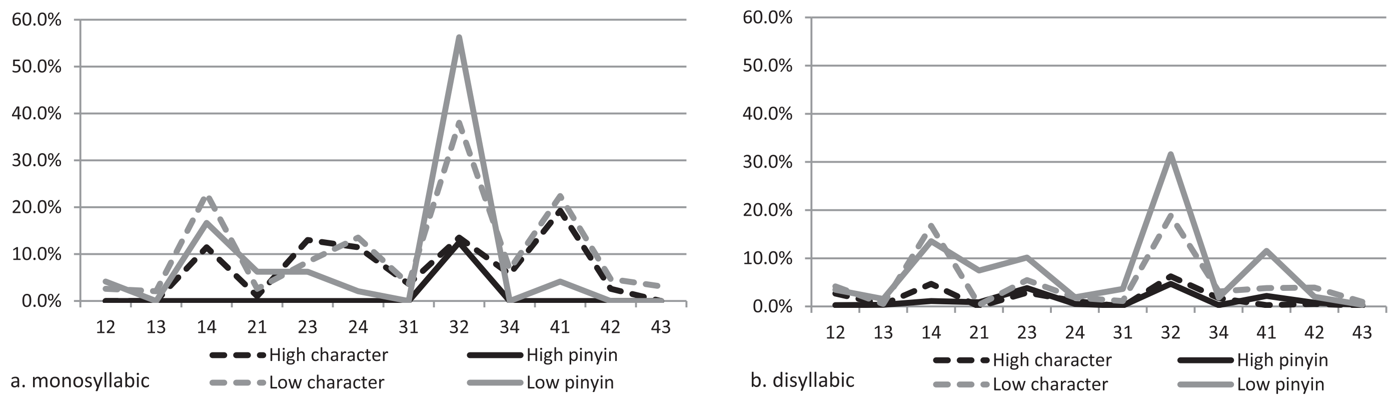

Fig. 4. Error rates in tone production by high and low performance speakers in (a) monosyllabic and (b) disyllabic words. Horizontal axis shows target and error tone pairs.

Table 7
Mean production accuracy (%) for monosyllabic and disyllabic words by performance group × tonal correspondence.

Word length		Monosyllabic				Disyllabic			
Performance		High		Low		High		Low	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Tonal correspondence	Regular	90.1	15.0	80.9	15.1	91.2	25.7	86.5	17.4
	Irregular	53.6	21.4	35.4	21.4	87.0	26.8	48.3	23.0

decrease L1 reading fluency because of the increased visual complexity (Bird, 1999). The faster conceptual route provided by Chinese characters can explain the advantage of characters over *pinyin* for low performance speakers. The production data suggest that in addition to being task dependent, orthographic effects can sometimes be stimuli and group dependent as well.

Figure 4 shows that the tone production errors concentrated on two contrasts, T1 versus T4, and T2 versus T3. This reflects findings from previous studies on Mandarin tones by Cantonese speakers (e.g. Hao, 2012). The Mandarin T1 is very similar to the Cantonese T1 (both a high-level tone [55]). As T1 in Cantonese has a high-falling allotone [53] (Bauer and Benedict, 1997), Cantonese speakers do not treat the two realizations as contrastive in their native language. Therefore, under the L1 influence, the participants had trouble distinguishing T1 [55] (high-level) and T4 [51] (high-falling) in Mandarin. Previous studies also confirmed that the T1-T4 confusion appears to be specific to Cantonese speakers (Hao, 2012; So and Best, 2010).

It is interesting to note that for T1 and T4, the error patterns were quite symmetrical with either tone being the target tone; whereas for T2 and T3, there were many more errors when T3 was the target tone than the other way round for low performance speakers. The T2-T3 contrast is hard for all L2 learners regardless of their L1 backgrounds (Hao, 2012; So and Best, 2010). In spite of this common difficulty, the large asymmetry in our results may be language-specific and can be explained by the tonal correspondence between Mandarin and Cantonese (Table 1). Mandarin T3 [214] are mostly mapped onto Cantonese T2 [25] or T5 [23], both of which are very similar to Mandarin T2 [35] as they are all rising tones. As L1 influence is particularly obvious in low performance speakers, the influence of tonal mappings on their production would be stronger than on that of high performance speakers. As a result, the low performance speakers made more errors producing Mandarin T3 as T2. In addition, as suggested by So and Best (2010), the presence of T3→T2 sandhi in Mandarin could also account for the T3-T2 confusion. T3 in Mandarin is more variable in realization both phonologically and phonetically, so the representation of T3 is possibly a less well established category in the learners' mind, and thus more challenging for L2 speakers to handle. Conversely, the Mandarin T2 [35] is acoustically very similar to Cantonese T2 [25], so even low performance speakers may not find it too difficult to produce.

Nevertheless, the language-specific tonal correspondence cannot fully explain why the asymmetry was worse in the *pinyin* condition for low performance speakers. It is suspected that their low fluency in reading *pinyin* spellings discussed above may have exacerbated their performance of this already confusing tone pair.

6. General discussion

Our study investigated the effects of orthography on the perception and the production of Mandarin tones by Cantonese learners. Specifically, it compared an opaque system, Chinese characters, with a transparent and schematic system, *pinyin*. Consistent results were found in that, generally, *pinyin* facilitated tone perception and production for monosyllabic words while characters were more beneficial in both tone perception and production for disyllabic words. The production data reveal that low performance speakers were more affected by orthographic input than high performance speakers were.

As discussed in the Introduction, previous studies on orthographic effects on L2 phonology can be grouped into three main categories: those showing a positive effect of orthography in helping learners to discriminate L2 phonological contrasts (Erdener and Burnham, 2005; Escudero et al., 2008); those showing that orthographic input could induce errors not explainable by the spoken input (Bassetti, 2006; Young-Scholten and Langer, 2015); and those showing that orthography might have no or only limited influence on L2 phonology (Escudero, 2015; Escudero et al., 2014). Previous studies were mostly on alphabetic writing systems and segmental contrasts. Our study is novel in including a logographic writing system in the study of orthographic effects, and is one of the few examining a suprasegmental feature. Although our results seem to have raised more questions than they can offer answers to, they do demonstrate that orthographic effects are non-uniform. They can be dependent on the nature of the tasks, stimuli and proficiency of the participants. Even a transparent system may not always be easier.

Our study differs from previous studies on orthographic effects in a number of important ways. First, many previous studies involved *ab initio* or first-exposure learners who were not familiar with the phonological contrasts at hand (e.g. Hayes-Harb et al., 2010), and examined how different orthographic input would affect their L2 phonological learning, although some studies also used experienced learners (Escudero et al., 2008, 2014). Our participants were learners already experienced with both the orthographic input (Chinese characters and *pinyin*) and the phonological contrasts (Mandarin tones). Second, because of the difference in target participants, previous studies often used a learning paradigm which included a novel word learning phase, a criterion test phase, and a final test phase, while our study examined how orthographic input affected tone perception and production of real words known to the participants already (i.e., no novel words given participants' native competence in written Chinese). Third, L1 phonology was involved mainly by manipulating the congruency of grapheme-phoneme conversion rules in orthographic input in previous studies. In our study, L1 phonology was evoked via a different route by activating the L1 lexicon using Chinese characters. Thus, different levels of processing were likely involved when the participants were presented with *pinyin* (lower and controlled) and with characters (higher and automatic). In addition, our study included different types of materials (monosyllabic and disyllabic words) while previous studies mainly used a mini lexicon of a few novel words with similar structure. Finally, previous studies tried to simulate the initial stage of L2 acquisition, while ours represents a more authentic experience in typical L2 learning.

Given these differences, it may appear that our study may not be entirely comparable to previous studies of orthography effects. As Showalter and Hayes-Harb (2013) rightly pointed out, in the context of Chinese language learning, the ultimate goal is for learners to acquire the system of Chinese characters, not just using *pinyin*. Also, they asked the question of whether the beneficial effects of tone marks on word learning (listening) found in their study with beginning learners could be found in learners' production as well. Our study thus builds on previous studies and extends the scope of investigation by using learners with real experience in both types of orthography, and incorporating both perception and production. Our findings can bring new insights into the variegated effects orthography can have in L2 phonology by providing data from different perspectives.

One seemingly unsurprising finding of our study is that the orthographic influence is modulated by learners' performance in the target language. Although learner proficiency was not determined by an independent measure in our study, the accuracies in both perception and production were significantly different between the two performance groups. Learners with higher L2 performance were less affected by orthographic differences. Nevertheless, our data also show that even very high performance learners (with an average perception accuracy of 97.4% and production accuracy of over 90%) were still not completely free from orthographic influence. Proficiency effects have not been widely investigated by previous studies on orthography. Escudero et al. (2014) examined if proficiency in the non-native language played a role in word learning accuracy with different orthography input, but they only found ambiguous results. Our study presents clear evidence that L2 proficiency is an important factor in the study of orthography effects, and that low performance speakers are more affected by orthographic information, and that even high performance speakers are not immune to orthographic influence.

One unexpected, yet consistent, finding in our study is that *pinyin* and characters have different effects on monosyllabic and disyllabic words. To be precise, in our data *pinyin* appeared to enhance the ability to perceive or produce tones in monosyllabic words. In disyllabic words, the differences between the two types of orthographic input were largely reduced, with characters 'catching up' with the accuracy of *pinyin*, as it were. Thus, it is not the case that the transparent *pinyin* system lost its appeal in facilitating tone perception and production in disyllabic words. Rather, the opaque Chinese characters were somehow not as difficult in disyllabic words as in monosyllabic words. It may well be the case that the learners have stored the tonal combinations of disyllabic words as fixed patterns in their mind. So instead of having all $4 \times 4 = 16$ tonal possibilities for a given disyllabic word, there are only a few possible competitors in the mental lexicon. Additionally, knowing exactly which lexical items were involved in the character condition could reduce uncertainty posed by homophony. The two orthographic systems have varying influence on the information retrieval of lexical items and provide different access to the L2 lexicon and L2 phonology. It is beyond the scope of this study to discuss the various issues involved in processing of Chinese characters. Reviews can be found in Williams and Bever (2010) and Sze et al. (2015). Chinese characters allow direct access to the meaning of words. Two consequences arose for the speakers upon reading such a system. First, in addition to activating the lexical concepts immediately, they would be able to activate the phonological representations of the words from the meaning, first in their L1 Cantonese, then the corresponding representations in L2 Mandarin (Perfetti and Zhang, 1991; Tan et al., 1995; Zhou and Marslen-Wilson, 1999). The tonal correspondence between the two languages (Table 1) also provides assistance (or hindrance) in tone perception and production, as shown by the significant effect of Tonal Correspondence in the results reported above. This is in contrast to reading or identifying *pinyin*. Although *pinyin* clearly signals the tones in Mandarin, it does not allow the learners to access the meaning of the words immediately (because they were not as familiar with *pinyin* as with characters, and there are many homophones), nor their L1 phonological representations. Such differences have a larger impact on low performance speakers than high performance speakers.

It is interesting to note that in addition to being more beneficial or less beneficial in general, *pinyin* and characters may evoke different types of tone errors. Table 2 shows that in perception listeners mainly confused T1 and T4 in the *pinyin* condition, while more confusion was found between the T2-T3 pair in the character condition. As a clear pattern like this is not found in the production data, we cannot be sure if this is a genuine effect. Perhaps the abilities to recognize and produce the different tonal categories are asymmetric. Still, both perception and production data (of low performance speakers) reveal that more errors were found for the T2-T3 pair in the *pinyin* than in the character condition. Such findings do point to the possibility that orthographic effects can even be contrast-specific. Clearly, further study is needed to further explore how orthography may influence learners' performance differently.

Some limitations of our study include the fact that the monosyllabic and disyllabic tasks were different in many aspects, and that our participants were more conversant with Chinese characters than with *pinyin*. A possible extension of the current study is to conduct a similar study with intermediate foreigner learners of Mandarin who also know enough Chinese characters. Cantonese learners in our study were much more familiar with Chinese characters than with *pinyin*, and they had very strong tonal influence from L1. Such imbalance between the two orthographic systems can be avoided with non-Chinese learners of Mandarin. It would be interesting to see if differing patterns of monosyllable and disyllabic words can still be found with learners who have equivalent exposure to both orthographic systems. The findings will be very useful in teasing apart the effects of orthographic input and lexical access.

The study of orthographic effects in L2 phonology is an emerging field. How orthographic effects can be incorporated into theories of L2 speech acquisition is still unclear, because the written forms only encode limited information about L2 phonology. Cutler (2015) pointed out that orthographic representation can be useful only if the relevant distinction can be perceived by L2 learners perceptually, otherwise orthographic encoding can result in worse perceptual performance by introducing more competitions in lexical processing. Current major L2 speech learning models like the SLM (Flege, 1995), PAM (Best, 1995) and PAM-L2 (Best and Tyler, 2007), and L2LP (Escudero, 2009) all proposed that difficulties in L2 learning hinge on the perceived phonetic similarity between L1 and L2 without considering the effects of orthography. In two related studies, Piske et al. (2002; 2011) demonstrated that even early Italian-English bilinguals with high English proficiency would produce certain English vowels as their L1 Italian counterparts in naming English nonwords despite the fact that correct English prompts were presented to them auditorily right before their production. Furthermore, such production errors were not found in their conversational speech, i.e., they could distinguish those vowels. Piske et al. (2002; 2011) argued that these errors were introduced by the orthography in the nonwords. Their results clearly indicate that perceived similarity between L1 and L2 can be affected by orthography as well. Thus, it is necessary that L2 speech learning models integrate orthographic effects in their predictions, as L2 learners most likely learn phonological contrasts and orthographic representations simultaneously.

Our study represents a step in understanding such effects from a perspective slightly different from previous studies: including a logographic system and using experienced learners. Our findings have raised some interesting questions and call for further investigation on this interesting topic which is relevant to many language learners around the world.

References

Bassetti, B., 2006. Orthographic input and phonological representations in learners of Chinese as a foreign language. Written Lang. Literacy 9 (1) 95–4.

Bassetti, B., 2008. Orthographic input and second language phonology. In: Piske, T., Young-Scholten, M. (Eds.), Input Matters in SLA. England: Multilingual Matters, Bristol, pp. 191–206.

Bates, D.M., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting Linear Mixed-Effects Models using {lme4}. J. Stat. Softw. 67 (1), 1–48.

Bauer, R.S., Benedict, P.K., 1997. Modern Cantonese Phonology. Mouton de Gruyter, Berlin.

Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. MD: York Press, Timonium, pp. 171–204.

Best, C.T., Tyler, M.D., 2007. Nonnative and second-language speech perception: commonalities and complementaries. In: Munro, M.J., Bohn, O.-S. (Eds.), Language Experience in Second Language Speech Learning: In Honor of James Emil Flege. John Benjamins, Amsterdam, pp. 13–34.

Bird, S., 1999. When marking tone reduces fluency: an orthography experiment in Cameroon. Lang. Speech 42 (1), 83–115.

Chao, Y.-R., 1930. A system of tone-letters. Le Maître Phonétique 45, 24–27.

Cheung, L.Y., Gao, S., 2000. 普通話字音認記訓練12講 [12 Lecture Series of the Recognition and Memorization of Mandarin pronunciation]. Joint Publishing Co. Ltd, Hong Kong.

Cutler, A., 2015. Representation of second language phonology. Appl. Psycholing. 36, 115–128.

Duanmu, S., 2007. The Phonology of Standard Chinese. NY: Oxford University Press, New York (2nd ed.).

Erdener, V.D., Burnham, D.K., 2005. The role of audiovisual speech and orthographic information in nonnative speech production. Language Learning 55 (2), 191–228.

Escudero, P.R., 2009. The linguistic perception of similar L2 sounds. In: Boersma, P.P.G., Hamann, S.R. (Eds.), Phonology in Perception. Mouton de Gruyter, Berlin, pp. 151–190.

Escudero, P.R., 2015. Orthography plays a limited role when learning the phonological forms of new words: the case of spanish and English learners of novel Dutch words. Appl. Psycholing. 36, 7–22.

Escudero, P.R., Hayes-Harb, R., Mitterer, H., 2008. Novel second-language words and asymmetric lexical access. J. Phonet. 36, 345–360.

Escudero, P.R., Simon, E., Mulak, K.E., 2014. Learning words in a new language: orthography doesn't always help. Bilingual.: Lang. Cognit. 17 (2), 384–395.

Flege, J.E., 1995. Second language speech learning: Theory, findings, and problems. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. MD: York Press, Timonium, pp. 233–277.

Hao, Y.-C., 2012. Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. J. Phonet. 40, 269–279.

Hayes-Harb, R., Brown, K., Smith, B.L., 2017. Orthographic input and the acquisition of German word-final devoicing by native English speakers. Lang. Speech.

Hayes-Harb, R., Cheng, H.W., 2016. The influence of the Pinyin and Zhuyin writing systems on the acquisition of Mandarin word forms by native English speakers. Front. Psychol. 7 (785), 1–13.

Hayes-Harb, R., Hacking, J.F., 2015. Beyond rating data: What do listeners believe underlies their accentedness judgments. J. Second Lang. Pronunc. 1 (1), 43–64.

Hayes-Harb, R., Nicol, J., Barker, J., 2010. Learning the phonological forms of new words: Effects of orthographic and auditory input. Lang. Speech 53 (3), 367–381.

Howie, J.M., 1976. Acoustical Studies of Mandarin vowels and Tones. Cambridge University Press, Cambridge.

McEnery, A.M., Xiao, Z., 2004. The Lancaster corpus of mandarin chinese: a corpus for monolingual and contrastive language study. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisboa, Portugal. pp. 1175–1178.

Perfetti, C.A., Tan, L.-H., 1998. The time course of graphic, phonological, and semantic activation in chinese character identification. J. Exper. Psychol.: Learn. Memory Cognit. 24 (1), 101–118.

Perfetti, C.A., Zhang, S., 1991. Phonological processes in reading Chinese characters. J. Exper. Psychol.: Learn. Memory Cognit. 17 (4), 633–643.

Piske, T., Flege, J.E., MacKay, I.R.A., 2002. The production of English vowels by fluent early and late Italian-English bilinguals. Phonetica 59, 49–71.

Piske, T., Flege, J.E., MacKay, I.R.A., Meador, D., 2011. Investigating native and non-native vowels produced in conversational speech. In: Wrembel, M., Kul, M., Dziubalska-Kołaczyk, K. (Eds.), Achievements and Perspectives in the Acquisition of Second Language Speech: New Sounds 2010. Bern: Peter Lang, pp. 195–205.

Showalter, C.E., Hayes-Harb, R., 2013. Unfamiliar orthographic information and second language word learning: A novel lexicon study. Second Lang. Res. 29 (2), 185–200.

Showalter, C.E., Hayes-Harb, R., 2015. Native English speakers learning arabic: the influence of novel orthographic information on second language phonological acquisition. Appl. Psycholing. 36, 23–42.

Simon, E., Chambless, D., Alves, U.K., 2010. Understanding the role of orthography in the acquisition of a non-native vowel contrast. Lang. Sci. 32 (3), 380–394.

So, K.L.C., Best, C.T., 2010. Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. Lang. Speech 53, 273–293.

Sze, W.P., Yap, M.J., Rickard Liow, S.J., 2015. The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. Q. J. Exper. Psychol. 68 (8), 1541–1570.

Tan, L.-H., Hoosain, R., Peng, D., 1995. Role of early presemantic phonological code in Chinese character identification. J. Exper. Psychol.: Learn. Memory Cognit. 21 (1), 43–54.

Tsang-Cheung, L.Y., 1988. The phonological correspondences between cognate morphemes in Cantonese and Mandarin. *MPhil Thesis*. University of Hong Kong, Hong Kong.

Williams, C., Bever, T.G., 2010. Chinese character decoding: a semantic bias. Read. Writing 23 (5), 589–605.

Young-Scholten, M., 2002. Orthographic input in L2 phonological development. In: Burmeister, P., Piske, T., Rohde, A. (Eds.), An Integrated View of Language Development: Papers in Honour of Henning Wode. Trier: Wissenschaftlicher Verlag Trier, pp. 263–279.

Young-Scholten, M., Langer, M., 2015. The role of orthographic input in second language German: Evidence from naturalistic adult learners' production. Appl. Psycholing. 36, 93–114.

Zhang, Q., Weekes, B.S., 2009. Orthographic facilitation effects on spoken word production: Evidence from Chinese. Lang. Cognit. Process. 24 (7–8), 1082–1096.

Zhou, X., Marslen-Wilson, W., 1999. Sublexical Processing in reading Chinese. In: Wang, J., Inhoff, A.W., Chen, H.C. (Eds.), Reading Chinese script: A cognitive Analysis. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 135–171.