ONOMÁZEIN
Journal of linguistics, philology and translation

PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE
FACULTAD DE LETRAS

English vowel perception by non-native speakers: impact of audio and visual training modalities

Yasna Pereira Reyes
Universidad de Concepción
Chile

Valerie Hazan
University College London
United Kingdom

ONOMÁZEIN 51 (March 2021): 111-136
DOI: 10.7764/onomazein.51.04
ISSN: 0718-5758

Yasna Pereira Reyes: Departamento de Idiomas Extranjeros, Facultad de Humanidades y Arte, Universidad de Concepción, Chile. | E-mail: yasnapereira@udec.cl
Valerie Hazan: Department of Speech, Hearing and Phonetics, Division of Psychology and Language Sciences, University College London, United Kingdom. | E-mail: v.hazan@ucl.ac.uk

Received: January 2019
Accepted: May 2019

51
March
2021

Abstract

Perception of sounds of a second language (L2) presents difficulties for non-native speakers which can be improved with training (Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997; Logan, Lively & Pisoni, 1991; Iverson & Evans, 2009). The aim of this study was to compare three different English vowel perceptual training programmes using audio (A), audiovisual (AV) and video (V) modes in non-native speakers with Spanish as native language (L1). 47 learners of English with Spanish as L1 were allocated to three different vowel training groups (AT, AVT, VT) and were given five training sessions to assess their improvement in English vowel perception. Additionally, participants were recorded before and after training to measure their improvement in the production of English vowels. Results showed that participants improved their perception and production of English vowels regardless of their training modality with no evidence of a benefit of visual information. These results also suggest that there is a lot of individual differences in perception and production of L2 vowels which may be related to a complex relation between speech perceptual and production mechanisms.

Keywords: visual information in L2 speech perception; English vowels; L2 speech perceptual training; perception and production link; individual differences.

1. Introduction

Speech perception in a second language (L2) is affected by the listener's native language (L1) (Best & Tyler, 2007; Flege, 1995), but improvement in perception can be achieved with perceptual training (Bradlow et al., 1997; Logan, Lively & Pisoni, 1991; Iverson & Evans, 2009). Although most L2 training studies have used auditory input, training using audio-visual material can have a positive impact on general L2 speech perception (Hardison, 1999; Navarra & Soto-Faraco, 2007; Wang, Behne & Jiang, 2008), particularly for sufficiently salient contrasts (Hardison, 2003; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006; Kawase, Hannah & Wang, 2014; Li, 2016). Yet, the number of studies on L2 English vowel perception using visual information is smaller (Aliaga-García & Mora, 2009; Flege, 1988; House, Beskow & Granström, 2001). The aim of this study is to compare the impact of three different training modalities: audio (A), audio-visual (AV) and video-alone (V), on the perception and production of English vowels by L2 learners with Spanish as L1.

1.1. Factors that intervene in L2 speech perception

Various factors have been shown to interfere with the perception of L2 sounds. The relation between L2 sounds and the learners' native categories affects the way in which novel L2 phonemes are perceived (Iverson & Evans, 2009). Perceptual distance of an L2 contrast from an L1 category may determine the degree of success in accurately perceiving the contrast and, if possible, in establishing new categories (Best, 1993; Best & Tyler, 2007; Flege, 1995). Spanish learners initially perceive English /iː/-/ɪ/ contrast as two instances of their L1 Spanish /i/. English /iː/ is perceived as closer to the Spanish /i/ category, whereas English /ɪ/ is perceived as a poor exemplar of Spanish /i/ (Flege & MacKay, 2004; Flege, Bohn & Jang, 1997; Fox, Flege & Munro, 1995; Morrison, 2008). Moreover, German learners of English achieved similar scores to American English native speakers in the perception of the English approximants /w/ and / j/, even though they lack this contrast in their L1, /w/ does not exist in German (Bohn & Best, 2010). An alternative account for L2 speech perception difficulties suggests that L1 language experience may prevent L2 sound category formation (Iverson, Khul, Akahane-Yamada, Diesch, Tohkura, Ketermmann & Siebert, 2001). The specialisation mechanism results in reduced perceptual sensitivity within the L1 phoneme inventory (Kuhl et al., 1992) and this may become an obstacle when learning non-native sounds in adulthood.

Besides the perceived distance between the L1 and L2 sounds, the size of the phoneme inventories may also play a role in L2 speech perception. L2 learners with smaller vowel inventories (e.g. Spanish L1) show poorer identification and learning performance than learners with larger vowel inventories (e.g. German L1) in Iverson and Evans' (2009) study. These results go against the prediction that L2 learners with a smaller vowel inventory would have more room to establish new categories in their perceptual space as predicted by the Speech Learning Model (Flege, 1995).

1.2. Issues in the use of visual cues to improve L2 speech perception

During L1 speech perception, both auditory and visual information are available and influence speech perception in face to face communication (Rosenblum, 2005, 2008). However, the weighting of these two sources of information may be affected by the visual salience of the talker's articulations, by the perceiver's visual speech-reading capacity, phonemic context of sounds and noise in the environment (McGurk & MacDonald, 1976; Massaro, Cohen, Gesi, Heredia & Tsuzaki, 1993; Owen & Blazek, 1985; Walden, Erdman, Montgomery, Schwartz & Prosek, 1981). Even though bimodality of speech in a native language has been well established (McGurk & MacDonald, 1976; Rosenblum, 2005, 2008), the focus of studies on L2 speech perception has centered mainly on the auditory channel. The effect of visual cues in L2 speech perception has been given less attention although some benefit has been found (Hardison, 1999, 2003; Hazan, Sennema, Iba & Faulkner, 2005; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006; Wang, Behne & Jiang, 2008).

One of the factors affecting the use of visual cues for L2 speech perception is the informational value of the cues. The extent to which L2 learners may attend to visual cues may vary depending on whether these cues add information to the contrast and as a function of their L1 perceptual categories (Hardison, 1996). The degree of salience of the visual cues is another influential factor; for example, visual information for bilabial or labio-dental contrasts is highly salient and affects perception more greatly than that for alveolar sounds (Hazan et al., 2006). This issue becomes relevant when studying to what extent visual cues can improve English vowel perception for non-native speakers.

Although visual cues have been shown to help the perception of difficult L2 contrasts, the weight of visual information may differ as a function of the phonemic status of the contrast. Consonant identification scores may improve in AV condition compared to the A, but confusions that are language dependant (like voicing or manner) do not improve with the addition of visual cues. For example, Spanish learners do not benefit from visual information to distinguish English contrasts /s/-/z/ that are allophonic in Spanish but phonemic in English (Ortega-Llebaria, Faulkner & Hazan, 2001).

Language experience can also influence the use of visual information for L2 speech perception (Wang, Behne & Jiang, 2009). Experience with the target language has an impact on learners; greater English experience (i.e. level of proficiency) has been found to correlate with greater use of visual cues (Wang, Behne & Jiang, 2008). Nonetheless, a possible age-related sensitive period for the learning of the use of visual information for L2 speech perception in adults suggests that the earlier the learners are in contact with the language, the more accurate interpretation of visual information is made (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastián-Gallés & Werken, 2013). Additionally, sensitivity to visual cues for speech perception may be lost if not used for perceiving L1 contrasts (Hazan, Sennema, Iba & Faulkner, 2005), thus making its use more difficult in the L2 context.

Further issues in the amount of use of visual information for L2 speech perception include whether the speaker is perceived as foreign (Chen & Hazan, 2007), familiarity with the talker (Hardison, 2006) and cultural aspects (Sekiyama, 1997; Massaro & Cohen, 1995). All the factors illustrate a wider picture of the complexity of using visual information when perceiving speech in a foreign language. It is important to highlight that most of the research on the use of visual cues in L2 speech perception has been based mainly on the perception of consonant contrasts, mostly using one-syllable words. Thus, more research is needed to establish to what extent visual information can help L2 learners to improve English vowels perception.

1.3. Perceptual training in L2 speech perception

Computer-based phonetic training can improve English vowel perception in L2 learners, with the most effective training method involving high-variability of speakers and tokens, the use of natural speech, immediate feedback and use of a training regime that trains a larger set of vowels (Lambacher et al., 2005; Iverson & Evans, 2009; Nishi & Kewley-Port, 2007). Such training may lead to generalisation to new tokens and new speakers, as well as retention of the learning after some months (Iverson & Evans, 2009). Perceptual learning has also been found to lead to improvements in the production of the trained sounds (Akahane-Yamada et al., 1996; Bradlow et al., 1997; Hazan et al., 2005; Lengeris & Hazan, 2010; Lambacher et al., 2005). Most of the participants in these studies have been L2 learners with little experience with the target language who can benefit fully from access to native speakers' input (Iverson & Evans, 2009). Moreover, even more experienced learners can benefit from high variability training for English vowel perception (Iverson, Pinet & Evans, 2012). One common feature of all these studies of vowel training is that they have been conducted using auditory input only. Yet, a different line of research using virtual tutors (Wik, 2011; Wik & Hjalmarsson, 2009) has explored the benefit of seeing an embodied conversational agent (ECA) in perceiving and producing Swedish vowel aspects like duration, and the impact of the feedback delivered by ECAs. It would be desirable to see if this type of training could be developed in the line of improving L2 learners' perception of English vowels.

The aim of the present study was to compare the impact of three vowel training modalities: auditory (A), audio-visual (AV) and a video-alone (V) training for English vowel perception of L2 learners with beginner level of proficiency and with Spanish as L1. The rationale behind the use of these three perceptual training approaches was that these training modalities may direct attention to different sources of information cueing English vowel perception. As a consequence, differences across training groups in the perception of English vowels in A, AV and V mode might appear at post-training testing. Additionally, the impact of perceptual training on the production of English vowels of L2 learners, judged by English native speakers, was also explored.

2. Method

2.1. Participants: L2 learners

Forty-seven L2 beginner learners (15 males, 32 females) were recruited and tested in Chile at Universidad de Concepción. Participants were first-year university students from an English teaching training programme who had just completed their first semester at university. They were aged between 18 and 24 years old (M: 19.6, *SD*: 1).

These participants had received intensive language training for a semester. All classes include a variety of speaking, listening, reading and writing activities with an emphasis on British English accent material. The participants' level of English proficiency was established by taking their final test results in the language course. This test includes measures of listening and reading comprehension, grammar and vocabulary knowledge and writing skills. Test scores were transformed into percentages. This information was provided by their teachers with the participants' consent. All participants reported to have normal hearing.

2.2. Procedure to form training groups

Participants were allocated to one of the three different training groups: audio training (AT), audio-visual training (AVT) or video-only training (VT) based on the results from the pre test audio (A) mode in the Vowel Identification test. This was done with the aim of balancing groups in terms of their pre test means and standard deviation. The balance across groups was however affected by the fact that eight participants who started the experiments did not complete all the training or missed the post-test. The AT group had 17 participants (M: 67, SD: 12), the AVT group had 14 (M: 61.3, SD: 13) and the VT group had 16 participants (M: 63, SD: 8) who completed all the pre- and post-tests, and training sessions.

2.3. Participants: English native speakers

A group of 20 (6 males, 14 females) English native speakers (ENS), with Standard Southern British English accent were tested in London at University College London in the Chandler House Speech Sciences Laboratory. They were university students, aged between 23 and 28 years old (*M*: 25.1; *SD*: 0.9). Participants received a small payment for their collaboration as regulated by UCL policies. All participants self-reported having normal hearing.

2.4. Vowel training programme

2.4.1. Materials

The Vowel Trainer developed by Iverson and Evans (2009) was chosen for the training phase as it had been shown to be highly successful in improving vowel perception in Spanish learners in previous studies. As the original trainer only used audio presentation, a new set of stimuli had to be developed in order to be able to include AV and V training. The same set of words and response labels was kept as in the original trainer to ensure comparability. The stimuli included in the training were real words containing 14 British English vowel sounds grouped into 4 clusters ([/e/, /ɑː/,/æ/, /ʌ/]; [/iː/, /ɪ/, /aɪ/, /eɪ/]; [/ɒ/, /əʊ/, /ɔː/]; [/uː/, /aʊ/, /ɜː/]) based on findings of confusions made by German and Spanish speakers in previous research (Iverson & Evans, 2007). Diphthongs were not part of this study but could not be removed from the original Iverson and Evans' Vowel Trainer software. That is why the clusters used in the Vowel Identification test will not be exactly the same as vowels (monophthongs) were grouped in a slightly different manner to exclude diphthongs. In the Vowel trainer, the first three clusters contain vowels that are mutually confusable and the remaining vowels form the last cluster. Test materials included 140 words (10 sets of minimal pairs per word/vowel). Video recordings were made by five SSBE native speakers (2 M, 3 F). The same video filming procedure was used as for the Vowel Identification test (Section 2.5.2). Each speaker recorded a randomised word list of 140 tokens twice. Video clips were compressed to .m4v files to be used as the material for the audio, audio-visual or video-only trainer. To obtain the audio, audio-visual and video only of each token, a script line was added to Iverson's Vowel Trainer to retrieve the A, AV or V component of the AV tokens.

2.4.2. Procedure

The training programme consisted of five sessions of high-variability phonetic training (HVPT, Logan, Lively & Pisoni, 1991) given to participants on a desktop computer and headphones. At each session, 225 training tokens were presented, a different talker was used per session, alternating from female to male. Each training session lasted between 45 to 60 minutes. At each session, the same 225 tokens were presented in audio (Audio Training, AT), audio-visual (audio-visual training, AVT) or video-only (video training, VT) mode, depending on the training group participants were allocated to.

The training software uses an adaptive procedure. In the first phase of the session, 70 fixed tokens are presented (five repetitions of words including the 14 vowel sounds). In the second (adaptive) phase, 85 tokens based on the most common mistakes from the first phase are presented. In the final phase, a further 70 fixed tokens are presented (five repetitions of words including the 14 vowel sounds). For further details of the training software, see Iverson and Evans (2009). No more than two sessions per week could be taken due to participants' availability, so the five sessions were completed over three weeks.

In the training session, participants heard or watched a speaker saying a word and had to click on the correct answer choosing from the alternatives (3 or 4 CVC-words differing in the vowel) that appeared on the screen some seconds after the stimulus was presented. All alternatives were accompanied by a "help-word" on the side; these were simple words that contained the same vowel as the stimulus tested and could help with the pronunciation of less familiar words in the stimulus list. Feedback was provided on the computer screen as to whether the answer was correct or incorrect. If correct, a “Yes” prompt appeared on the screen and a cash register sound was heard followed by the correct response—heard/seen, once more. A “Wrong” prompt was shown when an incorrect response was chosen, followed by two tones with descending pitch; then the correct word was heard/seen once, followed by a sequence of 4 stimuli: correct-incorrect-correct-incorrect word. Participants could see their final score at the end of each session.

2.5. Vowel Identification test

2.5.1. Materials

The test included audio-visual stimuli of 11 English vowels (/æ/, /ʌ/, /ɑ:/, /ɪ/, /i:/, /e/, /ɜ:/, /ɒ/, /ɔ:/, /ʊ/, /u:/) embedded in /bVt/ and /hVd/ words. Four native speakers of Southern British English (2 males, 2 females) recorded a list of 61 randomized words containing these English vowels in the two contexts (e.g. “bat”, “had”, “bet”, “head”). Non-words were excluded from the list (no word for the pronunciation /bʊt/ was found). Three repetitions per word were included and the clearest iteration of each word was selected for inclusion in the test materials. Another list of words containing the same 11 English vowels was filmed with a different SSBE male speaker reading a randomized list of 33 frequent words. These words were used as examples before the Vowel Identification test started (*cat, cup, card, sit, pet, feel, word, pot, caught, full* and *food*) and they were also used as the “response button word” in the test. This was done as some of the /bVt/ and /hVd/ words were of low frequency of occurrence and may not have been known by the L2 learners.

The video recordings were made in a sound-proof room using a Canon XL-1DV video camera; the speaker’s head was set against a blue background and was fully visible. Each individual video clip was edited so as to have a start and end point with a neutral facial expression. A selection of 21 tokens was made from each speaker; each vowel in two contexts (“hVd”, “bVt”-words), except for /ʊ/ (only “hVd”).

2.5.2. Procedure

A decision was made to present vowels in three separate sets with three or four response alternatives rather than as a single test with 11 response alternatives as it was felt that this would have been too confusing and taxing for participants. One disadvantage of this approach is that it constrains the choice of responses for a given vowel, but the response set was based on the most common confusions of monophthongs reported in studies on English vowel perception by Spanish speakers learning English (Garcia-Lecumberri & Cenoz, 1997; Ortega-Llebaria et al., 2001; Iverson & Evans 2007), thus mitigating this issue.

The three vowel sets used were: set 1 [/æ/, /ʌ/, /ɑ:/], set 2 [/ɪ/, /i:/, /e/, /ɜ:/] and set 3 [/ɒ/, /ɔ:/, /ʊ/, /u:/]. Familiar words were used as “response buttons” in the practice phase and in the test itself to facilitate the response procedure. The labels used were: [cat, cup, card] for set 1, [sit, pet, feel, word] for set 2 and [pot, caught, full, food] for set 3. The response buttons appeared on the screen after hearing or watching the /b-V-t/ and /h-V-d/ words. The researcher made sure participants had understood the test procedure after their practice session. No one seemed to have difficulties with following the instructions.

There were 84 tokens per mode presented in: audio (A), audio-visual (AV) and video-only (V), giving a total of 252 stimuli. The presentation order (A-AV-V or AV-A-V) was counterbalanced across participants, with the V condition always presented last as it was the most difficult condition. Before the test-phase started, participants were presented with tokens in A, AV and V mode with the English words (male speaker) used as the response labels for sets 1, 2 and 3. Participants practiced once with these token to get familiarised with the test procedure. After that, they would see/hear different words produced by 4 other people (2 males, 2 females) containing the same English vowels as in the response labels and they had to click on the response label with the word containing the vowel they had just heard/seen. No feedback on their answers was given. This test took around 30 minutes.

For presentation of the Vowel Identification test to English native speaker (ENS) participants, pink noise was added in order to avoid ceiling effects; the signal to noise ratio (SNR) was set at -10dB following piloting to aim for similar intelligibility levels in the A condition between ENS and L2 participants.

The order in which the tests were given to L2 learners at pre- and post-test was: a) first: Vowel Identification test and b) recordings of words in carrier sentences; after the pre-test, c) five Vowel Training sessions were given to participants.

3. Results

To evaluate the effect of training modality (AT, VT, AVT) and test presentation mode (A, V, AV) on vowel identification, the data for the pre- and post-training Vowel Identification tests for L2 learners is presented first in Section 3.1. After that, the vowel identification data for English native speakers (ENS) will be presented in Section 3.2 to evaluate to what degree visual information contributes to correct vowel identification for native listeners. Finally, the data for the goodness rating test for the L2 participants' production of English vowels pre- and post-training is shown in Section 3.3.

3.1. Vowel Identification by L2 learners

3.1.1. Effect of phonetic training on English vowel identification by L2 learners

The data from the Vowel Identification test (pre- and post-training) was used to analyse the impact of the three vowel training modality on vowel identification in L2 learners. A logistic regression analysis was used in R (glmmPQL) with time (pre-post-training), group (AT, VT and AVT training groups), test presentation mode (A, AV, V) and vowel (11 monophthongs) as fixed effects; participant and stimulus were treated as random factors. There was a significant effect of Time: vowel identification in the post-test (*M*: 64, *SD*: 24.2) was better than in the pre-test (*M*: 58, *SD*: 25.5), showing overall significant improvement in English vowel identification after training (Table 1). All significant effects were analysed with a logistic regression analysis (glmmPQL function). One key question is whether one training modality was more effective than the other two. There was no significant effect of group (Training Group) nor any other significant interaction that involved group. This suggests that there was no difference in the effectiveness of auditory, visual or audio-visual training and the vowel identification capacity that L2 learners achieved after training (Table 1).

The effect of mode was also significant, results revealed that there were no significant differences in scores between the A and AV mode, but scores in both were higher than in the V mode (Table 1). There was also a time*mode interaction; though all modes improved significantly after training, the V mode showed smaller amount of improvement (Table 1, Table 2).

TABLE 1

Vowel Identification test (pre and post). Fixed effects for vowel identification in A, AV and V mode for the three training groups (AT, AVT, VT).

FIXED EFFECTS	
Group	F(2,22984)= 1.249, p>.05
Time	F(1,22984)= 68.953, p<.001
Mode	F(2,22984)= 391.064, p<.001
Vowel	F(10,460)= 64.726, p<.001
Time*group	F(2,22984)= 1.308, p>.05
Time*mode	F(2,22984)= 8.973, p<.001
Time*vowel	F(10,22984)= 1.635, p>.05
Group*mode	F(4,22984)= 1.103, p>.05
Group*vowel	F(20,22984)= 1.437, p>.05
Vowel*mode	F(20,22984)= 20.054, p<.001
Time*group*mode	F(4,22984)= .335, p>.05
Time*group*vowel	F(20,22984)= .685, p>.05
Time*mode*vowel	F(20,22984)= .425, p>.05
Group*mode*vowel	F(40,22984)= 1.313, p>.05
Time*group*mode*vowel	F(40,22984)= .544, p>.05

Percentage of improvement relative to pre-test was estimated for A (M: 14.7%, SD: 14), AV (M: 14.2%, SD: 15) and V mode (M: 6.6%, SD: 13.4). Vowel identification improved to the same degree in A and AV modes and more so than in the V mode. There was a significant effect of vowel (Table 1), overall means varied from 45% correct identification for vowel /ʊ/ to 75% for /ɜː/. There was also a vowel*mode interaction (Table 1) which showed that vowel identification in V mode had, in general, lower scores than in A and AV mode. This interaction was due to different patterns found across vowels. For instance, the scores in A and AV mode were very similar for some vowels (/ɑː/, /iː/, /e/, /ɜː/, /ɒ/, /uː/). There were some vowels with overall higher identification scores in A mode (/ɪ/, /ɔ/, /ʊ/) and some other which were better identified in AV mode (/ʌ/) or showing no difference in scores per mode (/æ/).

TABLE 2

Vowel Identification test (pre- and post-test). Vowel identification means (%) per mode and standard deviations (M, SD) for the pre- and post-test, comparisons between overall mode scores and mode*time interaction

OVERALL (M, SD)	PRE AND POST (M, SD)	TIME*MODE
A (68.4; 9.6)	Pre (64.3; 11.1), Post (72.2; 8.1)	F(1,7324)= 50.482, p<.001
AV (70; 9.5)	Pre (63.2; 10.4), Post (71; 8.6)	F(1,7324)= 45.559, p<.001
V (47.4; 5.5)	Pre (45.3; 5.1), Post (48.2; 6.9)	F(1,7324)= 5.160, p<.05
A – AV: F(1,15154)=.069, p>.05 A – V: F(1,15154)=518.166, p<.001 AV – V: F(1,15154)=527.840, p<.001		

The overall percentages of pre-/post-training change for the Vowel Identification test for each training group were also estimated [((post-test - pre-test)/pre-test)*100] for each of the modes included (A, AV, V), see Table 3. The amount of improvement in the perception of English vowels in A mode for all groups after training (post-test) is similar to that reported in Iverson and Evans (2009) in which Spanish learners improved 10% and Germans 20% in their post-test vowel identification task, after five sessions of auditory vowel training.

TABLE 3

Percent change (improvement) relative to pre-test for the Vowel Identification (VID) test per mode (A, AV, V).

TRAINING GROUPS	VID TEST: A % CHANGE (SD)	VID TEST: AV % CHANGE (SD)	VID TEST: V % CHANGE (SD)
AT group	14.2% (16.7)	16% (15)	9.2% (14)
AVT group	17.5% (15)	15.8% (18.8)	4.4% (13.7)
VT group	12.6% (9.8)	10.8% (10.8)	5.8% (12.8)

In summary, the three training groups improved their overall English vowel identification to a similar degree following training, regardless of whether they were presented with audio-visual, audio alone or video alone stimuli during training. Concerning their use of auditory and visual cues, participants showed no overall greater benefit of visual information in AV mode compared to A, as their scores at pre- and post-test revealed no significant difference between A and AV mode (Table 2).

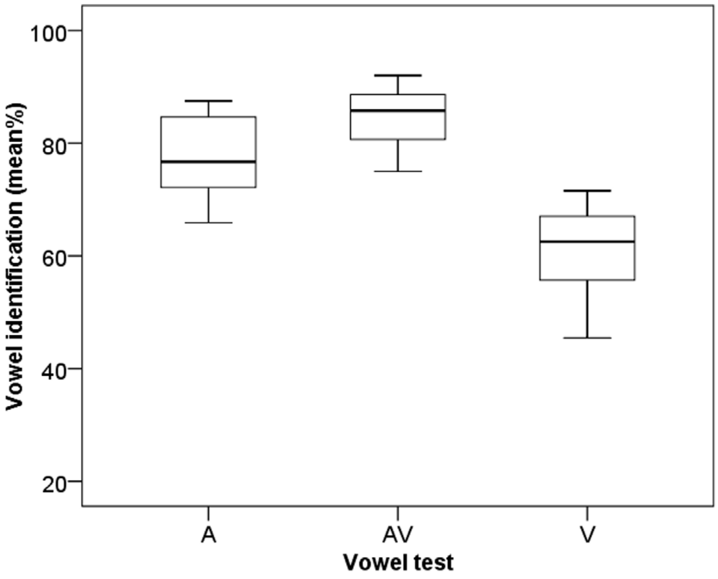
3.1.2. L2 vowel identification and learners' level of proficiency

Participants' level of proficiency was obtained by transforming their English module final mark into a percentage (1-100). This measure of proficiency level was used to run correlations with their vowel identification accuracy obtained from the scores of the Vowel Identification test before and after training (pre-, post-tests). The results showed a strong correlation between the level of proficiency and the vowel identification capacity before (r= .749, N=47, p<.001) and after training (r=.642, N=47, p<.001). This finding may indicate that L2 learners' vowel identification capacity is related to the learners' overall amount of knowledge and experience with the L2 language (English).

3.2. English vowel identification by English Native Speakers (ENS)

FIGURE 1

Vowel identification by English native speakers (ENS) in the Vowel Identification test in A, AV and V mode



For data from ENS participants, a logistic regression analysis was run in R (glmmPQL function) with mode, vowel and mode*vowel interaction as fixed effects; participants and stimulus were random factors. The results (Table 4) showed a significant effect of vowel with mean scores that ranged from 52% (/ʊ/) to 88.5% (/æ/). The mode effect was significant (Table 4); there were significantly higher results in AV mode (M: 87, SD: 4.7) than A (M: 78, SD: 6.7) and V mode (M: 61.4, SD: 7.2). However, this effect was modified by a mode*vowel interaction. An examination of the data revealed some vowels obtaining similar scores in two or the three modes. For example, vowel /ɒ/ had the same mean scores in A and AV mode (M: 89), and the scores for vowel /ʊ/ did not differ in A, AV and V mode (M: 52). Post hoc tests (logistic regression analysis in R, glmmPQL function) with vowel and mode, and pair-wise comparisons. Although results were generally higher in AV than in A mode, the AV-A difference was only significant for four vowels /ɪ/, /ɜː/, /uː/, /ʌ/.

Altogether, these results showed that native participants could benefit from the visual information available for vowels in AV mode compared to A (in noise), facilitating the identification of at least a subset of vowels. Scores in Video (V) mode were mostly above chance level; this information will be presented in relation to L2 learners' results below (Figure2).

TABLE 4

Results for the Vowel Identification test data for English native speakers (ENS) on 11 English vowels. Vowel identification was measured in three modes: audio, audio-visual and video mode

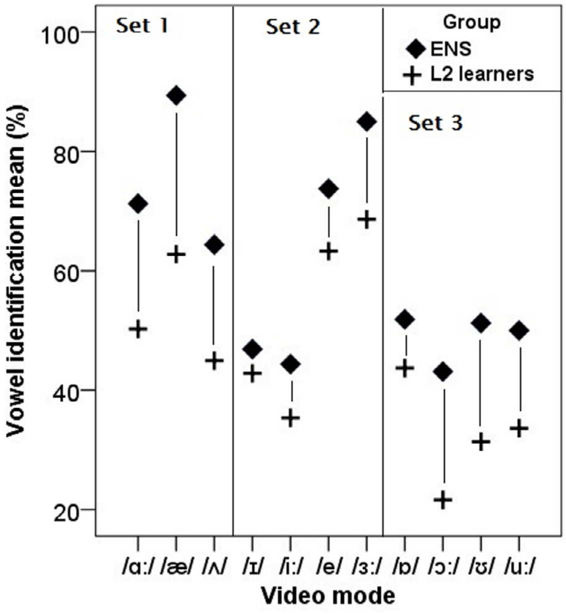
FIXED EFFECTS	
Mode	F(2,4988)= 128.864, p<.001
Vowel	F(10,4988)= 14.172, p<.001
Mode*vowel	F(20,4988)= 10.539, p<.001

To assess native and non-native perceivers' vowel identification based on visual information only (prior to any training for the L2 learners), the data in the V mode of the Vowel Identification test was analysed comparing ENS and L2 learners. A logistic regression analysis was run in R (glmmPQL function) with group, vowel and group*vowel interaction as fixed effects and participants as random factor. The results showed that the group effect was significant: vowel identification score for ENS participants (M: 61, SD: 7) was higher than for L2 learners' (M: 46, SD: 7). The vowel effect was also significant; overall means per vowel ranged from 28% (/ɔː/) to 73% (/ɜː/). The group*vowel interaction was significant; ENS' scores were significantly higher than the L2 learners' results for eight vowels. Similar scores were found for vowels /iː/, /ɪ/ and /ɒ/ (Figure 2). In spite of these differences, L2 learners showed a similar pattern for vowel identification to that of ENS participants in this V condition, as can be observed in Figure 2. All significant effects were analysed with a logistic mixed-effects model (glmmPQL function).

These results suggested that, within the constraint of the vowel sets within which they were presented, English vowels can be identified above chance level by L2 participants using visual information only, although identification scores were substantially lower than those obtained by ENS participants in the V condition. The vowels that appeared to be more visually distinctive for ENS participants were also the ones for which higher scores were obtained for L2 learners.

FIGURE 2

Vowel identification in Video (V) mode by ENS and L2 learners. The vertical lines in the graph show the separation per set, the way vowels were presented in the Vowel Identification test



3.3. L2 Learners' vowel production

3.3.1. Recording materials

The L2 learners who took part in the perceptual training sessions recorded 33 sentences before and after the training (pre- and post-test) to provide materials to be used in a goodness rating test given to English native listeners. There was one CVC keyword per sentence for each of the 11 English vowels included in this study (/iː/, /ɪ/, /e/, /ɜː/, /æ/, /ʌ/, /ɑː/, /ɔː/, /ɒ/, /uː/, /ʊ/). The keywords were presented in a carrier sentence (e.g. "The word in the box is *cap*"). The full set of words is as follows: [cap, flag, sand], [cut, luck, sun], [car, park, part], [seat, peach, beach], [kick, tin, bin], [blue, shoes, food], [book, push, foot], [surf, girl, word], [net, red, shell], [sock, dog, rock] and [shorts, shore, ball].

The recordings were made individually in a quiet room in the Spanish Phonetics Laboratory at Universidad de Concepción, Chile. The sentences were displayed randomly on a laptop computer using a Powerpoint presentation with a 10 second-interval between sentences. A digital voice recorder (Roland R05) and an external connected microphone (Behringer

XM1800S) were used for the recordings which were made at a sampling rate of 44.1 kHz and recorded in mono sound.

The vowel in each keyword was tagged using the Praat software version 5.1.24 (Boersma & Weenink, 2012). The start and end points of each vowel were chosen, including the transitions from the preceding and following consonant to avoid having too short a stimulus and interfere with their rating. Once vowels were tagged, a script was used to extract the vowel from the initial recording and create a new sound file; markers were placed at zero crossings to avoid discontinuities (using a script in Praat). The number of vowel tokens obtained per participant was 33 at pre- and 33 at post-test. The decision of using only the extracted vowel instead of the keyword as stimuli for the goodness rating test was made to avoid raters being influenced by the mispronunciation of the preceding or following consonants. As the participants had a beginner level of proficiency, they were highly likely to produce foreign accented consonants which could have made the judgement of vowel quality more difficult to English native speakers.

3.3.2. Procedure

To obtain a measure of the accuracy of the English vowels produced by the L2 learners, vowels produced by the learners before and after the vowel training sessions were presented to English native speakers (ENS) in a goodness rating test. Six tokens per vowel (vowels extracted from keywords) from each participant were selected (3 from pre-test, 3 from post-test). The 3102 tokens (6 tokens x 11 vowels, x 47 participants) were randomised to create three different tests given to the ENS raters in three separate sessions of approximately 35 minutes each. All raters heard all the tokens. This test was given individually using a laptop and headphones that allowed choosing the volume at a comfortable level. The test was presented in an MFC Praat experiment format. Listeners rated the vowel sounds they heard. A screen showed the prompt “How good an exemplar is the sound you hear of the vowel in *CAT*?” The rating scale went from 1 (very poor) to 7 (very good) and raters had the option to hear the token again if needed. A break option was introduced every 10 minutes.

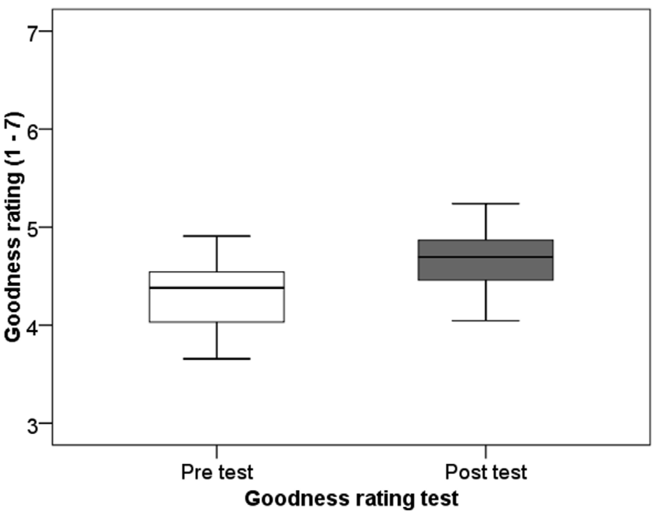
3.3.3. Results

The vowels produced by L2 learners before and after training were given to ENS in a goodness-rating test. To explore whether the 11 ENS raters had been consistent in their ratings, a reliability analysis using an Intra-class Correlation Coefficient (ICC) was run on the scores given by the 11 raters to the pre- and post-test vowel tokens. A two-way mixed model (with raters as fixed component) was chosen with a level of “absolute agreement” to be tested. The results showed a strong between-rater consistency with a Cronbach’s Alpha α=.844; values of. 70 and above are considered good indicators of consistency.

A linear mixed-effect model was run on the ratings data using the R software (lme function). The fixed-effects introduced in the model were training group (AT, AVT, VT), time (pre-/post-test)

FIGURE 3

Boxplots using the overall means in the goodness rating test before (pre-test) and after training (post-test). 11 English native speakers (ENS) rated L2 learners' vowel production with a scale from 1 (poor) to 7 (very good)



and vowel (/iː/, /ɪ/, /e/, /ɜː/, /æ/, /ʌ/, /ɑː/, /ɔː/, /ɒ/, /uː/, /ʊ/). Participant was treated as a random effect. The analysis (Table 5) revealed no significant effect of training group. There was an effect of Time (Figure 3): ratings were higher for the post test recordings (pre M: 4.3, SD: .048; post M: 4.66, SD: .044), indicating improvement in the goodness ratings of English vowel production by L2 learners. The vowel effect was also significant, but it was modified by a vowel*time interaction. The two-way interaction was due to some vowels showing no improvement after training (/ɪ/, /ʌ/, /ʊ/), though their overall mean scores were already high in the pre-test. The remaining eight vowels showed significant improvement in their ratings after training (Table 6).

TABLE 5

Results for the goodness rating test. L2 learners' vowel production from the pre- and post-test recordings from 11 English monophthongs were assessed by 11 ENS raters. Group in the table refers to training group (AT, AVT or VT)

EFFECT	
Group	F(2,44)= 1.443, p>.05
Time	F(1,924)= 67.115, p<.001
Vowel	F(10,924)= 23.814, p<.001
Group*time	F(2,924)= 2.020, p>.05
Vowel*group	F(2,924)= 0.991, p>.05
Vowel*time	F(10,924)= 2.582, p<.05
Vowel*group*time	F(20,924)= 0.769, p>.05

TABLE 6

Mean (M) rating (from 1 to 5) and standard deviation (SD) for vowel production in pre- and post-test per vowel (11) and level of significance for the vowel*time interaction (time effect)

VOWEL	PRE-TEST M (SD)	POST-TEST M(SD)	TIME EFFECT
/ɑː/	M: 4.7 (.69)	M: 5.4 (.67)	F(1,46)= 44.903, p<.001
/æ/	M: 3.9 (.94)	M: 4.5 (.82)	F(1,46,)= 13.780, p<.001
/e/	M: 3.7 (.68)	M: 4.1 (.87)	F(1,46,)= 9.824, p<.05
/ɜ/	M: 3.8 (.57)	M: 4.2 (.60)	F(1,46,)= 18.299, p<.001
/ɪ/	M: 4.7 (.69)	M: 4.7 (.82)	F(1,46,)= .001, p>.05
/iː/	M: 4.6 (.73)	M: 4.9 (.76)	F(1,46,)= 5.285, p<.05
/ɔː/	M: 4.3 (.94)	M: 5.0 (.91)	F(1,46,)= 23.243, p<.001
/ɒ/	M: 4.2 (.66)	M: 4.6 (.73)	F(1,46,)= 14.740, p<.001
/ʊ/	M: 4.5 (.57)	M: 4.6 (.62)	F(1,46,)= 1.051, p>.05
/uː/	M: 4.5 (.66)	M: 4.8 (.52)	F(1,46,)= 6.596, p<.05
/ʌ/	M: 4.1 (.69)	M: 4.2 (.72)	F(1,46,)= 2.445, p>.05

Overall, the results of the goodness rating test showed that L2 learners improved the quality of their English vowel production after training in a way that was perceptible by native listeners (ENS). There were three vowels (/ɪ/, /ʊ/, /ʌ/) which did not show significant improvement, but they were already rated as good productions before training in the pre-test.

3.4. Perception and production relation

To establish whether there was any relation between the L2 learners' English vowel perception and production either prior to or following training, the overall vowel identification scores (pre- and post-test) from the Vowel Identification test and the goodness-rating test (vowel production) were used in a Pearson product-moment correlation analysis. The results showed that L2 learners' vowel perception (vowel identification scores) was not correlated with the quality of their vowel production (goodness-rating scores) at pre-test (i.e. before training) and only a weak correlation was found at post-test (after training).

Initial English vowel production (pre-test) was significantly correlated with the post-test production ratings (r: .579, N: 47, p< .001). That is to say, the ranking of learners' quality of vowel production remained similar after training. The pre and post vowel identification scores did not seem to be related to the learners' capacity to produce English vowels. These findings suggest that those participants who showed better perception of English vowels did not necessarily obtain the highest ratings for their vowel production.

To find whether the participants' English vowel production was related to their overall level of language proficiency, a Pearson-moment correlation coefficient was estimated for pre- and post-test productions using the goodness ratings and the learner's proficiency scores, per training group. The results revealed a weak correlation between vowel production and proficiency level before (r: .339*, N: 47, p: .020) and also after training (r: .379**, N: 47, p: .009). Therefore, participants' level of proficiency could only account for a small amount of variability in the vowel production data.

4. Discussion

The main aim of this study was to evaluate the use of visual cues to vowel perception by L2 learners. For this purpose, three vowel training modalities were used to investigate the impact on the effectiveness of computer-based phonetic training aimed at improving L2 vowel identification accuracy. Besides, the impact of L2 perceptual training on L2 vowel production was also analysed.

Research using training has developed from using small sets of vowels to testing the whole vowel system and using multiple speakers mainly in auditory mode (Nishi & Kewley-Port, 2007; Iverson & Evans, 2007; Wang & Munro, 2004). To our knowledge, this study is novel as it used three different types of high-variability vowel training modalities (auditory, audio-visual y video-alone). Based on studies that have explored the contribution of visual information to improve speech perception (Hardison, 1999; Hazan et al., 2005; Wang et al., 2008), it was expected that the AV and the V training groups would show some benefit from the visual information available in their training. If they had learnt to use visual information for vowel perception, they would have obtained higher scores than the A training group in vowel perception in AV mode in the Vowel Identification test.

The three different training modalities (AT, VT or AVT) used in the present study focused the L2 learners' attention on either acoustic or visual cues to vowel distinctions. The results in vowel identification accuracy showed an overall improvement after training, although identification in AV mode did not show any advantage for the two groups that had access to visual information during their training sessions. There was no greater A-AV mode difference for the AVT and VT groups compared to the AT group. All three training groups showed greater reliance on the auditory input for vowel identification with similar amounts of improvement after training. This lack of a training modality effect has been reported on studies that examined the improvement in L2 speech perception after training using auditory input (Iverson et al., 2005; Hazan et al., 2005). The improvement in vowel identification observed after training is in line with previous studies which have reported improvement in perception capacity and evidence of retention even months after L2 perceptual training (Bradlow et al., 1999; Iverson & Evans, 2009; Lively et al., 1994; Nishi & Kewley-Port, 2007).

Learners trained with visual cues only (VT group) were also able to improve their auditory perception of vowels at a similar level of the other groups (AT, AVT) who did have access to auditory cues. This would suggest that training which provides access to the visual articulatory gestures for English vowel phonemes may also contribute to improve English vowel perception in L2 learners. To our knowledge, this type of training for the perception of English vowels has not been put to test before in previous studies and may need further research.

The learners' lack of integration of visual cues for English vowel perception found in the present study suggests that attending and learning to use visual information in an L2 seems more difficult than attending to the auditory cues and may require a different type of training. This difficulty in attending to visual speech for L2 vowels may be due to lack of experience with visual cues in their L1 (Hazan et al., 2006; Wang et al., 2008). As participants in this study (Spanish L1) have a vowel inventory with few and clearly discriminable vowels (/i/, /e/, /a/, /o/, /u/), visual cues to vowel identity in Spanish may carry little weight; visible features such as lip-rounding do not distinguish vowel contrasts in Spanish. As a consequence, when perceiving vowels in an L2, Spanish-L1 learners of English may focus on the channel that provides more information in the L1 for vowel perception—the audio channel.

The L2 learners' proficiency level showed significant relation with English vowel perception before training; this relation was weaker after training but still significant. However, the lack of training mode effect also revealed that language knowledge (i.e. higher proficiency level) was not related to more or less use of visual information for vowel perception. Unlike results in some previous studies which have found that lower proficiency learners were more likely to use visual information for L2 speech perception (Sueyoshi & Hardison, 2005). Other studies have shown just the opposite; Wang et al., (2008) found that more proficient learners made more use of visual cues for L2 speech perception. Unfortunately, in the current study no evidence of audio-visual benefit was found for beginner learners as perception in A or AV mode did not differ.

Concerning the impact of perceptual vowel training on the L2 learners' production of English vowels, the results showed an overall improvement in vowel production quality following five sessions of perception training, as confirmed by improved vowel ratings by native listeners (ENS). Similar findings have been reported after perceptual training for English consonants (Akahane-Yamada et al., 1996; Bradlow et al., 1997; Hazan et al., 2005) and vowels (Iverson et al., 2012; Lambacher et al., 2005; Lengeris & Hazan, 2010), even when no explicit production training was given. These findings have led researchers to suggest that the improvement in perception and production may reflect perceptual changes after training and a link between these two speech abilities. However, there is no agreement on how exactly this relation works (Iverson & Evans, 2009). Studies which have focused on production training have found that L2 learners' speech production actually improves but speech perception does not (Alshangiti & Evans, 2014; Hattori & Iverson, 2010), suggesting that the mechanisms underlying perceptual and production capacities may have a more complex link than just a mere direct relation.

In the present study, no difference between training methods and the learners' vowel production after training was found, suggesting that auditory (AT), audio-visual (AVT) and video-alone (VT) perceptual training may foster improvement in vowel production to a similar extent. It is important to notice that the VT group improved as much as the other two groups in production as well as in perception, though they did not have access to an auditory model for vowels during the training sessions. This finding suggested that it may be possible to improve vowel production by only training learners to attend to the articulatory gestures of vowels and that production improvement can be achieved not only through auditory training.

With regards to evidence for the link between perception and production in this study, no strong relation between the two abilities was found before or after training, although improvement was found in both areas. Similar findings have been reported in previous studies (Bradlow et al., 1997; Iverson et al., 2012; Fabra & Romero, 2012). It may be argued that more improvement in perception than production is a reflection of the ability targeted with the perceptual training sessions; thus, less improvement in production should not be a surprise. The effect of individual variability is a bit puzzling; participants who improved more in perception were not necessarily the ones who showed greater improvement in production. However, we must be cautious about the comparison between these two measures as two different procedures were usually used. The perceptual improvement is obtained from the actual participant's performance; typically on a scale from 1 to 100 per cent correct. While the production improvement comes from scores (ratings) given by native speakers to the learners' production (scale 1-7). So this is a more impressionistic measure. It could be arguable to what extent these two measures are comparable. Nevertheless, this is a common practice in L2 perception and production studies.

To account for this lack of correlation between perception and production, Bradlow et al. (1997) suggested that learners may show different rates at which they experience "the motor commands" change to improve pronunciation. The observed that improvement in perception and production may be taken as evidence for a unified mental representation for both speech processes. Furthermore, they advanced that the modifications that occur during training may alter the underlying representations of the L2 sounds; however, they may not be powerful enough to change the motor commands involved in the production of those sounds.

To our knowledge, there seems to be little attention to this lack of correlation between L2 perception and production mechanisms in the L2 speech models that exist so far (Best, 1993; Best & Tyler, 2007; Flege, 1995). Apart from the studies mentioned above, most of the research that focuses on L2 speech perception and production does not discuss the nature of the underlying representations and the processes that regulate the perception-production link. Neither do they propose a theoretical explanation for the mismatch between perceptual and production improvement found in most of the studies in the area.

5. Conclusion

The results of this study revealed that L2 learners may improve their vowel identification capacity and vowel production with perceptual training although their ability to use visual information was not improved. These findings suggested that L2 learners were not sensitive to visual cues that are available for the identification of English vowels, as observed in the overall results of ENS, and that they relied mainly on the auditory input for the perception of English vowels. The robustness of the vowel trainer effect has been proved in previous studies with different L1 background (Iverson & Evans, 2009). In a study which compared trained L2 learners and a control group (Lengeris & Hazan, 2010), the latter showed no improvement in their vowel identification capacity when tested at the post-test time, confirming the benefit of the training sessions for the experimental group.

This lack of use of visual information for L2 speech perception may be related to the age at which learners started learning the L2 (Weikum et al., 2013). Besides, learners' level of proficiency had little or no relation with their identification ability to identify English vowels.

As most training studies show improvement in perception of L2 phonemes in short words, the impact of perceptual training on more "ecological" contexts still needs to be explored.

6. Bibliographic references

Akahane-Yamada, Reiko, Yohichi Tohkura, Ann R. Bradlow & David B. Pisoni, 1996: "Does Training in Speech Perception Modify Speech Production?," paper presented at the *4th International Conference on Spoken Language Processing*, 606-609.

Alashangiti, Wafaa, & Bronwen Evans, 2014: "Investigating the domain-specificity of phonetic training for second language learning: Comparing the effects of production and perception training on the acquisition of English vowels by Arabic learners of English," paper presented at the *International Seminar for Speech Production, Cologne, Germany*.

Aliaga-García, Cristina, & Joan C., Mora, 2009: "Measuring Perceptual Cue Weighting after Training: A Comparison of Auditory vs. Articulatory Training Methods" in *New Sounds 2010, Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech*, 19-24.

Best, Catherine, 1993: "Learning to perceive the sound pattern of English," *Haskins Laboratories Status Report on Speech Research. SR-114*, 31-80.

Best, Catherine, & Michael Tyler, 2007: "Nonnative and Second-Language Speech Perception: Commonalities and Complementarities" in Ocke-Schwen Bohn & Murray J. Munro (eds.): *Language Experience in Second Language Speech Learning. In Honor of James Emil Flege*, 15-34.

Bohn, Ocke-Schwen, & Catherine Best, 2010: "Testing PAM and SLM: Perception of American English Approximants by Native German Listeners" in *New Sounds, Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech*, 43-48.

Boersma, Paul, & David Weenink, 2012: "Doing phonetics by computer [Computer program]," version 2012, vol. 5, 52.

Bradlow, Ann R., Reiko Akahane-Yamada, David B. Pisoni & Yohichi Tohkura, 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.

Bradlow, Ann R., David B. Pisoni, Reiko Akahane-Yamada & Yohichi Tohkura, 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.

Chen, Yuchun, & Valerie Hazan, 2007: "Language Effects on the Degree of Visual Influence in Audiovisual Speech Perception" in *16th International Congress of Phonetic sciences*, 2177-2180.

Fabra, Lucrecia, & Joaquín Romero, 2012: "Native Catalan learners' perception and production of English vowels," *Journal of Phonetics* 40 (3), 491-508.

Flege, James, & Ian MacKay, 2004: "Perceiving Vowels in a Second Language," *Studies in Second Language Acquisition* 26 (1), 1-34.

Flege, James, Ocke-Schwen Bohn & Sunyoung Jang, 1997: "Effects of Experience on Non-Native Speakers' Production and Perception of English Vowels," *Journal of Phonetics* 25 (4), 437-470.

Flege, James, 1995: "Second-language Speech Learning: Theory, Findings, and Problems" in Winifred Strange (ed.): *Speech Perception and Linguistic Experience: Issues in Cross-language research*, Timonium, MD: York Press, 233-277.

Flege, James, 1988: "Factors affecting degree of perceived foreign accent in English sentences," *The Journal of the Acoustical Society of America* 84 (1), 70-79.

Fox, Robert, James Flege & Murray Munro, 1995: "The Perception of English and Spanish Vowels by Native English and Spanish Listeners: A Multidimensional Scaling Analysis," *The Journal of the Acoustical Society of Amertica* 97 (4), 2540-2551.

García-Lecumberri, María Luisa, & Jasone Cenoz, 1997: "L2 perception of English vowels: Testing the validity of Kuhl's prototypes," *Revista alicantina de estudios ingleses* 10, 55-68.

Hardison, Debra, 2006: "Effects of Familiarity with Faces and Voices on Second-Language Speech Processing: Components of Memory Traces," *INTERSPEECH*, 2462-2465.

Hardison, Debra, 2003: "Acquisition of Second-Language Speech: Effects of Visual Cues, Context, and Talker Variability," *Applied Psycholinguistics* 24 (4), 495-522.

Hardison, Debra, 1999: "Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect," *Language Learning* 49 (1), 213-283.

Hardison, Debra, 1996: "Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect," *Language Learning* 46 (1), 3-73.

Hattori, Kota, & Paul Iverson, 2010: "Examination of the relationship between L2 perception and production: an investigation of English/r/-/l/perception and production by adult Japanese speakers," *Second Language Studies: Acquisition, Learning, Education and Technology*, 2-4.

Hazan, Valerie, Anke Sennema, Andrew Faulkner, Marta Ortega-Llebaria, Midori Iba & Hyunsong Chung, 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.

Hazan, Valerie, Anke Sennema, Midori Iba & Andrew Faulkner, 2005: "Effect of Audiovisual Perceptual Training on the Perception and Production of Consonants by Japanese Learners of English", *Speech Communication*, 47 (3), 360-378.

House, David, Jonas Beskow & Björn Granström, 2001: "Timing and interaction of visual cues for prominence in audiovisual speech perception" in *7th EUROSPEECH*, 387-390.

Iverson, Paul, Melanie Pinet & Bronwen Eevans, 2012: "Auditory Training for Experienced and Inexperienced Second-Language Learners: Native French Speakers Learning English Vowels," *Applied Psycholinguistics* 33 (1), 145-60.

Iverson, Paul, & Bronwen Evans, 2009: "Learning English Vowels with Different First-Language Vowel Systems II: Auditory Training for Native Spanish and German Speakers," *The Journal of the Acoustical Society of America* 126 (2), 866-77.

Iverson, Paul, & Bronwen Evans, 2007: "Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration," *The Journal of the Acoustical Society of America* 122 (5), 2842-2854.

Iverson, Paul, Valerie Hazan & Kerry Bannister, 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.

Iverson, Paul, Patricia K. Khul, Reiko Akahane-Yamada, Eugen Diesch, Yohich Tohkura, Andreas Kettermann & Claudia Siebert, 2001: “A perceptual interference account of acquisition difficulties for non-native phonemes,” *Speech, Hearing and Language* 13, 106-118.

Kawase, Saya, Beverly Hannah & Yue Wang, 2014: “The Influence of Visual Speech Information on the Intelligibility of English Consonants Produced by Non-Native Speakers,” *The Journal of the Acoustical Society of America*, 136 (3), 1352-1362.

Kuhl, Patricia. K., Karen Williams, Francisco Lacerda, Kenneth Stevens & Björn Lindblom, 1992: “Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age,” *Science* 255, 606-608.

Lambacher, Stephen G., William L. Martens, Kazuhiki Kakehi, Chandrajith A. Marasinghe & Garry Molholt, 2005: “The effects of identification training on the identification and production of American English vowels by native speakers of Japanese,” *Applied Psycholinguistics* 26 (2), 227-247.

Lengeris, Angelos, & Valerie Hazan, 2010: “The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek,” *The Journal of the Acoustical Society of America* 128 (6), 3757-3768.

Li, Ying, 2016: “Audiovisual Training Effects on L2 Speech Perception and Production,” *International Journal of English Language Teaching* 3 (2), 14-36.

Lively, Scott E., David B. Pisoni, Reiko A. Yamada, Yohichi Tohkura & Tsuneo Yamada, 1994: “Training Japanese listeners to identify English/r/and/l/. III. Long-term retention of new phonetic categories,” *The Journal of the acoustical society of America* 96 (4), 2076-2087.

Logan, John, Scott Lively & David Pisoni, 1991: “Training Japanese Listeners to Identify English /r/ and /l/: A First Report,” *The Journal of the Acoustical Society of America* 89 (2), 874-86.

Massaro, Dominc W., Michael M. Cohen, Antoinette Gesi, Roberto Heredia & Minori Tsuzaki, 1993: “Bimodal speech perception: An examination across languages,” *Journal of Phonetics* 21, 445-478.

Massaro, Dominic W., & Michael M. Cohen, 1995: “Perceiving Talking Faces,” *Current Directions in Psychologival Science* 4 (4), 104-109.

McGurk, Harry, & John MacDonald, 1976: “Hearing Lips and Seeing Voices,” *Nature* 264 (5588), 746-748.

Morrison, Geoffrey, 2008: “L1-Spanish speakers’ acquisition of the English /i/–/ɪ/ contrast: Duration-based perception is not the initial developmental stage”, *Language & Speech* 51, 285-315.

Navarra, Jordi, & Salvador Soto-Faraco, 2007: "Hearing Lips in a Second Language: Visual Articulatory Information Enables the Perception of Second Language Sounds," *Psychological Research* 71 (1), 4-12.

Nishi, Kanae, & Diane Kewley-Port, 2007: "Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets," *Journal of Speech, Language, and Hearing Research* 50 (6), 1496-1509.

Ortega-Llebaria, Marta, Andrew Faulkner & Valerie Hazan, 2001: "Auditory-Visual L2 Speech Perception: Effects of Visual Cues and Acoustic-Phonetic Context for Spanish Learners of English" in *International Conference on Auditory-Visual Speech Processing*, 149-154.

Owens, Elmer, & Barbara Blazek, 1985: "Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers," *Journal of Speech & Hearing Research* 28 (3), 381-393.

Rosemblum, Lawrence, 2008: "Speech Perception as a Multimodal Phenomenon," *Current Directions in Psychological Science* 17 (6), 405-409.

Rosemblum, Lawrence, 2005: "Primacy of multimodal speech perception" in David Pisoni & Robert Remez (eds.): *The Handbook of Speech Perception*, Oxford: Blackwell Publishing Ltd., 51-78.

Sekiyama, Kaoru, 1997: "Cultural and Linguistic Factors in Audiovisual Speech Processing: The McGurk Effect in Chinese Subjects," *Perception & Psychophysics* 59 (1), 73-80.

Sueyoshi, Ayano, & Debra Hardison, 2005: "The role of gestures and facial cues in second language listening comprehension," *Language Learning* 55 (4), 661-699.

Walden, Brian, Sue A. Erdman, Allen A. Montgomery, Daniel M. Schwartz & Robert A. Prosek 1981: "Some effects of training on speech recognition by hearing-impaired adults," *J. Speech & Hearing Research* 24, 207-216.

Wang, Xinchun, & Murray Munro, 2004: "Computer-based training for learning English vowel contrasts," *System* 32 (4), 539-552.

Wang, Yue, Dawn Behne & Hisheng Jiang, 2008: "Linguistic Experience and Audio-Visual Perception of Non-Native Fricatives," *The Journal of the Acoustical Society of America* 124 (3), 1716-1726.

Wang, Yue, Dawn Behne & Hisheng Jiang, 2009: "Influence of Native Language Phonetic System on Audio-Visual Speech Perception", *Journal of Phonetics* 37 (3), 344-356.

Weikum, Whitney M., Athena Vouloumanos, Jordi Navarra, Salvador Soto-Faraco, Núria Sebastián-Gallés & Janet F. Werker, 2013: "Age-Related Sensitive Periods Influence Visual Language Discrimination in Adults," *Frontiers in Systems Neuroscience* 7, 1-8.

Wik, Preben, 2011: *The Virtual Language Teacher: Models and applications for language learning using embodied conversational agents*. Doctoral thesis, KTH Royal Institute of Technology.

Wik, Preben, & Anna Hjalmarsson, 2009: "Embodied conversational agents in computer assisted language learning," *Speech Communication* 51 (10), 1024-1037.