Auditory-orthographic integration at the onset of L2 speech acquisition

Yasaman Rafat
Western University, Ontario, Canada

Ryan A Stevenson
Western University, Ontario, Canada

Abstract
Recent studies have provided evidence for both a positive and a negative effect of orthography on second language speech learning. However, not much is known about whether orthography can trigger a McGurk-like effect (McGurk & MacDonald, 1976) in second language speech learning. This study examined whether exposure to auditory and orthographic input may lead to a McGurk-like effect in naïve English-speaking participants learning a second language with Spanish phonology and orthography. Specifically, it reports on (a) production of non-target-like combinations such as [lj] as in [polji] for <pollio>-[pojo], where the auditory Spanish [j] and the first language English [l] that correspond to the shared digraph <ll> are integrated, and (b) fusion quantified in terms of [z] devoicing such as [z̥apito] for <zapito>-[zapito]. Moreover, the effects of (a) type of grapheme-to-sound correspondence, (b) position in the word, and (c) condition of training and testing were examined. Participants were assigned to four groups: (a) auditory only, (b) orthography at training and production, (c) orthography at training, and (d) orthography at production. The positions included word-initial and word-medial. The grapheme-to-sound correspondences consisted of <v>-[b], <d>-[ð], <z>-[s] and <ll>-[j]. Results were indicative of a McGurk-like effect only for the Spanish digraph <ll>. The highest rate of combination productions was attested in the orthography-training condition in the word-medial position.

Keywords
L2 speech learning, orthography, McGurk effect

Corresponding author:
Yasaman Rafat, Western University, Faculty of Arts and Humanities, Arts & Humanities Building, Rm 3G28D, London, ON N6A 3K7, Canada.
Email: yrafat@uwo.ca
Introduction

It is well known that many aspects of speech processing are multimodal, including speech perception (e.g., Sumby & Pollack, 1954; McGurk & MacDonald, 1976; Massaro, 1987; 2002; Soto-Faraco, Navarra, & Alsios, 2004), speech learning (Vigliocco, Perinnis, & Vinson, 2014), and second language (L2) speech learning (Hardison, 1999; Ortega-Llebaria, Faulkner, & Hazan, 2001; Erdener & Burnham, 2005; 2013). One of the most salient examples of the multimodal nature of speech is the McGurk effect (e.g., McGurk & MacDonald, 1976; Welch & Warren, 1980; Sekiyama & Tohkura, 1991; Munhall, Gribble, Sacco, & Ward, 1996; Sekiyama, 1997). The McGurk effect is elicited by the synchronous or simultaneous presentation of incongruent auditory (e.g., /ba/) and facial/visual cues (e.g., /ga/). The listener often integrates the auditory and visual information leading to either (a) a combination percept, such as /bga/ (McGurk & MacDonald, 1976; Green & Norrix, 1997) suggesting a strong influence of vision; or (b) a fused percept such as /da/ where the syllable perceived is not contained in either the auditory or the visual information (e.g., Green & Kuhl, 1989; Green, Kuhl, Meltzoff, & Stevens, 1991; MacDonald & McGurk, 1978; Manuel, Repp, Studdert-Kennedy, & Liberman, 1983; Massaro, 1987; Sekiyama & Tohkura, 1991; Summerfield & McGrath, 1984; Stevenson, Siemann, Schneider, Eberly, Woynaroski, Camarata, & Wallace, 2014). Both combination and fused perceptions are different from the individual original sounds presented separately in each modality not only in that they are different phonemes from the stimuli, but that they result from an interaction between sensory modalities.

There is also an abundance of research that has provided evidence for the orthographic (writing) channel interacting with auditory input in first language (L1) speech processing (Dijkstra, Roelofs, & Fieuws, 1995; Jakimik, Cole, & Rudnicky, 1985; Montant, Schön, Anton, & Ziegler, 2011; Seidenberg & McClelland, 1989; Seidenberg & Tannenhau, 1979; Van Orden & Goldinger, 1994; Taft, 2006; Treiman & Cassar, 1997; Ziegler & Ferrand, 1998; Ziegler & Muneaux, 2007, among others). For instance, Seidenberg and Tannenhau (1979) claimed that orthographic knowledge can affect spoken word processing. They conducted a rhyme judgment task and found that participants’ responses were faster in a rhyme when the pairs of words shared spellings (e.g., <toast> – <roast> vs. <toast> – <ghost>). Similarly, Jakimik et al. (1985) conducted priming tasks and found that the participants responded faster to the auditory target with the prior presentation of a phonologically similar prime that overlapped with the spelling of the target (e.g., message – mess) than to a prime with non-overlapping spellings (e.g., <definite> – <deaf>). Orthography has also been shown to affect underlying representations (Ranbom & Connine, 2007; 2011). Ranbom and Connine (2007), for example, provided some evidence that orthographic information affects mental representations of speech, specifically, the representation of lexically stored allophonic representations. They conducted a corpus analysis and found that the nasal flap realization in the /nt/ cluster of the word gentle is dominant in spoken US English, even though the production frequency of the nasal flap may vary within individual words. They then conducted a lexical decision task and showed that the highly frequent nasal flap was identified more quickly and accurately than the less frequent flap, but, crucially, [nt] productions resulted in faster and more accurate lexical decisions compared with the nasal flaps. The results of the lexical decision task demonstrated that orthographic information influences spoken word processing. There is also some evidence to suggest that orthography may exert an influence on spoken word production (Bentur, 1987; Ravid & Shlesinger, 2001; Temkin Martinez & Müllner, 2016; Han & Choi, 2016). For example, Han and Choi (2016) investigated the role of orthography in production and storage of spoken words by Korean speakers. The participants learned novel Korean words with different variants of /h/ including [fi] and [ø]. They were provided with the same auditory stimuli but different exposure to orthographic information in their written presentations.
orthography. There were two orthographic groups and an auditory-only group. One orthographic group was presented the letter for [ɨ] (<ɨ>) and the other with the letter for [o] (<ɔ>). The auditory group was presented with auditory input only. In picture-naming tasks, the participants presented with <ɔ> produced fewer words with [ɨ] than those presented with <ɨ>. In a spelling recall task, the participants who were not exposed to spelling displayed various types of spellings for variants, but after exposure to spelling, they began to produce spellings as provided in the task. These results were attributed to orthographic input influencing production because of its potential to restructure phonological representations.

The body of literature has also been expanding with respect to how orthography may interact with acoustic-phonetic input in L2 speech learning, although most research has focused on the Roman alphabet. When the learner’s L1 and L2 share the same alphabet, learners are faced with two main challenges: first, they have to learn that the L1 and the L2 mappings might be different. The correspondence of one grapheme to different L1 and the L2 sounds often leads to L1-based transfer (Rafat, 2011; 2013; 2015). Second, learners may need to learn one-to-many mappings (e.g., <x> in Spanish may map on to [ks] in the word <taxi> but to [x] in <México> and to [gz] in <examen>) or many-to-one mappings in the L2 (e.g., <x>, <gi>, <ge> and <j> correspond to [x] or its other variants in Spanish). The orthographic depth hypothesis (ODH; Katz & Frost, 1992) postulates that speakers of languages with shallow/regular/transparent orthographies tend to be more affected by orthographic input than speakers of languages with deep/irregular/opaque orthographies. However, we still do not know the extent to which L1 orthographic depth may modulate orthographic effects in L2 learners whose L1 orthographic system is alphabetic (e.g., Erdener & Burnham, 2005; Rafat, 2015; Escudero, 2015).

English and Spanish both have a Roman alphabetic system. Whereas the English orthographic system is characterized by one-to-many grapheme-to-phoneme correspondences and is therefore considered an irregular/deep orthography, the Spanish orthographic system is mainly characterized by one-to-one mappings. However, L2 speech learning by English-speaking learners of Spanish exhibits orthographic effects due to the differences between English and Spanish grapheme-to-phoneme correspondences (e.g., Rafat, 2011; 2015; 2016).

Orthographic effects have also been shown in several L2 perception and production studies. These studies have demonstrated that orthographic input may interact with auditory input and may promote (e.g., Erdener & Burnham, 2005; Steele, 2005; Showalter & Hayes-Harb, 2013; Bassetti, Escudero, & Hayes-Harb, 2015; Rafat, 2015), or hinder (e.g., Bassetti, 2007; Erdener & Burnham, 2005; Hayes-Harb, Nicol, & Baker, 2010; Young-Scholten, 2000; Young-Scholten, Akita, & Cross, 1999; Bassetti et al., 2015; Nimz, 2016; Rafat, 2011; 2015; 2016; Bassetti, 2017; Shea, 2017) the target-like production or correct perception of the target L2 sounds, or have no effect (Escudero, 2015; Showalter & Hayes-Harb, 2015).

Hayes-Harb et al. (2010), for example, examined the interfering effect of orthographic input in novel word learning by English-speaking participants when the grapheme-to-phonemes in the target language do not match. Participants were assigned to three different conditions at training: auditory-only, congruent, and congruent/incongruent orthography. The incongruent stimuli consisted of items spelled with a “wrong” letter (e.g., < faza > – [fαs]) and items with an extra letter (e.g., < kamand > – [kamad]). During testing, participants were shown an image and heard a word, and then asked whether the auditory word was the correct word for the image. The results yielded a significant effect of training condition on performance on the wrong-letter items, in which participants had a lower rate of accuracy in the incongruent/congruent orthography condition.

Mathieu (2016) also examined the effect of orthography at the onset of the acquisition of an L2. For the first time he reported the effects of three L2 scripts on the early acquisition of an Arabic consonantal contrast word-initially (e.g., /hɔal/–/xɔal/), showing that foreign written input can inhibit
learners’ ability to encode an L2 phonological contrast. He tested monolingual native speakers of English with no prior knowledge of Arabic. Participants took part in a word-learning experiment and were assigned to one of four learning conditions: no orthography, Arabic script, Cyrillic script, or Roman/Cyrillic blended script. The results showed that the degree of script unfamiliarity does not in itself seem to significantly affect the successful acquisition of the phonological contrast tested. However, the presence of certain foreign scripts in the phonological acquisition yielded significantly different learning outcomes in comparison to having no orthographic representation available. Specifically, the Arabic script exerted an inhibitory effect on L2 phonological acquisition, while the Cyrillic and Roman/Cyrillic blended scripts exercised different inhibitory effects based on whether grapheme-phoneme correspondences triggered L1-based phonological transfer. Mathieu (2016) proposed that L2 speech learning may be multimodal and subject to instantaneous and automatic processing similar to the McGurk effect. Specifically, he put forth that the processing of the visual input at the early stages of acquisition may prompt the auditory system to strengthen activation of the L1 phonological categories, in turn hindering the accurate perception of the novel L2 sounds.

The effect of orthographic input has also been shown in more advanced learners. Bassetti (2007) investigated the effect of orthographic inconsistency within the L2 on the production of triphthongs by Italian-speaking learners of Mandarin studying at a university in Italy. Participants used the alphabetic pinyin writing system and on average had studied Mandarin for 33 months. Although the participants had not been exposed to pinyin orthography during the character-reading task, the results yielded a 100% target-like realization of the vowel /o/ in the triphthong /iou/ when it was written with three graphemes as in <you>. However, erroneous productions were attested when the triphthongs were spelled with only two graphemes in pinyin (e.g., <iu> for /iou/). The author explained the results by proposing that pinyin generally is a transparent orthographic system, and the learners had overgeneralized this aspect of pinyin.

There is considerable evidence that when the L1 and L2 grapheme-to-phoneme relationships are incongruent, exposure to orthographic input may result in L1-based phonetic or phonological transfer (e.g., Young-Scholten, 2002; Rafat, 2011; 2015; 2016; Bassetti, 2017). An example can be observed in the production of digraphs (two graphemes such as <tt> in <kitty>) by highly proficient Italian-speaking learners of English. These learners produce digraphs as long/geminate consonants with a significant difference between their geminate (CC) vs. singleton (C) productions in English (Bassetti, 2017). The authors attributed this to transfer of L1 phonological rules, in this case gemination.

Although several studies have provided evidence of transfer effects, this effect is not categorical and may be modulated by various factors. For example, transfer in English-speaking learners’ devoicing of syllable-final consonants is argued to be modulated by the amount of exposure to orthographic input in German (Young-Scholten, 2000). In German, obstruents are devoiced in syllable-final position, although this is not cued in the orthography. For example, the word /bʊnd/ “federation” is written as <bund> but is realized as [bʊnt] creating a homophone with [bʊnt] “coloured,” which is written as <bunt>. When learning German as an L2, greater exposure to written text is related to a reduction in the acquisition of this obstruent final devoicing rule (Young-Scholten, 2000). Evidence has also been found for orthography-induced transfer in the word initial /z/ production of the same group of learners, where they produced <s>, which corresponds to /z/ in German as [s] (e.g., [sɪ] for <sɪɛ> “she” [zɪɛ]) (Young-Scholten & Langer, 2015). An acoustic analysis of the results also revealed evidence for some partially voiced versions of /z/, which the authors suggested might be a reflection of the variability in the auditory input to which the learners had been exposed.

Other factors such as type of grapheme-to-sound correspondence, position in the word, and condition of training and testing have been reported to control the rate of orthography-induced transfer in naïve English-speaking learners of Spanish (e.g., Rafat, 2011; 2016). Exposure to orthographic input at the time of learning yielded a significantly higher rate of transfer compared to when orthographic
input was presented at production or testing only. Moreover, different grapheme-to-sound correspondences resulted in significantly different rates of transfer. For example, whereas <ll>-[j] resulted in the lowest rate of transfer (0.01%), <v>-[b] and <d>-[δ] resulted in the highest rates of transfer (99% and 92%, respectively) in the orthography at training condition. The results suggested that the relative degree of acoustic-phonetic salience between an L2 and an L1 sound determines the rate of L1-based transfer (Rafat, 2011; 2016). Rafat (2011) also reported that combination productions for <ll>-[j] had been attested in the data, and attributed this to a process akin to the McGurk effect, although she did not conduct a quantitative analysis of this type of error.

A different type of acoustic-orthographic integration related to the effect of orthography has been found during the production of Spanish assibilated rhotics (发音前面的辅音响或“嘶鸣”声音) (Rafat, 2015), when naïve English-speaking learners are exposed to both auditory and orthographic input at training. Participants were assigned to two groups based on input: auditory only and auditory-orthographic. At training, participants in both groups heard auditory stimuli produced by a Mexican speaker of Spanish, whose rhotics were assibilated (e.g., <ahitaś>-[aitaś]). While the auditory-only participants were only exposed to auditory words accompanied by their meanings through images at training, participants in the auditory-orthographic group were exposed to both auditory and orthographic stimuli. Auditory stimuli and their corresponding images were accompanied by written words, which included the grapheme <r> (e.g., <ahitar>-[aitar]) in the auditory-orthographic group. At testing, participants in both groups were shown images indicating the words and asked to name them. Whereas assibilated rhotics were for the most part produced as sibilants such as [s] and [ʃ] (e.g., [aitaś] and [aitaʃ]) when learners were only exposed to auditory L2 speech, exposure to the grapheme <r> in the auditory-orthographic group promoted both the production of assibilated rhotics [aitaś] and approximant rhotics [aitaʃ]. The acoustic and orthographic cues thus interact in different ways, resulting either in the production of assibilated rhotics or English approximant rhotics. First, rhoticity is the less salient feature compared to assibilation in assibilated rhotics, and exposure to the grapheme <r> enhances the less salient feature in the input, leading to target-like productions. Based on an acoustic analysis of the degree of assibililation of the individual tokens in the input, together with the results in the auditory-orthographic condition, the author proposed that the degree of robustness of assibilation in the input modulates orthographic effects. That is, the more salient the degree of assibilation in the input, the more likely exposure to the grapheme <r> at training will lead to the production of an assibilated rhotic by the learners at testing. In the absence of strong assibilation in the auditory input, exposure to orthography at training may override the input and result in transfer or it might create a “perceptual illusion” of rhotic features, leading to approximant rhotic productions.

That orthography can lead to perceptual illusion has previously been proposed with respect to L1 processing (Hallé, Chéreau, & Segui 2000). Using a phoneme-monitoring task in French, the authors examined the effect of orthographic and phonological incongruence on the perception of /b/ and [p] in French-speaking adults. Because of voicing assimilation in French in words such as <absurd> (/bs/ and /bt/ words), the underlying /b/ written as corresponds to [p] rather than [b] in the prefix {ab-} (e.g., /absyrąd/ written as <absurde> is realized as [apsyrąd]). The authors found that the presentation of words, whose orthographic representation and phonetic realizations were incongruent, yielded a higher detection rate of [b] than [p] in <absurd>-[apsyrąd]. Hallé et al. (2000) attributed the results to a “perceptual illusion” effect, which overrides the input.

Auditory-orthographic interaction may also result in the production of a sound that is not identical to either the L1 or the L2 sound but rather exhibits characteristics of the L1 sound and approximates the L2 sound. A study on Polish-speaking learners’ perception and production of German vowels found that learners produced the German /e/ was written in German as <œ>, as a different sound, namely a diphthong [œ] (Nimz, 2016). The grapheme <œ> corresponds to /e/ in Polish but it is acoustically closer to /i/. The author explained the diphthongization by proposing that the learners incorporate
both the orthographic and perceptual interferences by starting with an orthography-induced /ɛ/ and satisfy the auditory input by moving towards the quality of the higher vowel /i/ (e.g., [ee]).

The above studies summarize some of the ways in which orthographic input may interact with the auditory input and result in either target-like or non-target-like productions, and give rise to the possibility of an orthographic McGurk-like effect in L2 learners. To date, there is only one study that has examined whether exposure to incongruent auditory and orthographic stimuli would also result in fused and combination responses in the L1 (Fowler & Dekle, 1991). This study tested whether the simultaneous presentation of incongruent auditory and orthographic /ba/ and /ga/ would result in a McGurk effect in 12 adult speakers of US English, leading to a /da/ or a /bga/ or /gba/ percept. Participants took part in an identification task including three conditions: auditory-only, auditory-orthographic, and auditory-haptic (Tadoma). In the auditory-orthographic condition, participants were seated in front of a computer screen. There was a total of 10 trials per condition and on each trial a printed <ba/ga> syllable simultaneously appeared with its synthesized syllable. Upon the presentation of each trial, the participants made their responses by circling B or G on the answer sheet. Although the results reported for the auditory-orthographic condition did not achieve significance, $F(1, 11) = 4.52, p = 0.055$, the associated effect size with this small sample ($d = 0.91$) suggests that such an orthographic McGurk effect may occur. With that said, currently no studies have examined whether exposure to incongruent auditory and orthographic input may result in a McGurk-like effect in L2 learners, although the idea of a McGurk-like effect has been previously suggested by both Rafat (2011) and Mathieu (2016).

The current paper aims to examine the multimodality of L2 speech learning further by determining whether exposure to auditory-orthographic input may lead to a McGurk-like effect in naïve English-speaking learners of an L2 with Spanish phonology and orthography. The McGurk-like effect could be seen in the case of a shared digraph (<ll> as in <pillow>-[pilo] in English and <polllero>-[pojero] in Spanish) that corresponds to two different sounds such as [l] in the L1 (i.e., English) and [j] the L2 (i.e., Mexican Spanish). The two conflicting L1 and the L2 sounds may be integrated and lead to non-target-like productions consisting of combinations such as *[pojero] or *[polllero] instead of the correct realization [pojero], or the result of pure L1 transfer-based effect *[polllero]).

1.1 Research questions

The research questions are as follows:

1. In the case of a shared grapheme, will exposure to incongruent auditory and orthographic input in naïve learners result in the integration of the auditory and orthographic input yielding a McGurk-like effect? It is important to study naïve learners to gain an understanding of the processes that may be involved in L2 speech learning right at the absolute initial stage/onset of acquisition.

2. Does condition of training and production, type of grapheme-to-sound, and position in the word modulate the rate of the production of integrated forms?

1.2 Hypotheses

To investigate the above research questions, we first examined whether four groups of participants exhibited differential production performance depending on the presence or absence of orthographic input available to them during the training and production phases. There were four groups: one auditory-only and three auditory-orthographic. All groups were exposed to auditory input at training and differed in terms of their exposure to orthographic input at training and production.
The auditory group was only exposed to auditory input, but the other three groups were exposed to the written forms either at training and/or at production. We also examined exposure to a number of grapheme-to-sound correspondences that differ between Spanish and English: \(<v>-[b]\), \(<d>-[\delta]\), \(<z>-[s]\) and \(<ll>-[j]\) (see Supplementary Table S2).

The predictions are as follows:

1. Integrated sounds will only be attested in the auditory-orthographic conditions based on two previous findings. First, exposure to orthographic input can induce L1-based phonological transfer (Young-Scholten, 2000; Young-Scholten & Langer, 2015; Rafat, 2011; 2015; 2016). Second, the integration of two conflicting auditory and facial/visual cues (e.g., /b/ and /g/) may result in an in-between sound (e.g., /d/) or a combination of sounds (e.g., /bg/).

2. Factors previously reported to increase the rate of orthography-induced transfer in Rafat (2011; 2016) will also promote the rate of the integration of incongruent auditory and orthographic input. These factors include: (a) condition of training and production, (b) type of grapheme-to-sound correspondence, and (c) position in the word.

2 Methodology

2.1 Participants

A total of 45 adult native speakers of Canadian English, born and raised in Toronto, were recruited. This study will only report on the 40 participants (19 males and 21 females) who met the linguistic profile sought initially. These participants did not report speaking any languages other than English: English was their parents’ native language, and they had been raised in English-only speaking households. Given all participants were being recruited in Ontario, where French is mandatory in the education system, minimal knowledge of French (self-reported beginner French level of proficiency) was accepted. Although we are not aware of any previous studies that have reported on the effect of minimal knowledge of an L2 grapheme-to-phoneme correspondence on the production of another language, we considered whether minimal knowledge of French would affect the participants’ productions. However, we did not believe that knowledge of French would affect the rate of integration in this experiment. For example, \(<ll>\) in French is realized as either a [l] or a [j] depending on the preceding vowel (e.g., [l] after \(<a,e,o,u,y>\) and a \(<a,e,o,u> + <i>\) combination) and not as a sequence *[lj]* or the Andean Spanish [ʎ], therefore it could not influence the rate of integration (e.g., [lj] production). A language background questionnaire was administered. All participants were over 18 (mean age was 21 years, 8 months), all literate, had at least 12 years of education, and reported that they did not have any speech or cognitive impairments. In total, 34 participants were students at the University of Toronto and six were working. They also declared they had never been exposed to Spanish or other languages through friends, media, and/or travelling.

2.2 Procedure

To explore the possibility of an auditory-orthographic integration, participants performed a picture-naming task in an artificial language with Spanish phonology and orthography, adapted from Steele (2002). The picture-naming task took place in a single session. The participants were recorded individually in Toronto by the first author, who speaks Farsi. The recording equipment used included an M-Audio Micro-track 24/96 professional two-channel mobile digital recorder and a lavaliere unidirectional microphone. The recordings were made at a sampling rate of 44.2
kHz and a quantization rate of 16 bits; the audio files containing the extracted tokens were downsampled at 22.1 kHz and saved in wave format. Participants were presented with the stimuli via a PowerPoint presentation and were required to wear headphones during the presentation. They were provided with triplets of Spanish words with an image paired to each word. Each triplet included two targets and one distractor. Each word was presented for three seconds, always accompanied by three auditory utterances of the word (one per second), with the first auditory token being presented synchronously with the onset of the visual token.

Within these audiovisual presentations, there were four separate conditions based on the inclusion of orthography, with each participant only being presented with one condition (n = 10 per condition; see Figure 1). Group A was presented with the written word simultaneously with the spoken word and the picture during both training and production (ortho-training and production). Group B was presented with the written word simultaneously with the spoken word and the picture during training but not production (ortho-training). Group C was presented with the written word only during production but not training (ortho-production). Group 4 was not presented with the written word at any point (auditory-only).

**Figure 1.** The presentation of auditory and orthographic input in the (a) ortho-training and production, (b) ortho-training, (c) ortho-production, and (d) auditory-only groups.
Immediately after the presentation of each triplet, the participants were presented with the image corresponding to each word for 3 seconds and were asked to verbally name the object in the image. The choice of an oral response relative to a key-press response was intentional. The type of response an individual is asked to make is known to influence which aspect of the stimulus (visual or auditory) a participant more heavily weights (Colin, Radeau, & Deltenre, 2005; Mallick, Magnotti, & Beauchamp, 2015; Orne, 1962). For example, by asking participants “What did you hear the speaker say?,” the participant may be biased towards reporting to have perceived the auditory component of the stimulus. Likewise, by avoiding an orthographic response (i.e., key press), the aim was to avoid biasing participants towards reporting to have perceived the token as represented orthographically. With that said, it should be clearly noted that collecting verbal responses introduces a speech production aspect into the task, which must be considered as a possible contributor to any significant findings. Participants repeated the entire task (108 words, 72 targets and 36 distractors) three times (216 targets in total) and were given breaks of two to three minutes or longer if needed in between each round. A familiarization session consisting of one trial of the same type as in the main experiment was also conducted to allow participants to adjust to the task.

### Stimuli

The data presented here were collected with a total of 72 Spanish-like word stimuli (see Table 1). The picture-naming task was designed to test the effect of condition of training and production, and the effect of grapheme-to-sound correspondences as well as position on L1-based phonological transfer (e.g., Rafat, 2011; 2016). The effect of several same and different grapheme-to-sound correspondences on transfer were previously considered. Here, only stimuli consisting of incongruent grapheme-to-sound correspondences are considered. The shared graphemes we considered in this study are <v>, <d>, <z> and the digraph <ll> and the digraph <ll> corresponds to both word-initially and word-medially in Spanish as in <vireca>-[bireka] but it maps on to [v] in English as in <vote>-[vot]. Whereas <d> word-initially corresponds to [d] in both Spanish and English, word-medially it corresponds to [δ] in Spanish, as in <codena>-[koðena], but to a [ɾ] in an unstressed position in North American English such as <madam>-['ma.dəm], or [d] in a stressed position such as <redact>-[ˈre.dəkt]. <z> corresponds to [s] word-initially in Spanish as in <zatara>-[satara] but to [z] word-initially in English <zoo>-[zu]. <ll> was the only digraph that corresponds to the same sounds in different positions in both English and Spanish: <l> corresponds to [l] both word-initially and word-medially in English (e.g., <lloyd>-[lojd] vs. <balloon>-[ˈbə.lun]) and to [j] both word-initially and word-medially in Spanish (e.g., <lloreta>-[lojera] vs. <pollero>-[po.jero] (see Table 1 and Supplementary Table S2). Therefore, a balanced number of words in both positions for this grapheme-to-sound

<table>
<thead>
<tr>
<th>Shared grapheme by position</th>
<th>Spanish sound</th>
<th>English sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;v&gt; I</td>
<td>[b], &lt;vireca&gt;-[bireka]</td>
<td>[v], &lt;vote&gt;-[vot]</td>
</tr>
<tr>
<td>&lt;d&gt; VCV</td>
<td>[d], &lt;codena&gt;-[koðena]</td>
<td>[ɾ], &lt;madam&gt;-['ma.dəm]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[d], &lt;redact&gt;-[ˈre.dəkt]</td>
</tr>
<tr>
<td>&lt;z&gt; I</td>
<td>[s], &lt;zatara&gt;-[satara]</td>
<td>[z], &lt;zoo&gt;-[zu]</td>
</tr>
<tr>
<td>&lt;ll&gt; I and VCV</td>
<td>[l], &lt;llanero&gt;-[lənero],</td>
<td>[l], &lt;lloyd&gt;-[lojd],</td>
</tr>
<tr>
<td></td>
<td>&lt;collete&gt;-[kojete]</td>
<td>&lt;balloon&gt;-[ˈbə.lun]</td>
</tr>
</tbody>
</table>

I: word initial; VCV, intervocalic.
correspondence were included (six per position). The rest of the stimuli occurred word-initially only (six in one position only). The stimuli were also controlled for stress.

Rafat (2011) also examined primacy and recency effects with respect to orthography-induced transfer. Therefore, the stimuli were also pseudo-randomized so that primacy and recency effects with respect to position in triplets could be controlled for. The positional permutations were as follows: first in training and first in production (1*1), first in training and last in production (1*3), last in training and first in production (3*1), second in training and second in production (2*2), and last in training and last in production (3*3). The stimuli were presented as follows: 36 triplets, each containing two target stimuli and one distracter, were formed from the 108 words (72 stimuli plus 36 distracters, see Rafat 2011; 2016 for more detail). The order of presentation of the stimuli was the same for all four conditions. All stimuli analyzed here were trisyllabic words with primary stress on the penultimate syllable (see Rafat, 2011 and 2016 for a more detailed description of the stimuli).

Whereas the target stimuli consisted of trisyllabic words, the 36 distractors (see Supplementary Table S3) were composed of 20 bisyllabic words such as <chorro>-[ʧoro] and 16 monosyllabic words such as <u>-[u]. Moreover, some of the bisyllabic stimuli differed from the target words in terms of their syllabic structure. Specifically, the former included clusters such as [kɾ] in [kɾoe] for <croe>. In addition, new grapheme-to-phoneme correspondences such as <rr> representing [r], and <gue> pronounced [ge] were included.

Although the stimuli are actual Spanish words, they were assigned new meanings via images of common picturable words, including household items, plants, and animals to increase the possibility of recall by participants (see Saint-Aubin & Poirier, 2005). For example, the image of “castle” was assigned to the word <ahotar>, which actually means “to contrast size and weight” in Spanish. This deception was explained to participants upon the completion of the experiment.

Two native speakers of English (who had training in linguistics) confirmed the words did not have English cognates. Although we were not aware of whether cognates could promote a higher rate of L1-based phonological transfer in naïve English speakers of Spanish, cognates have been shown to promote transfer in bilinguals (Patterson & Goldrick, 2007). Given that the target phonetic realizations are present in Mexican Spanish but not in all the other varieties of Spanish, the stimuli were produced by a 36-year-old female Mexican (Chihuahua) speaker of Spanish.

### Data analysis and results

The learners’ productions were transcribed by two individuals, namely the first author and another linguist with training in phonetics and L2 acquisition. Her native language is English and she has

<table>
<thead>
<tr>
<th>Grapheme-to-sound correspondence</th>
<th>% type of production</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auditory</td>
<td>Orthography-induced phonological transfer</td>
</tr>
<tr>
<td>&lt;d&gt;-[ð]</td>
<td>9.8%</td>
<td>90.2%</td>
</tr>
<tr>
<td>&lt;ll&gt;-[j]</td>
<td>89.1%</td>
<td>4.2%</td>
</tr>
<tr>
<td>&lt;v&gt;-[b]</td>
<td>8.4%</td>
<td>91.6%</td>
</tr>
<tr>
<td>&lt;z&gt;-[s]</td>
<td>37.1%</td>
<td>30.3%</td>
</tr>
</tbody>
</table>

*Note: Percentages result from the total number of instances relative to trials after removing “not produced,” “deleted,” and “other” codes, as described in the analysis.
near-native fluency in Spanish. The author is a native speaker of Farsi with near-native fluency in English and Spanish. The data were also inspected acoustically for fusion at the phonetic level as described below.

Responses were coded as: (a) “auditory” if the sound was the same as the target sound; (b) “combination” when a sequence of sounds was produced such as [lj] for [ll]-[j]; (c) “orthography-induced transfer” when a learner’s production consisted of the non-target-like substitution of an L1 sound for the target (such as production of [v] for [b], [d] for [ð], [z] for [s] and [l] for [j]); (d) “fusion”; (e) “not produced” if the entire word was not produced; (f) “deleted” if only the target sound was deleted; and (g) “other” for all other productions, such as a [r] for target [l]. Following the removal of responses coded “not produced,” “deleted,” or “other,” there were 2441 tokens (out of 2800) remaining for the primary analysis.

It was predicted that exposure to auditory-orthographic input would result in L1-based phonological transfer and combination productions and the type of grapheme-to-phoneme, condition, and position in the word would modulate the proportion of integrated productions. There were no instances of combination productions in the auditory-only condition, as predicted. As such, these tokens will not be analyzed further.

As shown in Table 2, there was evidence of both L1-based phonological transfer and combination productions in auditory-orthographic conditions, although there was generally a higher rate of the former than the latter. Moreover, whereas transfer was evident for all the grapheme-to-sound correspondences tested in this study such as <v>-[b] (e.g., [veneno] for [beneno]), <d>-[ð] (e.g., [koden]a for [kodena]), <ll>-[l] (e.g., [pajete] for [pajete]) and <z>-[s] (e.g., [zapito] for [zapito]), combination productions were only attested for <ll>-[j] (e.g., [pajete] for [pajete]). A chi-square test of independence was conducted to explore whether proportion of auditory responses, orthography-induced transfer, and combination productions was different for various grapheme-to-sound correspondences included in this study. The results of the test were significant, $\chi^2(6) = 1513.77, p < 0.001$, with a moderately strong effect size as measured by Cramer’s V (0.56). The proportions of each type of production for each grapheme-to-sound are displayed in Table 2. Given the presence of empty cells, a follow-up test was conducted collapsing across non-auditory responses (orthographic switches and combinations). These results were also significant, $\chi^2(3) = 1266.44, p < .001$, with an even stronger effect size as measured by Cramer’s V (0.72), suggesting that the significance of the original analysis was not a statistical anomaly due to empty cells.

As predicted condition also constrained the proportion of auditory-orthographic combination realizations for <ll>-[j] (see Figure 2). The highest percentage combination is produced in the ortho-training condition (12.3%; count = 37), followed by ortho-training and production condition (5.6%; count = 19), and ortho-production condition (2.7%; count = 9).

To investigate the relationship between combination productions and condition for <ll>-[j], $\chi^2$ test of independence was conducted. The results of this test were significant, $\chi^2(2) = 25.16, p < 0.001$, with small effect size, Cramer’s V = 0.16. Examination of standardized residuals were then assessed to determine if there was an orthographic influence during training ($\Delta$ RE > 2). Results indicated that the proportion of combination productions in the orthography-training condition were significantly higher than in the orthography-production condition (RE = 3.9 vs. -2.8) and marginally so in the orthography-training and production condition (RE = -0.8 vs. -2.8). To confirm these results, follow-up 2x2 $\chi^2$ tests were conducted, comparing combination rates during the ortho-training condition versus the ortho-production condition, as well as the ortho-training and production versus the ortho-production condition. These follow-up tests confirmed the analysis of residuals. Rates of combination were significantly higher in the ortho-training condition than the ortho-production condition, $\chi^2(1) = 22.25, p < 0.001$, and marginally higher in the ortho-training and production condition than the ortho-production condition, $\chi^2(1) = 3.51, p < 0.06$. 
With regards to the effect of position, chi-square test of independence was conducted to compare the proportion of combination production in intervocalic and initial position in words. The results of this test were significant, $\chi^2(1) = 25.52, p < 0.001$. However, the effect size of this difference was small, Cramer’s V = 0.16, with 11.0% combination productions in word-medial intervocalic position and 2.8% combination productions in word-initial position (see Figure 3). Therefore, we can conclude that proportion of combination productions is significantly higher in the intervocalic position.

Furthermore, in the word-initial position, combination occurs only in the ortho-training condition (7.9%; count = 12) and ortho-training and production conditions (1.2%; count = 2). In the word-medial position, it occurs in ortho-production (5.9%; count = 9) and ortho-training (17.6%; count = 17) and ortho-training and production condition (10.1%; count = 25). Therefore, the combination of condition and position that produces the largest proportion of combination production is orthography-training condition in the word-medial position (see Figure 4).

All but one pair of grapheme-to-sound correspondences were matched in voicing, except for the <z>-[s] pairing. In this instance, there is also the possibility of a fusion effect such that the production reflects neither an orthography-induced transfer ([z]) or an accurate reproduction of the auditory signal ([s]), but instead results in a devoiced [z̥] (see Table 3). To explore this possibility, the voicing status of this pairing was analyzed acoustically, following Jesus and Shadle (2002). Voicing is represented on a spectrogram by the presence of striation/vertical bars also referred to as the “voicing bar” (indicative of vocal fold vibration) in the lowest frequencies. In the waveform, voicing is represented by the presence of repetitive cycles or periodicity (i.e., the presence of pitch). Figures 5 and 6 show a voiced [z] and a devoiced [z̥] in the word <zapito>.

As with the previous analysis, condition also constrained the proportion of auditory-orthographic fusions for <z>-[s] (see Figure 5). The highest percentage fusion was produced in the ortho-training and production condition (45.7%; count = 81), followed by ortho-training condition (37.2%; count = 55), and ortho-production condition (21.1%; count = 38).

To investigate the relationship between frequency of fusion productions (see Figure 7) as defined by a change in voicing and condition, $\chi^2$ test of independence was conducted. The results of this test were significant, $\chi^2(2) = 34.98, p < 0.001$, with small effect size, Cramer’s V = 0.19. Examinations of standardized residuals were then assessed to determine if there was an orthographic influence during training (Δ RE > 2). Results indicated that the proportion of fusion productions in the orthography-training condition were significantly higher than in the orthography-production condition (RE = 0.6 vs. -3.1) and significantly higher in the orthography-training and production condition than in the orthography-production condition (RE = 2.6 vs. -3.1). To confirm these results,
follow-up 2x2 $\chi^2$ tests were conducted comparing fusion rates during the ortho-training condition versus the ortho-production condition, as well as the ortho-training and production versus the ortho-production condition. These follow-up tests confirmed the analysis of residuals. Rates of fusion were significantly higher in the ortho-training condition than the ortho-production condition, $\chi^2(1) = 9.63, p = 0.008$, Cramer’s $V = 0.17$, and significantly higher in the ortho-training and production condition than the ortho-production condition, $\chi^2(1) = 30.30, p < 0.001$, Cramer’s $V = 0.30$.

We also considered the possibility of fusion at the phonetic level for <d>-[δ]. Given the difference in the place of articulation between the English [d] and the Spanish dental approximant [δ], we looked for fusion in the form of dental plosive realizations. However, we did not find any evidence of it. Instead, there was a rate of transfer (90.2%), where 95% of erroneous realizations consisted of [d] productions and 5% of [ɾ] productions.

5 Discussion

The results in this paper demonstrate a novel way in which auditory and orthographic input may interact in naïve L2 learners and provide further evidence for the multi-modality of L2 speech.
We had predicted that in addition to transfer, integrated sounds would only be attested in the auditory-orthographic conditions (Hypothesis 1). We also predicted an effect of the condition of training and production, type of grapheme-to-sound correspondence and position in the word (Hypothesis 2). Both hypotheses were confirmed. Indeed, when learners were exposed to incongruent auditory-orthographic pairs, their L1 pronunciation of graphemes integrated with the auditory (L2) presentation, leading to both transfer and a McGurk-like combination effect. Similar to the McGurk effect, exposing naïve learners of an L2 to incongruent grapheme-to-sound correspondences

<table>
<thead>
<tr>
<th>Learning condition</th>
<th>% type of production</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auditory [s]</td>
<td>Orthography-induced transfer (voiced [z])</td>
</tr>
<tr>
<td>Ortho-training</td>
<td>41.2%</td>
<td>21.6%</td>
</tr>
<tr>
<td>Ortho-training and production</td>
<td>30.0%</td>
<td>24.3%</td>
</tr>
<tr>
<td>Ortho-production</td>
<td>35.6%</td>
<td>43.3%</td>
</tr>
</tbody>
</table>

*Note: Percentages result from the total number of instances relative to trials after removing “not produced,” “deleted,” and “other” codes, as described in the analysis.

Figure 5. Waveform and spectrogram of the devoiced instance of [zapito]. The absence of voicing is suggested by the fact that there is neither a voicing bar in the spectrogram or any periodicity present during the frication portion for word-initial [z].
might have triggered perceptual integration in which the conflicting/incongruent L2 and L1 sounds were perceived as a single percept, resulting in the production of combination sounds (e.g., [lj] as in [po\text{j}o] instead of [po\text{j}o] for <pol\text{l}o>). Moreover, there was evidence of fusion as a result of exposure to auditory [z] and orthographic <z> resulting in devoiced [\text{z}̥] realizations, as in [\text{zapito}] for <\text{z}apito>-[\text{z}apito].

So far, the McGurk phenomenon has been understood as the influence of visual or lip-read information on speech perception, although Fowler & Dekle (1991) have previously hinted at the possibility of an auditory-orthographic interaction during L1 perception, and Rafat (2011) and Mathieu (2016) have suggested the possibility of a process akin to the McGurk effect at the very onset of L2 acquisition. The results here suggest that the McGurk-like effect can be brought on by exposure to incongruent auditory-orthographic in a non-native context—a general tendency that has been previously observed by Chen and Hazan (2009) with respect to the effect of incongruent auditory-visual (facial) input. Additionally, to the best of our knowledge, the well-known McGurk phenomenon has been shown with respect to auditory-visual integration at the phonemic level and in terms of place of articulation; however, the results here show that exposure to conflicting information from two different streams (auditory-orthographic) may lead to the production of combination sounds as a result of integration of place of articulation at the phonemic level, as well as fusion, at the acoustic/phonetic level, where the [z] sounds were produced with a [-voice] feature, which is characteristic of [s]. It must be noted, however, that the results of the current study are based on a production experiment, and one must consider the possibility they may have been partly

\textbf{Figure 6.} Waveform and spectrogram of the voiced instance of [zapito]. Fricative voicing is indicated by the fact that a voicing bar is present in the spectrogram the frication portion, as well as by the accompanying periodic waveform.
driven by a perceptual effect. With that said, the significantly greater proportion of combination and fusion realizations in the condition in which orthography was present at training and not production relative to the condition in which orthography was presented at production and not training suggests that these effects, at the very least, include a perceptual component (see Figure 2). A follow-up perception study, where learners are exposed to incongruent auditory-orthographic input in an identification task, can further test the hypotheses put forth here regarding the effect of incongruent auditory-orthographic input integration in L2 speech learning.

Interestingly, auditory-orthographic integration in the form of combinations was found specifically for the grapheme-to-sound correspondence <ll>-[j]. A number of factors may have contributed to this. First, belonging to the same class of sounds and having a large enough distance in terms of place of articulation might be a prerequisite for a combination perception or production of the L2 and the L1 sounds. Both [l] and [j] are approximants and the distance between the two sounds in terms of place of articulation is larger than the other sounds examined in this study. The only other L1 and L2 sounds that belong to the same class of sounds are [s] and [z], which are voiceless and voiced sibilant fricatives, but they have the same place of articulation (i.e., alveolar). Additionally, the [l]-[j] pairing is the only one that occurs in English (e.g., <million>-/mɪljən/), whereas the other pairings do not, likely affecting both the perception and production of combinations.

The <ll>-[j] correspondence also resulted in the lowest rate of transfer in comparison with all the other grapheme-to-sound correspondences. Previously, Rafat (2011; 2016) proposed that a low rate of orthography-induced transfer is due to a larger acoustic/phonetic distance between an L1 and an L2 sound than a shared grapheme to which they may correspond. The results here also indicate that a sufficiently large acoustic-phonetic distance between the L1 and the L2 may also be a necessary condition for auditory-orthographic integration to take place and yield or result in combination productions. In the absence of a sufficiently large acoustic-phonetic distance between the L1 and the L2 sounds, the orthographic input may override the auditory input and/or vice-versa. An acoustic phonetic distance here might be related to distance in place of articulation.

Another reason why only the grapheme-to-sound pairing <ll>-[j] led to combination productions might have been due to <ll> being a digraph. Whether digraphs increase the probability of the occurrence of a McGurk-like effect when learners are exposed to auditory-orthographic input needs to be examined in future studies. Moreover, the orthographic McGurk-effect with the digraph <ll> seen

![Figure 7. Percentage fusion production with <z>-[s] grapheme-to-sound correspondence by auditory-orthographic condition. * Denotes significance at p < 0.05.](image)
here may be language-specific and may result in other effects in speakers of other languages. For example, exposure to digraphs induced L1-based phonological transfer leading to geminate production in highly advanced Italian learners (e.g., Bassetti & Atkinson, 2015; Bassetti, 2017) and Japanese learners of English (e.g., Sokolovic-Perovic, Dillon, & Bassetti, 2016). Furthermore, English has a deep orthographic system with multiple many-to-one and one-to-many grapheme-to-phoneme correspondences. Previously, it has been put forth that languages with deep orthographies are affected by orthography to a lesser degree (e.g., ODH, Katz & Frost, 1995; Erdener & Burnham, 2005). This prediction is also consistent with the predictions of the Dual Route Cascade Model of reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and some recent neuro-scientific evidence. The Dual Route Cascade Model proposes that after letter identification, word reading processing may follow two pathways in the brain: phonological and lexical (Binder, Medler, Desai, Conant, & Liebebthal, 2005). Reliance on one of these pathways is modulated by regularity, lexicality, and familiarity. There is also evidence showing that grapheme-to-phoneme mapping in languages with shallow orthographies mostly relies on regions involved in grapho-phonological processing (superior temporal, supra-marginal, and opercular inferior frontal regions), indicating an activation of non-lexical pathways, and associated with analysis at the grapheme-to-phoneme level (Buetler et al., 2014). In contrast, grapheme-to-phoneme mapping in languages with deep orthographies seems to mostly rely on regions involved in lexico-semantic processing (inferior and middle temporal and triangular inferior frontal regions), indicating an activation of lexical pathways and regions associated with the analysis at the word level. Thus, in addition to lexicality and familiarity, orthographic depth or language specific demands may indeed impact reading route selection or reading strategies (Buetler et al., 2014; 2015). It is also important to note that the engagement of a given pathway is not exclusive; that is, reading processing generally involves both routes, but one may be predominantly activated compared to the other depending on the orthographic depth index of the language (Buetler et al., 2014; Heim et al., 2005; Mousikou, Coltheart, Finkbeiner, & Saunders, 2010; Timmer, Vahid-Gharavi, & Schiller, 2012). Given there is some evidence to suggest that shallow and deep orthographies favor a differential degree of reliance on the phonological and lexical pathways, depending on the language-specific demands in learners (Buetler et al., 2014; 2015), the orthographic McGurk-like effect might be evinced at a higher rate in speakers whose L1 has a shallow orthographic system, such as Serbian, which is characterized by one-to-one grapheme-to-phoneme correspondences, than speakers with a deep orthographic system, such as English speakers.

As mentioned above, the hypothesis that condition of training and production would also modulate the McGurk-like effect was verified. The presentation of orthographic input at the training phase resulted in a significantly higher rate of combination productions. This is consistent with previous results where the presentation of orthographic input at training also led to a significantly higher rate of L1-based phonological transfer (Rafat 2011; 2016). That is, orthography appears to exert a significantly stronger effect when it is presented during the training phase together with the auditory input than when it is presented at production only. Previous research has also shown that the precision of temporal synchrony can have an impact on cross-modal interactions (Dixon & Spitz, 1980; Stevenson, Zemtsov, & Wallace, 2012; Stevenson and Wallace, 2013). Specifically, integration is more effective when multi-modal stimuli are presented simultaneously, as they were during the training phase in the current study, rather than consecutively (e.g., Senkowski, Talsma, Grigutsch, Herrmann, & Woldorff, 2007). The inclusion of orthography at production in addition to training did not increase the rate of combination productions, suggesting that exposure to orthography at training can have a more robust effect on increasing the possibility of auditory-orthographic integration than the type of input at production. In other words, the type of input at training appears to be more important in modulating auditory-orthographic integrations than the type of input at production. The fact that the inclusion of orthography at training and production did not
result in a higher rate of integration than in the orthography-training condition is also consistent with Rafat (2011; 2016).

As for the effect of position on combination productions, as predicted position in the word constrained the rate of combination productions. Specifically, the word-medial inter-vocalic position yielded the highest rate of combination productions for <lj>-[j]. The word-initial word-medial asymmetry has been attributed to the word-initial acoustic prominence and has been previously evoked to explain sound patterns of the world (Steriade, 1997; Beckman, 1998; Cho & Jun, 2000) and L2 speech learning (Colantoni & Steele, 2006; 2008). In this case, there is no evidence to suggest that the sound [l] is more acoustically salient in the word-initial position. In fact, it is more probable that another factor, namely L1 phonotactics, might have been at play here. That is, combination productions may have been constrained by the phonological and phonotactic structures of the L1, which may not truthfully reflect a perceptual reality of the auditory-visual integration that may allow for more illegitimate, or even fused, percepts than the ones restituted in production. L1 phonotactics have been shown to constrain sequence/cluster productions in L2 learners (e.g., Broselow, 1988; Hancin-Bhatt & Bhatt, 1997; Shademan, 2002; Colantoni & Steele, 2006; Cardoso, 2008). Crucially, L1 phonotactics can also affect L2 perception. In fact, Japanese learners actually perceive an illusory epenthetic vowel when presented with English onset clusters that are illegitimate in Japanese phonology (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999). It is possible that in the current study L1 phonotactics were also interfering at the perceptual and/or production levels and not allowing the percept or the production of [lj] in the word-initial position, where it is an illegal onset in the L1. Word-internally, however, words such as <million>-[mɪljən] exist in English, where [l] forms the coda of the first syllable and [j] the onset of the second one.

That L1 phonotactics may have modulated the production of combination productions is also evinced by the fact that exposure to none of the other grapheme-to-sound correspondences resulted in combination productions. Combinations such as *[bv], *[dɔ], and *[sz] are not legitimate onsets in English either. Moreover, these combinations are universally marked because they have a low sonority difference. According to the minimal sonority distance principle, the closer the adjacent segments in a cluster are to each other in sonority, the more marked the cluster (see Steriade, 1982, and Harris, 1993, among others for more on the Sonority Distance Principle). It has previously been shown that orthographic input may modulate the L2 production of Polish clusters by naïve English and Japanese speakers, where the presence of orthography leads to a lower rate of deletion and a higher rate of epenthesis (Young-Scholten et al., 1999). Here the relationship is reversed, where L1 phonotactics may modulate how the L1 and the L2 sounds are integrated as a result of exposure to orthographic input. It is also worth noting that [bv] and [sz] are homorganic (i.e., have the same place of articulation) but [dɔ] like [lj] is heterorganic (i.e., they have a different place of articulation). In the classical McGurk phenomenon, a new percept only results when the place of articulation of the visual and the auditory stimuli are different (Massaro & Simpson, 2014). Whether heterorganicity is one of the prerequisites for combination productions as a result of exposure to auditory-orthographic input needs to be further investigated.

If phonotactics play a role in constraining combination productions, this begs the question of why exposure to incongruent audio-orthographic input did not result in a [jl] sequence, because words such as <boiler>-[ˈbɔɪlər] are also legitimate and frequent in English. We tested the possibility that the [jl] sequence might be more frequent in English than the [lj] one. As such, we performed a frequency analysis using the English Lexicon Project Web Site database (Balota et al., 2007). We used the HAL word frequency database and isolated the words containing <lj> and <jl>. Whereas the <lj> versions have few items that are very frequent such as <milllion>, the average log frequency (5.72) was actually lower than the average log frequency of <lj> (5.94). Additionally, there are many more items with the <jl> sequence than the <lj> sequence (460 relative to 40). To
take this into account, we calculated the log of the summed frequencies for all words containing either <j> or <l>, and the log frequency of <jl> was much higher than <lj> (14.9 relative to 12.0). Being on a log scale, this difference of 2.9 is quite significant.

For consistency, we performed the same calculation for items where <l> was not preceded or followed by <j>. As expected, the log of summed frequencies for these items was again significantly greater, at 16.3. Given these analyses, we are confident that our effects were not driven by a higher frequency of <lj> in the English vocabulary, as <jl> turns out to be more frequent. We therefore propose that another confounding factor, namely primacy of the orthographic input (Bassetti, Escudero, & Hayes-Harb, 2015) might have played a role here. However, this hypothesis has to be further investigated.

We also question why phonotactic rules do not seem to play a role in audio-visual (facial) integration. That is, why did audio-visual /ba/ and /ga/ previously result in the percept of illegitimate combinations in such as /bga/ in English-speaking participants (Fowler & Dekle, 1991)? The difference might have various causes. First, combination productions are reported mostly when participants are presented with monosyllabic utterances such as /ba/ and /ga/ in experiments testing the McGurk effect, whereas in the present study the participants were presented with entire words. Second, in experiments examining the McGurk effect, participants are usually asked to report what they have heard whereas here the participants were asked to produce the words they were asked to learn. Third, this might be due to the differential degree of reliance of L2 learners on the orthographic input. In a non-native context, learners at the onset of acquisition, in particular, may be paying more attention to the orthographic input to parse unfamiliar speech than in the native context.

It must also be noted that the McGurk effect typically refers to both a fused (e.g., /d/ when presented with auditory /b/ and visual /g/) and a combination percept (e.g., /bg/), with the former typically occurring at a higher rate. In this study, in addition to combination productions (e.g., [lj]) there was also evidence of fused productions. However, fused productions were only tested for the grapheme-to-sound pair <z>-[s], where the resultant sound [z̥] exhibited the voicing characteristic of [s] (i.e., [-voice]). First, it is possible that place of articulation of the L2 and the L1 sounds that corresponded to the shared graphemes have played a role here. The pairs [b] and [v] and [z] and [s] are homorganic (have the same place of articulation), therefore a medial sound or place of articulation for these two pairs of sounds did not exist. However, [s] and [z] differ in voicing, making it possible that the percept exhibits some acoustic features of the L1 and the voicing feature of the L2 sound. As for [l] and [j], although they are heterorganic, the distance in the place of articulation between the components of each pair may not be large enough to lead to the percept and/or production of an in-between/fused sound. Some of the previous literature also suggests that the distance in the place of articulation between two incongruent sounds may also modulate the classical McGurk effect. Specifically, whereas the presentation of incongruent front and back visual and auditory consonants (e.g., /b,g/ and /p,k/) may result in a fused percept (e.g., /d/ and /t/), respectively) (McGurk & McDonald, 1976), the presentation of incongruent visual and auditory mid /d,t,n/ and back consonants /k,g/ does not result in a fused percept (Massaro & Simpson, 2014). In the absence of a voicing contrast, a larger distance in terms of place of articulation between the two incongruent sounds appears to be a prerequisite for the fusion of conflicting auditory-visual and auditory-orthography channels. Despite this, to the best of our knowledge there are no studies that report on a minimum required sufficient distance in terms of place of articulation for fusion. Two issues arise from this proposal. First, for [l] and [j], a sound in the same class that is in between the L1 and the L2 sounds in terms of place of articulation does not actually exist in the L1 English phonological inventory. Second, exposure to orthographic input may simply block the percept of a fused sound by creating a more robust illusion of the entire L1 sound instead of creating an illusion of only partial features.
of that L1 sound, which is then integrated with the L2 sound (auditory input) if the auditory input is
discernable (or can be clearly perceived). A future replication of the Fowler and Dekle (1991) study
needs to be conducted with L2 learners to directly test these hypotheses.

As for [d] and [ð], this is a sound that resulted in 90.2% transfer in the auditory-only condition.
This suggests it was categorized as a "similar" sound by the participants. In other words, in terms
of the Speech Learning Model (Flege, 1995), because the acoustic/phonetic distance between the
L1 and the L2 sounds was small, the L2 sound was mapped onto its nearest L1 category, removing
the possibility of any potential fusion effects and leading to a high rate of L1-based transfer.

6 Theoretical and pedagogical implications

Current dominant models of L2 speech acquisition, with the exception of Colantoni and Steele
(2008) are heavily based on perception (Flege, 1995; Brown, 1998, Escudero, 2005; Best & Tyler,
2007). Equivalence classification (Flege, 1995) and category assimilation (Best & Tyler, 2007)
explain some of the mechanisms that underlie L2 perception, and the role of universal phonetic and
articulatory constraints are highlighted in Colantoni and Steele’s (2008) more production-based
model. Bassetti (2017) raises the issue that none of the current models predict situations where (a)
the L2 has only one category that is mapped onto two different categories in the learner’s L1 pho-
nological system; or (b) two L2 categories map onto one category in the L1. The findings of our
study imply that L2 speech learning models in addition to addressing the issue of L1-based transfer
should also address situations where the L1 and the L2 categories are integrated, even though the
L2 sound is an old sound in the L1. Thus, there is a need for a model that would treat L2 speech
learning as a multi-modal event.

This study also has pedagogical implications. Previously, it has been suggested that exposure to
orthographic input may lead to the establishment of non-target-like L2 categories, even for sounds
that may exist in the learner’s L1 phonological inventory (Rafat, 2011; 2015; 2016). Therefore, it
is advisable that language instructors be mindful of the effect of orthography and consider the tim-
ing of presentation of orthographic input and the effect of individual grapheme-to-sound corre-
spondences. This work further demonstrates that the synchronous or simultaneous presentation of
auditory and orthographic input can have a more adverse effect on pronunciation than when ortho-
graphic input is presented asynchronously or consecutively, in this case, right after the presentation
of the auditory input at production. Moreover, it would be good practice for instructors who teach
L2 pronunciation to consider the different error types that different grapheme-to-sound corre-
spondences may trigger. The assumption that exposure to orthographic input may only lead to posi-
tive or negative L1-based phonological transfer is somewhat simplistic and in the absence of
awareness of the fact that exposure to auditory-orthographic input may result in other error types,
instructors may not notice, correct, or understand the root causes of variability and incorrect pro-
nunciation in the patterns of production of the L2 learners.

7 Conclusion

In sum, the current results show that exposure to orthographic and auditory input may induce a
McGurk-like effect that results in combination productions in L2 speech learning at the very onset
of acquisition, in addition to transfer. The same factors that have previously been shown to control
the rate of orthography-induced transfer, namely the type of grapheme, condition of training and
production, and position in the word (Rafat, 2011; 2016) also modulated the McGurk-like effect
here. Future studies may examine the effect of other grapheme-to-sound correspondences and
compare languages with shallow and deep orthographies. It would also be interesting to see whether
this effect persists in more advanced learners. Moreover, the current results call for new models of L2 speech learning that would paint a more complete picture of the processes that underlie L2 speech perception and production. Finally, we have identified a new error type induced by orthographic input, which instructors of Spanish as an L2 can consider in pronunciation teaching.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Please note that the digraph <ll> corresponds to different sounds in different varieties of Colombian Spanish (Lipski, 1994). However, it traditionally corresponded to [ʎ] in Peninsular Spanish (Navarro Tomás, 1990; Quilis, 1993) and has been preserved in more conservative varieties in certain regions such as Paraguay and the Andean region: parts of Colombia, Ecuador, Peru, Bolivia, and small areas of Chile and Argentina (Canfield, 1981). <ll> may also be realized as a [ʝ] and [ʃ] in Buenos Aires Spanish (Chang, 2008). <ll> may also be a stop [j] in some varieties (Hualde, 2005) or an affricate variant [ʤ] in Paraguay (Hualde, 2005).

2. We refer to what is typically referred to as “grapheme-to-phoneme” correspondences as “grapheme-to-sound” when we consider the shared graphemes that correspond to different sounds in Spanish and English in this study, because we are considering the incongruence between the L1 and the L2 sounds/allophonic realizations not the L1 and the L2 phonemes.

References


