Articulatory information helps encode lexical contrasts in a second language

Miquel Llompart & Eva Reinisch

Ludwig Maximilian University Munich

M.Llompart@phonetik.uni-muenchen.de, evarei@phonetik.uni-muenchen.de

this is a preprint of:
Llompart, M. & Reinisch, E. (2017). Articulatory information helps encoding lexical contrasts in a second language. Journal of Experimental Psychology: Human Perception and Performance, 43, 1040-1056. doi: 10.1037/xhp0000383

Running head: Articulatory information in L2 contrasts

Author note

Miquel Llompart and Eva Reinisch, Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Germany.

This project was funded by a grant from the German Research Foundation (DFG; grant nr. RE 3047/1-1) to the second author. This work will be part of the first author's PhD project. Parts of the work were presented at the 8th International Conference on Second Language Speech (New Sounds) 2016 in Aarhus, Denmark, and at the 22nd Annual Conference on Architectures and Mechanisms for Language Processing, 2016, in Bilbao, Spain. We would like to thank Rosa Franzke for her help with testing participants and Tomas Lentz for advice concerning statistics.

Correspondence concerning this article should be addressed to: Miquel Llompart, Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstraße 3, 80799 Munich, Germany. E-mail: M.Llompart@phonetik.uni-muenchen.de

Abstract

The present study examined whether obtaining additional articulatory information about the sounds of a difficult second language contrast (English /ɛ/-/æ/ for German speakers) could help non-native listeners to encode a lexical distinction between novel words containing these two categories. Novel words (e.g., *tenzer-tandek*) were trained with different types of input and their recognition was tested in a visual-world eye-tracking task. In Experiment 1, a baseline group was exposed to the words audio-only during training, while another group additionally saw videos of the speaker articulating the target words. In Experiment 2, listeners were asked to repeat the target words themselves as part of their training. It was found that both audiovisual input and word repetition during training resulted in asymmetric fixation patterns at test: words containing /ɛ/ were recognized more readily than those with /æ/, mirroring the recognition asymmetry reported for real English words. This asymmetry was not present for the audio-only group, where target words with the two vowels were fixated similarly. The results suggest that articulatory knowledge, acquired through both passive exposure to visual information (Exp. 1) and active production (Exp. 2), can help distinguishing words with difficult foreign sounds.

(191 words)

Keywords: second language, perception-production interface, speech perception, eye-tracking, lexical processing

Statement of public significance

When learning words in a second language, learners often face the problem that difficult-to-distinguish sound contrasts lead to confusions between newly learned words. This study showed that providing information about how new words are articulated - either by asking participants to watch a video of the speaker or repeat the words themselves after a native model - helps learners to better distinguish between these words. That is, additional production-related information during training facilitates word recognition in a second language.

Articulatory information helps encode lexical contrasts in a second language

Learning a second language (L2) entails the perception and production of sounds that are not present in the native language. Some of these sounds are easily mastered by L2 learners, while others may cause great difficulties. A central concern in the study of L2 phonetics is to what an extent the production and perception systems are linked, that is, how they interact with each other while learning the sounds of a new language. Proposals regarding this link differ substantially between models of L2 phonetics/phonology. They range from a tight and interdependent relationship between the two (PAM-L2; Best & Tyler, 2007) to a looser connection in which perception is at the core of L2 learning and the two modalities do not necessarily need to be in one-to-one correspondence (SLM; Flege, 1991, 1995).

Studies on the effect of phonetic training on L2 sounds provide valuable insights into this relationship, given that they often focus on how different types of training impact performance in the two modalities. A large body of research has examined how phonetic training improves the learners' ability to deal with what Best and Tyler (2007) labeled *single-category assimilations*, that is, cases in which two sounds that are contrastive in the second language are mapped onto the same native language category. Single-category assimilations are problematic for L2 learners, who have often been shown to be unable to perceptually discriminate between the two L2 categories and to produce them as two separate phonetic categories (Bohn & Flege, 1990; Díaz, Mitterer, Broersma & Sebastián-Gallés, 2012; Escudero & Boersma, 2004; Flege, Takagi & Mann, 1995; Goto, 1971; Sheldon & Strange, 1982, to name but a few). Evidence from training studies up to date suggests that L2 perception and production are not connected in a strict one-to-one fashion. This is mainly due to two recurrent findings: (i) phonetic training in

one modality triggers only small or even no improvement in the other modality (Akahane-Yamada, McDermott, Adachi, Kawahara & Pruitt, 1998; Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1997; Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997; Hirata, 2004; Lopez-Soto & Kewley-Port, 2009; Wong, 2013), and (ii) when the effects of phonetic training, either in perception or in production, are examined across modalities (i.e., from perception to production or the other way around), weak or nonexistent correlations between improvements in the two modalities have generally been found (Bent, 2005; De Jong, Hao & Park, 2009; Peperkamp & Bouchon, 2011; Wang, Jongman & Sereno, 2003). For instance, Bradlow et al. (1997) trained Japanese speakers on the perception of the English /r/-/l/ contrast and found that perceptual training resulted in an improvement in both identification and production accuracy. Nonetheless, gains in production were much smaller than gains in perception, and no within-individual correlation between the two was found.

Crucially, though, it has also been shown that access to articulatory information going beyond the information that can be deducted from exposure to the acoustic input can improve both L2 perception and production in a consistent manner (Hardison, 2005; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006; Herd, Jongman & Sereno, 2013; Inceoglu, 2015; Kartushina, Hervais-Adelman, Frauenfelder & Golestani, 2015, 2016). A first helpful source of information is feedback on the acoustics resulting from the learners' articulation (Akahane-Yamada et al, 1998; Herd et al., 2013; Hirata, 2004; Kartushina et al., 2015, 2016). Kartushina et al. (2015) presented participants with a display of a simplified vowel space of the target language based on the first two formants (F1, F2). In this display learners could compare the location of their produced vowels to native speakers' targets (both represented as dots in the vowel space). Since F1 and F2 are acoustic correlates of the degree of jaw opening and tongue position,

participants hence received feedback on their articulation. Results after one hour of such production training showed that L2 productions were closer to the native speaker's model in a post-test compared to a pre-test, and, importantly, perceptual discrimination accuracy for the newly-learned contrast also increased after training. Notably, these improvements could only be attributed to the participants having had access to additional visual information about their own productions, since a control group did not show any improvement in spite of having produced the same number of vowels and having been exposed to visual information about native productions, but not their own.

A second type of information about the production of sounds that has been shown to improve performance in both L2 production and perception is the use of audiovisual stimuli. For several L2 contrasts, the addition of a video of the speaker producing the critical stimuli resulted in enhanced accuracy in the production of non-native sounds (Hardison, 2005; Hazan et al., 2006; Inceoglu, 2015). Inceoglu (2015), for example, reports that American English listeners who were exposed to training with audiovisual stimuli improved more in their production of the French nasal vowels [ɔ̃, ɑ̃, ɛ̃] than listeners in an audio-only condition. Most importantly, L2 learners are also more accurate in perceptually categorizing L2 sounds when exposed to audiovisual stimuli than with audio-only stimuli (Hazan et al., 2006; Navarra & Soto-Faraco, 2007; Wang, Behne & Jiang, 2009). Audiovisual stimuli have also shown to increase the effects of long-term training relative to audio-only conditions (Hardison, 2003, 2005; Hazan, Sennema, Iba & Faulkner, 2005; Hirata & Kelly, 2010). Hazan et al. (2005) reported that native Japanese listeners who were exposed to audiovisual stimuli during training improved more on their audio-only perception of the English sounds /p/, /b/, and /v/ than a comparable group exposed to auditory stimuli without an accompanying video. Similar results have been found for the /r/-/l/

contrast (Hardison, 2003, 2005, but see Hazan et al., 2005). These findings indicate that exposure to additional information – here about (some of) the visual articulatory correlates of the sound contrast – results in improvements in both perception and production, which in turn suggests that learning about L2 production can positively impact L2 perception.

What all the studies discussed so far have in common, however, is that they focused exclusively on L2 category learning at the phonetic level. They all trained participants to identify or produce (or both) L2 categories in tasks that were explicitly designed to focus on these specific sounds. By contrast, little is known about the relationship between L2 production and the processing of words in a second language. It remains unclear whether additional information about the articulation of L2 sounds may have an impact on the ability of second language learners to encode distinctions in their lexicon; specifically, between words that differ in the sounds of a difficult L2 phonetic contrast (e.g., *pan-pen* for native speakers of Dutch, see below). Investigating how different sources of information – including articulation – impact perception at the word level (i.e., lexical encoding of L2 sounds) is especially important in the light of another recurrent finding: L2 learners are less accurate in tasks that involve the recognition of words with sounds of difficult L2 contrasts than in tasks involving the phonetic categorization of such contrasts (Amengual, 2015; Darcy, Daidone & Kojima, 2013; Díaz et al., 2012; Díaz, Mitterer, Broersma, Escera & Sebastián-Gallés, 2015; Sebastián-Gallés & Baus, 2005; Sebastián-Gallés, Echeverría & Bosch, 2005). A case in point is Díaz el al. (2012), who found that native speakers of Dutch were fairly good at identifying the two sounds in the difficult English /ɛ/-/æ/ vowel contrast (Cutler & Otake, 2004) in a categorization task. In fact, almost half of the listeners performed within the range of accuracy of native speakers of English. By contrast, fewer participants scored within the native range of accuracy on lexical tasks, where they had to accept

words and reject non-words that contained the critical sounds (e.g., *desk - *dask, *lemp - lamp*) This indicates that the listeners' performance was worse in tasks that directly tapped into their L2 lexicon. The present study builds on these previous findings and addresses whether the separation of sound categories at the lexical level (and hence the separation of words) can be improved by articulatory information that has been shown beneficial at the sound level. Specifically, we consider the roles of (i) visual information provided by a video of the speaker and (ii) articulatory information provided by the repetition of the to-be-learned words.

One explanation for the difficulties of L2 learners in lexical tasks is that, although they may be able to distinguish words containing two confusable categories of the second language, part of their lexical representations seems to differ from those of native speakers. That is, some L2 lexical representations are less well-defined, or "fuzzy" (Darcy et al., 2013). Relevant evidence of this non-native encoding of L2 lexical contrasts are asymmetries in word recognition, as demonstrated in visual-world eye-tracking tasks (Cutler, Weber & Otake, 2006; Escudero, Hayes-Harb & Mitterer, 2008; Weber & Cutler, 2004). Weber and Cutler (2004) presented Dutch listeners with English word pairs that overlapped in their first syllables except for the vowel, which was /ɛ/ in one and /æ/ in the other member of the pair (e.g., *pencil-panda*) and examined participants' fixations on the intended targets. They showed that Dutch listeners differed in how effectively they recognized target words with /ɛ/ versus /æ/ as first vowel, as indicated by their eye gaze patterns. When asked to *'click on the pencil'*, for instance, participants fixated the correct visual referent rapidly and with minimal interference from the competitor (i.e., the word with the confusable vowel; e.g., *panda*). When the target word was *panda*, however, they momentarily considered both *panda* and *pencil* as possible target words. This was not the case for a group of native speakers of English, who recognized both target types

equally fast, with hardly any interference from the competitor. This finding suggests that, for Dutch speakers, the representations of words with /ɛ/ are better defined with regard to the critical vowel than those of words with /æ/. The same asymmetric pattern was found for native Japanese speakers' perception of the English /r/-/l/ contrast (Cutler et al., 2006). For Japanese listeners, the picture of a *locker* was quickly fixated when the instruction was to click on the *locker*, but the target *rocket* triggered fixations on both *rocket* and *locker*.

One interpretation of these findings is that the asymmetries are driven by the acoustic and/or articulatory similarity to the native category to which the two sounds are assimilated (Cutler et al., 2006); English /ɛ/ is a closer match than /æ/ to the native Dutch category /ɛ/. Likewise, English /l/ is a better fit to Japanese /ɾ/ than English /r/ (Iverson, Hazan & Bannister, 2005). Therefore, in each case, there is a dominant category, closer to the L1 category, and a non-dominant category that is a not-so-good fit to the native category. This is in turn reflected in the asymmetric target fixation patterns during online word recognition (see Cutler et al., 2006 and Darcy et al., 2013 for a more detailed discussion). Similar asymmetric patterns in lexical tasks have also been reported for native English speakers responding to words with singleton vs. geminate stop consonants in Japanese (Hayes-Harb & Masuda, 2008) and front vs. back rounded vowels in German (Darcy et al., 2013). Importantly, what these asymmetries clearly indicate is that L2 learners could somehow distinguish between the two categories in the words they heard. If listeners had not established a contrast between the two sounds in their lexicons (i.e., between words with the two sounds), temporary competition would have been expected to be strong and symmetric for the two categories because the two sounds would have simply been encoded as the same for both types of targets. Nonetheless, the fact that participants were not equally good at recognizing words with the two categories indicates that these L2 contrasts were not encoded in

a native-like fashion. Instead, they seem to be instantiated as a distinction between 'category X' (dominant) vs. 'category not-X' (non-dominant) (Hayes-Harb & Masuda, 2008).

There is evidence that the asymmetric encoding of difficult contrasts in the L2 lexicon is not exclusively modulated by how the acoustic input is perceived. Note that these contrasts normally entail difficulties in phonetic perception for L2 learners, which makes the acoustic input less reliable for them than it is for native speakers of the language (Sebastián-Gallés, 2005). In the light of this, other factors, such as explicit instruction in the L2 and, especially, the orthographic mapping of the two sounds in the foreign contrast, have been suggested to help establishing a lexical distinction (Cutler et al., 2006; Weber & Cutler, 2004). A key feature of the L2 contrasts examined by the eye-tracking studies discussed above is that the two categories that form the L2 contrast are systematically mapped onto two different letters; English /ɛ/ and /æ/ are (mostly) represented by *e* and *a,* /r/ and /l/ are spelled *r* and *l*, respectively. This means that the orthographic representations of *rocket* and *locker*, for instance, already indicate that there most likely is a difference between the pronunciation of the first consonant of these two words. The same holds for the rest of contrasts examined up to date; the two sounds were always orthographically distinguishable (Darcy et al., 2013; Hayes-Harb & Masuda, 2008). Therefore, even though listeners' perception of the critical sounds might not always be sufficiently reliable, it can be hypothesized that explicit knowledge about these words' spelling facilitates the encoding of a contrast in their lexicons between words that differ in these sounds.

Escudero et al. (2008) provided evidence in support of the role of orthography in the encoding of L2 word contrasts. They examined lexical competition patterns for the English /ɛ/-/æ/ contrast in Dutch listeners, like Weber and Cutler (2004), with the difference that novel words (e.g., *tenzer, tandek*) were used instead of real English words. This ensured that listeners

did not know the spelling of the words before the experiment. The main question was whether Dutch listeners would extend the asymmetric encoding of the contrast (/ɛ/ as dominant category) they exhibited for real words to the novel words, with and without the help of orthographic representations. The novel words were learned through an intensive training session where participants were assigned to one of two input conditions: (i) audio-only and (ii) audio + orthographic representation. After training, listeners were tested on the novel words using the visual-world paradigm, with the aim of assessing the recognition of novel words containing /ɛ/ vs. /æ/. Crucially, only the performance of the group that received training including orthography mirrored the asymmetric pattern of participants in the studies with real words (Weber & Cutler, 2004). For them, /ɛ/ targets were readily recognized, whereas /æ/ targets led to the momentary consideration of both /æ/-words and /ɛ/-words. This indicated that listeners had been able to establish the same /ɛ/ vs. not-/ɛ/ distinction that they exhibit for real English words in the novel words containing the two target sounds. The group who received audio-only training did not show such an asymmetry, with comparable fixation patterns for referents with the two vowel sounds. Therefore, the information contained in the acoustic input alone was not enough for these listeners to distinguish reliably between novel words with /ɛ/ and /æ/ as first vowel. These findings suggest that listeners can, and indeed may need to, make use of a cue that is external to the acoustic signal (i.e., orthography) to be able to establish a contrast in their second language lexicon.

In the present study we asked whether other sources of information could also help distinguish between new L2 words containing sounds of difficult L2 contrasts. Specifically, we focused on articulatory information that comes from one of two different sources: (i) passive exposure to visual articulatory information (i.e., mouth and jaw movements) by means of video

stimuli, and (ii), own active articulation of target words through delayed repetition of the critical stimuli. Note that throughout the paper we will be talking about two different types of articulatory information, however, it is undisputed that these two types also involve differences in acoustics (i.e., the native speaker's production in the video vs. the learners' own non-native productions during repetition). This issue will be taken into account for the discussion of the results. Critically, by examining the impact of these two sources of articulatory information on lexical processing, this study contributes to characterizing the relationship between L2 production and speech perception at the lexical (i.e., word) level.

Using the same type of training-test paradigm as in Escudero et al. (2008), we present two experiments in which native speakers of German were trained on novel English words containing /ɛ/ or /æ/, a contrast that is difficult for them. This is because German has only one category (/ɛ/) onto which both English categories are mapped. However, as for the Dutch participants in the studies described above, English /ɛ/ is a better fit than /æ/ (Bohn & Flege, 1990; Flege, Bohn & Jang, 1997). In the present study, listeners were trained to associate pairs of novel English words with pictures of novel objects (i.e., the same as in Escudero et al, 2008; see also Shatzman & McQueen, 2006). Experimental words were quasi-minimal pairs that overlapped phonetically on their first syllable except for the vowels that formed the English /ɛ/-/æ/ contrast (e.g., ***ten****zer*-***tan****dek*). The unambiguous second syllables ensured that learners could learn the names of the objects even without reference to the difficult vowels. However, the eye-tracking paradigm used at test allowed us to tap the earliest moments of word recognition to show whether, depending on the training condition, listeners had learned to effectively use the vowel contrast for target recognition.

During training, participants were presented with pictures on the screen and received spoken (audio-only) instructions to click on the picture that matched the novel word. Once they made a decision, they got corrective feedback on their response (correct/incorrect) and then the novel word was repeated. Crucially, the way in which the novel word was presented the second time differed between conditions. In Experiment 1, novel words were repeated by the same native speaker of English and were presented either audio-only (Audio condition; baseline) or together with a video of the speaker articulating the word (Video condition). In Experiment 2, participants were asked to repeat the target words themselves. After training, all participants performed the same visual-world eye-tracking task where words were presented audio-only and without feedback.

If listeners can make use of the additional articulatory information they have access to, either by exposure to videos (Experiment 1, Video) or by their own articulation of the target words (Experiment 2), we expect them to show an asymmetry in fixations in favor of words with the category that is closer to the L1 (/ɛ/). This would then show that the novel words have been learned including a differentiation of the difficult L2 vowels, as indicated by the asymmetric distinction between /ɛ/ and /æ/ that learners have been shown to make for real English words (*pencil-panda*; see Weber & Cutler, 2004). In line with Escudero et al. (2008), no asymmetry is expected for participants in the baseline Audio condition, where listeners do not receive any information in addition to that contained in the acoustic signal.

EXPERIMENT 1

Experiment 1 was devised to test whether visual articulatory information, as provided by videos of a native speaker showing her lip and jaw movements, would facilitate the encoding of a lexical contrast between novel words that contain sounds from a difficult L2 contrast. As

mentioned above, one group of listeners was exposed to auditory stimuli only during training (Audio condition) and a second group was exposed to videos (with audio) of a native speaker articulating the novel words (Video condition). Importantly, the /ɛ/-/æ/ contrast is visually cued by a difference in the degree of jaw opening (Carey, 2004). The jaw is lowered more for /æ/ than for /ɛ/. Hence, listeners in the Video condition received additional information about the articulation of the target sounds via jaw movements in the videos. The main question was whether they would be able to pick up on this information while learning the novel words. If so, they should become aware of the existence of two categories and establish a contrast between the vowels in their lexical representations of the novel words. It is expected that, at test, participants in the Video condition will show a fixation asymmetry between words containing /ɛ/ vs. /æ/ – as has been found for learners listening to real English words (Weber & Cutler, 2004) or novel words trained via orthography (Escudero et al. 2008). Such an asymmetry would be evidence of an early differentiation between target word pairs, that is, during the first syllable containing the difficult L2 contrast. As mentioned above, no fixation asymmetry is expected for listeners in the Audio condition, which would be a replication of the baseline in Escudero et al.

Method

Participants

Forty-one native speakers of German (23 females; age = 23.76, sd = 3.03), students at the University of Munich, took part for a small payment. None reported any hearing problems and all had normal or corrected-to-normal vision. Participants were recruited so that they had not learned any language other than German in their childhood; had not spent more than 6 months in an English-speaking country and were not enrolled in a language-program at the university.

Twenty participants were assigned to the baseline Audio condition and twenty-one to the Video condition. All participants filled in a background questionnaire assessing a number of self-estimated measures of English competence. Importantly, participants in the Video and Audio condition did not differ on these measures, which are shown in Appendix A. In addition, all participants reported that they were frequently exposed to spoken English (on a scale from 1 (very frequently) to 7 (never); mean = 2.88, sd =1.78) but they did not speak it themselves as frequently (mean = 4.22, sd =1.90). The variety of English that served as the desired model for their pronunciation differed across participants but not between groups (overall: British, *n* = 13; American, *n* = 13; none in particular, *n* = 15).

Auditory and visual materials

The word materials in the present study were the twenty disyllabic English nonwords (henceforth referred to as "novel words" or simply "words") used by Escudero et al. (2008). Half of the novel words (10) formed five target pairs. In each pair, the two words overlapped on their first syllables except for the vowel, which was /ɛ/ in one member of the pair and /æ/ in the other. The five target pairs were: *bestet-baskle*, *gebbet-gabble* [1], *hestel-haskum*, *meskle-mastik* and *tenzer-tandek*. The remaining half of the words formed 5 control/filler pairs that were created by

[1] A reviewer pointed out that *gabble* exists as a real English word ('to talk quickly'). To ensure that our results were not modulated by the accidental inclusion of a real word in our novel word list (following Escudero et al. 2008), all analyses were also conducted without the *gebbet-gabble* word pair. Even with reduced statistical power, the analyses for the remaining subset of data did not differ from the analyses including the *gebbet-gabble* word pair. Therefore, only the latter are reported.

replacing the first vowel of each target word with /ʊ/ (i.e., FOOT vowel; e.g., t[ʊ]nzer-t[ʊ]ndek). The words were assigned drawings of novel objects with the same pairings as used in Escudero et al. (2008).

Recordings

A female native speaker of Australian English was recorded in high-quality audio while she articulated the novel words both in isolation and at the end of the sentence 'click on the __'. Simultaneously she was videotaped (head and shoulders) on a digital camera in front of a light-grey background. Each word was recorded multiple times. The talker was instructed to produce the words with stress on the first syllable and with a similar speech rate and intonation contour across items. All audio recordings were equalized in amplitude and the first vowel (/ɛ/, /æ/, /ʊ/) of each target word was manually annotated. F1 and F2 values (LPC 25ms Gaussian window at midpoint as implemented by Praat; Boersma & Weenink, 2010) were extracted and used as reference for the final selection of stimuli to ensure consistent spectral values across words. To further confirm that the selected tokens for the critical contrast (/ɛ/-/æ/) were good examples of the native English categories, a pretest was conducted. Two native speakers of Australian English (other than the talker) categorized five presentations of each selected stimuli as containing the vowel in *pen* or *pan*. They were always correct except for one of them in one trial, which likely was an accidental wrong button press. Their performance indicated that the two vowels in the chosen stimuli were clearly distinguishable for native speakers of English. The selected high-quality audio stimuli produced in isolation were then paired with their respective video recordings for 17 out of the 20 words. For the remaining three words, a different video recording was selected since in the original video the speaker blinked or moved her head, which

may have been distracting for participants. For these three stimuli, it was ensured that the timing of audio and the new video matched as closely as possible[2].

Procedure

In order to ensure that participants were in an English language mode throughout the experiment, they were addressed in English by the experimenter and received written instructions on their task in English. They were told that they would be learning new words in English that would appear at the end of the English carrier sentence 'click on the __'. The experiment consisted of two phases: Participants first completed a training phase in which the novel words had to be learned with the help of feedback. Training was then immediately followed by a test phase, without feedback, in which participants were asked to identify the words in a visual-world paradigm while their eye-movements were tracked (Allopenna, Magnuson & Tanenhaus, 1998).

Participants were tested individually in a sound-attenuated booth. The experiment was conducted running Psychopy2 (v.1.83.01; Peirce, 2007). Images (and videos for Video training condition) were shown on a 19" screen and auditory stimuli were presented over headphones at a comfortable listening level. Eye-tracking during the test phase was conducted by means of an Eye Tribe portable eye-tracker (The Eye Tribe Aps, Copenhagen, Denmark) at a rate of 60 Hz.

---

[2] Listeners have been shown to integrate audiovisual signals with up to 200 ms of mismatch between the visual and the audio components (van Wassenhove, Grant, & Poeppel, 2007)

The experiment including training, test, and background questionnaires lasted approximately 1 h 30 min.

a) Training phase

Training was modeled after Escudero et al. (2008) and Shatzman and McQueen (2006). It consisted of 480 trials divided into eight blocks of 60 trials each. The twenty novel words were always presented once before being repeated. Within each repetition, words appeared in a fully randomized order. Participants were presented with pictures on the screen and had to choose which one matched the target at the end of the instruction "Click on the __". In blocks 1 to 4 they had to pick the correct item from two possible pictures; in blocks 5 to 8, difficulty was increased to four possible pictures. The pictures were located in the two upper quadrants of the screen for blocks 1-4 and in all four quadrants for blocks 5-8. The non-target pictures could be any of the other 19 words/pictures and were drawn randomly on every trial for each participant. The position in which the target appeared was counterbalanced over the training phase. Between blocks participants were allowed to take a short break.

Throughout the whole training phase participants received visual feedback on the correctness of their responses as indicated by a green tick for correct or a red cross for incorrect responses. Symbols (tick or cross) were presented in the middle of the screen. At the same time all pictures except the correct one disappeared from the screen allowing participants to re-view the correct picture. After 700 ms the target word was repeated: participants in the Audio condition heard the target word produced audio-only, while the target picture remained in its original position on the screen for another 2 s. Participants in the Video condition were shown a video of the speaker articulating the target word. The video was centered on the screen and did not overlap with the target picture, which remained in its original position during video

presentation. The target picture stayed on the screen until approximately 1-1.5 s after video offset (3 s total) so as to allow participants to focus on the video and re-view the correct picture.

b) Test phase

Immediately after training, the eye-tracker was connected and participants were calibrated at a distance of approximately 60 cm from the screen. The procedure during the test phase was the same as in the training blocks 5-8 (with four alternative response options to choose from) except for two key modifications. First, no feedback was given on responses during the test phase. Once participants had chosen the picture they identified as the target, the experiment moved on to the next trial. Second, while in the training phase the non-target pictures were chosen randomly, in the test phase every target appeared together with the other member of its pair, that is, its direct competitor, and an unrelated distractor pair. For critical target pairs, the competitor was the word with which the target shared the first syllable except for the first vowel (e.g. t[æ]ndek - t[ɛ]nzer). For filler pairs, the competitor was the word the target shared the whole first syllable with (e.g. t[ʊ]ndek - t[ʊ]nzer).

The test phase consisted of five repetitions of the twenty novel words for a total of 100 trials[3]. All words were presented once before they were repeated. In the experimental trials, participants heard the words with the critical vowels. In half of these trials the target contained

[3] Previous research has found that item repetition does not affect fixation patterns in the visual-world paradigm (Allopenna et al., 1998); fixation patterns have been shown to reflect gradual goodness of category fit even in phonetic categorization tasks using many continuum steps and token repetitions (e.g., McMurray, Tanenhaus & Aslin, 2002; McMurray, Aslin, Tanenhaus, Spivey & Subik, 2008; Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013).

/ɛ/ as first vowel, while the competitor contained /æ/. In the other half, targets contained /æ/ and competitors /ɛ/. In filler trials, target and competitor came from a filler pair with /ʊ/ as first vowel. For all test trials, the two unrelated distractors were another word pair that was drawn randomly on every trial with the restriction that it could not be the pair starting with the same consonant as the target-competitor pair. For example, when *tenzer* was the target and *tandek* the competitor, the distractors could be any of the other target pairs, like *hestel-haskum*, or one of the other filler pairs, like *mooskle-moostik*, but not the related filler pair *toonzer-toondek*. Target and competitor pictures were set to appear equally often in the four positions on the screen and the two distractor pictures were assigned to the two remaining positions in each trial at random.

Results

Training phase

Two participants had data missing due to equipment malfunction, one in each training part. Their data were retained in all parts where output files were available (including the test). The analyses included all data from experimental and filler pairs. Accuracy rates were analyzed separately for the parts with two and four alternatives. Two generalized linear mixed-effects models were fitted with a logistic linking function (lme4 package 1.1–10 in RStudio version 0.99.486) with Response (correct/incorrect) as categorical dependent variable. The predictor variables were Condition (Audio/Video), Trial Type (experimental/filler) and Block (1-4 and 5-8, respectively for the parts with two and four alternatives), as well as all interactions. Block was centered on zero in each part (e.g., Block 1 = -1.5, Block 2 = -0.5, Block 3 = 0.5, Block 4 = 1.5) and Condition was contrast coded such that Audio was coded as -0.5 and Video as 0.5. Trial Type was contrast coded with filler as -0.5 and experimental as 0.5. Contrast coding was used in

all analyses reported in the present study for an easier interpretation of effects and their interactions. When all factors are contrast coded and centered on zero, the grand mean is mapped onto the intercept and the effects (and their interactions) can be interpreted as main effects, similarly to traditional ANOVA. That is, effects can be interpreted relative to the grand mean rather than the factor level that is mapped onto the intercept. The regression weights (negative vs. positive values) indicate the direction of the effects. The random-effects structures were chosen by model fitting and random slopes were not included if they did not improve the model's fit, as measured by a log-likelihood ratio test. That is, we first fitted a model with only random intercepts for participants and items and then stepwise added random slopes for within participant or within item fixed effects and their interactions. At each step the simpler model was tested against the more complex one. If the more complex model fitted the data better, the random slope was retained and the whole procedure was repeated. The final models included random intercepts for Participants and Items, and a random slope for Block over Participants and Items. A random slope for Condition over Items was additionally included in the model for the four-alternatives data.

(Insert Figure 1 about here)

Figure 1 (left panel) shows the mean proportion correct responses for the first part of the training phase, where listeners had two alternatives to choose from. The model revealed a main effect of Block (b = 1.42; z = 15.81; p < .001) indicating that performance improved over training. No effect of Condition (b = -0.11; z = -0.37; p = .71) or Trial Type (b = 0.03; z = 0.21; p = .83) was found. All two-way interactions and the three-way interaction between Condition,

Trial Type and Block were non-significant (all $p > .1$). This shows that the rate at which participants learned the novel words did not differ between the two training conditions (Audio = 87.85% correct; Video = 86.29% correct; over the 4 blocks). In addition, accuracy rates over the four blocks did not differ between the two trial types, that is, words containing the difficult vowel contrast (experimental = 86.94% correct) or the same vowels (filler = 87.13% correct).

In the second part of the training, with four options on the screen to choose from, both groups of participants were already extremely accurate (+90% correct) throughout the whole part (see Figure 1, right panel). However, the statistical analysis showed an effect of Block ($b = 0.43$; $z = 6.18$; $p < .001$) indicating that over blocks participants still became more accurate. Neither the effect of Condition ($b = -0.31$; $z = -0.69$; $p = .49$; Audio = 93.38% correct; Video = 91.29% correct) nor Trial Type ($b = 0.03$; $z = 0.11$; $p = .91$; experimental = 92.38% correct; filler = 92.29%) was significant. Again, none of the interactions was significant ($p > .5$).

Test phase

Data from two participants, one from each training condition, had to be removed from the dataset due to eye-tracker malfunction. These were different from the participants whose training files were missing. Since these were included in the test, this left us with data from 39 participants. Overall, test performance was extremely accurate for all participants (minimum % correct = 72%; mean % correct = 94.73%). It was comparable between the two conditions (Audio = 96.63%; Video = 92.83%), between experimental and filler trials (experimental = 95.82%, filler = 93.62%), and, importantly, between words with each critical vowel within the experimental trials (/ɛ/-targets = 96.13%, /æ/-targets = 95.50%).

For the analyses of the eye-tracking data only trials including the critical vowel contrast (/ɛ/-/æ/) and trials in which participants clicked on the correct picture were taken into account.

202 trials (5.28%) were excluded because listeners clicked on pictures other than the target picture. Figure 2 shows fixation proportions on target, competitor, and the average of the two distractors over time plotted for the Audio condition (left panel) and Video condition (right panel) for targets with the two critical vowels (/ɛ/-targets grey lines, /æ/-targets black lines). The vertical lines indicate the time window of interest, from 200 to 800 ms after target onset. The onset of the time window was motivated by the finding that listeners need roughly 200 ms to program and launch a saccade (e.g., Allopenna et al., 1998). This onset hence allows us to capture eye movements as early as they can be driven by acoustic information from the target words. The end of the window (800 ms) was determined by inspection of the grand average of fixations over time (i.e., pooled over all conditions) and taken as the point in time when competitor fixations had approximately dropped to the level of fixations on the distractor objects (see Salverda, Dahan & McQueen, 2003). The resulting time window was thus sufficiently large to allow us to use Growth Curve Analyses (Mirman, Dixon, & Magnuson, 2008; for details see below) to model fixation trajectories over time in order to assess the influence or interaction of training condition (Audio/Video) and target vowel (/ɛ/-/æ/) during word recognition.

(Insert Figure 2 about here)

A first inspection of Figure 2 suggests that listeners in both training conditions rapidly recognized the target pictures as indicated by a rise in fixations on the targets. However, as expected, listeners suffered from lexical competition: the pictures of the objects whose name started with the same consonant and had a perceptually similar vowel to the target were fixated on more than the distractor pictures. Critically, within the time window of interest, the rate of

increase in fixations on the targets with /ɛ/ and /æ/ was about equal in the Audio condition (left panel) but differed in the Video condition (right panel). It appears that participants who received additional visual information on the identity of the critical vowels were delayed in their recognition of targets including the vowel /æ/. Notably, this pattern appears to be mirrored in competitor fixations for this group of listeners, with slightly stronger competition from /ɛ/-competitors on /æ/-targets than the other way around. However, our analyses focus exclusively on target fixations, since it has been suggested that the amount of training that participants undergo in a study with only one training session may not be sufficient to reliably assess how novel words engage in lexical competition (Dumay & Gaskell, 2007; Escudero et al. 2008; Gaskell & Dumay, 2003).

The observations from Figure 2 were confirmed by statistical analyses. For statistical analysis, the eye-position data over time were fitted and conditions were compared using Growth Curve Analyses (GCA; Mirman et al. 2008). GCA is explicitly geared towards modeling change over time and is therefore well suited to address the time course of activation patterns among words as measured with the visual world paradigm (cf. Mirman et al. 2008). It allows the capture of differences between conditions in the rise (and fall) of fixations over time that an overall analysis of fixation proportions within the critical time window would miss. More specifically, GCA models the shape of the probability distributions of target fixations in the different conditions over time, and these fitted curves are then compared by a multi-level statistic that assesses differences in the parameters describing these curves. For a detailed argumentation on the advantages and a detailed description of the workings of GCA we would like to refer the reader to Mirman et al. (2008) and repeat here only the most critical issues, specifically regarding the modeling of time. The parameters to model time are orthogonal polynomials. In the

present case the model with polynomials including a linear and quadratic term was best fitting. These orders of "Time" can be directly related to the shape of the curves such that the intercept term matches the overall height of the curves in the time window, the linear (first-order) term refers to the slope of the rise in target fixations over the time window, and the quadratic (second-order) term refers to the symmetric rise and fall (or fall and rise) around a central inflection point (shape or curvature). Orthogonal polynomials have the advantage over other possible terms for Time that they are independent and hence differences in height, slope, and shape of the fixation curves can be assessed independently. Importantly, the multi-level approach of GCA allows assessing these differences clustered within participants and items. That is, our model matched common mixed-effects regression (including random-effects for participant and item; see below) with the addition of fixed and random terms for Time.

The dependent variable was whether at any point in time a fixation was on the target or on one of the other pictures on the screen. Given the dichotomous nature of this variable, a logistic linking function was used in our model (Jaeger, 2008). Fixed effects included Condition (Audio/Video), Vowel (/ɛ/-/æ/) and their interaction, as well as interactions with the first and second order polynomials representing Time. All factors were contrast coded such that Condition was coded as Audio = -0.5 and Video = 0.5; and Vowel as /ɛ/ = -0.5 and /æ/ = 0.5. Orthogonal polynomials are per definition centered on zero. Therefore, with our coding the grand mean is mapped onto the intercept and the direction of the regression weights indicate the direction of the effects. For ease of interpretation, effects of Condition and Vowel will be reported separately for each order of Time (i.e., intercept, linear, and quadratic). The random effects structure was built such that our fixed effects were clustered within individual participants and items, allowing the Time components as well as the effect of Vowel to vary between participants, and the Time

components between items (note that vowel is manipulated between items). This model was the best fitting one with the largest random effects structure (Barr, Levy, Scheepers & Tily, 2013) that converged. Figure 3 shows the fitted model.

(Insert Figure 3 about here)

Table 1 shows the statistical results of our model, which fit with our observations from Figure 3. We found significant interactions between Vowel and Condition on the linear and quadratic terms of Time. That is, the slope and curvature of target fixations over time for the two vowels differed between the two conditions. To follow up on these interactions separate models using GCA were fit for data from the Audio and Video conditions. Models were identical to the model described above with the difference that in addition to our Time components (again linear and quadratic terms fit best) only Vowel (/ɛ/-/æ/) was entered as a fixed factor and as a random effect over participants. The results for these analyses are reported in Table 2 and confirm the differences observed in Figure 3. First, fixations on targets in the Audio condition did not differ between the two target vowels in any order of time. In short, the height, slope and curvature of the fixation trajectories were nearly identical between vowels. In the Video condition, by contrast, the effect of Vowel was significant on the quadratic time term. Figure 3 shows that the pattern of fixations over time for /ɛ/ targets in the Video condition is almost linear, similarly to the two fixation curves in the Audio condition. However, for /æ/ targets, we can observe a concave pattern. This indicates that /æ/ targets received fewer looks throughout the most part of the analyzed time window, but this difference was made up for towards its end. In contrast to the model reported above, the effect of Vowel on the linear time term failed to reach significance

also for the Video Condition alone. This is likely because the overall slopes of target fixations for the two vowels do not differ despite their difference in curvature.

(Insert Tables 1 and 2 about here)

Discussion

Experiment 1 showed that listeners are able to pick up visual differences in the articulation of a difficult-to-distinguish sound contrast (i.e., differences in jaw opening as seen in videos of the speaker's productions) and use them to establish a lexical contrast between words containing these sounds. During training, listeners were exposed either to audio-only stimuli (Audio condition) or to audiovisual stimuli that provided additional articulatory information (Video condition). Importantly, the two groups showed similarly high learning rates for the novel word-picture associations. This ensured that listeners would be able to identify the correct targets during the subsequent test phase, where it was assessed, by means of the eye-tracking data, whether the different groups of listeners were able to make use of the difficult L2 vowels (/ɛ/-/æ/) to distinguish between the novel words pairs during word recognition. Analyses on target fixations showed that in the Audio condition looks to the target pictures did not differ as a function of target vowel. Listeners in this group were as likely to fixate the target when it contained /ɛ/ (e.g., *tenzer*) as /æ/ (e.g., *tandek*). Participants in the Audio condition hence treated the first syllables of the two words in the target pairs (e.g., *ten*-, *tan*-) as the same. Participants in the Video condition, by contrast, did fixate targets with the two vowels differently. They showed a delay when trying to contact the newly-formed representations of /æ/-targets in comparison with /ɛ/-targets. This suggests that they were able to encode a contrast between the two vowels in

their representations of the novel words. Their fixation patterns mirrored the asymmetric, non-native-like patterns found for L2 learners with real English words and novel words cued by orthography (Escudero et al., 2008; Weber & Cutler, 2004).

EXPERIMENT 2

In Experiment 2 we further assessed the role of articulatory information in establishing L2 lexical contrasts. However, in contrast to the visual information provided for the Video condition in Experiment 1, Experiment 2 tested the impact of the participants' own articulation. Therefore, the only extra information participants had in addition to the audio model instructing them to click on a target were the movements involved in their own articulation and the concurrent perceptual exposure to their own productions. In this case, no external (i.e., visual) information about a native speaker's articulation was provided. Experiment 2 allowed us to test (i) whether participants would be able to produce an acoustic difference between /ɛ/ and /æ/ when repeating the words after hearing them from a native speaker, and (ii) whether the act of producing words with /ɛ/ and /æ/ and hearing their own productions, would lead to the establishment of a lexical contrast between the critical vowels in novel words. This would be expected if asking participants to repeat after the native speaker would be sufficient to draw their attention to the phonetic detail relevant to the encoding of a contrast. If so, we would again expect the fixation asymmetry in the eye-tracking task at test that was found in previous studies (Escudero et al., 2008; Weber & Cutler, 2004) and also for the Video condition in Experiment 1. Experiment 2 will thus be able to answer a second question of interest, that is, whether it is only external cues providing contrastive information (e.g., orthography, videos) that are helpful in the encoding of lexical contrasts, or whether L2 learners can learn about the contrastive nature of two L2 categories from their own non-native productions.

Method

Participants

Thirty-one native speakers of German, students at the University of Munich (19 females; age 24.87; sd = 3.58), took part for a small monetary compensation. They all had normal or corrected-to-normal vision and did not report any hearing problems. None had participated in Experiment 1. Recruiting requirements and background questionnaires were the same as in Experiment 1. Participants rated their proficiency in English similarly to listeners in the Video and Audio conditions in Experiment 1. Measures are reported in Appendix A. Values for use of English in their daily lives were also similar to those of Experiment 1 (exposure: mean = 2.87, sd = 1.46; spoken: mean = 4.87, sd = 1.43). Again, different varieties of English served as the desired pronunciation model for participants in this group (British, n = 10; American, n = 10; none in particular, n = 10).

Materials Design and Procedure

All materials (words, pictures, recordings) were the same as in Experiment 1, as was the experimental design and procedure. However, in Experiment 2 no second repetition of the novel word by the native speaker was provided. Therefore, the auditory stimuli that were used were only those in which target words were embedded in the carrier sentence 'click on the____'. Neither the video recordings nor the audio-only recordings of the target words in isolation were used. The nature of the task to be performed after corrective feedback constituted the only change with respect to Experiment 1. After clicking on the picture that participants thought that corresponded to the novel word they again were first informed about the correctness of their responses by means of a green tick or a red cross (700 ms). However, instead of hearing the

target word a second time, 250 ms after the offset of corrective feedback (hence 950 ms after they made a decision on the target word) they were prompted to repeat out loud the target word they had heard. Participants in this experiment will be henceforth referred to as being in the "Repetition" condition. Their repetitions were recorded and recording time was cued by an iconic representation of a microphone, which appeared centered on the screen and remained there for 2.75 s. Importantly, the correct target picture remained on its original position throughout corrective feedback and word repetition, so that participants were able to re-view the correct picture comfortably. The test phase did not differ from that of the previous experiment. The setup and equipment were the same used for Experiment 1 plus an AT3031 condenser microphone (Audio-Technica, Tokyo, Japan) and an M-Audio MobilePre USB device (M-Audio, Rhode Island, USA) to record the participants' productions during the training phase. Recordings were made using the *microphone component* of Psychopy2. The speech signal was sampled at 48 kHz with 16-bit quantization. The experiment, including training, test, and background questionnaires lasted approximately 1 h 30 min.

Results

Training phase: accuracy

Performance in Experiment 2 was analyzed in comparison with the baseline Audio condition from Experiment 1. One participant was excluded from all analyses on training and test data, due to high error rates during training and test (below 70% correct in all parts). Due to missing output files, data from three further participants had to be excluded from the analyses of the training phase (one in the two-alternatives part and two in the four-alternatives part). Their data was retained in the analyses of the remaining training part and the test phase. The final

dataset thus contained data from 29 participants for the two-alternatives training part and 28 for the four-alternatives part.

As in Experiment 1, training data were analyzed separately for the training parts with two and four alternatives and the analyses included responses to both experimental pairs and filler pairs. Data were submitted to two generalized linear mixed-effects models fitted with a logistic linking function. The dependent variable was Response (correct/incorrect) and Condition (Audio/Repetition), Trial Type (Experimental/Filler), Block (1-4 and 5-8, respectively), and their interactions were entered as predictors. Block was centered on zero and Condition was contrast coded such that Audio was coded as -0.5 and Repetition as 0.5. Trial Type was contrast coded with Filler as -0.5 and Experimental as 0.5. The models included random intercepts for Participants and Items, and a random slope for Block over Participants and Items. No random slopes for Condition over Items or for Trial Type over Participants were included in the analyses because they did not improve the model's fit for any of the two parts of the training phase.

Figure 4 shows the identification accuracy rates for participants in Experiment 2 in comparison with the baseline Audio condition. The statistical model on the training with two alternatives revealed significant effects of Block (b = 1.30; z = 18.74; p < .001) and Condition (b = -0.53; z = -2.48; p < .05; Audio = 87.85% correct; Repetition = 83.13% correct). No effect of Trial Type (b = -0.02; z = -0.12; p =. 90; experimental = 85.23% correct, filler = 84.77% correct) was found and none of the interactions were significant (p > .2). The effects of Block and condition can be seen in Figure 4, left panel: participants in the Repetition condition were less accurate than participants in the Audio condition in Experiment 1. This is most likely due to the fact that their task after feedback was initially more demanding, since in order to repeat the word, they had to keep the word in memory until they were prompted to speak it out loud. This may

initially have drawn their attention away from the visual referent and hence slowed down the reliable establishment of word-object associations. This is in contrast to the passive listening required of participants in the Audio condition.

(Insert Figure 4 about here)

In the model on training with four alternatives, by contrast, only Block was found to have a significant effect ($b = 0.46$; $z = 7.23$; $p < .001$), while the effects of Condition ($b = 0.12$; $z = 0.36$; $p = .72$; Audio = 93.38% correct, Repetition = 95.01% correct) and Trial Type ($b = 0.01$; $z = 0.03$; $p = .98$; Experimental = 94.39% correct, Filler = 94.30% correct) as well as all interactions were not significant (all $p > .3$). While all participants kept improving in their recognition of the novel words, as illustrated by the effect of Block, the two conditions did not differ in their accuracy scores for this part of the training. This shows that even though participants in the Repetition condition were less accurate than those in the Audio condition during the first part of the training, the ultimate attainment of the word-picture associations for the Repetition condition was just as good as that of the baseline group.

Training phase: acoustic analysis

Each participant produced 240 recordings per training part, which resulted in a dataset of 13680 recordings. Half of them were tokens of the experimental items with /ɛ/ or /æ/ as first vowel and only these were subjected to acoustic analyses. Recordings were sorted by participant, trial number and target word and were segmented and phonetically annotated by means of the Munich Automated Annotation Service (*WebMAuS,* Kisler, Schiel & Sloetjes, 2012; Schiel, 1999). For each token values for the first and second formants (F1 and F2) at vowel midpoint

(LPC 25ms Gaussian window) were measured using *Praat* (Boersma & Weenink, 2010). Vowel tokens whose formant structure could not be tracked by *Praat* due to either a segmentation error or a recording malfunction were discarded from analyses (139 words; 2.07% of the dataset). Data were further trimmed in order to exclude outlier values of F1 and F2 likely resulting from inaccurate segmentation or measurements. After visual inspection of the dataset's formant values, cutoff points for F1 were established at 300 Hz and 1250 Hz, with values under and over these values, respectively, being excluded. For F2, values under 1000 Hz and over 2500 Hz were excluded. Note that the cutoff points were placed at a considerable distance from the values generally reported for /ɛ/ and /æ/ in English (Deterding, 1997; Hillenbrand, Getty, Clark & Wheeler, 1995; Watson, Harrington & Evans, 1998), always ensuring that they were more than three standard deviations from the mean for the two vowels in our dataset. Based on this criterion, 20 more tokens (0.29%) were excluded. The final dataset consisted of 6681 vowel tokens.

The metric used to analyze the acoustics of the two critical vowels was, for each vowel, the difference score between F2 and F1 (F2-F1) in Hertz. Using the difference F2-F1 allowed us to perform statistical analyses on only one value, instead of having to resort to separate analyses for F2 and F1. In English, /ɛ/ has a lower F1 and a higher F2 than /æ/ (Deterding, 1997; Watson et al., 1998). Consequently, if a contrast between the vowels is produced, the F2-F1 difference is expected to be higher for /ɛ/ than /æ/. If, on the contrary, participants of the present study, who are native speakers of German, have difficulties with the /ɛ/-/æ/ contrast in production, no or only a small difference between the F2-F1 for /ɛ/ and /æ/ is expected.

Data were submitted to a generalized linear mixed-effects model with F2-F1 as dependent variable and Vowel (/ɛ/ - /æ/) and Block (1-8), as well as their interaction, as predictors. Block

was centered on zero and Vowel was contrast coded with /ɛ/ as -0.5 and /æ/ as 0.5. The model included random intercepts for Participants and Items, and random slopes for Block over Participants and Items and Vowel over Participants. The model revealed a significant effect of Vowel (b = -132.03; t = -2.73; p < .05). No effect of Block was found (b = 1.23; t = 0.65; p = .52) and the interaction between Vowel and Block was not significant (b = 2.64; t = 1.40; p =. 20). Results therefore show that the difference between F2 and F1 was bigger for /ɛ/ than for /æ/, indicating that participants were on average able to produce the two vowel sounds as acoustically different. Moreover, the lack of an effect of Block or its interaction with Vowel suggests that the participants' productions did not change throughout the training session. Figure 5 shows the difference between F2 and F1 values of the two vowel categories as produced by the native German speakers and the median values of the native English vowels in the target words (dashed lines) as a reference. This figure serves to illustrate that there was substantial variation in the difference between the two vowels that participants produced, which is apparent through the considerable overlap between the distributions for the two categories. In addition, it can also be observed that the differentiation between vowels by non-native speakers was on average much smaller than that produced by the native speaker. In sum, the contrast that participants produced was not always clear-cut, or at least not as clear-cut as in the model they were exposed to.

(Insert Figure 5 about here)

Test phase

Data from 30 participants in the Repetition condition were considered in the test phase analyses. As in the previous experiment, test performance was high (minimum % correct = 78%;

mean % correct = 94.44%), was comparable between experimental and filler trials (experimental = 94.41%, filler = 94.48%), and, importantly, between critical vowels in the experimental trials (/ɛ/-targets = 93.37%, /æ/-targets = 95.45%). Only experimental trials in which participants clicked on the correct picture were included in the analyses of the eye-tracking data. 161 trials (5.56%) were discarded because listeners clicked on the wrong picture. The eye-tracking data for participants in Experiment 2, the Repetition condition, were analyzed in comparison with both conditions tested in Experiment 1. For ease of interpretation, comparisons with the Audio and Video condition were done separately. A comparison with the Audio condition assesses the potential benefit of repetition relative to our baseline, while a comparison with the Video condition allows us to test differences between the effects of passive observation of articulatory movements vs. active production of articulatory movements on the recognition of words with the critical vowel contrast.

Figure 6 shows fixation proportions on target, competitor, and the average of the two distractors over time plotted for the Repetition condition. This figure suggests that the rate of increase in fixations to the target picture differed for novel words containing the two vowels: fixations to target words with /ɛ/ increased more rapidly than fixations to target words with /æ/ consistently over the time window of interest. This pattern is therefore similar to that found in Experiment 1 for the Video condition and contrasts with the results for the Audio condition, where targets with the two vowels showed very similar rises in fixation proportions over time (see Figure 2).

(Insert Figure 6 about here)

These observations were confirmed by statistical analyses. Data from the Repetition and Audio conditions were submitted to a Growth Curve Analysis with two orthogonal polynomials, that is, a linear and a quadratic time term, as in Experiment 1. The dependent variable was whether at any moment in time the target picture was fixated, as opposed to any other picture on the screen. Predictors included Condition (Audio coded as -0.5 and Repetition as 0.5), Vowel (/ɛ/ = -0.5 and /æ/ = 0.5) and their interaction, as well as interactions with the first and second order polynomials representing Time (that are centered on zero; see Experiment 1 for details). The random effects included random intercepts for Participants and Items and random slopes for Vowel and the Time components over participants and the Time components over Items. This model was again the best fitting one, as assessed through model comparisons, with the largest random effects structure that converged. Figure 7 (left panel) shows the fitted lines for target fixations and Table 3 (left) shows the statistical results of the model. A significant interaction between Vowel and Condition on the linear term of Time was found. This indicates that the slope of target fixations over time for the two vowels differed between the two conditions. While fixations on targets in the Audio condition were nearly identical in words with the two vowels, in the Repetition condition the slope of the curve for /ɛ/-targets was steeper than for /æ/-targets. Words with /ɛ/ were recognized faster than words with /æ/. The latter showed a considerable delay in recognition (see Figures 6 and 7).

(Insert Figure 7 about here)

A second model was built in the same fashion to compare the Repetition and the Video conditions. The only difference was that Condition was coded as Repetition = -0.5 and Video =

0.5. Figure 7 (right panel) shows the regression lines fitted by the model. The results of the statistical analysis are shown in Table 3 (right). The critical significant effect is the interaction between Vowel and Condition on the quadratic term of Time. As can be seen in Figure 7 (right panel), this is due to the fact that, although in both conditions /æ/-targets were fixated less frequently than /ɛ/-targets, the regression lines for fixations to /æ/-targets differ between conditions in terms of their degree of curvature. In the Video condition we observe a clearly concave pattern, as already described in Experiment 1, while fixations to targets with /æ/ in the Repetition condition increased linearly, ending clearly below fixations to /ɛ/-targets.

(Insert Table 3 about here)

Discussion

Experiment 2 showed that the active articulation of words, together with perceptual exposure to one's own productions, can lead to the encoding of a difficult L2 contrast at the word level. During training, participants in Experiment 2 received corrective feedback on their choice of the object representing the novel words – just as in Experiment 1. However, instead of hearing a native speaker repeat the word, they were asked to repeat the words themselves (after a delay of at least 950 ms). First, listeners in Experiment 2 were able to learn the intended word-picture associations as accurately as participants in the baseline (Audio) condition, even though they were less accurate in the first part of the training. This was likely because word repetition was a more demanding task than passive listening. Second, participants in Experiment 2 were able to differentiate between /ɛ/ and /æ/ in production, even though on average the acoustic difference

was smaller than that of the native speaker model and there was substantial variation within as well as across participants.

As for our main question on the patterns of recognition for the novel words in the test phase, participants in the Repetition condition showed an asymmetry in target recognition. Targets with /ɛ/ were fixated faster and more consistently than targets with /æ/. As in the Video condition in Experiment 1, listeners were efficient at contacting the lexical representations of newly-learned words with the sound that is closer to an L1 category (/ɛ/), but were delayed when accessing representations with the vowel /æ/, which is a worse fit to the native vowel inventory. The asymmetric processing of the two target types can again be taken as evidence of the encoding of a contrast between /ɛ/ and /æ/, where /æ/ could be represented as not-/ɛ/ (Hayes-Harb & Masuda, 2008), in the newly learned words. Active production triggered thus lexical separation in a similar way as passive exposure to the different types of external contrastive information examined up to date (i.e., visual articulatory information, orthographic representations). This shows that additional articulatory information about difficult L2 sounds, acquired through active production, can help the establishment of a contrast between words containing those sounds, even in the absence of external, contrastive information.

Interestingly, though, from our data it seems that the delay in fixating /æ/ targets was longer-lasting in the Repetition condition than in the Video condition. Even though any attempt to characterize this difference remains speculation, a tentative explanation could be that listeners in the Video condition were better able to make up for the delay caused by the difficult non-native category because they listened to native productions throughout the whole training phase, while in the Repetition condition participants were exposed not only to native but also to their own non-native productions. Extensive exposure to consistent native productions during training

could have resulted in a better mapping of the acoustics of the target stimuli onto the newly-established lexical representations for the novel words than simultaneous exposure to native and non-native productions. This would then be expected to translate into a faster recovery from processing the "fuzzy" category once more information was presented.

Eventually, a question that follows from the design of Experiment 2 is whether there is a relationship between the participants' accuracy in producing /ɛ/ and /æ/ as two different sounds during word repetition and the patterns of target fixations they exhibited in the visual-world eye-tracking task. In order to address this issue, a linear regression model was run with Asymmetry (proportion of looks to /ɛ/-targets minus proportion of looks to /æ/-targets) as dependent variable and Acoustic Difference (mean F2-F1 for /ɛ/ - mean F2-F1 for /æ/) as predictor. The model showed that the acoustic difference between the two sounds during the training phase did not predict the fixation patterns in the test phase (b = -0.01; t = -0.55; p = .59). This indicates that there is no one-to-one correspondence between how accurate participants were at producing the two vowels as two different categories and how successful they were at encoding a distinction between the two vowels in the newly-formed lexical entries. This lack of an effect appears to be in line with most findings of training studies that examined how improvements in L2 production affect L2 perception and vice versa (Bent, 2005; De Jong et al., 2009; Peperkamp & Bouchon, 2011; Wang et al., 2003). However, this lack of connection should be interpreted with caution. The production data we collected were a by-product of the task of repeating after a native model. Note also that participants were not explicitly told to "imitate" model. Therefore, we refrain from making strong claims on how participants would produce the target sounds in their everyday speech.

General discussion

The present study tested whether German learners of English can use articulatory information to establish lexical contrasts between novel words containing the difficult /ɛ/- /æ/ vowel contrast. Lexical access was measured using a visual-world eye-tracking task and a distinction between novel lexical items was attested by an asymmetry in fixation patterns as had been shown for real words containing this contrast (Weber & Cutler, 2004). Results showed that both passive exposure to visual articulatory information through audiovisual stimuli (i.e., videos of the speaker) and active target word articulation during word learning resulted in asymmetric fixation patterns during word recognition: /ɛ/-targets were fixated more readily than /æ/-targets during the time window of interest (200-800 ms). This can be taken as evidence that these participants encoded a difference between the difficult L2 vowels in the novel words they learned, as learners have been shown to do with real English words (cf. e.g., Weber & Cutler, 2004). Critically, the encoding of such contrast enabled listeners in the Video and Repetition condition to distinguish between the novel word pairs already by their first vowels (t[ɛ]nzer – t[æ]ndek). Participants in the baseline Audio-only condition, in contrast, did not show an asymmetric fixation pattern. They were as fast in recognizing /ɛ/-targets as /æ/-targets, without showing a preference for the category that better fits with the L1 (/ɛ/). This provides further support to the claim by Escudero et al. (2008) that novel words with confusable L2 sounds that are learned exclusively through auditory exposure cannot be reliably separated by their critical vowels. Taking this into consideration, we therefore suggest that added articulatory information during word learning can serve as a cue to establish lexical contrasts in newly-learned second language words.

In Experiment 1, listeners in the Video condition were provided with information on visual differences in one articulatory dimension of the /ɛ/-/æ/ contrast (i.e., jaw opening). Unlike the orthographic representations in an earlier study (Escudero, et al., 2008) visual articulatory information in the videos was dynamic and unfolded over time. Although on any given trial the maximum jaw opening was brief, the critical information appeared sufficient for listeners to pick up on, since the recognition pattern of the novel words during test was asymmetric. This indicates that listeners were able to make use of articulatory information that complemented the auditory cues in the signal to separate new lexical entries with the sounds of a difficult L2 contrast. The Video condition also adds to previous studies on the benefit of audiovisual information in L2 sound classification (Hazan et al., 2005; Hirata & Kelly, 2010; Navarra & Soto-Faraco, 2007; Wang et al., 2009) by demonstrating that additional visual information does also impact how lexical representations are established in the lexicon during L2 word learning.

Experiment 2 showed that active articulation of novel words with /ɛ/ and /æ/ during training allowed learners to encode a contrast between the two vowels in these words, just as did visual articulatory information in Experiment 1. However, there are some crucial differences between the two experiments that need to be considered. Participants in Experiment 2 were not exposed to any kind of additional external (e.g., visual) source of information during training. They just heard the native speaker produce the novel words and were asked to repeat them after a certain delay. Critically, an acoustic analysis of their productions of the words with the relevant L2 vowels showed that they were overall able to produce an acoustic difference between the two categories, even if not as big as the difference produced by the native speaker. Hence, lexical separation for target words with /ɛ/ and /æ/ for this group needs to be attributed to their ability to pick up contrastive information from their own productions (since for the Audio condition in

Experiment 1 exposure to an audio-only model of the native speaker was not enough to establish lexical contrast). Learners could have become aware that they were producing two different categories either by noticing that they systematically articulated them differently, by hearing themselves produce an acoustic difference between them, or likely by a combination of the two. Even though the present study is not able to tease these three possibilities apart, it clearly shows that listeners in the Repetition condition learned from the process of repeating the words that contained the two different sound categories. This subsequently led them to encode a contrast between /ɛ/ and /æ/ in the newly-formed lexical representations of the novel words, as evidenced by their asymmetric pattern of target fixations.

A point to be considered is that visual articulatory information and active production are likely to both focus the learners' attention on the identity of the sounds that form the words to be learned. This could be one underlying cause of the L2 learners becoming aware of the existence of two vowel categories. By focusing on the individual sounds in the novel words, with the help of videos or through repetition, their attention to the properties of the critical sounds was also enhanced. As a consequence, this could lead L2 learners to notice these differences, and use them for the recognition of the targets. Note that this account would also capture the findings concerning orthography in Escudero et al. (2008), since orthographic representations provide information about the novel words' individual sounds via sound-to-orthography mapping (Van Orden, 1987). Listeners in the baseline condition, on the contrary, were exposed to auditory stimuli only and were thus not guided to focus on the segments that formed the novel words. Consequently, they may have simply failed to gather reliable evidence on the existence of two categories because, without additional attention to the critical sounds, the perceptual difference between them was not salient or reliable enough to be picked up.

However, while there was no difference in how items with the two vowels were recognized in the Audio condition, note that lexical separation for /ɛ/ and /æ/ in the Video and Repetition conditions was instantiated as a delay to contact novel words with the new (weak) /æ/ category. Such a delay may ultimately constitute a disadvantage, although only a temporary one, for items with this vowel in terms of word recognition (Cutler, 2015). Considering this, the question is thus whether encoding the vowel contrast can indeed be seen as beneficial for L2 processing. We would argue "yes", at least from a developmental point of view, since the asymmetries in the Video and Repetition conditions constitute evidence for differential lexical encoding for the novel words. A separation of words and sounds in the lexicon should be a long-term goal in order to approximate native performance. The asymmetries indicate that listeners in the two experimental conditions encoded the vowels of /ɛ/-targets and /æ/-targets as two different categories, a strong /ɛ/ category and a weaker, fuzzier /æ/ category. Given that the native model is one where there are two categories that are both equally robust (Cutler et al., 2006; Darcy et al., 2013; Hayes-Harb & Masuda, 2008; Weber & Cutler, 2004), this separation is already one step further towards a native-like response to words with the English /ɛ/-/æ/ contrast than the no-distinction pattern observed in the Audio condition.

By examining the role of articulatory information in the lexical encoding of L2 contrasts, the present study contributes to our understanding of the relationship between L2 production and L2 perception at the word level. Both passive exposure to visual articulatory information in a video of the speaker and active word repetition during training were successful in triggering a distinction between trained novel words with /ɛ/ and trained novel words with /æ/. This finding shows that acquiring more accurate articulatory knowledge about difficult L2 sounds results in a better, although still non-native, encoding of these sounds in newly-learned words. In sum, we

conclude that production-related information has a positive impact on the perception of confusable L2 categories, and, most precisely, on their encoding to representations in the learners' second language lexicon.

References

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. S. (1998). Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores. In *Proceedings of ICSLP-1998*, 1-4, Sydney, Australia.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419-439. http://dx.doi.org/10.1006/jmla.1997.2558

Amengual, M. (2015). The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics*, 1-31. http://dx.doi.org/10.1017/s0142716415000557

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bent, T. (2005). *Perception and production of non-native prosodic categories* (Doctoral dissertation, Northwestern University).

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.S. Bohn & M. J. Munro (Eds.) *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/lllt.17.07bes

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer. Version 5.4.22, http://www.praat.org/.

Bohn, O. S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, *11*(03), 303-328. http://dx.doi.org/10.1017/s0142716400008912

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977-985. http://dx.doi.org/10.3758/bf03206911

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310. http://dx.doi.org/10.1121/1.418276

Carey, M. (2004). CALL visual feedback for pronunciation of vowels: Kay Sona-Match. *CALICO Journal*, 571-601.

Cutler, A. (2015). Representation of second language phonology. *Applied Psycholinguistics*, *36*(01), 115-128. http://dx.doi.org/10.1017/s0142716414000459

Cutler, A., & Otake, T. (2004). Pseudo-homophony in non-native listening. *The Journal of the Acoustical Society of America*, *115*(5), 2392-2392. http://dx.doi.org/10.1121/1.4780547

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269-284. http://dx.doi.org/10.1016/j.wocn.2005.06.002

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*(3), 372-420. http://dx.doi.org/10.1075/ml.8.3.06dar

de Jong, K., Hao, Y. C., & Park, H. (2009). Evidence for featural units in the acquisition of speech production skills: Linguistic structure in foreign accent. *Journal of Phonetics*, *37*(4), 357-373. http://dx.doi.org/10.1016/j.wocn.2009.06.001

Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association*, *27*(1-2), 47-55. http://dx.doi.org/10.1017/s0025100300005417

Díaz, B., Mitterer, H., Broersma, M., Escera, C., & Sebastián-Gallés, N. (2015). Variability in L2 phonemic learning originates from speech-specific capabilities: An MMN study on late bilinguals. *Bilingualism: Language and Cognition*, 1-16. http://dx.doi.org/10.1017/s1366728915000450

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*(6), 680-689. http://dx.doi.org/10.1017/s1366728915000450

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

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551-585. http://dx.doi.org/10.1017/s0272263104040021

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

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. *Crosscurrents in Second Language Acquisition and Linguistic Theories*, *2*, 249-289. http://dx.doi.org/10.1075/lald.2.15fle

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. 233-277). Timonium, MD: York Press.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437-470. http://dx.doi.org/10.1006/jpho.1997.0052

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /ɹ/ and /l/ accurately. *Language and Speech*, *38*(1), 25-55.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105-132. http://dx.doi.org/10.1016/s0010-0277(03)00070-2

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds “L” and “R”. *Neuropsychologia*, *9*(3), 317-323. http://dx.doi.org/10.1016/0028-3932(71)90027-3

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495-522. http://dx.doi.org/10.1017/s0142716403000250

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics*, *26*(4), 579. http://dx.doi.org/10.1017/s0142716405050319

Hayes-Harb, R., & Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research*, *24*(1), 5-33. http://dx.doi.org/10.1177/0267658307082980

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740-1751. http://dx.doi.org/10.1121/1.2166611

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. http://dx.doi.org/10.1016/j.specom.2005.04.007

Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. *The Journal of the Acoustical Society of America*, *133*(6), 4247-4255. http://dx.doi.org/10.1121/1.4802902

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099-3111. http://dx.doi.org/10.1121/1.411872

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning*, *17*(3-4), 357-376. http://dx.doi.org/10.1080/0958822042000319629

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. http://dx.doi.org/10.1044/1092-4388(2009/08-0243)

Inceoglu, S. (2015). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, 1-25. http://dx.doi.org/10.1017/s0142716415000533

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267-3278. http://dx.doi.org/10.1121/1.2062307

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446. http://dx.doi.org/10.1016/j.jml.2007.11.007

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. http://dx.doi.org/10.1121/1.4926561

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21-39. http://dx.doi.org/10.1016/j.wocn.2016.05.001

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: the use case WebMAUS. In *Proceedings of Digital Humanities Conference 2012*, 30-34, Hamburg, Germany.

Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a second language. In *Proceedings of Meetings on Acoustics, 6*(1), 1-15, Montreal, Canada. http://dx.doi.org/10.1121/1.3262006

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2), B33–B42. http://dx.doi.org/10.1016/s0010-0277(02)00157-9

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6), 1609–1631. http://dx.doi.org/10.1037/a0011747

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475-494. http://dx.doi.org/10.1016/j.jml.2007.11.006

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, 69(4), 527–545. http://dx.doi.org/10.1016/j.jml.2013.07.002

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. http://dx.doi.org/10.1007/s00426-005-0031-5

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8-13. http://dx.doi.org/10.1016/j.jneumeth.2006.11.017

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *Proceedings of the 12th Annual Conference of the*

International Speech Communication Association (Interspeech 2011), 161-164, Florence, Italy.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics,* 41(2), 101–116.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90(1), 51–89. http://dx.doi.org/10.1016/s0010-0277(03)00139-2

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the International Congress of Phonetic Sciences*, 607-610, San Francisco, USA.

Sebastián-Gallés, N. (2005). Cross-language speech perception. In D. Pisoni and R. Remez (Eds.) *The Handbook of Speech Perception* (pp. 546-566). Oxford, UK: Blackwell Publishing Ltd. http://dx.doi.org/10.1002/9780470757024.ch22

Sebastián-Gallés, N., & Baus, C. (2005). On the relationship between perception and production in L2 categories. In A. Cutler (Ed.) *Twenty-first century psycholinguistics: Four cornerstones* (pp. 279-292). Hillsdale: LEA.

Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language*, *52*(2), 240-255. http://dx.doi.org/10.1016/j.jml.2004.11.001

Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372-377. http://dx.doi.org/10.1111/j.1467-9280.2006.01714.x

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(03), 243-261. http://dx.doi.org/10.1017/s0142716400001417

van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, *15*(3), 181-198. http://dx.doi.org/10.3758/bf03197716

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607. http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.001

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. http://dx.doi.org/10.1016/j.wocn.2009.04.002

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033-1043. http://dx.doi.org/10.1121/1.1531176

Watson, C. I., Harrington, J., & Evans, Z. (1998). An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics*, *18*(2), 185-207. http://dx.doi.org/10.1080/07268609808599567

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1-25. http://dx.doi.org/10.1016/s0749-596x(03)00105-0

Wong, J. (2013). The effects of perceptual and or productive training on the perception and production of English vowels /ɪ/ and /i:/ by Cantonese ESL learners. In *Proceedings of*

the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), 2113-2117, Lyon, France.

Tables

Table 1

Results of the Growth Curve Analysis on the effects of Condition (Audio/Video) and Vowel (/ɛ/-/æ/) on Target fixations for each order of Time (Intercept, Linear, Quadratic).

	b	z	*p*
Intercept			
Condition	-0.18	-1.54	.12
Vowel	-0.09	-0.66	.51
Condition x Vowel	-0.15	-0.79	.43
Linear			
Condition	-0.37	0.19	.43
Vowel	0.15	0.35	.37
Condition x Vowel	-0.74	-3.52	< .001
Quadratic			
Condition	0.35	1.61	.11
Vowel	0.33	1.48	.14
Condition x Vowel	0.56	2.68	< .01

Table 2

Results of the separate Growth Curve Analyses for the Audio (left) and Video conditions (right) on the effects of Vowel (/ɛ/-/æ/) on Target fixations for each order of Time (Intercept, Linear, Quadratic).

	Audio condition			**Video condition**		
Vowel	*b*	z	*p*	*b*	z	*p*
Intercept	-0.01	-0.09	.93	-0.16	-0.77	.44
Linear	0.53	0.91	.36	-0.21	-0.39	.70
Quadratic	0.05	0.16	.87	0.62	2.16	< .05

Table 3

Results of the Growth Curve Analyses on the effects of Condition (Audio-Repetition on the left; Repetition-Video on the right) and Vowel (/ɛ/-/æ/) on Target fixations for each order of Time (Intercept, Linear, Quadratic).

	Audio-Repetition			**Repetition-Video**		
	b	z	*p*	*b*	z	*p*
Intercept						
Condition	-0.18	-1.70	.08	0.002	0.03	.98
Vowel	-0.10	-0.91	.37	-0.17	-1.38	.51
Condition x Vowel	-0.18	-1.18	.24	-0.03	-0.14	.88
Linear						
Condition	-0.33	-0.86	.39	0.03	0.08	.93
Vowel	0.23	0.53	.60	-0.16	-0.34	.73
Condition x Vowel	-0.60	-3.16	< .01	0.16	0.81	.42
Quadratic						
Condition	-0.01	-0.05	.96	-0.37	-1.54	.12
Vowel	-0.06	-0.21	.83	0.26	1.48	.14
Condition x Vowel	-0.13	-0.71	.48	-0.69	-3.52	< .001

Figure captions

Figure 1. Proportion of correct responses by Condition (Audio, Video) and Block Number (1-8) for training parts with two (left panel) and four alternatives (right panel).

Figure 2. Proportion of fixations over time to target, competitor and averaged distractors as a function of Condition (Audio –left panel, Video –right panel) and Vowel (/ɛ/ –in grey, /æ/ –in black). Vertical bars indicate the time window of interest (200-800 ms).

Figure 3. Fitted probability of fixating the target picture as a function of Condition (Audio –solid lines, Video –dashed lines) and Vowel (/ɛ/ –in grey, /æ/ –in black).

Figure 4. Proportion of correct responses by Condition (Audio, Repetition) and Block Number (1-8) for training parts with two (left panel) and four alternatives (right panel).

Figure 5. Difference between F2 and F1 in Hertz for vowel productions by the non-native speakers in Experiment 2 for /ɛ/ (in grey; median = 1087 Hz (sd = 180 Hz)) and /æ/ (in black; median = 951 Hz (sd = 165)). Dashed lines show the median of the native English speaker who recorded the stimuli (/ɛ/: 1177 Hz (sd = 70 Hz); /æ/: 832 Hz (sd = 66 Hz)).

Figure 6. Proportion of fixations over time to target, competitor and averaged distractors in the Repetition condition as a function of Vowel (/ɛ/ –in grey, /æ/ –in black). Vertical bars indicate the time window of interest (200-800 ms).

Figure 7. Fitted probability of fixating the target picture as a function of Condition (Left panel: Audio –solid lines, Repetition –dashed lines; Right panel: Video –solid lines, Repetition –dashed lines) and Vowel (/ɛ/ –in grey, /æ/ –in black).

Figures

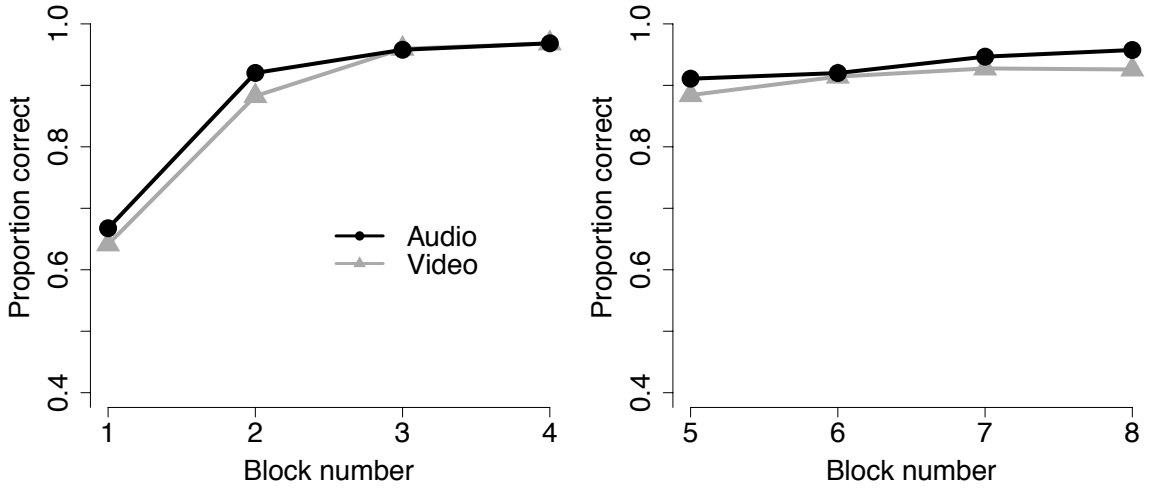

Figure 1

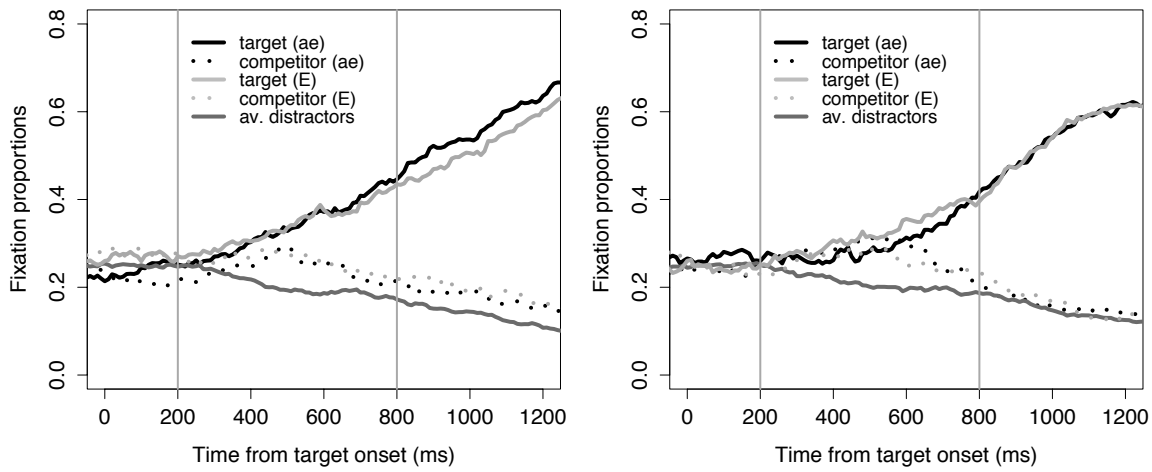
target (ae)
competitor (ae)
target (E)
competitor (E)
av. distractors
Fixation proportions
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
1200
Time from target onset (ms)
target (ae)
competitor (ae)
target (E)
competitor (E)
av. distractors
Fixation proportions
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
1200
Time from target onset (ms)

Figure 2

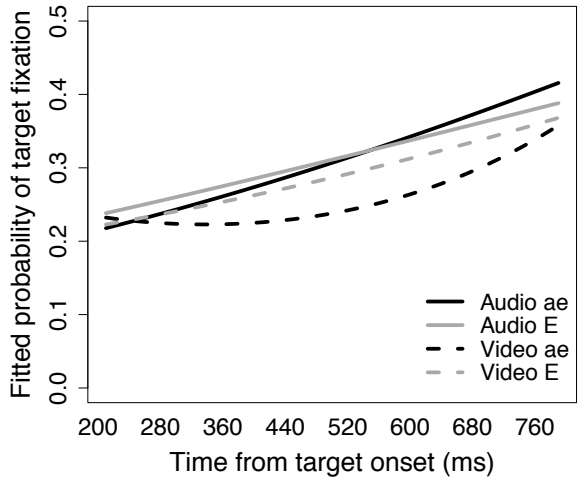
Fitted probability of target fixation
0.0
0.1
0.2
0.3
0.4
0.5
200
280
360
440
520
600
680
760
Time from target onset (ms)
Audio ae
Audio E
Video ae
Video E

Figure 3

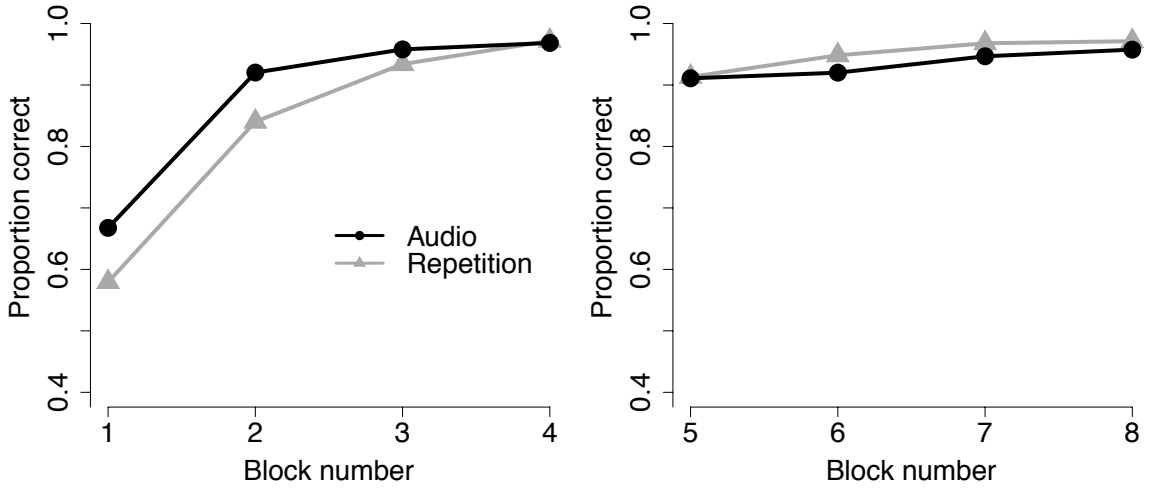
Proportion correct
0.4
0.6
0.8
1.0
1
2
3
4
Block number
Audio
Repetition
Proportion correct
0.4
0.6
0.8
1.0
5
6
7
8
Block number

Figure 4

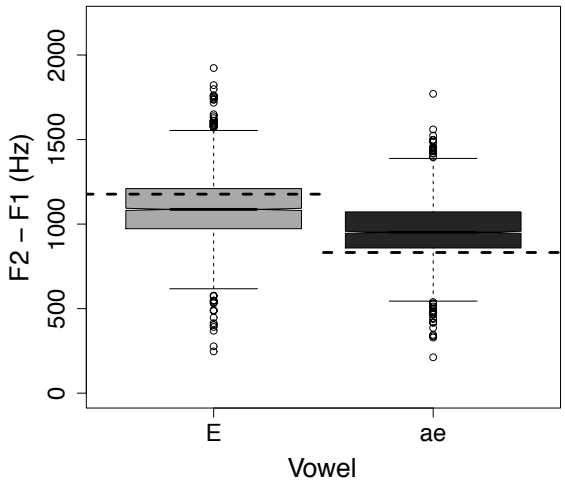

Figure 5

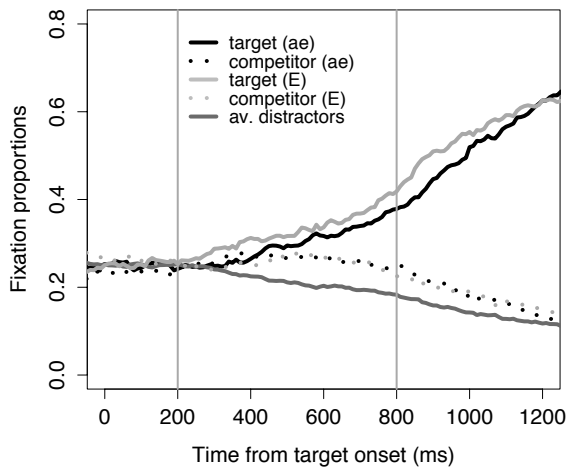
target (ae)
competitor (ae)
target (E)
competitor (E)
av. distractors
Fixation proportions
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
1200
Time from target onset (ms)

Figure 6

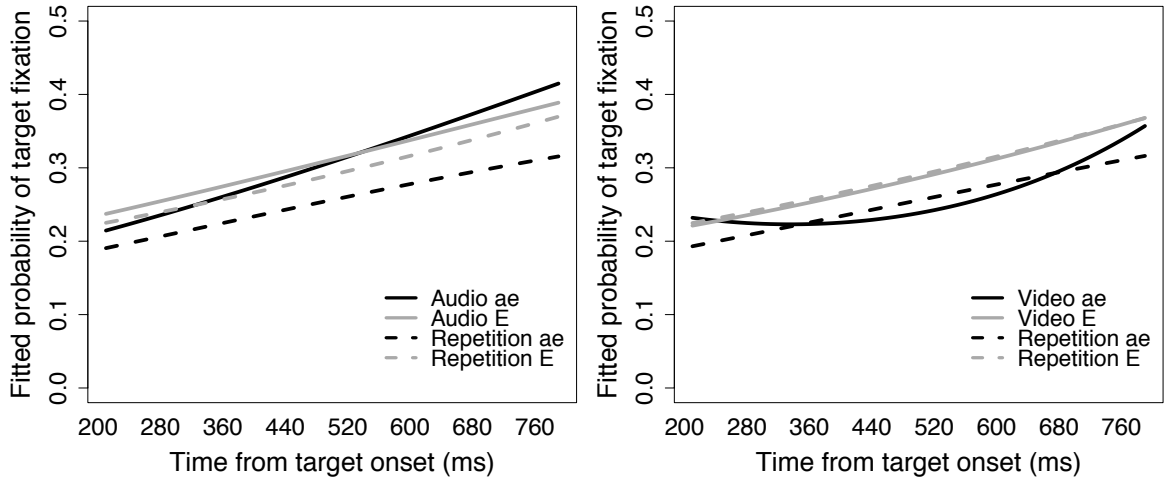
Fitted probability of target fixation
0.0
0.1
0.2
0.3
0.4
0.5
200 280 360 440 520 600 680 760
Time from target onset (ms)
Audio ae
Audio E
Repetition ae
Repetition E
Fitted probability of target fixation
0.0
0.1
0.2
0.3
0.4
0.5
200 280 360 440 520 600 680 760
Time from target onset (ms)
Video ae
Video E
Repetition ae
Repetition E

Figure 7

Appendix A

Mean self-rated English proficiency measures for the three groups of participants in Experiments 1 and 2. Ratings were elicited in response to the questions 'How good are you at Listening/Speaking/Reading/Writing in English?*' and* 'How good is your English overall?' *on a scale from 1 (native-like) to 7 (very poor). Standard deviations are in brackets.*

English Proficiency

Group	Listening	Speaking	Reading	Writing	Overall
Audio (Exp. 1)	2.45 (1.00)	3.30 (1.17)	2.25 (0.97)	3.10 (1.12)	2.80 (1.06)
Video (Exp. 1)	2.52 (0.93)	3.43 (1.43)	2.48 (1.08)	3.38 (1.28)	3.00 (1.30)
Repetition (Exp. 2)	2.47 (1.17)	3.37 (1.07)	2.47 (1.19)	3.10 (0.96)	2.83 (1.02)