The contribution of visual articulatory gestures and orthography to speech processing: Evidence from novel word learning

Chotiga Pattamadilok, Pauline Welby, Michael D Tyler

Short title: Visual information and speech processing

The contribution of visual articulatory gestures and orthography to speech processing: Evidence from novel word learning

Chotiga Pattamadilok[a], Pauline Welby[a], Michael D. Tyler[b]

[a] Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

[b] School of Psychology and the MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

Corresponding author

Chotiga Pattamadilok

Laboratoire Parole et Langage

Centre National de la Recherche Scientifique (UMR 7309)

5, Av. Pasteur

13100 Aix-en-Provence

France

Email: chotiga.pattamadilok@lpl-aix.fr

Tel : +33 601323435

ABSTRACT

Auditory speech appears to be linked to visual articulatory gestures and orthography through different mechanisms. Yet, both types of visual information have a strong influence on speech processing. The present study directly compared their contributions to speech processing using a novel-word-learning paradigm. Native speakers of French, who were familiar with English, learned minimal pairs of novel English words containing the English /θ/-/f/ phonemic contrast under one of three exposure conditions: 1) the auditory forms of novel words alone, 2) the auditory forms associated with articulatory gestures, or 3) the auditory forms associated with orthography. The benefits of the three methods were compared during training and at two post-training time points where the visual cues were no longer available. We also assessed participants' auditory-only discrimination of the /θ/-/f/ contrast pre- and post-training. During training, the visual cues facilitated novel word learning beyond the benefit of the auditory input alone. However, these additional benefits did not persist when participants' discrimination and novel-word-learning performance were assessed immediately after training. Most interestingly, after a night's sleep, participants who were exposed to orthography during training showed significant improvement in both discrimination and novel-word learning compared to the previous day. The findings are discussed in terms of online versus residual impacts of articulatory gestures and orthography on speech processing: While both visual cues are beneficial when they are simultaneously presented with speech, only orthography shows residual impacts leading to a sleep-dependent enhancement of lexical knowledge through memory consolidation and retuning of the second language /θ/-/f/ contrast.

Keywords: visual speech; spelling knowledge; L2 phonemic contrast; memory consolidation; phonological representation

INTRODUCTION

Speech is primarily investigated as an auditory phenomenon. Nevertheless, it is widely acknowledged that visible articulatory gestures and orthography are two major sources of visual input that strongly affect the way speech is processed and represented in the cognitive system (Grainger & Ziegler, 2007; Harm & Seidenberg, 1999; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Stone & Van Orden, 1994). The association between speech and articulatory gestures is natural and relies on a concrete, biological link between action and perception. Infants are sensitive to this association at as early as two months old, when they are able to accurately match an auditory input with an articulating face (Bristow et al., 2009; Dodd, 1979). A few months later, infants can integrate two frequently co-occurring auditory and visual speech stimuli into a single percept (Bristow et al., 2009; Burnham & Dodd, 2004). This ability is demonstrated by the emergence of the "McGurk effect" (McGurk & MacDonald, 1976). It is one of the most powerful perceptual illusions, accounting for the impression of perceiving, for example, the syllable /da/ when simultaneously exposed to the spoken syllable /ba/ and a face articulating the syllable /ga/. This effect, which also exists in adults, demonstrates that speech sounds are integrated with the articulatory gestures during speech perception. Due to their sensitivity to this visual input, at around six months old, infants are able to extract phonetic information from their interlocutors' lip movements, allowing them to perceive and learn phonemic contrasts more efficiently (Teinonen, Aslin, Alku, & Csibra, 2008).

It has been argued that during speech processing the presence of articulatory gestures contributes to speech processing at two levels, by providing temporal markers corresponding to acoustic properties of the speech signal and by providing specific information about the identity of individual phonemes. The latter contribution has a direct consequence on spoken word recognition since it allows both anticipatory auditory processing and the reduction of

competition among plausible lexical candidates (Fort et al., 2013; Grant & Seitz, 2000; see Peelle & Sommers, 2015 for a review). The benefits of visual speech are particularly significant in adverse speech processing situations, such as when the acoustic signal is degraded, in the presence of noise, or in hearing-impaired perceivers (Grant & Seitz, 2000; Schwartz, Berthommier, & Savariaux, 2004; Sumby & Pollack, 1954; Summerfield, 1987).

Unlike articulatory gestures, the orthography of a specific language or variant is connected to speech through artificial and arbitrary links; for example, other than convention, there is no organic reason why, in many languages using the Roman alphabet, the letter 'B' is associated with the phoneme /b/. Nevertheless, orthography provides a visual code that allows the classification of variable, transient, and often ambiguous speech signals into more stable abstract categories (Ehri, 1984, 1985). Given the nature of the association, it is not surprising that the influence of orthography has mostly (although not exclusively) been reported in high-level speech processing situations that require recognition, memorization or metaphonological analyses (Muneaux & Ziegler, 2004; Pattamadilok, Kolinsky, Luksaneeyanawin, & Morais, 2008; Pattamadilok, Lafontaine, Morais, & Kolinsky, 2010; Seidenberg & Tanenhaus, 1979; Tyler & Burnham, 2006; Ventura, Morais, Pattamadilok, & Kolinsky, 2004). In addition to these artificial experimental conditions, the influence of the written code on speech comprehension has also been reported in real-life situations such as watching movies in a foreign language. Subtitles that matched speech sounds were reported to improve second language (L2) speech understanding by providing lexical cues that helped listeners to retune their perceptual system (Mitterer & McQueen, 2009). Subtitles in the listener's native language (L1), on the other hand, induced lexical interference that hindered perceptual learning. In line with the findings obtained in these speech processing situations, several studies also reported a higher success rate of oral vocabulary acquisition when the pronunciations of novel words were presented with their spellings, which suggests a contribution of orthography in maintaining

newly acquired knowledge (on L1 in young children: Ehri & Wilce, 1979; Ricketts, Bishop, & Nation, 2009; Rosenthal & Ehri, 2008; on L2 in adults: Bürki, Welby, Clément, & Spinelli, 2019).

Most impacts of orthographic knowledge on speech processing have been explained by an activation of the written code upon hearing speech (Grainger & Ziegler, 2007; Harm & Seidenberg, 1999; 2004). However, several studies conducted on the L1 also argued that learning to read could modify the nature of the phonological representations themselves. It could lead to, for instance, a reduction of the grain-size of phonological representations, a better specification of phoneme boundaries, a modulation of the activation threshold of spoken words or a transformation of phonological into "phonographic" representations (Burnham, 2003; Hoonhorst et al., 2011; Morais, Cary, Alegria, & Bertelson, 1979; Pattamadilok, Morais, Colin, & Kolinsky, 2014; Pattamadilok, Perre, Dufau, & Ziegler, 2009; Perre, Pattamadilok, Montant, & Ziegler, 2009; Serniclaes, Ventura, Morais, & Kolinsky, 2005; Taft, 2006, 2011; Veivo & Järvikivi, 2013). Thus, while the links between orthography and phonology may be arbitrary, there are a number of possible mechanisms by which orthography could have a direct influence on phonology.

Another research field where the contribution of visual articulatory gestures and orthography has been investigated is the acquisition of L2 speech. It is clearly established that acquiring the specific phonological system of a native language leads to poor discrimination accuracy for certain non-native phonological distinctions (Best, 1995; Kuhl, 1992; Werker & Tees, 1984; for recent reviews, see Bohn, 2019; Tyler, 2021). For example, English and Greek monolinguals are known to differ from each other in their discrimination of contrasts from the Ma'di language, according to the phonological system of their L1 (Antoniou, Best, & Tyler, 2013). Both Greek and English have the phonemes /d/ and /t/ but they have different phonetic realizations in each language. In Greek, /d/ is prevoiced [d] and /t/ is voiceless unaspirated [t],

whereas in English /d/ is [t] and /t/ is voiceless aspirated [tʰ]. The Greek monolinguals outperformed the English monolinguals on the discrimination of Ma'di prevoiced /d/ ([d]) versus voiceless unaspirated /t/ ([t]), as the phonetic difference signalled a phonological contrast in Greek but not in English. Both groups had difficulty discriminating Ma'di prevoiced plosive versus implosive coronal stops (/d/-/ɗ/), which did not signal a phonological contrast for either group. Similar influences of the L1 on perception have been found for vowels (Faris, Best, & Tyler, 2016, 2018; Tyler, Best, Faber, & Levitt, 2014) and lexical tone (Chen, Best, & Antoniou, 2020; Reid et al., 2015).

Models of L2 speech learning have been devised to predict the likelihood of acquiring new L2 categories and of improving L2 contrast discrimination (e.g., the Second Language Linguistic Perception Model: van Leussen & Escudero, 2015; the Perceptual Assimilation Model of Second Language Speech Learning: Best & Tyler, 2007; the Speech Learning Model: Flege, 1995). For all current models, factors that increase the perceived dissimilarity between L1 and L2 sounds are predicted to improve the ability to perceive L2 sounds and to form new phonemic categories. There is a large body of research on the use of high-variability phonetic training with auditory-only speech to improve L2 speech perception (e.g., Carlet & Cebrian, 2019; Logan, Lively, & Pisoni, 1991), but more recently researchers have turned their attention to how visual articulatory gestures and orthography may be used as visual cues to distinguish difficult phonemic contrasts.

With respect to the contribution of visual articulatory gestures, Fenwick et al. (2017) examined the categorization of non-native consonants by Australian English monolinguals and found a benefit of audiovisual input when visual and auditory modalities showed a converging phoneme categorization pattern. That is, categorization consistency for audiovisual presentation was higher than for auditory-only presentation when the consonant was labeled as the same category in both auditory-only and visual-only conditions. When the categories were

mismatched in auditory-only and visual-only conditions, however, the categorization consistency was lower for audiovisual than auditory-only presentation. An audiovisual benefit has also been reported for Spanish-Catalan bilinguals' perception of the Catalan /ɛ/-/e/ contrast (Navarra & Soto-Faraco, 2007) and in Korean and Mandarin Chinese listeners' perception of English interdental fricatives (/θ/ and /ð/, Wang, Behne, & Jiang, 2009). In a speech perception in noise task, native speakers of Korean benefited from audiovisual speech in L2 English (Xie, Yi, & Chandrasekaran, 2014).

Overall, the literature suggests that the contribution of articulatory gestures on L2 speech processing is robust, at least in low-level speech perception tasks. This contribution has mainly been explained by an automatic integration of information from visual and auditory sources whenever they are present. The resulting percept provides richer information than each of the sensory modalities in isolation (Navarra & Soto-Faraco, 2007). Nonetheless, the literature suggests that the benefit of visual articulatory gestures also depends on multiple factors that are related to the characteristics of both sensory input and the listener's perceptual system, such as speech intelligibility, acoustic and visual salience, cross- and within-talker variation, the distance between L1 and L2 phonemic categories, and the L1 background of listeners. The advantage of multimodal over unimodal input is not restricted to the situations where articulatory gestures are presented in synchrony with speech. Hirata and Kelly (2010) trained their native English speakers to perceive Japanese vowel length contrasts in both unimodal and multimodal contexts. They provided evidence suggesting that "this natural coupling [between speech and articulatory gestures] may create stronger perceptual traces of the phonemes (Calvert et al., 1997) which may make the speech sounds more salient and clear for later processing even in the absence of the visual information"(p. 305) (see also Hardison, 2003; Hazan, Sennema, Iba, & Faulkner, 2005, for similar observations).

Orthography has been shown to contribute to the processing of L2 speech sounds and to the learning of novel L2 words. There is growing evidence that its effects are shaped by a combination of factors, whose relative contributions are not yet fully understood. These factors may include the nature of the L1 and L2 grapheme-to-phoneme correspondences (GPCs), the difficulty of perceiving the L2 sound categories or contrasts, and the L1 and L2 writing systems. Some authors propose that the effect of orthography is mediated by the phonology, through building or reinforcing phonological categories or grapheme-phoneme mappings. For instance, Escudero, Hayes-Harb, and Mitterer (2008) argued that orthographic forms provide abstract knowledge that allows L2 perceivers to establish phonological representations, which can be used in novel word learning. In a study where L1 Spanish speakers learned novel words in L2 Dutch, Escudero, Simon, & Mulak (2014) found that orthographic forms with GPCs that were the same or similar across the two languages facilitated the learning of the association between an auditorily presented word and its pictured meaning, while those with GPC mismatches hindered learning. Ota, Hartsuiker, and Haywood (2010) found that L1 Japanese/L2 English speakers confused "near homophone" pairs of written English words containing vowels that do not contrast in their L1, but that are represented by different L2 graphemes (e.g., *fan* [fæn], *fun* [fʌn]). L1 Spanish/L2 English speakers, however, did not confuse such pairs. Although Spanish does not contrast the critical vowel phonemes (e.g., in *fan* [fæn], *fun* [fʌn]), the orthography of the language uses the critical graphemes to represent distinct vowels (<a> ~ /a/, <u> ~ /u/). Ota et al. attributed this pattern of results to differences in the writing systems. Spanish, like English, uses the Roman alphabet, and, in addition, has a shallow orthography. L1 Spanish orthography can therefore influence the L2 phonology and "block" the near homophony, while the non-alphabetic writing systems of Japanese cannot.

Unlike articulatory gestures, which mainly influence speech processing at the low-level perceptual stage (but see Llompart & Reinisch, 2017), the role of orthography seems to be most

prominent at later stages of L2 speech processing or when discrimination performance is explicitly assessed. Evidence for a late influence was provided by Han and Oh (2018) who investigated the joint impacts of L1-L2 phonetic similarity and orthography on the ability of native speakers of Korean to process Arabic phonemes at perceptual and post-perceptual levels. While phonetic similarity affected performance in a lexical decision task that required participants to distinguish new words containing an Arabic phoneme from their confusable minimal pairs, the impact of orthography was observed only in an offline task where they had to write the novel words. According to the authors, the visual cue provided by orthography may not contribute to the speech recognition process at the initial stage of L2 learning but may nevertheless allow L2 learners to build more accurate representations of novel words that can be retrieved in the absence of time pressure. Similarly, Eger, Mitterer and Reinisch (2019) reported that the benefit of orthography was indeed more obvious in tasks that required an explicit judgement of speech sounds (e.g., how well the words containing a confusable phoneme are pronounced) than when the perception of the phonemic contrast was assessed implicitly (e.g., in a visual-world eye-tracking paradigm where listeners spontaneously fixate on visual referents when presented with speech input). In the latter situation, it was observed that both non-native phonemes with and without written correspondences led to comparable outcomes.

The present study

The aim of the present study was to compare the contribution of two main sources of speech-related visual information, that is, articulatory gestures and orthography, to different stages of speech processing. Based on the literature reviewed above, it could be assumed that visual articulatory gestures and orthography affect speech at different processing levels and through different mechanisms. On the one hand, articulatory gestures provide natural and concrete visemic information that allows distinguishing ambiguous phonemic contrasts through

positions and movements of articulators occurring during speech production (Fisher, 1968). The role of speech production gestures on speech processing has mainly been explained within the framework of the Motor Theory of Speech Perception considering that, already at the earliest perceptual stage, processing speech implies an auditory-to-articulatory mapping process (Browman & Goldstein, 1990; Fowler, 1986; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985, but see Hickok, Holt & Lotto, 2009 and Lotto, Hickok & Holt, 2009). By contrast, orthography provides artificial and abstract graphemic information that allows the participants to construct representations of distinct phonemic categories even from non-distinct speech signals. Its relationship with speech has mainly been accounted for by the connectionist models assuming communication between different forms of language knowledge in both bottom-up and top-down directions, thus, allowing the abstract orthographic knowledge to "reshape" speech representations at different processing levels (Harm & Seidenberg 1999, 2004).

Based on current theoretical assumptions, these two types of audio-visual association have been considered as two distinct language phenomena and examined in separate studies using different experimental protocols. Although this methodological approach is reasonable given the current theories, it allows us neither to directly contrast the contributions of the two types of visual information to different stages of speech processing, nor to explore the possibility that they may share some characteristics. A stronger test of their dissociation requires an alternative methodological approach in which both types of audio-visual association are examined within a single protocol, using the same spoken material. The present study aims to fill this gap in the literature by adopting such an approach.

To this aim, we examined how the two types of visual information contribute to the acquisition of an L2 phonemic contrast using a novel word learning task. Our learning protocol focused on the English voiceless dental fricative /θ/ (e.g., *th*ought), a phoneme not present in

the French consonant inventory and labio-dental fricative /f/ (e.g., *fought*). These two fricatives known to be both acoustically similar and auditorily confusable, but are produced with distinct articulatory gestures. The acoustic similarity between /f/-/θ/ and /v/-/ð/ is well established. Ladefoged and Maddieson (1996) went so far as to claim that it was "profitless to try to characterize the acoustic spectra" (p. 173) of these fricatives. Jongman, Wayland, and Wong (2000) did find evidence of spectral and amplitude differences differences between /f/-/θ/ and /v/-/ð/, but their discriminant analysis using acoustic measures as predictors showed a low classification rate for these non-sibilant fricatives (66% vs. 88% for sibilants). Harris (1958) proposed that the identity of labiodental and dental fricatives was cued not by properties of the fricative noise itself, but rather by vowel formant transitions. Support for this claim, however, has been mixed. For example, Jongman et al. (2000) found no evidence of a role for formant transitions, and McGuire and Babel (2012) found that the informativeness of formant transitions depended contextual factors such as vowel identity. Wagner, Ernestus, and Cutler (2006) reported that the usefulness of formant transition information may depend on the fricative inventory of the native language. The perceptual confusability of /f/-/θ/ and /v/-/ð/ in auditory presentation is also well known (Balise & Diehl, 1994; McGuire & Babel, 2012; Miller & Nicely, 1955), even for native speakers of a language for which the contrast is phonemic. The contribution of visual articulatory information to the perception of the labio-dental fricatives /f/, /v/ (articulated with the upper teeth against the lower lip) and the dental fricatives /θ/, /ð/ (articulated with the tongue tip between the teeth or against the upper teeth) was proposed as early as Miller & Nicely (1955) and has been established both for hearing-impaired individuals (Walden, Prosek, Montgomery, Scherr, & Jones, 1977) and for individuals with normal hearing (Jongman, Wang, & Kim, 2003; McGuire & Babel, 2012).

The present study relied on our previous observation that for native listeners of European French the voiceless English dental fricative /θ/, absent in the inventory of their native

language, is frequently perceptually assimilated to /f/ (Tyler et al., 2019; see also Brannen, 2002, as well as McGuire & Babel, 2012 for a review of similar phenomena across languages). Based on this finding, we used a novel word learning paradigm in which this population was required to learn four minimal pairs of English pseudowords associated with eight unknown objects (e.g., Antoniou, Liang, Ettlinger, & Wong, 2015). The novel words within each minimal pair differed in their initial consonant, /θ/ or /f/. During learning, two minimal-pair novel words were systematically associated with two unknown objects. This ensured that the two critical phonemes were always presented in distinct lexical contexts that indicated the presence of a phonemic contrast (Stephens & Holt, 2010; Thiessen, 2007). The key point of the training was to examine whether this initial learning situation would benefit from the presence of two speech-related visual cues, articulatory gestures and orthography, that further emphasized the dissociation between the two phonemes.[1] This manipulation resulted in an assignment of participants to three training methods. In the "auditory training" group (*Aud*), participants were exposed to only the auditory form of the novel words during the training phase. In the "auditory-articulatory training" group (*AudArtic*), the auditory form was presented along with speaker's articulatory gestures. Finally, in the "auditory-orthography training" group (*AudOrtho*), the auditory form was associated with its orthographic form.

The efficiency of the three training methods was assessed both during and after training. At the post-training phase, participants' performance was assessed both on novel word acquisition and on their perceptual ability. To this end, an AX discrimination task was conducted to examine participants' ability to discriminate /θ/ and /f/ from each other. This low-level task allowed us to examine how the input modalities provided during the different training conditions influenced participants' perceptual ability, despite the fact that the training tasks

[1] Eger, Mitterer & Reinisch (2019) found an influence of the explicit vs. implicit nature of the task on performance in L2 speech sound learning. Although we did not manipulate this dimension, our participants' awareness that the /θ/ vs. /f/ distinction was being investigated pushes the tasks of the current study toward the explicit end of the implicit-explicit continuum.

exclusively focused on learning novel lexical items. The performance obtained in this low-level perceptual task was assessed along with that obtained in the picture-word matching task, which reflected high-level abilities in recognizing and memorizing the newly acquired spoken words. Given the nature of the relationship between speech and each type of visual cue, we hypothesized that the articulatory gestures, which are part of the sensorimotor component of speech, would have a stronger impact on the perceptual stage of speech processing. On the other hand, abstract orthographic representations should play a more crucial role at higher processing levels.

Finally, in order to examine whether the impacts of the two types of visual information changed over time, participants' performances in both AX discrimination and picture-word matching tasks were assessed in two consecutive days, that is, immediately after the training phase and after a night's sleep. The literature on the impact of overnight sleep on the consolidation of newly acquired spoken words suggests that word acquisition occurs in two stages (Dumay & Gaskell, 2007; Henderson et al., 2012; Tamminen et al., 2010; see also Davis & Gaskell, 2009 and Palma & Titone, 2020 for reviews). The first stage is a rapid, initial acquisition that enables learners to obtain good recognition performance soon after having been exposed to novel information. The second stage corresponds to a slower learning process achieved by offline consolidation of previously acquired information. This process allows the integration of a novel word into the mental lexicon, reflecting a transformation of episodic memory traces into lexicalized representations. Here, we hypothesized that the initial, episodic representations of the speech signal would become more abstract, that is, contaminated by the associated knowledge or representations (orthography and the articulatory gestures) after the consolidation period. Thus, comparing the performance obtained immediately after training and after a night's sleep for each training method should allow us to better understand the role of auditory, articulatory gestures, and orthography at the different stages of word learning.

METHODS

Participants

The participants were 101 native speakers of French, all university students (70 women, 93 right-handers, M_{age} = 23 years, *SD* = 3.4, range: 19–47). Roughly half of the participants had grown up in the south of France, and the others in a variety of other regions of France. All participants had studied English and at least one other language at school (as required by the French national school curriculum), typically Spanish or German. The mean age of the start of English language study was 9 years (*SD* = 2.2, range 3-13 years). The vast majority of participants (94/101) started learning English as L2 after the onset of reading acquisition. All participants had previously participated in another experiment, at least 12 months prior, where they categorized English consonants into French (L1) and English (L2) phonological categories and rated their goodness of fit, and completed six AXB discrimination tasks.[2] The results of the previous experiment were used to assign participants to one of the three training groups, *Aud* (*n* = 34), *AudOrtho* (*n* = 33), or *AudArtic* (*n* = 34), in such a way that they were matched on their categorization of /θ/ and /f/.[3] This was done to minimise any differences between the groups on perception of /θ/ and /f/ prior to training. This study was carried out in accordance with French law, with written informed consent being obtained from all the participants in accordance with the Declaration of Helsinki. An ethics approval of a national French "Comité

[2] All 151 participants from the previous study were invited to participate in the present study, and 101 accepted. Results for the L1 categorisation are presented in Tyler et al. (2019).

[3] For readers familiar with the Perceptual Assimilation Model (PAM; Best, 1995; Best & Tyler, 2007), the groups were matched on the basis of a composite assimilation type for /θ/-/f/ from the L1 and L2 categorization results (following Faris, 2017). For each participant, the L1 or the L2 assimilation type was chosen according to which one was predicted by PAM to yield the most accurate discrimination (two category [TC] > uncategorized-categorized [UC] > category goodness [CG] > single category [SC] > uncategorized-uncategorized [UU]). For example, if /θ/-/f/ was CG based on the L1 task and TC based on the L2 task, then TC was chosen for that participant. Participants were allocated to groups in this study ensuring that there was an even spread of assimilation types across the three groups (*Aud*: 8 TC, 16 UC, 3 CG, 7 SC; *AudOrth*: 9 TC, 15 UC, 4 CG, 5 SC; *AudArtic*: 9 TC, 15 UC, 4 CG, 6 SC).

de Protection des Personnes" was not required for the current study in accordance with the applicable institutional and national guidelines and regulations (Jardé law n°2012-300).

Stimuli

As described in detail below, two types of language materials were generated. The first one was used in the AX discrimination task. The second one was used in the training phase and the post-training picture-word matching task.

AX discrimination task. Four tokens of the syllables /θa/ and /fa/ were taken from the stimulus set developed for Tyler et al. (2019). They were produced by a phonetically trained female native speaker of Australian English. As illustrated in Figure 1, this speaker consistently produced /θ/ with an interdental articulation and /f/ with the upper teeth against the lower lip. The stimuli were recorded at the Centre d'Expérimentation sur la Parole, Laboratoire Parole et Langage, Aix Marseille University, CNRS, in Aix-en-Provence, France, in a sound-attenuated booth using a Beyerdynamic TG H55c microphone and an Edirol UA-25EX USB audio capture device. The audio was recorded using Audacity at a 44.1 kHz sampling rate with 16-bit resolution. The session recording was high-pass filtered at 70 Hz to remove low-frequency rumble and to correct for the DC component. The best four tokens of /θa/ and /fa/ were selected on the basis of auditory and visual inspection of the waveform and spectrogram of all the tokens initially recorded. The tokens were selected on that basis, rather than on the basis of expected acoustic characteristics, to allow natural acoustic variability among tokens. The duration of the vowel was truncated to 80 ms, and a 5 ms ramp was applied to minimize the influence of the vowel on categorization (Guion, Flege, Akahane-Yamada, & Pruitt, 2000).

Novel word learning paradigm and picture-word matching task: Four minimal pairs of monosyllabic English pseudowords were constructed: *fint* /fɪnt/, *thint* /θɪnt/; *fedge* /fedʒ/, *thedge*

/θedʒ/; *felk* /felk/, *thelk* /θelk/; and *fald* /fɐːld/, *thald* /θɐːld/.[4] The members of each pair differed only in their onset consonant: /f/ or /θ/. The other consonants and the vowels all had close counterparts in French.

The utterances were recorded by the same female speaker in a sound-attenuated booth at the MARCS Institute for Brain, Behaviour and Development at Western Sydney University, Australia. The speaker sat in front of a white backdrop, illuminated with two studio lights covered in diffusion paper (Studio-Lite Photon Beard Highlight 110). Her head and shoulders were recorded using a Sony HXR-NX30P video camera recorder (1080p resolution, 25 frames/s, H.264 codec). The audio was simultaneously recorded using a Sony ECM-XM1 shotgun microphone mounted on the camera at a 44.1 kHz sampling rate via a MOTU ultra-lite MK3 sound card connected to Adobe Audition on a Windows PC. At the beginning of the video, a hand clap was used for subsequent synchronisation of the audio and video tracks.

A laptop computer was used to display the orthographic form of each of the pseudowords, with presentation controlled by Psyscope X B77 (http://psy.ck.sissa.it/). Before beginning, the speaker read each item aloud from a sheet of paper to ensure that she knew how each item should be pronounced. The items were then recorded one-by-one, with a variety of other items not used in the current experiment, in five randomized blocks.

After recording, the audio file was high-pass filtered at 100 Hz using Praat software (Boersma & Weenink, 2019) to remove any low-frequency rumble and to correct for the DC component. The four best tokens of each item were selected on the basis of auditory inspection and visual analysis of the waveform and spectrogram. A 5-ms fade-in and fade-out was applied to the beginning and end of each utterance.

[4] When presented to the Australian speaker, this minimal pair was spelled *farled/tharled*, in order to elicit an /ɐː/ vowel similar to French /a/, in line with the grapheme-to-phoneme correspondences (GPCs) of this variety of English. The other minimal pairs also respect the GPCs of Australian English. When presented to the French participants of the *AudOrtho* group, however, this pair was spelled *fald/thald*, in line with the GPCs of French.

It was necessary for the video stimulus to begin before the onset of the audio to ensure that each item began with the speaker's mouth in a neutral position. To eliminate the possibility that any non-speech auditory information could influence recognition prior to the onset of speech, 1 s before and after each utterance was converted to silence. The audio files were segmented using a Praat script at 800 ms before the onset and 200 ms after the offset.

The high-pass filtered audio recording was imported into Adobe Premiere Pro CC 2018 on MacOS 10.14 and aligned to the audio track from the video. The video camera audio was then removed. The time codes for each utterance, obtained from Praat, were converted to minutes, seconds, and audio samples from 1 to 44100 cycles per second. Those values were used to precisely locate the onset and offset of each video in Adobe Premiere Pro CC 2018. Each video was cropped as illustrated in Figure 1, and then saved in Quicktime format (720 × 576 resolution, 25 frames/s), ensuring that the raw audio format (PCM wave) was preserved.

Figure 1: Screen shots of two of the video clips during the pronunciation of a novel word starting with the phonemes /θ/ (left) and /f/ (right).

The eight novel words were paired with black and white images of eight invented objects selected from the Horst and Hout (2016) image database.[5]

[5] The identities of the selected objects were: #2004 (/fɪnt/), #2029 (/θɪnt/), #2015 (/fɐːld/), #2024 (/θɐːld/), #2025 /fedʒ/, #2057 (/θedʒ/), #2039 (/felk/), #2063 (/θelk/).

Procedure

Participants were tested individually in a quiet room. Auditory stimuli were presented through headphones. Visual materials (orthographic form of novel words, video presenting articulatory gestures, images of unknown objects) were presented on a laptop computer screen. Stimuli presentation and data collection were controlled by E-Prime 3.0 software (Psychology Software Tools, Pittsburgh, PA).

As illustrated in Figure 2, the experiment was conducted in two sessions that took place on two consecutive days; thus, there was, for each participant, one night's sleep between the two sessions.[6] The first session (Session 1) consisted of three phases: 1) Pre-training AX discrimination allowing an evaluation of participants' pre-training ability to discriminate /θ/ and /f/, 2) Novel word training phase during which the participants learned to associate the four minimal pairs of novel words with the eight objects using one of the three training methods, and 3) Post-test AX discrimination and picture-word matching tasks which allowed us to (re)evaluate participants' ability to discriminate /θ/ and /f/ and to successfully learn the novel words. During the second session (Session 2), the participants performed the same post-test AX discrimination and picture-word matching tasks as in the last phase of the first session, with no additional training. A detailed description of each task is presented below.[7]

[6] The average number of hours between the first and the second experimental session was comparable across the three training groups [$p > .05$; 22.5hrs (*SD* = 4.1), 23.5hrs (*SD* = 4.1) and 24hrs (*SD* = 3.7) for the *Aud, AudOrtho* and *AudArtic* training, respectively). It was also the case for the average number of hours between training and bedtime (using midnight as reference: $p > .05$; 8.5hrs (*SD* = 3.1), 9hrs (*SD* = 3) and 9.5 hrs (*SD* = 3.2) for the *Aud, AudOrtho* and *AudArtic* training, respectively). Statistical analyses considering each of these values as covariable led the same result pattern as in the analyses presented in the Results section. No significant effect of these factors or of their interactions with the other variables was observed.

[7] Participants in all groups also completed a production task of real English words (based on a French translation of the English word – *soleil* for *sun* – plus a picture for picturable words) after the pre-training AX discrimination in Session 1, and a production task of real English words and the novel words (based on images of unknown objects) after the training session. No training on speech production and no feedback on participants' productions was provided during the experiment. According to the literature, without intensive training with feedback, the impact of production on perception remains controversial (Kartushina et al., 2015; Leach & Samuel, 2007; Llompart & Reinisch, 2017). As the focus of this paper is on perception, the production data will not be discussed here.

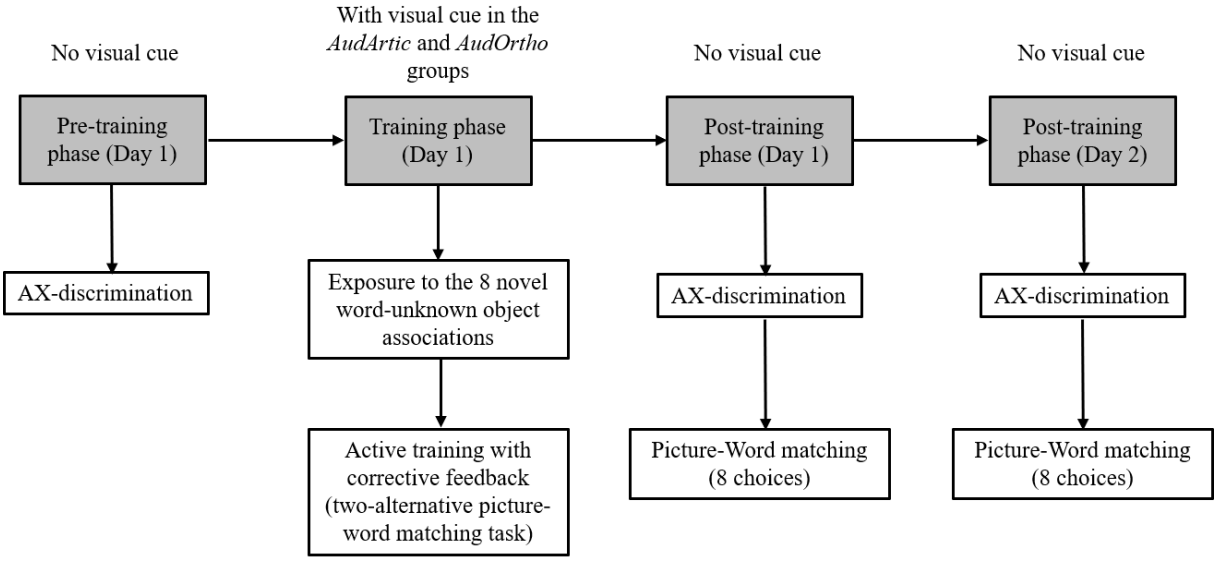

Figure 2: Summary of the experimental protocol.

AX discrimination: The /θa/ and/or /fa/ syllables were presented consecutively on each trial. Each of the resulting four trial types (/θa/-/θa/, /θa/-/fa/, /fa/-/θa/ and /fa/-/fa/) was repeated 12 times and, among these trials, each of the four tokens was repeated three times. In the "same" trials, where the same syllable was repeated, two different tokens were always used. The 48 trials were presented in one experimental block in a random order. In each trial, participants had to indicate, by pressing '1' or '5' on the computer keyboard, whether the consonant of the second syllable belonged to the same category as the consonant of the first syllable or to a different category. A concrete example using L1 sounds was provided (i.e., the initial consonants of /ta/-/ta/ belong to the same category while the initial consonants of /da/-/ta/ belong to different categories). There was a 1-s interval between the two stimuli of a pair and participants had to respond within 2 s of the offset of the second token. If participants failed to respond within this time frame, they were prompted by a warning message on the screen to respond faster and the trial was repeated later in the task. Otherwise, once a response had been registered for a trial, the following trial began after 1 s. A fixation cross was presented on the

centre of the screen during this inter-trial interval. To familiarise participants with the task, 10 practice trials using non-critical syllables were presented before the actual experiment. No feedback was provided at any point in the task. The task lasted about 5 min.

Novel word training paradigm: The training paradigm contained two parts. During *initial exposure*, participants were presented with the association between the eight novel words and the eight objects. They were explicitly instructed to memorize the associations between the novel words and the objects. In the audio-visual training groups, they were instructed to pay attention to both the pronunciation of each word and the associated visual input. Each novel word-object pair was presented 24 times, corresponding to six repetitions of the four tokens of each novel word. The 192 trials were presented in a random order. Each trial began with a fixation cross presented at the centre of the screen for 500 ms. For the *Aud training* group, this was followed by 800 ms of blank screen. The auditory version of a novel word was then presented over headphones. The screen remained blank during the presentation of the auditory stimulus and for 200 ms afterwards. For the *AudOrtho* training group, the orthographic form of the novel word was presented at the centre of the screen at the same time and for the same duration as the auditory form. For the *AudArtic training* group, the video file of the speaker producing the novel word was presented immediately after the fixation cross. The video started 800 ms before the onset of the sound and ended 200 ms after its offset. After the presentation of the novel word, the associated object was presented at the centre of the screen for 2 s. A 500-ms blank screen separated the offset of the object and the onset of the next trial. The exposure phase lasted about 15 minutes.

Following the initial exposure phase, participants completed *active training with corrective feedback* through a two-alternative picture-word matching task. In this second phase of training, each of the eight novel words were again presented 24 times in a random order and in the modality specific to each training method (*Aud, AudOrtho, AudArtic*). After the word

was presented, both the correct object and the object corresponding to the other member of the novel word's minimal pair were presented side-by-side on the screen for 3 s. The position of the objects on the screen was counterbalanced across trials. During this interval, participants were required to click on the correct object. Once a response had been registered, or when 3 s had elapsed, the screen turned blank for 1 s. Trials without a response were not repeated. After each trial, a feedback message ("correct", "incorrect" or "please respond faster", in French) and the correct object were presented simultaneously on the screen for 2 s, even on trials for which participants gave no response. During that time, the novel word was also presented in the modality specific to each training method. The task lasted about 20 min.

Picture-word matching task: The aim of this task was to evaluate participants' ability to learn the novel words, which required them to correctly identify the minimal-pair pseudoword and to match each word with the correct object. The eight novel words were repeated eight times (twice per token), for a total of 64 trials, presented in a random order. The picture-word matching task completed after training differed from the one during active training in two aspects: 1) only the auditory version of the words was provided without any visual cue and 2) on each trial, participants had to match the auditory word with one of the eight (rather than two) objects presented on the screen. The eight objects were presented in a horizontal 4 × 2 grid. The object position varied randomly from one trial to another. Participants had 10 s to click on the object that corresponded to the auditory word. Once a response was registered, or when the 10 s had elapsed, the screen turned blank for 1.5 s. No feedback was provided and trials without a response were not repeated. The task lasted approximately 6 min.

RESULTS

Pre-training AX discrimination

To ensure that participants in the three groups were matched on their initial ability to perceive /θ/, a preliminary analysis was performed on the results obtained in the pre-training

AX discrimination task. Figure 3 shows the response accuracy (raw score) obtained in each training group at the different phases of the protocol. Rather than considering in the analysis the response accuracy of each individual trial, we followed the method recommended by Snodgrass et al. (1985) of converting the discrimination accuracy scores to A-prime (A') values, a non-parametric index of sensitivity. For each participant and each phase of the protocol, the A' value was computed based on the proportions of "hits" ("H", i.e., the participant responded "different" when the two syllables of a pair did not share the same initial phoneme) and "false alarms" ("FA", i.e., the participant responded "different" when the two syllables of a pair shared the same initial phoneme) according to the following formulas:

If H = FA, then A' = 0.5.

If H > FA, then A' = 0.5+((H−FA)×(1+H−FA))/((4×H)×(1−FA)).

If FA > H, then A' = 0.5−((FA−H)×(1+FA−H))/((4×FA)×(1−H)).

This computation provides a more sensitive measure of participants' discrimination ability than accuracy because it corrects for the false alarm rate. An A' score of 1 indicates perfect discrimination sensitivity, whereas an A' score of .5 indicates a lack of sensitivity (i.e., responding at chance). Although there was some within-group variability, the mean A' scores obtained in the *Aud*, M = .83, SD = .10, *AudOrtho*, M = .82, SD = 0.13, and *AudArtic*, M = .82, SD = .09, training groups were equivalent, as confirmed by the results of pairwise comparisons, $t(64) = 0.45$, $p = .65$; $t(65) = 0.43$, $p = .67$; $t(64) = 0.04$, $p = .97$ for the *Aud vs. AudOrtho*, *Aud vs. AudArtic, and AudOrtho* vs. *AudArtic* comparisons, respectively. Thus, before training, the three groups of participants did not differ significantly in their levels of sensitivity to the /θ/-/f/ contrast.

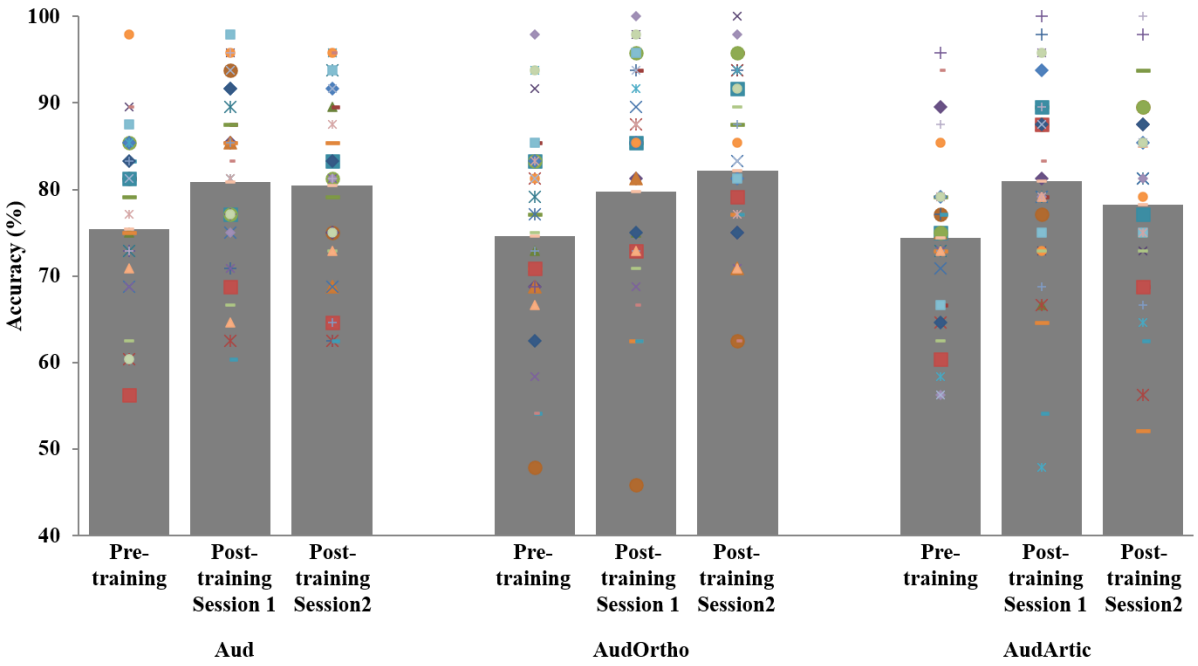

Figure 3: Percentage of accuracy obtained in the AX discrimination task at the different phases of the protocol. The markers represent individual data.

Performance during the active training with corrective feedback task

During active training with corrective feedback, participants performed a two-alternative picture-word matching task to match each novel word with one of the two objects, which corresponded to the novel word and the other member of the minimal pair. For analysis of this task, performance on each individual trial was considered. The raw accuracy scores were analyzed with R software (R Core Team, 2017), using a generalized linear mixed-effects model (glmer) with a binomial link function. Training method (*Aud*, *AudOrtho*, and *AudArtic*) was treated as a fixed factor and participants and items as random intercepts (Baayen, Davidson, & Bates, 2008). The analysis showed that, although the same amount of training was provided in the three groups, the accuracy scores were significantly lower in the group of participants who were exposed to the auditory input alone (71%) compared to those who were concurrently exposed to auditory and visual input (*AudOrtho*: 97.6%, Estimate = 2.88, *SE* = 0.26, $z = 11.25$, $p < .001$; *AudArtic*: 93.4%, Estimate = 2.12, *SE* = 0.27, $z = 7.78$, $p < .001$). The advantage of

the *AudOrtho* training over the *AudArtic* training was also statistically significant (Estimate = -0.76, *SE* = 0.29, *z* = -2.64, *p* = .008).

Impact of training on the perceptual ability

To examine the impact of training on AX discrimination, we computed, for each participant and for each of the two post-training sessions, the percentage of change in A' values obtained in the AX discrimination task conducted after training compared to the values measured before training.

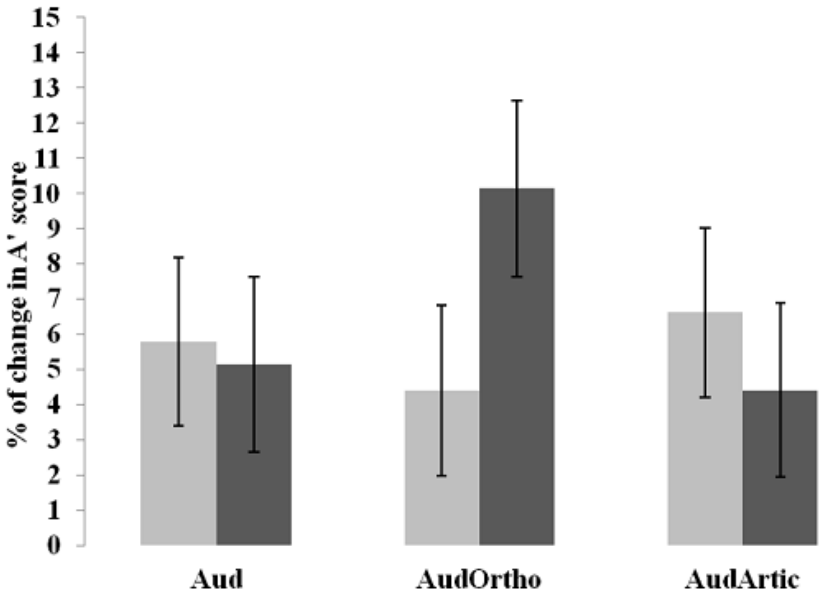

Figure 4; Percentage of change in A' score obtained immediately after training (Session 1, light bars) and after the consolidation period (Session 2, dark bars) compared to the pre-training baseline across the three training conditions. A value of 0% indicates no change in A' score between the pre- and post-training sessions. Error bars represent standard error of the mean.

As illustrated in Figure 4, there was an overall improvement in discrimination of /θ/-/f/ compared to the individual baseline level measured before training. One-sample *t*-tests (compared against 0) performed on the improvement rates measured immediately after training

showed significant improvement in all groups: $t(33) = 2.40$, $p = .022$; $t(32) = 2.53$, $p = .017$; $t(33) = 2.28$, $p = .028$ for the *Aud, AudOrtho* and *AudArtic* group, respectively. To examine the benefits of the different training methods, the impact of overnight consolidation on the increase of discrimination sensitivity and their interaction, a mixed ANOVA was conducted on the percentages of change in A' score, with session (Session 1, Session 2) as a within-participant factor and training method (*Aud, AudOrtho, AudArtic*) as a between-participants factor. There was no main effect of session, $F(1, 98) = 0.50$, $p = .48$, $\eta_p^2 = .005$ or training method, $F(2, 98) = 0.23$, $p = .798$, $\eta_p^2 = .004$. However, we observed a significant interaction between the two factors, $F(2, 98) = 3.18$, $p = .046$, $\eta_p^2 = .061$. This interaction remained significant even when participants' initial discrimination performance measured at the pre-training stage was considered as covariate, $F(2, 97) = 3.21$, $p = .045$, $\eta_p^2 = .062$. Further investigation of the interaction did not reveal a significant difference between the percentages of change in A' score obtained in the three training methods in either session (all $Fs < 1$ for all pair-wise comparisons except for those between "*AudOrtho* vs. *Aud*" and "*AudOrtho* vs. *AudArtic*" in session 2 where $F(1,98) = 2.004$, $p = .16$ and $F(1,98) = 2.64$, $p = .10$, respectively). However, the impact of session clearly depended on the training method. In comparison with the performance obtained in Session 1, the one night's sleep, which corresponded to the consolidation period, led to improvement in the *AudOrtho* group, $F(1, 98) = 5.84$, $p = .017$, but not in the *Aud*, $F(1, 98) = 0.08$, $p = .782$, or *AudArtic* group $F(1, 98) = 0.09$, $p = .348$. This result suggests that being exposed to the orthography of the novel words in addition to their auditory forms not only led to an immediate performance improvement compared to the baseline level but also induced overnight improvement.

Based on the literature, the absence of the additional benefit of the articulatory training on perceptual ability, compared to the auditory alone training, is rather surprising. One factor that could contribute to this observation is the fact that some participants already had a relatively

good level of perceptual ability even before training, and therefore had a limited margin for improvement. Although this factor could not explain the main interaction between the improvement rate in the two sessions and the training method, we further examined whether our initial result pattern was still observed when the data from participants who showed near-ceiling perceptual ability were removed. To this aim, we conducted an additional analysis in which we excluded the data from participants with AX pre-training A' scores greater than or equal to .90. This criterion was selected since it left us with a greater margin for improvement and guaranteed that none of the participants had a maximum score even after training. It also left us with a reasonable sample size to conduct a meaningful analysis (n = 24, 23, and 28 in the *Aud*, *AudOrtho*, and *AudArtic* group respectively, i.e., 75 of 101 participants).

This new analysis replicated the initial finding and confirmed the outcome of the analysis considering the pre-training A' scores as a covariate. We observed no significant main effect of session, $F(1, 72) = 1.02$, $p = .32$, $\eta_p^2 = .01$, or of training method, $F(2, 72) = 0.19$, $p = .83$, $\eta_p^2 = .005$. Again, there was a significant interaction between the two factors, $F(2, 72) = 4.41$, $p = .016$, $\eta_p^2 = .109$. Without the participants who might have induced a ceiling effect due to their high perceptual ability, further investigation of the interaction showed the same result pattern as in the initial analysis: In comparison with the performance obtained in Session 1, the consolidation period led to improvement only in the *AudOrtho* group, $F(1, 72) = 8.30$, $p = .005$. No hint of improvement was found either in the *Aud*, $F(1, 72) = 0.06$, $p = .805$, or in the *AudArtic* group $F(1, 72) = 1.11$, $p = .294$.

Impact of training on novel word learning

In these analyses, we examined the impact of the three training methods on novel word learning, which involved participants' ability to correctly identify the minimal pairs of words and to match the words that they recognized with the correct objects. The raw accuracy scores

obtained in the picture-word matching task conducted after training were analysed in two steps. In the first step, one sample *t*-tests were applied to examine the efficiency of each training method, by comparing the obtained score to 12.5%, which reflected the chance level of choosing the correct object among the eight objects. In the second step, the relative benefits of the different training methods and the impact of overnight consolidation were investigated by applying a generalized linear mixed-effects model (glmer) with a binomial link function on the raw scores obtained in the individual trials. Training method (*Aud, AudOrtho, and AudArtic*), session (Session 1, Session 2) and their interaction were treated as fixed factors, and both participants and items as random intercepts (Baayen, Davidson, & Bates, 2008). Note that a more complex model including participants' initial discrimination performance measured at the pre-training stage as a covariate was also conducted. Since both models led to the same conclusion and adding an additional variable did not improve the model fit ($p > .05$) only the results of the original design are presented here.

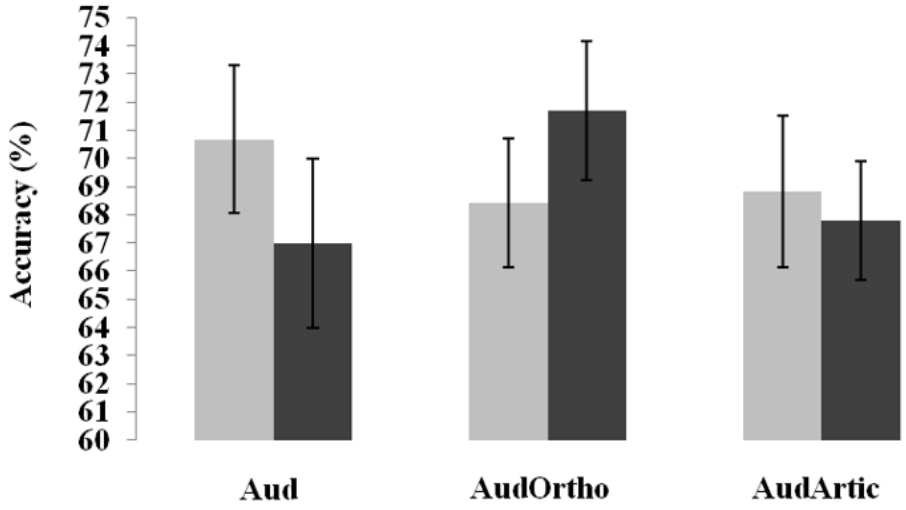

Figure 5: Percentage of accuracy obtained in the picture-word matching task immediately after training (Session 1, light bars) and after the consolidation period (Session 2, dark bars) across the three training conditions. Error bars represent standard error of the mean.

The percent accuracy obtained in the picture-word matching task are shown in Figure 5. One sample *t*-tests conducted on these scores showed that immediately after training the participants in the three groups were able to dissociate the minimal pair words and associate them to the objects well above chance level, $t(33) = 11.17$, $p < .001$; $t(32) = 24.37$, $p < .001$; $t(33) = 20.96$, $p < .001$ for the *Aud, AudOrtho* and *AudArtic* group, respectively. Although the glmer did not reveal an overall significant difference between the accuracy scores obtained in the three training methods in either session, the impact of session clearly depended on the training method. Specifically, while the benefit of the *Aud* training significantly dropped from Session 1 to Session 2, estimate = -0.197, $SE = 0.07$, $z = -2.81$, $p = .005$, being exposed to the articulatory gestures in addition to the auditory input during training resulted in a maintenance of post-training performance across the two sessions, estimate = -0.05, $SE = 0.07$, $z = -0.79$, $p = .429$. Most importantly, and in line with the finding obtained in the AX discrimination task, the presence of orthography during training was the most beneficial to the consolidation process since it led to a significant increase in learning performance after an overnight sleep even in the absence of additional training, estimate = 0.173, $SE = 0.07$, $z = 2.44$, $p = 0.01$.

DISCUSSION

Our previous study showed that European French speakers have difficulty with the English dental fricative /θ/ and frequently assimilate it to /f/ in perception (Tyler et al., 2019). Here, we built on this observation and examined whether and how providing articulatory gestures or orthography as visual cues in addition to auditory input during the training phase could improve the ability of French speakers to perceive the /θ/-/f/ contrast and to learn novel words containing these two phonemes. To this end, participants from the Tyler et al. study were recruited and divided into three different training groups. In the *Aud* group, the participants were exposed to auditory input alone during the training. In the *AudArtic* group, the auditory

input was associated with the articulatory gestures. In the *AudOrtho* group, the auditory input was associated with the orthographic form. The post-training benefit of the two types of visual cues presented during the training phase was assessed in both low-level perceptual and high-level novel word learning tasks during which only the auditory input was provided (see Figure 2).

An analysis of perceptual ability before training showed that our allocation of participants to groups on the basis of their previous categorization results was effective. The participants of the three groups were statistically matched on their ability to discriminate /θ/ and /f/. During the training phase, all participants received the same amount of exposure and active training. However, the level of performance measured during the active training phase, using a two-alternative picture-word matching task, varied significantly across training methods. The two groups of participants who were exposed to both auditory and visual inputs outperformed those who were exposed to auditory input alone. This observation reflected a clear advantage of multiple input modalities and indicated that both orthography and visual articulatory gestures provided valid and useful visual cues for learning two words forming a minimal pair. This observation replicated what has been reported in the literature regarding the contribution of visual cues during speech processing (Frost, Repp, & Katz, 1988; Grant & Seitz, 1998; Grant, Walden, & Seitz, 1998; Navarra & Soto-Faraco, 2007; Wang et al., 2008). Interestingly, at least in this specific learning situation, the abstract visual cue provided by orthography showed a stronger benefit than the natural visual cue provided by articulatory gestures. The weaker contribution of articulatory gestures to the identification of minimal-pair words could be due to the fact that the visual cue that they provide are transient and, for L2 listeners, they probably provide more ambiguous information about the phonemes contained in the novel words than that provided by orthography. It is also possible that literate adult population is generally more sensitive to the written code than to articulatory gestures,

especially in L2 learning. However, it is premature to confirm whether the stronger advantage of orthographic training was specific to known graphemes representing specific phonemes or could be obtained with any visual cue that provides unambiguous information about a phoneme's identity. More extensive research is needed to explore the features that are inherent to the two forms of visual cue (for instance, the using movements vs. abstract symbols to convey information), whether these features are language specific, and how they contribute to spoken language processing. Nevertheless, as discussed below, the performance obtained immediately after training and after a one-night interval does not merely reflect the initial differences between the three groups in their ability to attend to the additional context provided by the two visuals cues during the training phase.

Impact of training on AX discrimination

The training methods used in the present study aimed at teaching participants novel lexical items that were associated with unknown objects. Since the lexical items were minimal pairs, successful learning relied on participants' ability to identify the critical phonemes. The findings obtained in the AX discrimination task conducted after training showed that, compared to the pre-training baseline, the participants in all training groups showed an immediate improvement even though they were assessed on untrained monosyllabic stimuli (/θa/-/fa/).

However, contrary to our expectations and to the pattern of performance obtained during training, we did not find any significant difference in the overall benefits of the three training methods. In particular, the link between speech sounds and articulatory gestures that had been rendered explicit during the *AudArtic* training did not facilitate the discrimination of the two phonemes beyond the *Aud* training alone. This observation may seem at odds with the claim that speech perception performance is generally improved when the articulatory gestures are provided. However, it could be explained by the fact that the benefits of the articulatory gestures

in previous studies have mostly been assessed by comparing speech processing performance obtained in the conditions where the auditory input was either presented alone or in synchrony with the articulatory gestures (Fenwick et al., 2017; Grant & Seitz, 1998; Navarra & Soto-Faraco, 2007; Wang et al., 2008). This is precisely the circumstance in which participants' performance was measured during the active training phase, and the additional benefit of the articulatory gestures (and of orthography) was indeed observed. The finding obtained during the post-training phase complements this observation. At least in the current paradigm, which used only one short training session with no specific training on phoneme perceptual discrimination, the benefit of the articulatory gestures did not persist beyond the training period, that is, when the visual cue was no longer present (see Vroomen, van Linden, de Gelder, & Bertelson, 2007 for a similar observation on perceptual learning and Samuel & Dumay, in press, for a review).

Our most intriguing observation is the fact that, although none of the visual cues showed an additional benefit immediately after training, being exposed to novel word spellings during the *AudOrtho* training led to a significant overnight improvement in perceptual ability, unlike the other two training methods. Existing findings suggest that the impact of the consolidation (that was assumed to take place after a night's sleep in the present study) on perceptual learning of non-native sounds might depend on several factors, such as task demands, variability of speech tokens and relationship between training and assessment tasks (Earle & Myers, 2014; Earle & Myers, 2015; Eisner & McQueen, 2006; Fenn, Margoliash, & Nusbaum, 2013; Fenn et al., 2003; Qin & Zhang, 2019). Here, we further suggest that the modality of language inputs provided during training also plays a significant role: The benefit of the abstract orthographic code that enabled learners to distinguish minimal pair words as two separate lexical entries seems to generalize to a lower processing stage and help in disambiguating the /θ/ and /f/ phoneme categories, at least those that are pronounced by the same talker. The fact that

overnight improvement was generalized across tasks and materials, that is, to a low-level AX discrimination task using untrained spoken items, suggests a particularly powerful impact of orthographic information on the consolidation of the critical phoneme categories. The mechanisms leading to such improvement in low-level perception following acquisition of lexical knowledge are discussed further below.

Impact of training on novel word acquisition

As in the AX discrimination task, all training methods were effective. Picture-word matching scores measured immediately after training were significantly higher than chance level in all groups. Clearly, the short training method used (with both simple exposure and an active two-alternative picture-word matching training task) allowed the participants to learn the eight new words and associate them with eight unknown objects. Participants successfully accomplished word learning, despite the possibility that the two-alternative picture-word matching task employed during the active training session could have been completed through a simple mapping between the initial phonemes and each of the two objects.

According to the Complementary Learning Systems framework proposed by McClelland et al. (1995) and specifically applied to spoken word learning by Davis and Gaskell (2009), this post-training performance would reflect the initial stage of acquisition that enables learners to obtain good recognition performance soon after having been exposed to novel information. Once again, in the absence of visual cues during the post-training test phase, the participants from all three groups showed a comparable level of performance. We therefore conclude that, at this early stage of novel word acquisition, neither the information from articulatory gestures nor that from orthography that had been provided during training led to an additional benefit compared to the auditory input alone.

A comparison of the learning performance obtained immediately after training and after a night's sleep provides further insight into how different input modalities contribute to the dynamics of novel spoken word learning. In the group of participants who were presented with the auditory input alone, the percentage of correct picture-word matching significantly dropped after a night's sleep, thus reflecting a decay of episodic memory representations over time or an interference from native or non-native language inputs that the participants might have been exposed to during the two experimental sessions (Earle & Myers, 2015; Fenn et al., 2003). On average, the intervals between the two sessions and between the training and bedtime is equivalent in the three groups and including these interval values in the analyses did not change the main result pattern (see Footnote 6). Therefore, the drop in novel word recognition performance in the *Aud* group was likely due to the characteristics of the language input that the participants received during training. A more rapid decay and stronger interference is expected when speech signals do not match any existing speech representations, or are degraded or ambiguous, as was the case for the L2 speech input used here (Wagner, Torgeson, Laughton, Simmons, & Rashotte, 1993). Indeed, we observed different patterns of results in the groups of participants who had also been exposed to the visual cues during training, in addition to the auditory speech input. These visual inputs allowed the participants to distinguish minimal pair words and, therefore, to establish more accurate and stable representations of the novel words. Additionally, once the problem of the initial phoneme discrimination had been resolved, participants had more attentional resources available to learn the word-object associations. In the *AudArtic* group, performance remained constant across the two post-training sessions. This stable outcome could be explained by the fact that, during the training, the articulatory gestures provided a visual cue that contributed to disambiguating words of the minimal pairs to some extent (as shown in the performance obtained during active training). These seemingly more accurate representations of the novel words were more resistant against memory decay or

interference, as compared to the *Aud* condition. Note that, the *absence of forgetting* (memory stabilization) was historically considered as a sign of memory consolidation (Müller & Pilzecker, 1990; Duncan, 1949). Finally, the most favorable learning outcome was again observed in the *AudOrtho* training method: We observed an *enhancement* of the initial memory representations of the novel word-object associations after one night's sleep, which suggests a stronger benefit of the orthographic visual cue to the consolidation of new knowledge. Using a training protocol in which unknown spoken words were presented either alone or with their orthography, Escudero et al. (2008) reported that Dutch participants who received the auditory training confused novel words containing English /æ/ (a phoneme not present in Dutch) and /ɛ/ symmetrically. By contrast, those who were also presented with the orthographic forms of the novel words showed an asymmetric confusion pattern, that is, while the /ɛ/ tokens were correctly perceived as /ɛ/, the /æ/ tokens were equally perceived as /æ/ and /ɛ/. This asymmetric pattern observed on novel words mirrors the pattern that Dutch speakers typically show when they perceive known English words that contains these phonemes (Weber & Cutler, 2004). The observation led Escudero et al. to conclude that the orthographic code that provided explicit information over the contrastive nature of two speech sounds contributed to building up separate lexical representations for similar-sounding L2 novel words. Our finding provides supporting evidence along these lines with the difference that, in the present learning protocol, the building up of the lexicalized representations did not take place immediately after learning but only after a night's sleep.

Online versus residual impact of articulatory gestures and orthography on speech processing

The present study aimed at comparing the contribution of articulatory gestures and orthography to different levels of speech processing. Overall, the findings obtained at the different phases of the protocol could be described in terms of the "online" or "residual" impact

of a visual cue, which refer respectively to situations where the visual cue was or was not concurrently present with auditory input.

Given the strong link between speech perception and production, articulatory gestures were expected to affect speech processing at a low-level, perceptual stage. Existing studies that examined their online influence in speech perception tasks have indeed provided a great deal of evidence supporting this claim (Grant & Seitz, 2000; McGurk & MacDonald, 1976; Schwartz et al., 2004; Sumby & Pollack, 1954). Although no such evaluation of the online influence on a perception task was conducted here, the results that we obtained during the active training phase further showed their contribution in a high-level spoken word learning task, which is in line with a controversial view that articulatory gestures might also play a significant role during recognition of lexical items (Fort, Spinelli, Savariaux, & Kandel, 2010, 2012).

Nevertheless, this robust online influence did not induce any residual effect on speech processing at the perceptual level and induced limited effect at the higher word recognition level (maintenance but no enhancement of word recognition performance), once the visual input was no longer available. This observation is somewhat surprising given the literature in infants' native language development consistently claiming a long term benefit of articulatory gestures on spoken language acquisition, and some existing findings from training studies in adults that showed a post-training benefit of audiovisual speech compared to auditory speech (Hardison, 2003; Hazan et al., 2005; Hirata & Kelly, 2010; Llompart & Reinisch, 2017). Several factors could explain this result. One might be the range of participants' age of starting to acquire L2 English: Almost all participants in the present study started learning English as L2 after the onset of reading acquisition. As a result, unlike in L1 acquisition, their L2 learning or practice might rely more on written than on spoken language and articulatory gestures. Also, L2 learners at different proficiency levels may use visual information differently in the process of learning novel words. Here, we strictly controlled for participants' perceptual ability, which is the key

factor for the present protocol. Future studies taking into account participants' general English proficiency as well as vocabulary size could contribute to further understanding of how L2 speech representations are acquired, processed, and stored in the mental lexicon. Finally, it is also possible that the absence of the residual impact of the *AudArtic* training was to some extent due to the nature of the present protocol: Using longer or multiple training sessions (e.g., Hirata and Kelly, 2010) or assessing participants' behavior in implicit speech processing tasks with more fine-grained measures (e.g., Llompart & Reinisch, 2017) could reveal the subtle contribution of the articulatory gestures.

With respect to the online influence of orthography, the presence of word spellings also facilitated identification of minimal-pair words and their acquisition. Most interestingly, and in contrast to articulatory gestures, the abstract visual information provided by orthography clearly induced a residual effect that was detectable both in the picture-word matching task, using trained stimuli, and in the perception task, using untrained stimuli. As mentioned above, the residual impact of orthography did not occur immediately after training but only after a night's sleep. This delay suggests that orthography did not play a significant role during the initial stage of spoken word acquisition where the words' episodic memories were formed. The pattern of result obtained at this early stage, which is first qualified as episodic or phonological learning by Gaskell and Dumay (2003; see also Dumay & Gaskell, 2007), and then as lexical configuration by Leach and Samuel (2007), is different from what happened after the consolidation period. After one night's sleep, the benefit of orthographic knowledge emerged. This information most likely contributed to a formation of well-defined, abstract and stable lexical representations that facilitated long term retention of the newly acquired spoken words (Davis & Gaskell, 2009; Palma & Titone, 2020). It is worth specifying that the benefit of overnight sleep on lexicalization process is consistently reported in the literature (e.g., Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Henderson et al., 2012; Tamminen et al., 2010) and

that the picture-word association learning paradigm, as the one used here, is argued be an effective tool to boost this process (Leach & Samuel, 2007). As we discuss below, in addition to the formation of more abstract speech representations, our finding also fits with the description proposed by Leach and Samuel (2007): "these new words could do what "real" words do – they can support perceptual learning in which the boundaries of phonetic categories get reshaped, with the lexical representations guiding the respecification of the sublexical units" (p. 13). As argued by the authors, this suggests the existence of lexical engagement. However, this conclusion would be strengthened by more direct evidence that the acquired knowledge is truly integrated in the existing lexico-semantic network. Such evidence could be obtained, for example, with an experimental design that allows examining the link between the novel and the existing words and whether learning novel words interferes with the recognition of existing words (Gaskell & Dumay 2003; Leach & Samuel, 2007; see also Bakker et al., 2014, which addressed a similar issue).

The remaining issue is to understand how the acquired lexical knowledge reshapes the phonological categories recruited during the AX discrimination task. Earle and Myers (2015; see also Earle & Myers, 2014 for an extensive review on the role of sleep on the construction of phonetic categories) argued that overnight improvement in speech sound discrimination performance following training might reflect a procedural learning which results from an implicitly acquired ability to attend selectively to relevant acoustic-phonetic details of the speech signal. Yet, in their study, the overnight consolidation was reported only on trained items while we observed this effect across tasks and on untrained items. Moreover, if this selective attention mechanism were the critical factor, it would have led to an overnight improvement in the *Aud* and, more specifically, the *AudArtic* training where participants' attention was explicitly drawn to the articulatory gestures of the critical phonemes. Although this mechanism might have contributed to some extent to the overall increase of discrimination performance for

all training methods, the overnight improvement that was specific to the *AudOrtho* training requires further explanation.

Interestingly, Earle and Myers (2015) mentioned a potential role of abstract information on the construction of sound categories and, thus, the sleep-related generalization of performance to untrained items and tasks. In their study, the authors mainly considered increasing training time and more exposure to phonetic variation as a means of generating abstract representations. Here, orthography provided an excellent source of abstract phonological information that allowed the categorization of variable and transient speech signals. More specifically, during training, it allowed participants to associate the auditory inputs with the correct phoneme categories. These abstract representations would enable the participants to overcome the ambiguity in the L2 acoustic speech signals. The resulting well-defined lexical knowledge seems to exert a top-down influence leading to a consolidation of phonological categories that correspond to the onset consonants of the words in the minimal pairs. However, this process takes time, and thus could not be observed immediately after exposure to the new knowledge.

This residual impact of orthography on speech perception is consistent with the idea that lexical information can play a role in modifying phonetic categorization by sending feedback to adjust speech processing at the perceptual, pre-lexical stage (Norris, McQueen, & Cutler, 2003; Mirman, McClelland, & Holt, 2006). Some previous studies indeed provided evidence that reading acquisition reshapes the nature of speech representations. For instance, Serniclaes et al. (2005) reported that literate participants displayed a more precise categorical boundary between phonemes in their native language than did illiterate participants, even though both populations showed categorical perception. Similar observations were also reported by Burnham (2003) and Hoonhorst et al. (2011) where the identification of native phonemic contrasts improved with children's reading experience. These findings suggest that acquiring

alphabetic literacy might contribute to further refining phoneme perception even in the L1 where phonemes are already correctly perceived (Ziegler & Goswami, 2005). Our findings are in accordance with this view and suggest that the same mechanism may be involved in the acquisition of L2 speech sounds. Finally, there is also evidence that acquiring new lexical items helps to build more robust phonological categories for L2 phonemes. Previous studies have shown that having a larger L2 vocabulary in the early stages of learning may be associated with better phonological acquisition in the L2, independently of other factors such as amount of exposure to the L2 or years of language instruction. Bundgaard-Nielsen, Best, & Tyler (2011) found that learners with a larger L2 vocabulary were more consistent in their vowel assimilation patterns (although this effect may be reversed when discrimination is poor; for a discussion, see Tyler, 2019). This result is in line with similar findings for L1 acquisition. As Edwards et al. (2004) put it, "knowledge of more word forms is associated with more robustly generalized knowledge of how to learn to hear and say new word forms" (p. 434). For L2 learners, the availability of orthographic information may facilitate novel word learning and help build a larger vocabulary conducive to developing a more native-like phonology.

In conclusion, a direct comparison of the contributions of articulatory gestures and orthography to spoken word learning revealed the benefit of both visual cues when they were simultaneously presented with speech. These types of contributions could be explained by the use of available input modalities in recognizing and memorizing novel spoken words. However, it must be differentiated from what we qualify as a residual contribution of visual cues, that is, whether having learned or been exposed to visual cues during training induces a long-lasting impact on speech processing even when the visual cue is no longer present. The present study suggests that this kind of contribution is mainly facilitated by the abstract orthographic code which enables learners to consolidate their lexical knowledge and, in turn, leads to an

enhancement of their perceptual ability, through a consolidation of distinct phonological categories of an L2 contrast.

ACKNOWLEDGEMENTS

This work was supported by the French National Research Agency: ANR-19-CE28-0001-01 (to C.P.), ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix Marseille University (A*MIDEX). We warmly thank Kylie Tyler for stimulus recording, Amélie Brassart and Caroline Gruffy for data collection, and Elsa Spinelli for sharing some E-Prime scripts. Nicolas Dumay and two anonymous reviewers provided very insightful suggestions during the review process.

REFERENCES

Antoniou, M., Best, C. T., & Tyler, M. D. (2013). Focusing the lens of language experience: Perception of Ma'di stops by Greek and English bilinguals and monolinguals. *Journal of the Acoustical Society of America*, *133*(4), 2397–2411. https://doi.org/10.1121/1.4792358

Antoniou, M., Liang, E., Ettlinger, M., & Wong, P. C. M. (2015). The bilingual advantage in phonetic learning. *Bilingualism*, *18*(4), 683–695. https://doi.org/10.1017/S1366728914000777

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed

random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bakker, I., Takashima, A., Hell, J. G. Van, Janzen, G., & Mcqueen, J. M. (2014). Competition from unseen or unheard novel words : Lexical consolidation across modalities. *Journal of Memory and Language*, *73*, 116–130. https://doi.org/10.1016/j.jml.2014.03.002

Balise, R. R., & Diehl, R. L. (1994). Some distributional facts about fricatives and a perceptual explanation. *Phonetica*, *51*, 99–110.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* (pp. 171–206.). York Press, Baltimore.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer, Version 6.0.21.

Bohn, O.-S. (2017). Cross-language and second language speech perception. In E. M. Fernández & H. S. Cairns (Eds.), *Handbook of Psycholinguistics* (pp. 213-239). Hoboken, NJ: Wiley.

Brannen, K. (2002). The role of perception in differential substitution. *Canadian Journal of Linguistics*, *47*(1–2), 1–46.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. https://doi.org/10.1162/jocn.2009.21076

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined

articulatory structures. *Journal of Phonetics, 18(3),* 299–320. https://doi.org/10.1016/s0095-4470(19)30376-6

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *33*(3), 433–461. https://doi.org/10.1017/S0272263111000040

Bürki, A., Welby, P., Clément, M., & Spinelli, E. (2019). Orthography and second language word learning: Moving beyond "friend or foe?" *Journal of the Acoustical Society of America*, *145*(4), EL265–EL271. https://doi.org/10.1121/1.5094923

Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing*, *16*(6), 573–609. https://doi.org/10.1023/A:1025593911070

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*, 204-220. https://doi.org/10.1002/dev.20032

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., … David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Carlet, A., & Cebrian, J. (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A sound approach to language matters - In honor of Ocke-Schwen Bohn* (pp. 91-119). Aarhus, Denmark: Dept. of English, School of Communication & Culture, Aarhus University. https://doi.org/10.7146/aul.322.218

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013.

Chen, W.-F., Chao, P.-C., Chang, Y.-N., Hsu, C.-H., & Lee, C.-Y. (2016). Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain and Language*, *157–158*, 51–62. https://doi.org/10.1016/j.bandl.2016.04.005

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Duncan, C.P. (1949) The retroactive effect of electroshock on learning. *Journal of Comparative and Physiological Psychology, 42*, 32–44.

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. *Frontiers in Psychology*, *5*, 1192. https://doi.org/10.3389/fpsyg.2014.01192

Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1680–1695. https://doi.org/10.1037/xhp0000113

Edwards, J., Beckman, M. E., & Munson, B. (2004). Vocabulary size and phonotactic production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

Eger, N. A., Mitterer, H., & Reinisch, E. (2019). Learning a new sound pair in a second language: Italian learners and German glottal consonants. *Journal of Phonetics*, *77*, 100917. https://doi.org/10.1016/j.wocn.2019.100917

Ehri, L. C. (1984). How orthography alters spoken language competencies in children learning to read and spell. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read* (pp. 119–147). New York: Springer-Verlag.

Ehri, L. C. (1985). Effects of printed language acquisition on speech. In D. R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 333–367). Cambridge, UK: Cambridge University Press.

Ehri, L. C., & Wilce, L. S. (1979). The mnemonic value of orthography among beginning readers. *Journal of Educational Psychology*, *71*(1), 26–40. https://doi.org/10.1037/0022-0663.71.1.26

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953. https://doi.org/10.1121/1.2178721

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

Escudero, P., Simon, E., & Mulak, K. E. (2014). Learning words in a new language: Orthography doesn't always help. *Bilingualism: Language and Cognition, 17*(2), 384-395. https://doi.org/ 10.1017/S1366728913000436

Faris, M. M. (2017). Perceptual assimilation, discrimination, and acquisition of non-native and second-language vowels assimilated as uncategorised [Doctoral dissertation]. Western Sydney University Thesis Collection. http://hdl.handle.net/1959.7/uws:45190

Faris, M. M., Best, C. T., & Tyler, M. D. (2016). An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *Journal of the Acoustical Society of America*, *139*(1), EL1–EL5. https://doi.org/10.1121/1.4939608

Faris, M. M., Best, C. T., & Tyler, M. D. (2018). Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories. *Journal of Phonetics, 70*, 1–19. https://doi.org/10.1016/j.wocn.2018.05.003

Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition, 128*(3), 280–286. https://doi.org/10.1016/j.cognition.2013.04.007

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*(6958), 614–616. https://doi.org/10.1038/nature01951

Fenwick, S. E., Best, C. T., Davis, C., & Tyler, M. D. (2017). The influence of auditory-visual speech and clear speech on cross-language perceptual assimilation. *Speech Communication, 92*, 114–124. https://doi.org/10.1016/j.specom.2017.06.001

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*, 796–804.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–276). Baltimore: York Press.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics, 14(1)*, 3–28. https://doi.org/10.1016/s0095-4470(19)30607-2

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes, 28*(8), 1207–1223. https://doi.org/10.1080/01690965.2012.701758

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication, 52*(6), 525–532.

https://doi.org/10.1016/j.specom.2010.02.005

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *International Journal of Behavioral Development*, *36*(6), 457–467. https://doi.org/10.1177/0165025412447752

Frost, R., Repp, B. H., & Katz, L. (1988). Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language*, *27*(6), 741–755. https://doi.org/10.1016/0749-596X(88)90018-6

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Grainger, J., & Ziegler, J. C. (2007). Cross-code consistency effects in visual word recognition. In E. L. Grigorenko & A. J. Naples (Eds.), *Single-word reading: Biological and behavioral perspectives* (Lawrence E, pp. 129–157). Mahwah, NJ.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, *104*, 2438–2450. https://doi.org/10.1121/1.423751

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.1288668

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, *103*(5), 2677–2690. https://doi.org/10.1121/1.422788

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*,

107(5), 2711–2724. https://doi.org/10.1121/1.428657

Han, J. I., & Oh, S. (2018). The role of phonetic similarity and orthographic information in asymmetrical lexical encoding in second language. *Journal of Psycholinguistic Research, 47*(5), 1015–1033. https://doi.org/10.1007/s10936-018-9574-7

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, M. G. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science, 15*(5), 674–687. https://doi.org/10.1111/j.1467-7687.2012.01172.x

Hardison, D. (2003). Acquisition of second language speech: Effects of visual cues, context and talker variability. *Applied Psycholinguistics*, *24*, 495–522. https://doi.org/10.1017/S0142716403000250

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528. https://doi.org/10.1037/0033-295X.106.3.491

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech, 1*(1), 1-7

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

Hickok, G., Holt, L. L., & Lotto, A. J. (2009). Response to Wilson: What does motor cortex contribute to speech perception? *Trends in Cognitive Sciences, 13*(8), 330–331. https://doi.org/10.1016/j.tics.2009.05.002

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-

language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Hoonhorst, I., Medina, V., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2011). Categorical perception of voicing, colors and facial expressions: A developmental study. *Speech Communication*, *53*(3), 417–430. https://doi.org/10.1016/j.specom.2010.11.005

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1367–1377. https://doi.org/10.1044/1092-4388(2003/106)

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252. https://doi.org/10.1121/1.1288413

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America, 138(2)*, 817–832.

Kuhl, P. K. (1992). Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 293–332). Washington: American Psychological Association.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell.

Leach, L., & Samuel, A. (2007). Lexical configuration and cexical Engagement: When adults learn new words. *Cognitive Psychology, 55*(4), 306–353.

https://doi.org/10.1016/j.cogpsych.2007.01.001

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 413–461. https://doi.org/10.1037/h0020279

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Llompart, M., & Reinisch, E. (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 1040–1056. https://doi.org/10.1037/xhp0000383

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1038/jid.2014.371

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences, 13*(3), 110–114. https://doi.org/10.1016/j.tics.2008.11.008.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McGuire, G., & Babel, M. (2012). A cross-modal account for synchronic and diachronic patterns of /f/ and /θ/ in English. *Laboratory Phonology*, *3*(2), 251–272. https://doi.org/10.1515/lp-2012-0014

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, (264), 746–748. https://doi.org/10.1038/264746a0

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some

English consonants. *Journal of the Acoustical Society of America*, *27*(2), 338–352. https://doi.org/10.1121/1.1907526

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin and Review*, *13*(6), 958–965. https://doi.org/10.3758/BF03213909

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS ONE*, *4*(11), e7785. https://doi.org/10.1371/journal.pone.0007785

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, *7*(4), 323–331. https://doi.org/10.1016/0010-0277(79)90020-9

Müller, G. E. & Pilzecker, A. (1900): Experimentelle Beiträge zur Lehre vom Gedächtnis, Zeitschrift für Psychologie und Physiologie der Sinnesorgane, Ergänzungsband 1, Leipzig: Verlag von Johann Ambrosius Barth.

Muneaux, M., & Ziegler, J. C. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, *19*(5), 641–660. https://doi.org/10.1080/01690960444000052

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2010). Is a FAN always FUN? Phonological and orthographic effects in bilingual visual word recognition. *Language and Speech*, *53*(3), 383-403. https://doi.org/ 10.1177/0023830910371462

Palma, P., & Titone, D. (2020). A review of the literature on sleep-related lexicalization of novel words in adults. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01809-5.

Pattamadilok, C, Morais, J., Colin, C., & Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language, 137*, 103–111. https://doi.org/10.1016/j.bandl.2014.08.005

Pattamadilok, C., Kolinsky, R., Luksaneeyanawin, S., & Morais, J. (2008). Orthographic congruency effects in the suprasegmental domain: Evidence from Thai. *Quarterly Journal of Experimental Psychology, 61*(10), 1515–1537. https://doi.org/10.1080/17470210701587305

Pattamadilok, C., Lafontaine, H., Morais, J., & Kolinsky, R. (2010). Auditory word serial recall benefits from orthographic dissimilarity. *Language and Speech, 53*(3), 321–341. https://doi.org/10.1177/0023830910371450

Pattamadilok, C., Perre, L., Dufau, S., & Ziegler, J. C. (2009). On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience, 21*(1), 169–179. https://doi.org/10.1162/jocn.2009.21014

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring? *Brain Research, 1275*, 73–80.

Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLoS ONE, 14*(12), 1–23. https://doi.org/10.1371/journal.pone.0221498

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation

for Statistical Computing, Vienna, Austria.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Xu Rattanasone, N., & Best, C. T. (2015). Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Attention, Perception, and Psychophysics*, *77*(2), 571–591. https://doi.org/10.3758/s13414-014-0791-3

Ricketts, J., Bishop, D. V. M., & Nation, K. (2009). Orthographic facilitation in oral vocabulary acquisition. *Quarterly Journal of Experimental Psychology*, *62*(10), 1948–1966. https://doi.org/10.1080/17470210802696104

Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology*, *100*, 175–191. https://doi.org/10.1037/0022-0663.100.1.175

Samuel, A. G., & Dumay, N. (in press). Auditory selective adaptation moment by moment, at multiple timescales. *Journal of Experimental Psychology: Human Perception and Performance*.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 546–554. https://doi.org/10.1037/0278-7393.5.6.546

Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition*, *98*(2), B35–B44. https://doi.org/10.1016/j.cognition.2005.03.002

Snodgrass, J. G., Levy-Berger, G., & Haydon, M. (1985). *Human experimental psychology*. New York: Oxford university press.

Stephens, J. D. W., & Holt, L. L. (2010). Learning to use an artificial visual cue in speech identification. *The Journal of the Acoustical Society of America, 128*(4), 2138–2149.

Stone, G. O., & Van Orden, G. C. (1994). Building a resonance framework for word recognition using design and system principles. *Journal of Experimental Psychology. Human Perception and Performance*, *20*(6), 1248–1268. https://doi.org/10.1037/0096-1523.20.6.1248

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing By Eye: The Psychology of Lip-reading* (pp. 3–51). Hove, UK: Lawrence Erlbaum Associates Ltd.

Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of Psycholinguistic Research*, *35*(1), 67–78. https://doi.org/10.1007/s10936-005-9004-5

Taft, M. (2011). Orthographic influences when processing spoken pseudowords: Theoretical implications. *Frontiers in Psychology*, *2*, 1–7. https://doi.org/10.3389/fpsyg.2011.00140

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, *24*(3), 320–335. https://doi.org/10.1016/0749-596X(85)90031-2

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *Journal of Neuroscience*, *30*(43), 14356–14360. https://doi.org/10.1523/JNEUROSCI.3028-10.2010

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to

phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850–855. https://doi.org/10.1016/j.cognition.2008.05.009

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56*(1), 16–34. https://doi.org/10.1016/j.jml.2006.07.002

Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A sound approach to language matters - In honor of Ocke-Schwen Bohn* (pp. 607–630). Aarhus, Denmark: Dept. of English, School of Communication & Culture, Aarhus University. https://doi.org/10.7146/aul.322.218

Tyler, M. D. (2021). Phonetic and phonological influences on the discrimination of non-native phones. In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 157-174). Cambridge, UK: Cambridge University Press.

Tyler, M. D., Clot, E., Villain--Bailly, M.-S., & Pattamadilok, C. (2019). Perceptual assimilation of English dental fricatives by native speakers of European French. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (pp. 2580–2584). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica, 71*(1), 4–21. https://doi.org/10.1159/000356237

Tyler, M. D, & Burnham, D. K. (2006). Orthographic influences on phoneme deletion response times. *Quarterly Journal of Experimental Psychology (2006), 59*(11), 2010–2031. https://doi.org/10.1080/17470210500521828

van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second

language: The L2LP model revised. *Frontiers in Psychology*, *6*(August), 1–12. https://doi.org/10.3389/fpsyg.2015.01000

Veivo, O., & Järvikivi, J. (2013). Proficiency modulates early orthographic and phonological processing in L2 spoken word recognition. *Bilingualism: Language and Cognition*, *16*(4), 864–883. https://doi.org/10.1017/S1366728912000600

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004). The locus of the orthographic consistency effect in auditory word recognition. *Language & Cognitive Processes*, *19*(1), 57–95. https://doi.org/10.1080/01690960344000134

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45(3)*, 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Wagner, A., Ernestus, M., & Cutler, A. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *The Journal of the Acoustical Society of America, 120*(4), 2267-2277.

Wagner, R. K., Torgeson, J. K., Laughton, P., Simmons, K., & Rashotte, C. A. (1993). Development of young readers' phonological processing abilities. *Journal of Educational Psychology, 85*, 83–103. https://doi.org/10.1037/0022-0663.85.1.83

Walden, B., Prosek, R., Montgomery, A., Scherr, C., & Jones, C. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*, 130–145. https://doi.org/10.1044/jshr.2001.130

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *Journal of the Acoustical Society of America*, *124*(3), 1716–1726. https://doi.org/10.1121/1.2956483

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on

audio-visual speech perception. *Journal of Phonetics, 37*(3), 344–356. https://doi.org/10.1016/j.wocn.2009.04.002

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Werker, J., & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75*, 1866–1878. https://doi.org/10.1121/1.390988

Xie, Z., Yi, H. G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PLoS ONE, 9*(12), 1–17. https://doi.org/10.1371/journal.pone.0114439

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*, 3–29. https://doi.org/10.1037/0033-2909.131.1.3